Cardiff University

School of Biosciences

Cloning and Engineering of Novel Coleopteran Luciferase

Jack Paul Bate

30th of March 2022

Cardiff School of Biosciences
The Sir Martin Evans Building
Museum Avenue
Cardiff
CF10 3AX

A thesis submitted for the degree of Doctor of Philosophy for Cardiff University

Abstract

Firefly luciferases produce visible light by catalysing a reaction utilizing *D*-luciferin, Mg^{2+}, ATP, and molecular oxygen, in a widely studied bioluminescent system which has been developed into a range of applications relevant to industry and academia.

Advancement of firefly luciferase applications depends upon the continued discovery of novel luciferase phenotypes and their respective sequences, regardless of whether these are derived from nature or mutagenic means.

The feasibility of bioprospecting novel luciferase gene sequences from dry preserved Coleoptera provided by Amgueddfa Cymru – National Museum Wales was explored. A non-destructive DNA extraction method was followed to enable the preservation of all specimens. A cross-species capture hybridisation method was developed, using biotinylated probes of the luciferase gene from *Photinus pyralis*, to enrich for luciferase gene fragments prior to Illumina sequencing and the recovery of luciferase gene sequences with a bioinformatics strategy. With this approach, a gene sequence was recovered from an unidentified Costa Rican firefly, which encoded a novel luciferase capable of catalysing a bioluminescence reaction, termed CRLuc.

Simultaneously, engineering of the luciferase from *Phosphaenus hemipterus* (*Phem*Luc) was attempted to generate two discrete variants which possessed improved compatibility with the synthetic substrate analogue infraluciferin, and increased thermostability. PhemLuc was selected as it presented the opportunity to discover novel mutagenic functionality in a previously uncharacterized enzyme.

Homology models of *Phem*Luc were constructed to identify residues in proximity to the bound substrate for targeted mutagenesis. From 32 targets mutagenized, the mutations H245W and A313G produced considerably increased bioluminescence with infraluciferin, and conferred cumulative effects when combined into a dual mutant termed x2 Infra.

Thermostabilisation of *Phem*Luc was attempted using 15 mutations reported to thermostabilise *Photinus pyralis* luciferase. Although this mutant was inadequate, DNA shuffling identified an x14 reversion mutant with restored bioluminescence and improved thermostability. This mutant was further improved using epPCR to include mutations I231V and L306H to complete a final x16 thermostable mutant, later demonstrated as capable of functioning in LAMP-BART.

The work conducted here demonstrates the potential of targeted bioprospecting for genetic discoveries using museum materials, in addition to the identification of several novel mutations in PhemLuc which warrant further investigation in related luciferase variants.

Contents

List of Figures

List of Tables

List of Equations

Frequent Abbreviations

Å	Angstrom
λ_{max}	Peak emission wavelength
ALH_2	Amino luciferin (6'-amino-*D*-luciferin)
AMP	Adenosine monophosphate
ATP	Adenosine triphosphate
BART	Bioluminescent Assay in Real Time
BLI	Bioluminescent imaging
BOLD	Barcode of Life Database
BP	DNA base pair(s)
cDNA	Complementary DNA
CDS	Complete coding DNA sequence
CODEHOP	Consensus-Degenerate Hybrid Oligonucleotide Primers
COI	Mitochondrial cytochrome *c* oxidase I gene
CRLuc	Unidentified Costa Rican firefly luciferase
d.p.	Decimal place
D_2O	Heavy water (deuterium oxide)
dH_2O	Molecular grade water
DLSA	5'-O-[N-(dehydroluciferyl)-sulfamoyl]-adenosine
DMSO	Dimethylsulphoxide
DNA	Deoxyribonucleic acid
dNTP	Nucleoside triphosphate
DTT	Dithiothreitol
epPCR	Error prone polymerase chain reaction
Fluc	Firefly luciferase
FWHM	Full width half maximum
gDNA	Genomic DNA
H-bond	Hydrogen bond

iDLSA	5'-O-[N-(dehydroinfraluciferyl)-sulfamoyl]-adenosine
iLH$_2$	*DL*-Infraluciferin
I$_{max}$	Maximum observed intensity
IMD	Imidazole
IPTG	isopropylthiogalactoside
LAMP	Loop-mediated amplification
LB	Luria Bertani medium
LH$_2$	Beetle *D*-luciferin
LH$_2$-AMP	Luciferyl adenylate
Lnoc	*Lampyris noctiluca*
Luc	Luciferase
MM	Michaelis-Menten
MnCl$_2$	Manganese chloride
NGS	Next generation sequencing
NTC	No template control
*Phem*Luc	*Phosphaenus hemipterus* luciferase
PMT	Photomultiplier tube
PPi	Inorganic pyrophosphate
Ppy	*Photinus pyralis*
qPCR	Quantitative polymerase chain reaction
RLU	Relative luminescence units
RT	Room temperature
s.f.	Significant figure
SDM	Site directed mutagenesis
SEM	Standard error of the mean
T$_{max}$	Time to peak

Acknowledgements

This work was funded by the KESS2 Knowledge Economy Skills Scholarship in partnership with Amgueddfa Cymru – National Museum Wales. A special thank you must go to Dr Michael Wilson of the National Museum Wales for his involvement, support, and assistance acquiring Coleopteran specimens, without which much of this work would not have been possible.

Thank you to the Murray lab research group for facilitating this work and creating a welcoming and supportive research environment like no other. Additional thanks must go to Joanne Kilby for her patience with my many questions and ensuring that all materials I required were available.

A special thank you goes to my mentor Dr Amit Jathoul for the creation of this PhD project and for encouraging my initial interest in pursuing scientific research. Amit's overwhelming knowledge within and beyond the fields of bioluminescence and protein engineering has been a constant inspiration and a reminder that there is always more to be learnt. His enthusiasm for his work is infectious, and I hope to carry that attitude forward in all that I do. I am also grateful for his guidance with the writing and proofreading of this thesis.

Another special thank you must go to Dr Patrick Hardinge for his support and insight, and who's example of scientific knowledge, skill, precision, and admirable work ethic I have attempted to mirror over the years of my development as a research scientist. I am also grateful for his guidance with the writing and proofreading of this thesis.

A special thank you also goes to my supervisor Prof Jim Murray. Thank you for the continued support, guidance, and understanding over the duration of this project. Thank you for the many insightful questions which have always pushed me to contemplate my work at greater depth, and for always having a relevant suggestion for any complications encountered. I am also grateful for his guidance with the writing and proofreading of this thesis.

I am also grateful to Prof Jim Anderson for his provision of Infraluciferin, Dr John Day for his contribution of the Phosphaenus hemipterus luciferase gene sequence, and Dr Daniel Pass for bioinformatics training and support throughout this project.

Thank you to my family for their encouragement, and lastly to my wonderful partner Chloe for the endless support, patience, and understanding. It's been a long journey and I couldn't

have done any of it without you by my side. I look forward to the next chapter of our life together, and all that are to follow.

Chapter 1

Introduction

1.1. Bioluminescence from Antiquity to the 20th Century

Bioluminescence is enzyme catalysed chemiluminescence occurring naturally in living organisms selected through evolution to enable the conversion of stored chemical energy into the controlled emission of light. The word bioluminescence was first introduced in the 20th century by E. Newton Harvey to describe all forms of 'living light'. The first documented use of the word luminescence can be accredited to the German physicist Eilhard Wiedemann in 1888 to describe all systems of light without an observable radiation of heat as a secondary energy from the system (Harvey 1957). Human observation and fascination with bioluminescence predates this by millennia, with the earliest known references to fireflies and Glow-worms believed to be from ancient writings and poetry from the region that would now be recognized as modern day Southern China (Lee 2008). Oral traditions and folk stories are harder to date, but include myths on the origin of fire from the burning seas, an event documented as far back as the ancient Greeks (circa. 500 BC) and famously by Christopher Columbus in 1492. We now know that these events are common, and since the 1830's we have been able to attribute the luminescence of the oceans to marine Dinoflagellates due to the extensive microscopy studies of ocean water samples by the German naturalist Christian Gottfried Ehrenberg (Harvey 1957).

The first detailed observations of bioluminescence were made by Aristotle (384-322 BC) who amongst a wide spectrum of naturalist works managed to document 180 marine specimens along with descriptions of fireflies and Glow-worms and was the first to recognize bioluminescence as cold-light, without the heat associated with the Sun or a flame. The earliest documented experiments attempting to understand the mechanism responsible for bioluminescence were performed in 1667 by Robert Boyle, who was able to identify that the process was dependent on air (Boyle 1668). In truth, the bioluminescent reaction is dependent on oxygen, but oxygen would not be discovered for more than a hundred years in 1772 by a Swedish chemist, Carl Wilhelm Scheele. In 1885 Raphael Dubois was able to conduct the first experiments investigating the biochemistry of Coleopteran bioluminescence by extracting two key components of the bioluminescent system that when combined would generate the expected light emission. Dubois noted that the first of his extracts was heat resistant and the second was heat sensitive, from this work the extracts were named ‘luciferin’ and ‘luciferase’, respectively (Fraga 2008).

1.2. Living Light in the Natural World

Bioluminescent organisms exist in various forms across almost all ecosystems in nature. For example, in the deep ocean, disparate bioluminescence systems can be found serving a multitude of functions (Cormier and Karkhanis 1971; Thompson *et al.* 1989; Vysotski and Lee 2004.; Markova *et al.* 2015). Bioluminescence is much less prevalent in terrestrial environments than the deep ocean, but can still be observed across the kingdoms of bacteria, fungi and animals (Day *et al.* 2004), with plantae being the notable exception. There exist numerous discrete functions for bioluminescence which benefit the host organism and a variety of ways in which it is displayed and controlled, including as a means of camouflage, sexual communication, and aposematism, a term describing the concept of warning colouration or in the case of bioluminescence, light emission or flashing (Sivinski J, 1981).

Unsurprisingly, bioluminescence in terrestrial organisms is often brighter due to their selection against higher levels of background light. The colour of the emitted light varies between systems, with identified examples traversing the expanse of the visible electromagnetic spectrum (Wood *et al.* 1989; Viviani *et al.* 2006; Trowell *et al.* 2016).

Curiously, whilst bioluminescence can commonly be autogenic, being produced by the host organism independently, bioluminescence may also be bacteriogenic in origin where a symbiotic relationship allows a host organism to express the phenotype of bioluminescence that they would not otherwise be able to produce autogenically. This relationship can be common in marine environments; a familiar example exists in deep sea Anglerfish which host bioluminescent bacteria at the end of a modified dorsal ray in a specialized organ called an esca (Hulet and Musil 1968). This adaption provides the Anglerfish with a luminescent lure for luring prey and mate attraction.

1.3. Disparate Bioluminescence Systems

The morphology of bioluminescence can be as diverse as the organisms in which it is possessed. Some organisms have developed specialised structures for the containment and control of luminescence by the nervous system allowing the regulated stimulation of light emission. These structures can vary from the eye-like structures of luminous fishes and squids to the thoracic lantern organ of fireflies (Shimomura 2006). In single celled bioluminescent organisms such as bacteria and fungi, the entire system for light emission is present without any structures or processes of regulation, and thus luminosity can be a constant phenotypic state. The use of chained interrelated chemical reactions allows for intermediates between these two groups of control, where stimulation from the environment can trigger a cascade of signalling events that can lead up to the final light emitting reaction.

It is estimated that 80% of all bioluminescent organisms exist in marine environments, where such systems are more likely the rule rather than the exception. Bioluminescent systems can and do exist in freshwater environments, but their presence is insignificant compared with the abundance of marine luminescence. This discrepancy has been hypothesized to be due to lack of optical transparency that is more common in fresh water (Haddock *et al.* 2010). The remaining 20% exists in terrestrial organisms and although it is not fully understood why there is such a divergence in abundance between the two environments it is agreed that the oceans are likely a more favourable environment for the evolutionary development of bioluminescence due to their significantly reduced levels of background light and comparatively stable environmental conditions.

The majority of bioluminescent marine organisms emit blue light (410-550 nm) which permits the greatest optical transparency in the ocean whilst also aligning to the peak sensitivities of specific opsin proteins, which mediate the conversion of photons of light into electrochemical signals which are interpreted by the brain as vision (Kahlke and Umbers 2016). Largely due to differences in optical transparency in terrestrial environments, terrestrial organisms light emission wavelengths are commonly green (550 nm), but can extend in to the far-red with the most redshifted emission spectra being recorded at the peak wavelength of 628 nm in the click beetle *Phrixothrix hirtus* (Viviani *et al.* 1999). The emission wavelengths of any bioluminescent system are dependent on multiple factors including the structure of the luciferase protein and the luciferin substrate of the reaction. Light emission wavelength can also be considerably modified through the use of secondary emitters such as fluorescent proteins which absorb light of a particular wavelength to convert into the emission of an often red-shifted wavelength. This process is known as bioluminescence resonance energy transfer (Shimomura 1995).

Perhaps contending for the most unusual of bioluminescent systems are those found in certain marine copepods including *Gaussia princeps, Metridia longa,* and *Metridia pacifica* which all possess the ability to secrete light emitting molecules in contrast to the more common intracellular bioluminescence in which organisms concentrate the emission within an internal environment (Verhaegen and Christopoulos 2002; Markova *et al.* 2004; Takenaka *et al.* 2008). Generally, the luciferases of these copepod systems are well conserved small proteins (20-30 kDa) which catalyse light emission from the shared substrate coelenterazine without the requirement for additional cofactors in the reaction (Takenaka *et al.* 2016; Markova *et al.* 2019).

The ability to produce a luciferin-like compound is dependent on multi-enzymatic reactions which are likely the most complex requirement to many natural bioluminescent systems. This does not however imply that all bioluminescent organisms have the ability to synthesize their own bioluminogenic substrate molecules. To explore a single example, the Midshipman fish, *Porichthys notatus* is divided into two genetically identical but geographically discrete groups. The consumption of *Vargula tsujii*, a bioluminescent crustacean present in only one regional population, permits the absorption of crustacean synthesized luciferin in the gut which is then circulated in the blood stream allowing bioluminescence in the otherwise non-luminescent *Porichthys.* In addition to producing their own luciferase enzyme, *Porichthys* is capable of maintaining a low steady-state level of luciferin in the blood by recycling luciferin molecules that have previously undergone the enzymatic reaction (Thompson *et al.* 1988). This dietary supplementation of substrate is in direct contrast to the fully autogenic bacterial bioluminescent systems which concentrate all of the required components of their bioluminescent system, from enzyme to substrate synthesis, to be encoded under the lux operon (Tu and Mager 1995).

1.4. The Evolution of Bioluminescence

Intriguingly, the origin of the bioluminogenic substrates and their respective protein catalysts can vary such that it is thought bioluminescent systems have originated independently possibly in excess of forty times, making bioluminescence a remarkable case of convergent evolution (Haddock *et al.* 2010). The luciferase enzymes catalysing the light-emitting reactions of fireflies, coelenterates, and bacteria, for example, show little sequence homology to each other, and the luciferin substrates of their reactions are chemically unrelated (Hastings and Wilson 1998). To date only nine natural luciferins have had their structures solved (Kaskova *et al.* 2016).

For so many convergent evolution events to occur and result in the phylogenetic distribution seen for bioluminescence models, a wide range of selective pressures for these systems must exist in nature. The proposed benefits that bioluminescence confers to host organisms include attraction, repulsion, communication, camouflage, and illumination. However, it may be that bioluminescence is a secondary property of these paralogous biochemical systems, and that the true primary selection was for an ancient oxygen detoxification mechanism (Timmins *et al.* 2001).

In the primordial stages of life on Earth the emergence of oxygenic photosynthesis would have drastically increased the proportionate levels of oxygen in the environment which in turn could photochemically generate toxic reactive oxygen species including H_2O_2 and O_2^-. This would have created a new pressure for supplementary antioxidative systems in susceptible primitive organisms (Rees *et al.* 1998; Timmins *et al.* 2001). To combat the increased oxidative stress, it has been proposed that the evolutionary foundation of bioluminescent systems are the luciferin molecules as opposed to the luciferase enzymes responsible for the reactions catalysis. In this proposal, luciferin would have first been selected to act as toxic oxidant scavenging molecules, and only later serving as light emitting substrates for early luciferase-like enzymes (Dubuisson *et al.* 2004). This is evidenced by the known antioxidant properties of coelenterazine which is the functioning luciferin molecule in many marine bioluminescent systems (De Wergifosse *et al.* 2004).

The action of early luciferase-like enzymes in catalysing the reduction of toxic oxidants by luciferin would have initially produced inconsequential levels of light emissions, due to the low total atmospheric oxygen. It is only as oxygen levels increased and the intensity of the antioxidant light emissions proportionately increased that the light signals would have been sufficient for detection by primitive photoreceptors and evolution of bioluminescence could finally be driven primarily by light emission (Timmins *et al.* 2001).

1.5. Beetle Bioluminescence

1.5.1. The roles of beetle bioluminescence

Within the world of terrestrial bioluminescence, the distribution across species is not as diverse as within marine environments. One collection of organisms comprises such a significant proportion of terrestrial bioluminescence that it is perhaps the most familiar of all biological light, the bioluminescent beetles. Beetle bioluminescence is distributed across the families Lampyridae, containing fireflies and Glow-worms, Phengodidae, known also as Glow-worms, and Elateridae which are commonly known as click beetles. The geographical distribution in the Lampyridae family alone comprises all continents excluding the Antarctic, as indicated by the specimen distribution on the BOLD database (www.boldsystems.org).

Bioluminescent beetles produce light in the peroxisome of photocytes in external luminescent organs of such incredible diversity that they can range from small pin-head structures found anywhere from the head to the tip of the abdomen, up to tail lanterns which can occupy the entire ventral surface of several abdomen segments (Buck 1948). In fireflies, bioluminescence is paramount to communication with the environment, whether that be for sexual communication within the species, or for aposematism, i.e. prey attraction and predation deterrence. Bioluminescence in many species of Coleoptera is not only displayed in adults but also throughout the larval stage of development. This larval bioluminescence has been demonstrated to serve an aposematic role discrete from the sexual communication of bioluminescence in adults, whereby naïve predators are rapidly able to associate the light signals with the distasteful lucibufagin compounds synthesized within larvae (Underwood *et al.* 1997). In adults, bioluminescence is most strongly associated with sexual communication and mating behaviour with each species adhering to strict parameters of light intensity, flash patterns, and flash synchrony in order to best attract a high fitness reproductive mate (Lloyd 1983).

1.5.2. The evolution of beetle luciferases

The roles of aposematism and bioluminescent courtship behaviour have been subject to long evolutionary selection, with the oldest known fossil record of the Lampyridae family belonging to *Protoluciola*, a specimen found in Burmite amber and dated to ~99 million years ago (Mya) (Kazantsev 2015). Recent estimates suggest that the ancestor of the luciferase gene in Lampyridae may have diverged ~205 Mya (Zhang *et al.* 2020a). From the pervasive distribution of fireflies both geographically and chronologically, there exists significant interspecies variation in all aspects of bioluminescence characteristics through diversification of the luciferase enzyme. Whilst the broad variation and study of natural bioluminescent systems is key to the development of bioluminescent technologies in academia and industry, Coleopteran luciferases have been awarded significant attention due to the flexibility of their characteristics providing a catalogue of enzymes traversing a vast spectrum of emission colours and kinetics, all with a high luminescent intensity and utilizing the same bioluminescent system derived from a common ancestor.

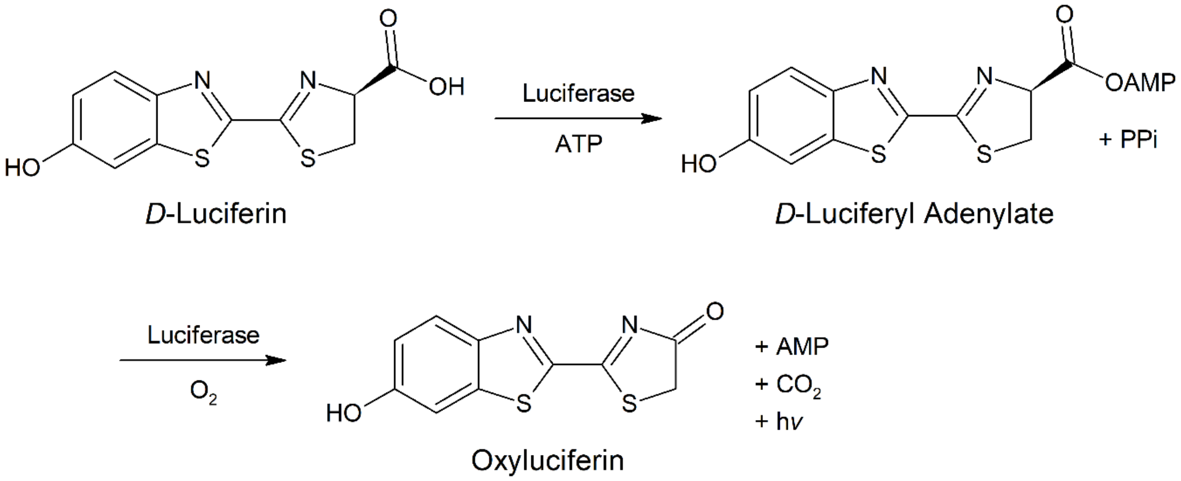
Interestingly, luciferase exhibits bifunctionality alongside its role as an ATP-dependent monooxygenase in the bioluminescence reaction, as it also performs in a CoA-ligase reaction akin to other long chain fatty acyl-CoA synthetase enzymes. The role of luciferase in both pathways requires its catalytic adenylation action using Mg-ATP. Where the CoA-ligase pathways differ is in the formation of luciferyl CoA by the substitution of AMP by CoA (Oba *et al.* 2003).

Firefly luciferases have a high degree of structural homology with, and fall into the enzyme superfamily of, acyl-CoA synthetases, with around 60% amino acid conservation. The homologous gene of firefly luciferase in *Drosophila melanogaster* has also been shown to function as a fatty acyl-CoA synthetase (Pubchem: CG6178), but with no activity as a luciferase in a bioluminescence reaction (Oba *et al.* 2004; Oba *et al.* 2005). The indication of this biochemical and phylogenetic analysis is that firefly luciferase evolved from a fatty acyl-coenzyme A synthetase following an initial gene duplication event (Oba *et al.* 2006).

1.5.3. The beetle bioluminescence reaction

As represented in Figure 1.1., the beetle bioluminescence reaction is catalysed by the bi-functionality of luciferase in the two processes of adenylation and subsequent oxidation. The first stage requires the activation of the native substrate in bioluminescent Coleoptera, *D*-luciferin (LH_2 – (S)-2-(6-hydroxy-2-benzothiazolyl)-2-thiazoline-4-carboxylic acid) by adenylation using adenosine triphosphate (ATP) to produce *D*-luciferyl adenylate (LH_2-AMP) and a molecule of inorganic pyrophosphate (PPi). Following the production of *D*-luciferyl adenylate, oxidation occurs in the presence of molecular oxygen to generate a dioxetanone ring that undergoes rearrangement and subsequent decarboxylation to produce oxyluciferin in an excited state (LO*). In order for this excited oxyluciferin to return to the ground energy state (LO), energy must be rapidly lost by the emission of photons (White *et al.* 1971).

Figure 1.1.



The two-step reaction catalysed by firefly luciferase. *D*-luciferin is first converted to a *D*-luciferyl adenylate intermediate and subsequently oxidated and decarboxylated to an excited state oxyluciferin that releases the bioluminescence emission, represented here as h*v* for light. Diagram produced using ChemSketch, available at https://www.acdlabs.com.

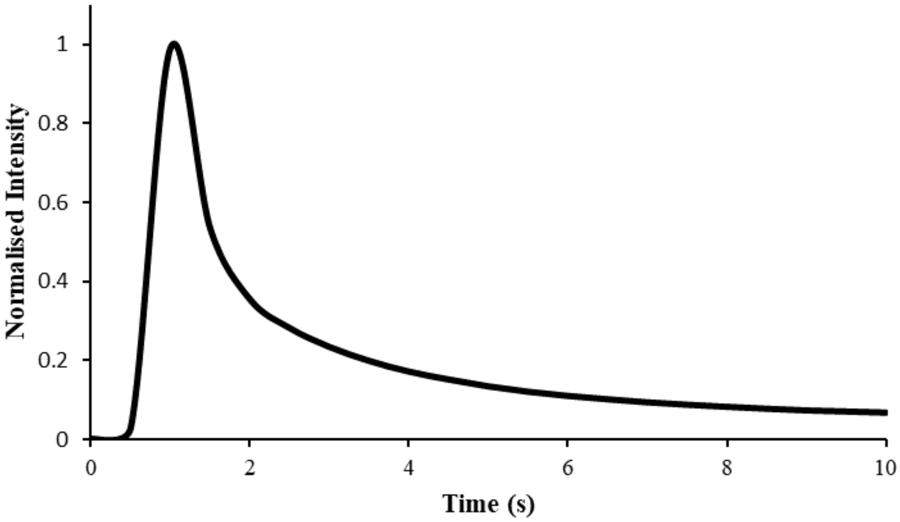
1.5.4. Bioluminescence emission kinetics

To measure the maximal light emission of any firefly luciferase *in vitro*, the assay is performed under saturating conditions of ATP and luciferin. When the reaction is performed with North American firefly *Photinus pyralis* (*Ppy*) luciferase (Fluc) at 25 ˚C, there is a characteristic flash kinetic (rise to maximum activity) over 300 ms that follows an initial short lag phase (DeLuca and McElroy 1974) and illustrates the occurrence of two rate limiting conformational changes (Sandalova and Ugarova 1999). Figure 1.2. demonstrates this characteristic Fluc emission profile. Following the lag period, the light emission achieves a peak output of maximal light intensity (I_{max}) as a result of the first turnover of luciferase with substrate (DeLuca and McElroy 1974). Once I_{max} is achieved, light emission rapidly decays due to the significant product inhibition of the reaction by dehydroluciferyl-AMP, and this rapid decay completes the characteristic flash kinetic (Lemasters and Hackenbrock 1977). Although not directly involved in the bioluminescence reaction, dehydroluciferyl-AMP accumulates as the co-product of a side reaction between luciferyl adenylate and molecular oxygen, which react to produce hydrogen peroxide. About 80% of the luciferyl adenylate intermediate participates in the bioluminescence reaction, whilst the remaining 20% is oxidized in this secondary pathway (Fraga et al. 2006). The reaction is further competitively inhibited by the presence of oxyluciferin, but to a lesser extent than dehydroluciferyl-AMP (Ribeiro and Esteves da Silva 2008). Subsequent to the rapid decay to a lower basal light emission is a considerably decreased rate of decay whereby light emission appears relatively stable as substrate is being utilized and protein aggregates form (Brovko *et al.* 1994).

In order for the firefly luciferase bioluminescence reaction to be maintained in Coleoptera a process of LH_2 regeneration is required. Although this research topic has attracted much attention over decades of bioluminescence research, no mechanism has yet been proven. However, the current suggestion from density functional theory calculations is that luciferin regeneration consists of three sequential steps where oxyluciferin produced in the bioluminescence reaction is first hydrolysed by the luciferin regenerating enzyme to generate 2-cyano-6-hydroxybenzothiazole (CHBT). CHBT is then thought to combine with L-cysteine to produce L-luciferin via a condensation reaction, which then inverts into D-luciferin (LH_2) in luciferase and thioesterase (Cheng and Liu 2019).

Photinus pyralis (the North American firefly) possesses what is perhaps the most well studied Fluc. At optimum pH of ~7.8 the *Ppy* Fluc emits light in the visible green region of the electromagnetic spectrum with a wavelength maximum (λ_{max}) of 550-560 nm (White *et al.* 1980). Due to the sensitivity of this reaction, changes in pH or increases in temperature result in a red shifting and broadening of the emitted light, a bathochromic shift (Mcelroy *et al.* 1969). Without modifying the enzyme itself it is also possible to alter the light emission spectra by introducing secondary factors to the reaction environment such as heavy metals and further ions.

Figure 1.2.



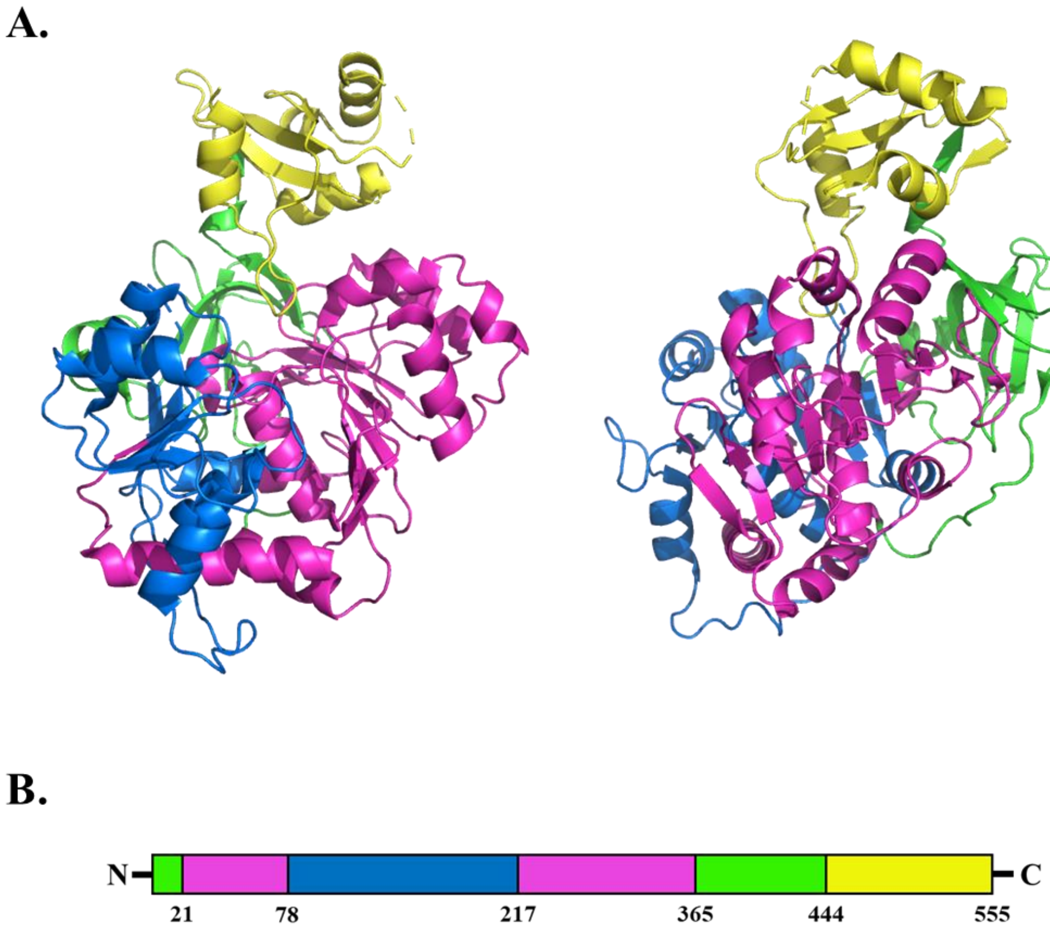
Bioluminescence emission profile. Representation of the light emission overtime from the initiation of the bioluminescence reaction. Normalised light intensity relative to the peak is plotted against time. Emission curve is for illustrative purposes only and not derived from existing emission data.

1.6. The Structure of Firefly Luciferase

1.6.1. Overall structure

The firefly luciferase gene is structured as 7 exons with 6 introns, all of which are less than 60 nucleotides in length (De Wet *et al.* 1987). Analysis of the 5' untranslated regions has revealed evidence for a conserved putative core promoter region from -190 through to -155 upstream of the luciferase start codon (Day *et al.* 2006). X-ray crystallography reveals the encoded *Ppy* firefly luciferase to be a 62kDa monomer organized into two domains (Sundlov *et al.* 2012). A larger N-terminal domain (residues 1-444) is comprised of three subdomains: a compact distorted antiparallel beta-barrel which connects with two separate beta-sheets represented in Figure 1.3. by green, purple, and blue colouring, respectively. The beta-sheet subdomains of the N-domain are each flanked by alpha-helices on either side to form a five-layered αβαβα tertiary structure (Conti *et al.* 1996). The smaller second domain is at the C-terminus (residues 445-555), the structure of which is an alpha and beta domain shown in Figure 1.3. as yellow.

Figure 1.3.



Orthogonal views of the *Ppy* Fluc as ribbon representations. (**A**) Colour coded depictions of the three subdomains of the large N-terminal domain and the smaller C-terminal domain. Subdomain β-sheet A is shown in blue and comprises residues 79-217. Subdomain β-sheet B is shown in purple and comprises residues 22-78 and 218-365. Subdomain β-barrel is shown in green and comprises residues 1-21, and 366-444. The small C-terminal domain shown in yellow comprises residues 445-555. (**B**) Schematic representation of the subdomain boundaries in the amino acid sequence, using the same colouring scheme. Boundary numbers indicate the final residue of the left hand region. Protein structure and domain/subdomain boundary information are taken from PDB ID: 4G36 (Sundlov *et al.* 2012). Protein images produced in PyMOL.

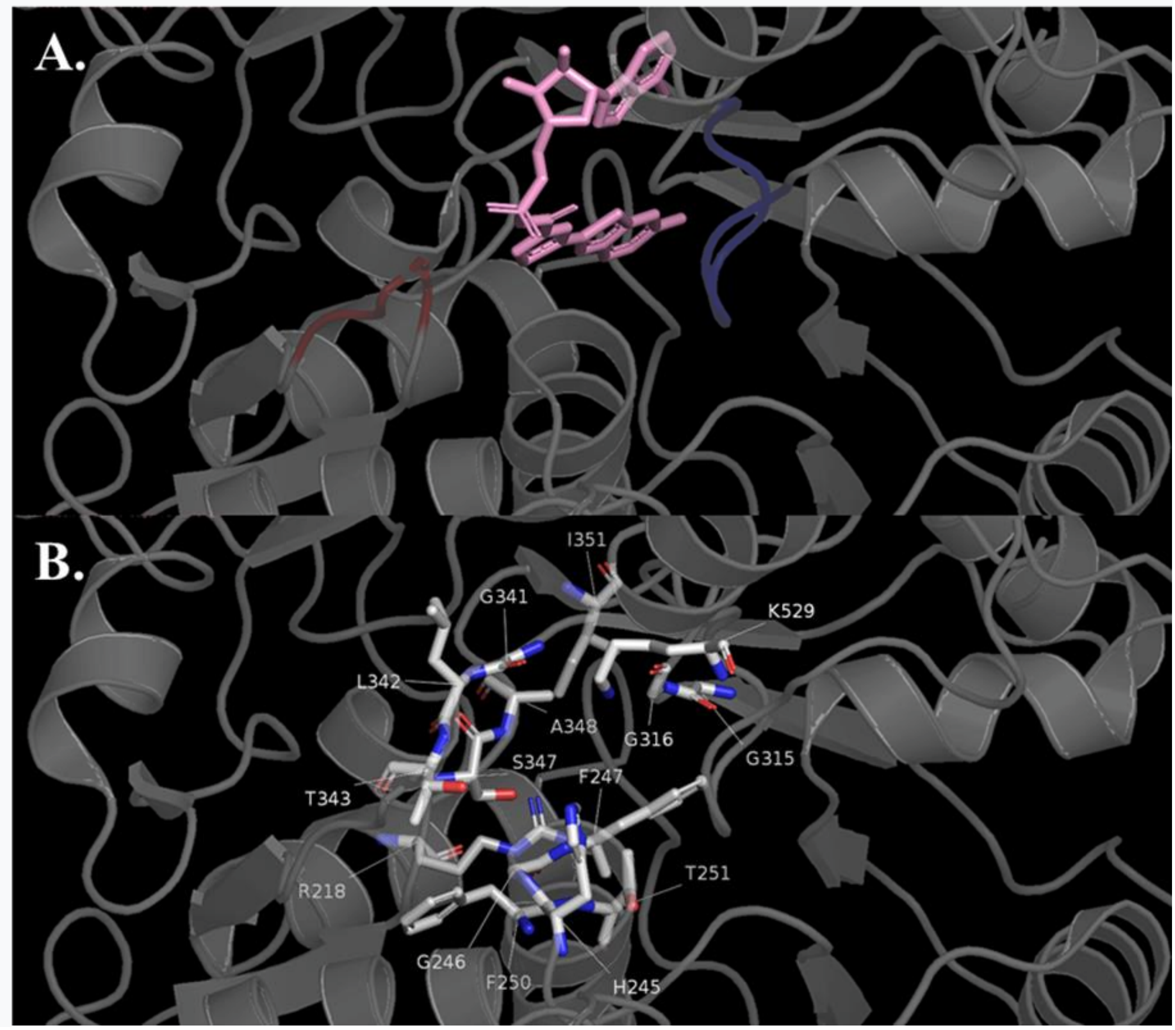
1.6.2. The active site

The smaller C-domain is connected to the N-domain by a flexible hinge structure to produce a large cleft in which the most conserved active site residues reside, occupying positions on either side of the cleft that are too distant to interact (Conti *et al.* 1996). The implication of this is that the two domains must come together to envelope the substrate. This concept has however matured in more recent years from evidence that upon substrate binding the open structure of the two domains is initially brought together to form a hydrophobic pocket which tightly sandwiches the substrate benzothiazole ring in a closed conformation which catalyses the first adenylation step of the reaction. Following adenylation and the release of PPi, a 140˚ rotation of the C-domain occurs to allow the protein to adopt a new conformation capable of catalysing the subsequent oxidation of the *D*-luciferyl adenylate produced by the primary closed conformation (Branchini *et al.* 2005b; Nakatsu *et al.* 2006; Gulick 2009). Two lysine residues have been implicated for their roles in this two-part adenylation and oxidation. K529 is thought to catalyse the initial adenylation, whereas K443 catalyses oxidation. Mutagenesis at either position will disrupt their specific role without affecting the proteins ability to perform the other half of the reaction (Sundlov *et al.* 2012).

Two loop structures residing in the active site region have been identified for their involvement in the substrate binding of firefly luciferases (Figure 1.4A). The phosphate binding loop (P loop) is a universal motif in ATP binding enzymes, and comprises residues S198 – K206 in *Ppy* Fluc. The second loop known as the active site loop comprises residues K524 – L530. It is speculated that these loops interact to form the substrate binding site, and subsequently have a vital role in substrate binding and enzyme reactivity (Jazayeri *et al.* 2017).

Key residues of the active site are not thought to be limited only to those which reside in these two loop structures. A model of the active site constructed by Branchini *et al* (1998) sought to identify key residues of the luciferin binding site, and identified fifteen putative residues within 5 Å of the substrate (Figure 1.4B). Site-directed mutagenesis of these positions identified that substitution at twelve of these positions resulted in a ≥4-fold K_M difference for the luciferin substrate binding affinity. Of these twelve, the seven residues spanning the region R218 – A348 had ≥30 nm red-shifted bioluminescence emission maxima when mutated. This investigation and the work which followed speculated that similar experimental approaches may provide a foundation to alter the substrate specificity of firefly luciferases (Branchini *et al.* 2003).

Figure 1.4.



Overview of the *Ppy* Fluc active site. (**A**) The two loop structures of the active site. The P loop is shown in Red and the active site loop in blue. The structural analogue of LH_2 (DLSA) is shown in Pink. (**B**) The 15 putative active site residues from Branchini *et al*. 2008. View is maintained relative to **A**. DLSA is omitted for improved visualisation. Model produced using 4G36.pdb (Sundlov et al. 2012) in PyMOL.

1.7. Bioluminescence Emission Spectra

1.7.1. Mechanisms of emission spectra variation

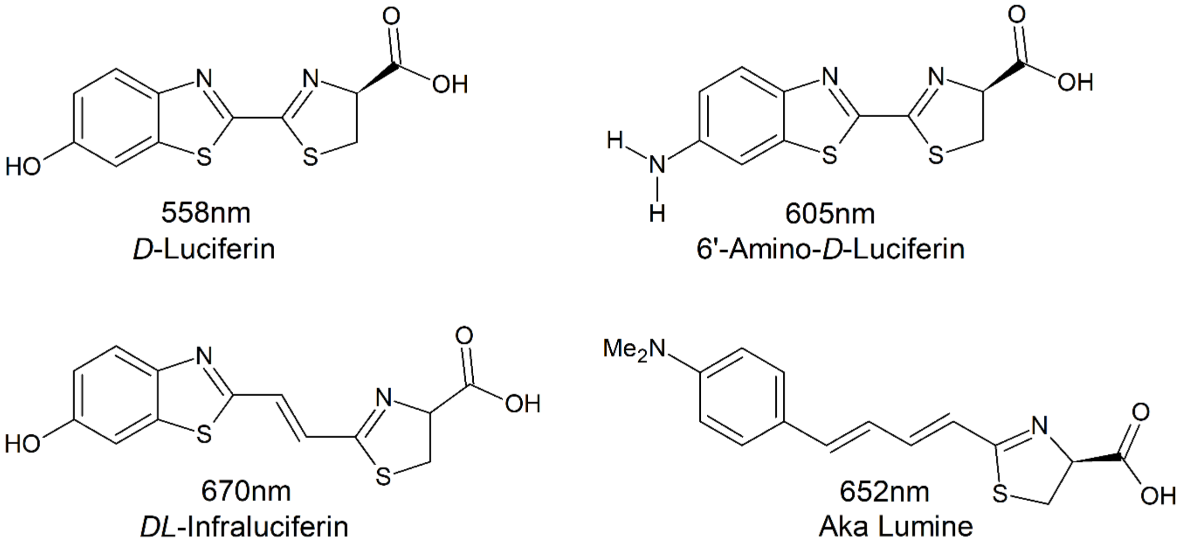
Of the luminous beetles, firefly luciferases are known to be uniquely pH-sensitive and under acidic conditions a typical red-shift of the bioluminescence spectra will occur known as bathochromic shift (Viviani *et al.* 2008). This effect can additionally be brought about by further destabilising conditions including high temperature, the presence of heavy metals, denaturants, and other various ions. The extent of bathochromic shift is pH-dependent such that as the reaction pH is dropped from pH 8.0 to pH 6.0 the bioluminescent spectra can shift in excess of 50 nm. As the luciferases from click beetles and railroad worms are not known to exhibit such a bathochromic shift, the implication is that this effect is protein mediated (Tisi *et al.* 2002b).

Regardless of destabilising conditions which contribute to bathochromic shift, the emission spectra of bioluminescence can vary significantly between all beetle luciferases. Whilst there is currently no clear consensus on a mechanism which can fully explain the observed variation, a recent suggestion is that the luciferin binding and catalytic amino acids of the active site may favour three different conformations of the excited oxyluciferin, which each possess unique emission properties. The three conformations include a red (λ_{max} ~615 nm) monoanion keto form, a green (λ_{max} ~540 nm) dianon enolate form, and an intermediate emitting monoanion enol form, and therefore the emission spectra for a given luciferase is dictated by the profile of equilibrium between the oxyluciferin forms produced during catalysis (Naumov and Kochunnoonny 2010; Bechara and Stevani 2018).

1.7.2. Emission spectra in response to substrate analogue substitution

The colour of firefly luciferase bioluminescence can be significantly altered via the substitution of different luciferin analogues (Figure 1.5.), such as 6'-amino-*D*-luciferin (ALH_2). The difference between LH_2 and ALH_2 resides within the 6'-group which is not considered to be the light emitting source. When ALH_2 is paired with the *Ppy* Fluc, the colour of the bioluminescent output is shifted from 555 nm to λ_{max} 605 nm (White *et al.* 1966). Study of protein function with ALH_2 has led to the development of synthetic bioluminogenic analogues, with an intent to improve substrate to protein specificity for the purposes of bioluminescent imaging *in vivo* (Adams and Miller 2014).

Figure 1.5.



Structure and λ_{max} of *D*-luciferin and a selection of analogues. Values of λ_{max} relate to measurements obtained with the *Ppy* Fluc. Aka Lumine is a 4-(dimethylamino)phenyl derivative conjugated to a thiazoline group (Iwano *et al.* 2013). Measurements of λ_{max} were taken from White *et al.* (1966) for 6-amino-*D*-luciferin and Jathoul *et al.* (2014) for remaining substrates. Structures produced using ChemSketch, available at https://www.acdlabs.com.

In recent years, a number of synthetic analogues of *D*-luciferin have been developed for different processes. The substrates shown in Figure 1.5. represent a small sample of all that have been developed to date, but even the provided selection demonstrates how limited structural variation can allow adaptations to many characteristics of the reaction including total light yields and the corresponding emission wavelength maxima (Jathoul *et al.* 2014). The ability to synthesize luciferin analogues for use in conjunction with novel engineered firefly luciferases provides the opportunity to explore new applications that were not previously achievable with native substrates. For example, of the substrate analogues detailed in Figure 1.5. pairing with an engineered variant of *Ppy* Fluc termed x5 S284T was demonstrated to shift markedly the emission spectra λ_{max} of *D*-luciferin and *DL*-infraluciferin, in the order of 50 nm to 605 nm and 706 nm, respectively. However, Aka Lumine shifted only marginally to 658 nm.

Engineering substrates to the conformity of the luciferase active site can also be performed reciprocally, to alter protein active site to accept the new analogue pairing. This has previously been done with orthogonal luciferase-luciferin pairs, to produce pairings with high substrate specificity and significantly reduced activity with non-paired analogues (Adams and Miller 2014; Jones *et al.* 2017).

1.8. Applications of Bioluminescence

Luciferases have found several applications, not only restricted to research purposes. Perhaps the most well-known application in academia is luciferase's function as a reporter gene by attaching the luciferase gene to a regulatory sequence of a gene of interest. The resulting expression of luciferase allows visualisation and study of a gene of interest using bioluminescence as a proxy for expression levels and patterns (Noguchi and Golden 2017). Commercially, Fluc is used systemically as high sensitivity ATP detection systems. Complete assay kits are available that instruct the user to take a swab of a contaminated surface. This swab will be introduced to a reagent mix containing luciferase and luciferin alongside other substances that will lyse any bacteria present. This is all performed in hand-held luminometers that will record the bioluminescent signal generated as the liberated ATP from the bacteria ignites the luciferase reaction. A high recording of bioluminescent intensity indicates increased ATP concentrations, and therefore high levels of bacteria on the surface the swab originated from (Stanley 1989; Selan *et al.* 1992; Kuzikov *et al.* 2003).

There are however two specific applications that have become key areas of interest to this project. These are how firefly luciferases can be adapted to optimise their characteristic for bioluminescence imaging in numerous biomedical applications, and for *in vitro* diagnostics, such as ATP assays or the bioluminescence assay in real time (BART).

1.8.1. Bioluminescence imaging

Similarly to its function as a reporter gene, firefly luciferase has found a novel application in the emerging field of bioluminescence imaging (BLI). BLI allows the construction of disease models that can be accurately tracked longitudinally in living animals. BLI is a noninvasive imaging method for live animals that requires only the expression of luciferase enzyme and provision of the luciferin substrate, as other components of the reaction are provided by the tissues being imaged (Adams and Miller 2014). A typical example of BLI utilization would be in a tumorigenesis study where virus containing a luciferase gene and a fluorescent protein reporter or other cell surface marker can be used to transfect a cell line for a specific cancer model. Cells that have been successfully transfected can be sorted from the rest of their population by visualising the fluorescent protein in flow cytometry. Sorted cells can then be implanted into a model organism such as mice. Implanted cells will reliably grow tumours and potentially metastasise, with all emerging cell populations expressing the transfected luciferase. Injection of the luciferin substrate into the mouse will trigger bioluminescence emission in the growing tumours, allowing visualisation and study of the disease model over time, as repeated over different time points.

BLI provides higher levels of sensitivity over related imaging technologies such as fluorescence-based technologies due to a lower background signal and the lack of requirement for an external light signal for protein excitation, which allows luciferase imaging to provide a high signal to noise ratio over fluorescence imaging with fluorescent proteins. Previously, luciferase technologies have been demonstrated capable of detecting tumour cells 1 day after cell inoculation, verses 7 days for fluorescent protein methods (Choy *et al.* 2003).

For BLI to provide its full potential in mammal tissues the luciferase system must first be engineered for high activity and thermostability and to emit bioluminescence in the ideal wavelength range of 600-800 nm, otherwise known as the bio-optical window for imaging *in vivo* (Iwano *et al.* 2013). Emission within this range prevents a large loss of signal due to absorption of visible light below 600 nm by pigmented macromolecules haemoglobin and myoglobin, which are present in the blood of mammalian tissues (Rice and Contag 2009), and to improve signal rendering due to less scatter at higher wavelengths (Rice *et al.* 2002).

1.8.2. The Bioluminescent assay in real-time

Quantification of low copy number DNA has routinely been performed in the past using quantitative PCR. This technology uses fluorescence to indicate amplification events and is dependent on sensitive cycling of heating and cooling events and complex optical hardware capable of capturing and interpreting the assay fluorescence (Gandelman *et al.* 2010). This specialist hardware requirement can restrict access to the technology, alongside further complications that can arise from PCR inhibition (Wilson 1997). Ordinarily, PCR will be affected by interaction between the inhibitors and the template DNA or polymerase enzymes. These inhibitors are often co-collected due to lack of optimization or specificity in the sample collection method, or inability to be removed using traditional DNA purification strategies (Bessetti 2007).

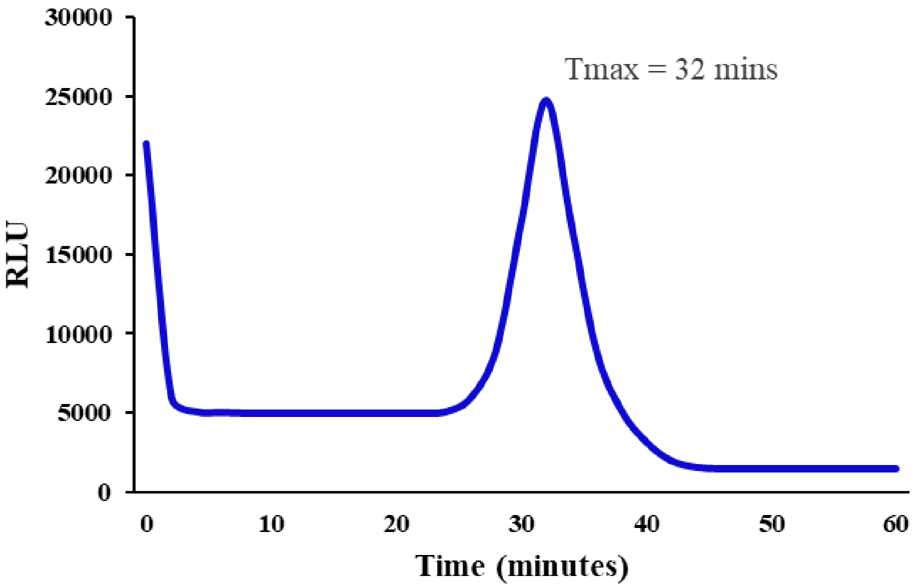
The complex cycling of PCR is not required for isothermal nucleic acid amplification technologies (iNAATs) of which there are numerous examples. The most prevalent in academic press is loop-mediated amplification (LAMP) (Notomi *et al.* 2000). LAMP is a rapid, high sensitivity amplification technology which uses four distinct primers designed to recognize six unique regions of a target DNA template. Amplification occurs at a constant temperature driven by a strand displacing DNA polymerase which generates hairpin loops of amplified DNA that act as further template to accelerate subsequent amplification (Notomi *et al.* 2000; Tomita *et al.* 2008).

A need for an economical DNA quantification method to pair with LAMP which was able to be performed with limited hardware across a diverse range of collected field samples drove the innovation of the Bioluminescent assay in real-time (BART). During DNA synthesis, an inorganic pyrophosphate (PPi) is produced. As the level of DNA synthesised increases exponentially, an equal production of PPi occurs. In these forms neither the DNA or PPi can be detected. However, in BART the production of PPi from LAMP is exploited by the addition of ATP sulfurylase which acts to convert PPi into a molecule of ATP. The newly synthesized ATP molecule can then initiate a bioluminescence reaction, provided luciferase enzyme and luciferin substrate are also present (Gandelman *et al.* 2007; Gandelman *et al.* 2010).

Unlike qPCR, LAMP-BART is performed at a fixed temperature of around 60 ˚C – 65 ˚C, requiring no specialist equipment for consistent cycling. In qPCR, fluorescent molecules such as SYBR green or alternatives (Taqman probes, molecular beacons, etc) serve as the detection method. LAMP-BART is dependent on bioluminescence in the place of fluorescence, requiring less complex optical systems to capture and integrate light. However, the luciferase must be capable of maintaining activity at elevated assay temperatures which are optimized for polymerase activity.

As the luciferase bioluminescence reaction is dependent on ATP concentration, the enzymatic conversion of PPi generated during the exponential amplification of DNA produces a unique kinetic signature (Figure 1.6.). This kinetic signature is characterised by an initial high level of bioluminescence emission at the start of the assay which rapidly decays to a baseline level due to the presence of deoxyadenosine triphosphate (dATP) which is required for DNA synthesis and serves as an alternative but less efficient substrate for luciferase. During DNA synthesis, the exponential generation of PPi produces a sharp BART peak after which the emission signal rapidly decays to almost undetectable levels as ATP sulfurylase is depleted and the activity of luciferase is inhibited by the increasing concentration of PPi. Consequently, the duration of time to reach the peak light output (T_{max}) in LAMP-BART is proportional to the original concentration of DNA in the assay, and thus can be used for quantitative calculations (Hardinge 2014).

Figure 1.6.



LAMP-BART emission profile. Representation of the bioluminescent light output from nucleic acid amplification and detection with LAMP-BART. Light intensity is plotted against the assay time, and the time to peak maximum light intensity (T_{max}) is labelled. Emission curve is for illustrative purposes only and not derived from existing emission data.

1.9. Luciferase Engineering

1.9.1. Natural variation and engineering potential

The applications of firefly luciferases are dependent upon engineering to provide improved or desirable characteristics. Engineering efforts have been performed by a number of approaches (Koksharov and Ugarova 2012), both rationally and semi-rationally to introduce predicted beneficial mutations, and more commonly in directed evolution studies where various mutagenesis methods are used to produce mutagenized libraries for characterisation. These mutations are most commonly produced and studied in the North American firefly *Photinus pyralis*, but mutations producing a certain effect can often predictably produce the same result when introduced to a luciferase derived from a distinct firefly species (Kitayama *et al.* 2003). For example, mutagenesis at position E354K in *Ppy* Fluc is conserved for thermostabilising effects in the corresponding positions of E356K in *Luciola mingrelica*

Fluc, and E354Q in *Lampyris turkestanicus* Fluc (White *et al.* 1996; Koksharov and Ugarova 2011a; Mortazavi and Hosseinkhani 2011).

However, the advantageous effects of mutations are not always conserved across all homologous Flucs. The mutation E356R in *Luciola (Hotaria) parvula* Fluc produces no significant influence on thermostability alone, but when paired with V368A confers a 12-fold increase to enzyme half-life at 45 ˚C. Further to this, *wild-type* Flucs possess diverse bioluminescence characteristics and enzyme stability properties, hence the extent to which conserved advantageous mutations confer improvements can vary significantly between enzyme variants. Therefore, natural variation can serve as an important source of novel protein templates which possess different properties that make them more suitable for use as bioluminescent tools.

Additionally, novel enzymes hold the possibility to discover uniquely advantageous mutations that have the potential to function cumulatively with the inclusion of established mutations discovered across various species in existing studies. This combination of both existing mutations and the discovery of uniquely advantageous mutations in an enzyme variant is fundamental to the development of fully optimised luciferase systems, where the enzyme target activity or property exceeds the threshold requirement of the relevant end application. Fortunately, existing protein engineering methods offer multiple strategies for the discovery and incorporation of such beneficial mutation subsets, enabling the pursuit of fully optimised enzyme variants.

1.9.2. Bioprospecting

Bioprospecting is the exploration of biodiversity for new biological resources of social and economic value. It is most commonly associated with the pharmaceutical industry, but has been a source of technological advancement for a wide variety of industries including agriculture, manufacturing, crop protection, and cosmetics, to name only a few (Beattie *et al.* 2011). From drug discovery alone, bioprospecting has arguably been a key force in the advancement of modern medicine, with examples including the discovery of the first antibiotic Penicillin from the *Penicillium* mould by Alexander Fleming in 1928, and the immune system suppressant anti-inflammatory drug Cyclosporin from the soil fungus *Tolypocladium inflatum*, which is used in the treatment of autoimmune diseases such as rheumatoid arthritis and Crohn's disease (Borel et al. 1995). By applying this same rational to search for novel luciferase gene sequences in fireflies, it may be possible to discover luciferase enzymes which possess advantageous properties for the purpose of bioluminescence applications.

Whilst bioprospecting has renewed interest in protecting biodiversity and enabled the discovery of a wide variety of products including chemicals, proteins, genes, and complete metabolic pathways, it is not without criticism. The term biopiracy describes an instance where a region's biological resources or associated traditional knowledge are appropriated or exploited for commercial interests. To mitigate these concerns, a legal framework was devised called the Nagoya Protocol on Access and Benefit Sharing (ABS), which aims to enable the fair and equitable sharing of benefits arising out of the utilization of genetic resources, and ultimately contribute to the conservation and sustainable use of biodiversity (https://www.gov.uk/guidance/abs). The implications and compliance of this project to the Nagoya protocol are discussed further in Chapter 3.

1.9.3. Protein engineering strategies

1.9.3.1 Directed evolution

Across decades of work to advance the understanding of protein structure and function, few methods have been as ubiquitously utilised as directed evolution. Directed evolution mimics the principles of Darwinian evolution with a two-step process of genetic diversification and screening under a user-defined selection pressure in order to achieve biological entities with desired traits (Cobb *et al.* 2013). The source and extent of genetic diversification can be varied, but in its origin and still most commonly, PCR-driven random mutagenesis is performed to acquire randomly diversified libraries from a gene of interest. Library members which possess improvements in the desired phenotype can then be identified by heterogeneous expression in transformed cells and subsequent high-throughput screening or selection (Lutz 2010). Firefly luciferases are particularly amenable to directed evolution as their primary characteristic of bioluminescence is readily recorded with modern imaging devices, and offers a convenient indication of the proteins ability to function under a particular selection pressure.

However, whilst directed evolution is capable of generating random molecular diversity, herein lies the problem. In a theoretical protein sequence of only ten residues, there exists in excess of $1x10^{13}$ combinations possible from the twenty natural amino acids, which is further complicated by the unbalanced degenerate nature of amino acid codons. Therefore, screening of randomized protein libraries can only ever sample a minuscule fraction of the complete sequence space possible (Wong *et al.* 2006). For this reason, much modern day protein engineering has moved beyond the 'broad strokes' of directed evolution, and in its place explored strategies to generate targeted diversity based on hypotheses of structure-function relationships, which in theory contributes to smaller, higher quality libraries (Lutz 2010).

1.9.3.2 Rational design

In protein engineering, rational design is the generation of enzymes with a desired functionality using available structural and functional data to predict how alterations to protein structure will affect the corresponding behaviour. As of 2022, the RCSB Protein Data Bank has in excess of 187,844 atomic-resolution structures of proteins (https://www.rcsb.org/). With the increasing availability of protein structural data, biophysical information, protein function, sequence-based data, and corresponding *in silico* analysis tools, the rational design approach to engineering has become increasingly common in efforts to improve the biochemical properties of enzymes, including the kinetic behaviours, thermostability, organic solvent tolerance, and substrate specificity (Pongsupasa *et al.* 2022).

Whilst rational design can often involve the substitution of individual amino acids or motifs with a user-defined specific selection, a strategy termed semi-rational design combines the approaches of directed evolution and rational design. Commonly, semi-rational design involves site saturation mutagenesis, where a single codon or set of codons are substituted to encode all possible twenty amino acids at that position, and subsequently a screening library is constructed to enable the assessment of effects conferred from all possible amino acids, such that the variant with the preferred functionality can be selected (Georgescu *et al.* 2003).

1.9.4. Engineering for thermostability

A significant limitation of *wild-type* Flucs is that they are often highly thermolabile and can undergo near-total inactivation even at room temperature, which negates much of their value in bioluminescence applications due to the difficulty in reproducing results (Tisi *et al.* 2002a; Baggett *et al.* 2004). As this necessitates luciferases with increased resistance to thermal inactivation, protein engineering has routinely been explored to produce thermostable variants. For example, through mutagenesis in *Ppy* Fluc, substitution of the position E354 with either lysine or arginine was found to confer significant improvements to thermostability (White *et al.* 1996). Later studies of this enzyme region identified a flexible loop structure of the N-terminal domain (residues T352 to F368) as an omega loop within which mutations can produce an array of effects including changes to the emitted light spectrum, emission kinetics, and crucially thermostability (Willey *et al.* 2001; Halliwell *et al.* 2018). Further investigation of the omega loop's involvement in overall protein thermostability suggested that the mutation E356R acts to improve stability within the local loop structure of *Ppy* Fluc by reducing the disorder in the region, which in turn restricts flexibility and regulates the structural integrity of the polypeptide chain (Moradi *et al.* 2009). Restriction of flexibility is considered a large factor affecting the stability of any firefly luciferase protein.

Based on observations of higher arginine frequencies occurring in thermostable proteins, a correlation has been demonstrated between substituting arginine residues for hydrophobic solvent-exposed residues of engineered firefly luciferases and the resulting overall thermostability (Mortazavi and Hosseinkhani 2011). The introduction of disulphide bridges to luciferase proteins can also improve thermal stability by increasing active site rigidity, as well as inducing a colour shift to red (Nazari and Hosseinkhani 2011). More recently, a highly thermostable Fluc was constructed using a SpyTag-SpyCatcher dual system that fuses the N-domain to the C-domain through an irreversible covalent bond to create a circular Fluc with no loss of functional efficiency (Si *et al.* 2016).

The greatest improvement to thermal stability are made by employing all known methods for improvements and generating engineered firefly luciferases with the cumulative effect of multiple mutations. The current frontrunner in this technology is the Ultra-Glo protein (UG) (Promega Corp., Madison, WI, USA) that has been proven to be stable at 65 ˚C for more than 5 hours and contains up to 27 separate mutations (Hall *et al.* 1999; Jathoul 2008).

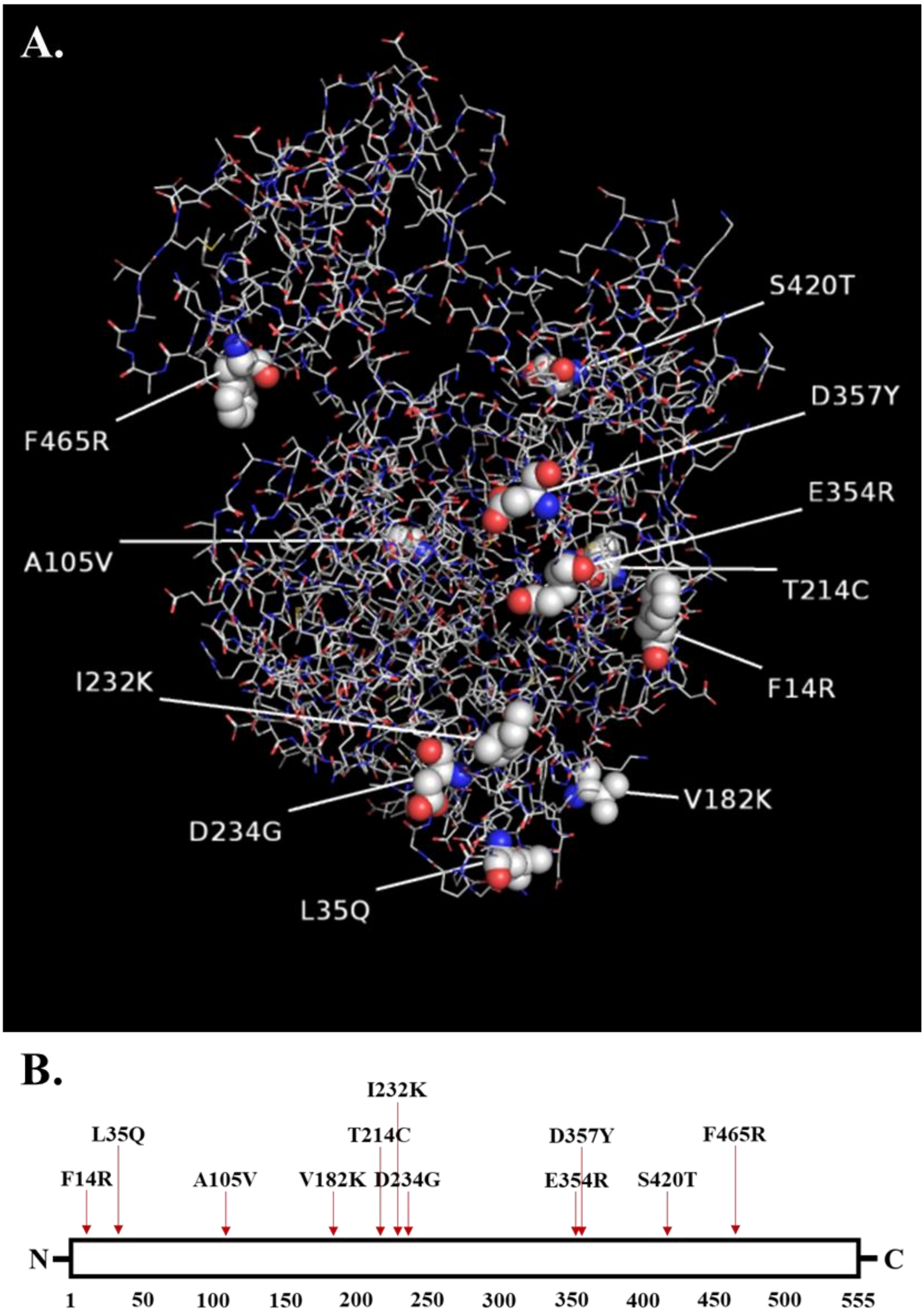
Other non-commercial thermostable firefly luciferases exist which display thermostability coupled with high bioluminescence intensity, such as the x11 Fluc (Jathoul *et al.* 2012).

Table 1.1.

x11 Fluc		
Mutation	**Location**	**Contribution to Thermostability**
F14R	Surface	Reduces the amount of buried hydrophilic surface (Prebble et al. 2001).
L35Q	Internal	Reduces the amount of buried hydrophilic surface (Prebble et al. 2001).
A105V	Surface	Reduces the amount of buried hydrophilic surface (Prebble et al. 2001).
V182K	Surface	Substitutes to a more hydrophilic solvent-exposed residue (Law et al. 2006).
T214C	Internal	Reduces the amount of buried hydrophilic surface (Prebble et al. 2001).
I232K	Surface	Reduces the amount of buried hydrophilic surface (Prebble et al. 2001).
D234G	Surface	Reduces the amount of buried hydrophilic surface (Prebble et al. 2001).
E354R	Surface	Loss of negative charge at this position/region (White et al. 1996).
D357Y	Surface	Substitutes to a more hydrophilic solvent-exposed residue (Kim-Choi et al. 2006).
S420T	Surface	Reduces the amount of buried hydrophilic surface (Prebble et al. 2001).
F465R	Surface	Substitutes to a more hydrophilic solvent-exposed residue (Law et al. 2006).

The eleven mutations of the x11 Fluc. Mutations are listed alongside their relative positions in the protein structure. The hypothesized contribution to thermostability for each mutation is provided, as stated in each corresponding publication. Location information taken from (Jathoul *et al.* 2012).

Figure 1.7



Positions of x11 mutations in *Ppy* Fluc. (**A**) 3D stick model of *Ppy* Fluc with the positions mutated in x11 represented as spheres and labelled. Model produced using PyMOL. (**B**) Linear block diagram indicating the relative locations of x11 mutations in *Ppy* Fluc.

x11 Fluc is an engineered derivative of the luciferase from the North American firefly *Photinus pyralis*, containing the eleven mutations detailed in Table 1.1. The position of these eleven mutations in the *Ppy* Fluc can be visualised in Figure 1.7. The wavelength of emission peaks in the region of 550 – 560 nm, and the enzyme displays significant stability against environmental pressures including elevated temperature and fluctuations of pH. Of the eleven total mutations, nine are located on the surface of the protein, making x11 a good example demonstrating that surface mutations can cumulatively improve the stability of Fluc. An example of a solely surface engineered firefly luciferase exists in the x5 Fluc which contains a subset of five solvent exposed mutant sites included in x11, which cumulate to an increased resistance to thermal inactivation and improved pH-tolerance (Law *et al.* 2006).

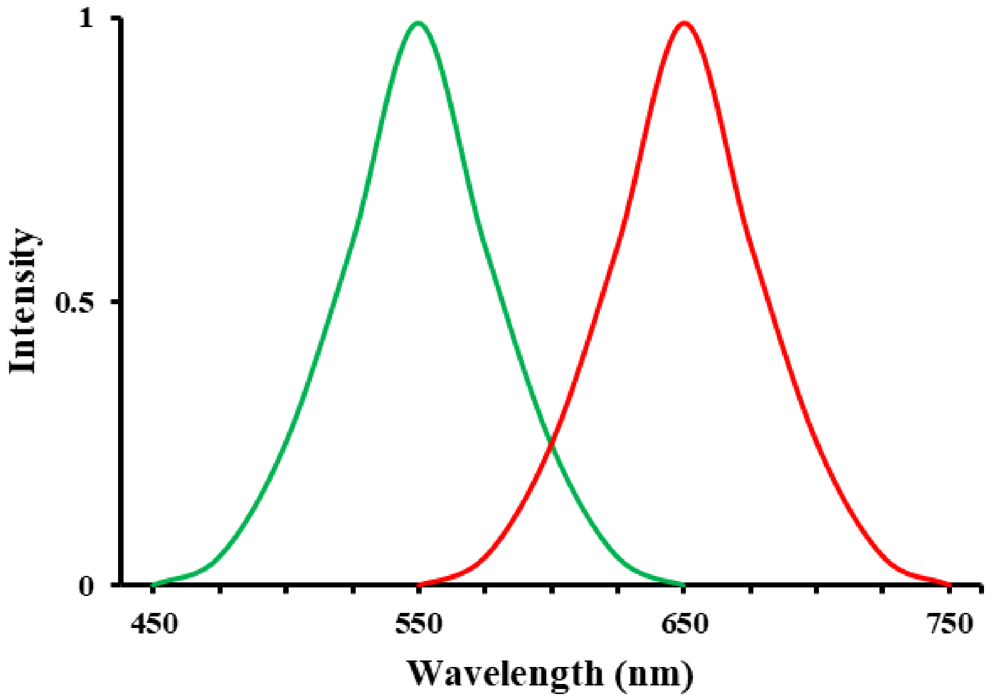
An often overlooked method to produce effective and entirely novel mutants is deletion mutagenesis. A recent study explored the effect of consecutive single amino acid deletions in the x11 Fluc. Six loop structures were originally explored, out of which an omega-loop structure was identified (residues T352-G360) in which single consecutive deletion mutants exhibited properties including significantly enhanced activity with *D*-luciferin, improved resistance to thermal inactivation, and altered substrate specificity for red-shifted substrate analogue *DL*-infraluciferin (Halliwell *et al.* 2018). These thermostable x11 Fluc variants were designed to be applied as bioreporters in BLI.

1.9.5. Engineering for BLI using infraluciferin

Contemporary efforts to advance the development of BLI technologies have shifted focus from engineered variants of luciferase to investigate the potential of synthetic substrate analogues which provide more desirable reaction properties. In recent years, the first dual-colour, far-red to near-infrared emitting analogue of the Coleoptera beetle luciferin has been described (Jathoul *et al.* 2014). This near-infrared emitting luciferin (infraluciferin - iLH_2) is produced by chemical synthesis from commercially available precursors to efficiently yield far-red to near-infrared luciferins (Anderson *et al.* 2017). The distinctive advantage of iLH_2 for the purposes of bioluminescence imaging is the emission maxima generated, which extend up to λ_{max} = 670 nm with *Ppy* Fluc, and λ_{max} = 706 nm with engineered variants. This potential for significantly red-shifted emission spectra alongside the capability to emit different colours of bioluminescence when catalysed by different engineered *Photinus* *pyralis* derived luciferases is crucial for simultaneous imaging of more than one target in haemoglobinised tissues or environments.

As BLI technologies continue to mature, dual reporter multicolour luciferase assays have become increasingly common as they reduce the experiment variability and provide more information than single reporter approaches (Nakajima *et al.* 2005) In such an assay a red emitting firefly luciferase tagged to a gene or cell of interest can be paired with a contrasting green emitting luciferase which acts as an internal control. When both enzymes are catalysing the same luciferin substrate, if the emission peak of the green emitter is separated from the emission peak of the red-shifted emitter sufficiently to reduce the level of overlap between the two emission spectra (Figure 1.8.), the two bioluminescence signals from a single experiment can be distinguished between with filtered emission acquisitions.

Figure 1.8.



Dual-parameter imaging with Coleopteran-based luciferases. Representation of two distinct red and green emitting luciferase bioluminescence spectra signals. Separation of the emission peaks is sufficient that filtered bioluminescence acquisitions could distinguish between the two signals. Emission curves are for illustrative purposes only and not derived from existing emission data.

As previously discussed, imaging of mammalian tissues benefits from red-shifted emission spectra which fall in the 600 nm – 800 nm bio-optical window. For this reason, much of the work to develop synthetic luciferin analogues has focused on developing substrates which red-shift the emission of all firefly luciferase pairings beyond 600 nm. The previous best attempt at producing a near-infrared substrate analogue achieved up to λ_{max} = 675 nm but possessed no capability of emitting different wavelengths of bioluminescence when catalysed by different engineered firefly luciferase variants (Iwano *et al.* 2013; Jathoul *et al.* 2014). Where the design of iLH_2 differed to the synthesis of previous red-shifted luciferins was the vital retention of the 6'-hydroxy group which has previously been shown to be important for colour modulation (White *et al.* 1966). Therefore, this synthetic luciferin analogue was uniquely constructed with careful consideration of how the differing active site microenvironments of disparate firefly luciferase variants can be crucial for multiparametric imaging due to their influence on the bioluminescence emission wavelength.

1.10. Aims and Objectives

The advancement of bioluminescence technologies is dependent on the acquisition of novel and improved enzyme functionality. In recent years, this has been a multifaceted effort ranging from the discovery of novel firefly luciferases from nature, protein engineering for desired functionality, or the development of synthetic substrate analogue bioluminescence systems. The work undertaken here explored these three areas of bioluminescence discovery, and aimed to demonstrate the acquisition of novel functionality using a combination of prevalent and innovative strategical approaches.

With the intention of demonstrating how existing biodiversity can remain undisturbed in favour of the wealth of biological resources which are available within compliance to the Nagoya protocol, isolation of novel firefly luciferase gene sequences was attempted by bioprospecting dry-preserved Coleoptera from the collections of Amgueddfa Cymru – National Museum Wales or the California Academy of Sciences in San Francisco, via the National Museum Wales.

In parallel to bioprospecting, the previously unreported luciferase from the lesser British Glow-worm *Phosphaenus hemipterus* was engineered in two discrete ventures to develop a thermostable enzyme variant using directed evolution, and the investigation of a rational design strategy to generate a variant possessing improved activity with the synthetic substrate analogue infraluciferin. The uncharacterised state of this luciferase gene and enzyme presented the unique opportunity to discover novel mutagenic functionality which may ultimately contribute to the pool of known conserved mutagenic targets across related enzyme variants.

The aims of this work is therefore:

1. To explore the feasibility of cloning novel luciferase genes from dried museum specimens using non-destructive sampling coupled with next generation sequencing.
2. To clone and characterise the luciferase from the Lesser British Glow-worm *Phosphaenus hemipterus*.
3. To engineer the luciferase from *Phosphaenus hemipterus* for improved bioluminescence activity with the synthetic substrate analogue infraluciferin.
4. To engineer the luciferase from *Phosphaenus hemipterus* for significantly improved thermostability.

Chapter 2

Materials and Methods

2.1. Materials

2.1.1. Chemicals

Milli-Q Ultrapure water was used throughout as the source of distilled water (dH_2O) for molecular biology methods and reagent and buffer preparation. Unless otherwise stated, all general chemicals for reagent and buffer preparation were obtained from Sigma-Aldrich (MA, USA). Stock solution of sterile Carbenicillin (Melford, Suffolk, UK) were prepared as 100 mg/ml solutions and filter sterilized with a 0.22 µM filter unit (Thermo Fisher Scientific, MA, USA) and stored as 500 µl aliquots at -20 ˚C. Stock solutions of sterile isopropyl β-D-thiogalactopyranoside (IPTG) (Melford, Suffolk, UK) were prepared as 1 M solutions and stored as 500 µl aliquots at -20 ˚C. Stock solutions of ATP (Roche Diagnostics, IN, USA) were prepared to 100 mM (pH 7.8) and stored as 100 µl aliquots at -20 ˚C. *D*-luciferin (LH_2) potassium salt was obtained from Regis Technologies, Inc. (IL, USA), and prepared as 10 mg/ml stock solutions. *DL*-Infraluciferin (iLH_2) was provided by Dr Amit Jathoul, and obtained from Jim Anderson's group as University College London. Due to the unknown rate of degradation of iLH_2 in solution, 5 mg/ml stocks were prepared fresh from powder as required. All luciferin powders and prepared stocks were stored in light-proof containers at -20 ˚C until use. Luciferin and ATP stocks were prepared in pH 7.8 TEM buffer (see Buffers).

2.1.2. Buffers

Where buffers required adjustment to pH, buffers were initially made up to ≈80% final volume, and pH appropriately adjusted. Once the desired pH had been reached, buffers would be topped up to final volumes. Where provided, percentages and molar concentrations relate to the final concentration in the buffer. Saline-Sodium Citrate buffer (SSC) was purchased as a 20x concentrate from Sigma-Aldrich (MA, USA) and used in the composition of further buffers.

2.1.2.1 General buffers

10x TEM

Tris-acetate	1 M
Ethylenediaminetetraacetic acid (EDTA)	20 mM
Magnesium sulphate ($MgSO_4$)	100 mM
dH_2O	up to final volume

adjusted to pH 7.8 with NaOH and acetic acid

50x TAE electrophoresis buffer

Tris	2 M
Acetic acid	1 M
EDTA	50 mM
dH_2O	up to final volume

1 M Tris-HCl Buffer (100 ml)

Tris-HCl	12.11 g
dH_2O	Up to 100 ml

adjusted to pH 8 with hydrochloric acid (HCL)

0.5 M EDTA pH 8 (100 ml)

EDTA disodium salt, dihydrate	18.61 g
dH_2O	up to 100 ml

adjusted to pH 8 with NaOH

0.1 M Sodium Citrate (1L)

Sodium citrate dehydrate	12.5 g
Citric acid	11.3 g
TEM	800 ml
dH_2O	up to 1L

adjusted to pH 5 with NaOH and citric acid

10x DNase1 Buffer

Tris-HCl buffer pH 7.5	500 mM
$MnCl_2$	100 mM

2.1.2.2 Bioprospecting buffers

Non-Destructive DNA Extraction Buffer

$CaCl_2$	3 mM
Sodium dodecyl sulphate (SDS)	2% (w/v)
Dithiothreitol (DTT)	40 mM
Proteinase K	250 µg/ml
Tris-HCl buffer pH 8	100 mM
NaCl	100 mM
dH_2O	up to final volume

100x Denhardt's Solution

Bovine serum albumin (Fraction V)	2% (w/v)
Ficoll 400	2% (w/v)
Polyvinylpyrrolidone (PVP)	2% (w/v)
dH_2O	up to final volume

2x Hybridisation Solution

NaCl	1.5 mM
Sodium phosphate buffer pH 7.2	40 mM
EDTA pH 8.0	10 mM
100x Denhardt’s solution	10% (v/v)
SDS	0.2% (w/v)
dH_2O	up to final volume

TEN Buffer

Tris-HCl pH 7.5	10 mM
EDTA	1 mM
NaCl	1 M
dH_2O	up to final volume

Low-Stringency Wash Buffer

SSC buffer 20x concentrate	1x
SDS	0.1% (w/v)
dH_2O	up to final volume

High-Stringency Wash Buffer

SSC buffer 20x concentrate	0.1x
SDS	0.1% (w/v)
dH_2O	up to final volume

2.1.2.3 Protein purification buffers

Storage Buffer (200 ml)

10 x TEM	20 ml
50% glycerol	40 ml
H_2O	up to 200 ml
1 M dithiothreitol (DTT) solution	400 µl

adjusted to pH 7.8 with NaOH and acetic acid prior to addition of DTT

Buffer A (400 ml)

Phosphate buffered saline (PBS) tablets (Thermo Fisher Scientific, MA, USA)	2 tablets
NaCl	3.8 g
50% glycerol	160 ml
H_2O	up to 400 ml

adjusted to pH 8.0 with NaOH and acetic acid

4 M IMD (5 ml)

Imidazole (IMD)	1.36 g
Buffer A	5 ml

adjusted to pH 8.0 with H_3PO_4

Lysis Buffer (10 ml)

Buffer A	10 ml
4 M IMD	500 µl
25 x EDTA-free protease inhibitor (PI) *(1 EDTA-free PI tablet* dissolved in 2 ml H_2O)*	400 µl
adjusted to pH 8.0 with NaOH and acetic acid	
1 M 2-Mercaptoethanol (β-ME) *(added to chilled solution)*	100 µl
Resuspend bacteria pellet in 10 ml	
Triton X-100	100 µl in 10 ml resuspension
100 mg/ml Lysozyme	100 µl in 10 ml resuspension
Benzonase 250 units/µl	10 µl in 10 ml resuspension

*Roche Diagnostics IN, USA

IMD solutions

Constituents	**IMD concentration (mM)**				
	20	**50**	**200**	**300**	**500**
Buffer A	10 ml	10 ml	15 ml	15 ml	15 ml
4 M IMD	50 µl	125 µl	750 µl	1125 µl	1875 µl
1 M β-ME *(added to chilled solution)*	100 µl	100 µl	150 µl	150 µl	150 µl
adjusted to pH 8.0 with NaOH and acetic acid prior to the addition of β -ME					

2.1.2.4 Reagents for SDS-PAGE gels and Coomassie staining

10x SDS-PAGE Running Buffer

Tris-HCl	250 mM
Glycine	1.92 M
SDS	1% (w/v)
dH_2O	up to final volume

2x Laemmli Loading Buffer

Bromophenol blue	0.004% (w/v)
2-Mercaptoethanol (β-ME)	10% (v/v)
Glycerol	20% (v/v)
SDS	4% (w/v)
Tris-HCl	125 mM
dH_2O	up to final volume

Methanol Fixer Solution (1L)

Methanol	400 ml
Acetic acid	100 ml
dH_2O	500 ml

Coomassie Staining Solution (1L)

Coomassie Blue R-250	2.5 g
Methanol	400 ml
Acetic acid	100 ml
dH_2O	500 ml

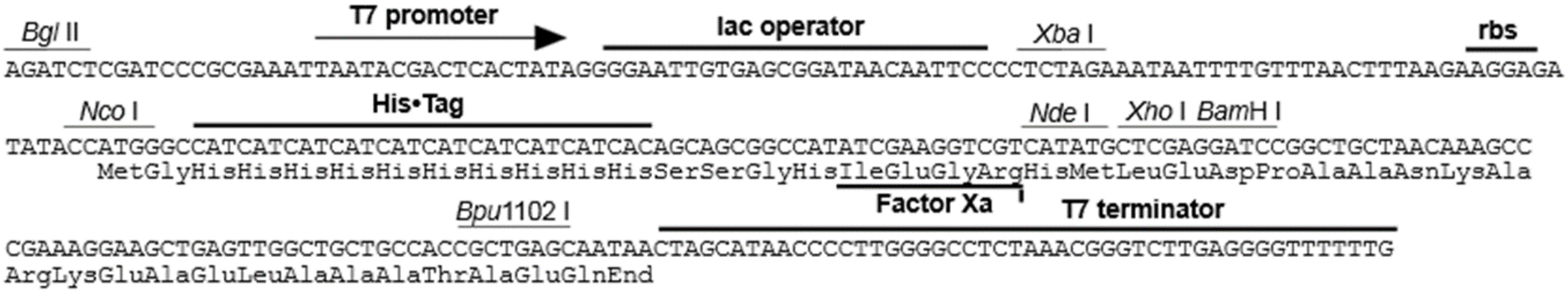
Destain Solution (1L)

Methanol	50 ml
Acetic acid	75 ml
dH_2O	875 ml

2.1.3. Bacterial cell strains and plasmids

Glycerol stocks of *Escherichia coli* BL21 (DE3) carrying the pET16b vector encoding *wild-type Photinus pyralis* and the thermostable x11 Fluc gene were provided by Dr Amit Jathoul (Cardiff University, UK). Further luciferase constructs were provided by Dr Jathoul as purified plasmid of CBR and ELuc both in pET16b. The pET16b plasmid encodes ampicillin resistance as a selection marker, with gene expression under the control of the T7 promoter. Expression of T7 RNA polymerase (and hence the T7 promoter) is induced by IPTG in an appropriate *E. coli* strain carrying the T7 RNA polymerase construct such as BL21 (DE3). Details of the cloning region in pET16B are available in Figure 2.1. NEB® 5-alpha Competent *E. coli* (High Efficiency) were obtained from New England BioLabs Inc. (MA, USA) and BL21 (DE3) pLysS Competent Cells for protein expression were obtained from Promega (WI. USA).

Figure 2.1.


Bgl II
T7 promoter
lac operator
Xba I
rbs
AGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGA
Nco I
His•Tag
Nde I
Xho I
BamH I
TATACCATGGGCCATCATCATCATCATCATCATCATCATCACAGCAGCGGCCATATCGAAGGTCGTCATATGCTCGAGGATCCGGCTGCTAACAAAGCC
MetGlyHisHisHisHisHisHisHisHisHisHisSerSerGlyHisIleGluGlyArgHisMetLeuGluAspProAlaAlaAsnLysAla
Factor Xa
Bpu1102 I
T7 terminator
CGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTG
ArgLysGluAlaGluLeuAlaAlaAlaThrAlaGluGlnEnd

Map of cloning expression region in pET16b. Modified from Novagen pET16b manual (EMD Millipore Corporation, Darmstadt, Germany), indicating the position and nucleotide sequences of the T7 promoter, lac operator, rbs, the N- terminal 10x His tag and T7 terminator. Restriction sites within this region are further indicated, and include *BglII, XbaI, NcoI, NdeI, XhoI, BamHI, Bpu1102I.*

2.1.4. Bacterial growth media

Lysogeny Broth (LB) (Melford, Suffolk, UK) was prepared by dissolving 20 g/l in dH_2O. LB Agar was prepared by dissolving 35 g/l in dH_2O. SOC Outgrowth Medium was obtained from New England BioLabs Inc. (MA, USA). Prepared LB medias were sterilized by autoclave at 121 ˚C, and supplemented with 100 µg/ml carbenicillin immediately prior to use.

2.1.5. Reagents

2.1.5.1 Molecular reagents

Restriction enzymes and their complementary buffers were obtained from NEB using high-fidelity enzymes where available, most commonly including *NdeI* and *BamHI* in CutSmart buffer. Agarose powder for gel electrophoresis was obtained from Bioline (London, UK) and SafeView Nucleic Acid Stain from NBS Biologicals (Cambridgeshire, UK). SmartLadder molecular weight marker was obtained from Eurogentec (Seraing, Belgium). DNA Gel Loading Dye, Purple (6x) was additionally sourced from NEB. *Taq* PCR Master Mix was obtained from Qiagen (Venlo, Netherlands) in addition to the Qiagen dNTP Set and dNTP Mix, which was used to prepare 10 mM stock solutions of deoxyribonucleotide triphosphates (dNTPs). SYBR Green JumpStart™ Taq ReadyMix™ for qPCR was obtained from Sigma-Aldrich (MA, USA).

2.1.5.2 Protein work reagents

Nylon Hybond-N membrane was obtained from GE Life Sciences (PA, USA). HisPur Ni-NTA Resin was obtained from Thermo Fisher Scientific (MA, USA) and disposable PD-10 desalting columns from GE Healthcare (WI, USA). Prestained protein markers for SDS-PAGE were obtained from Fisher Scientific (MA, USA).

2.1.6. Oligonucleotide primers

Oligonucleotide primers were obtained from Integrated DNA Technologies (IA, USA) and Sigma-Aldrich (MA, USA) as lyophilised stocks to be reconstituted in dH_2O to 100 µM. Primary stocks were diluted into 10 µM working stocks to reduce degradation from freeze-thaw events, and both stored at -20 ˚C. Sanger sequencing of constructs within the pET16b vector was performed using T7 primers provided as part of the Eurofins sequencing service (Luxembourg, Luxembourg).

Terminal primers detailed in Table 2.1. were designed using Primer3, available at https://primer3.org/ (Untergasser *et al.* 2012).

A primer set designed by Zeale *et al* (2011) to amplify 'mini-barcode' regions of the mitochondrial cytochrome oxidase subunit 1 (COI) gene is detailed in Table 2.2.

A luciferase gene short fragment targeting degenerate primer set was designed using the CODEHOP method (Rose *et al.* 1998; Rose *et al.* 2003; Boyce *et al.* 2009). CODEHOP is an acronym of Consensus Degenerate Hybrid Oligonucleotide Primers, which are a pool of large degenerate primers constructed from conserved blocks of amino acid sequence alignments. The luciferase gene sequences used in the CODEHOP design are available in Appendices Table 9.2., and the details of the primer pair can be found in Table 2.3.

Thirty-two site-directed mutagenesis primer pairs were designed for mutagenesis targets using the QuikChange® Primer Design Program (Agilent, CA, USA), and are detailed along with their target in Table 2.4.

Six LAMP-BART primers were provided by Dr Patrick Hardinge targeting the N2 gene region of the SARS-CoV-2 RNA sequence (Zhang *et al.* 2020b), and are detailed in Table 2.5.

Table 2.1.

Name	**Orientation**	**Primer Sequence (5’-3’)**
*Phem*Luc_F	Forward	GCTTGGGTCGTCATATGGAAG
*Phem*Luc_R	Reverse	GCACGACCCAAGGATCCTTA
*Ppy*_F	Forward	AGGTCGTCATATGGAAGACGCCAAAA
*Ppy*_R	Reverse	GCAGCCGGATCCAGTTACATTTTACA

Sequences of terminal primers used. Primer names and orientation are listed alongside the corresponding sequence.

Table 2.2.

Name	**Orientation**	**Primer Sequence (5’-3’)**
ZBJ-ArtF1c	Forward	AGATATTGGAACWTTATATTTTATTTTTGG
ZBJ-ArtR2c	Reverse	WACTAATCAATTWCCAAATCCTCC

Sequences of ‘mini-barcode’ primers. Primer names and orientation are listed alongside the corresponding sequence.

Table 2.3.

Name	**Orientation**	**Primer Sequence (5’-3’)**
DKYD-F 8x	Forward	CTTCTTCGCCAAGTCCACGCTGGTCGAYAARTAYGA
GYG-R 32x	Reverse	TGATAGCGGAGGTGGTCTCGGTCAGNCCRTANCC

Sequences of CODEHOP primers. Primer names and orientation are listed alongside the corresponding sequence.

Table 2.4.

Target	**Name**	**Primer Sequence (5'-3')**
R218	Phem R218X F	AT ACG TCA GTG TGT GTT **NNK** TTT AGC CAC TGT CGC G
	Phem R218X R	C GCG ACA GTG GCT AAA **MNN** AAC ACA CAC TGA CGT AT
H245	Phem H245X F	A TCC GTC ATT CCG TTC CAC **NNK** GGT TTT GGA ATG TTC ACA
	Phem H245X R	TGT GAA CAT TCC AAA ACC **MNN** GTG GAA CGG AAT GAC GGA T
G246	Phem G246X F	TC ATT CCG TTC CAC CAT **NNK** TTT GGA ATG TTC ACA AC
	Phem G246X R	GT TGT GAA CAT TCC AAA **MNN** ATG GTG GAA CGG AAT GA
F247X	Phem F247X F	C GTC ATT CCG TTC CAC CAT GGT **NNK** GGA ATG TTC ACA ACA T
	Phem F247X R	A TGT TGT GAA CAT TCC **MNN** ACC ATG GTG GAA CGG AAT GAC G
T251	Phem T251X F	CAC CAT GGT TTT GGA ATG TTC **NNK** ACA TTG GGC TAC TTA ATT T
	Phem T251X R	A AAT TAA GTA GCC CAA TGT **MNN** GAA CAT TCC AAA ACC ATG GTG
E311	Phem E311X F	C GAT CTT AGT AAT TTA CAC **NNK** ATC GCT TCG GGC GGT
	Phem E311X R	ACC GCC CGA AGC GAT **MNN** GTG TAA ATT ACT AAG ATC G
A313	Phem A313X F	GAT CTT AGT AAT TTA CAC GAA ATC **NNK** TCG GGC GGT GCA CC

	Phem A313X R	GG TGC ACC GCC CGA **MNN** GAT TTC GTG TAA ATT ACT AAG ATC
S314	Phem S314X F	TT AGT AAT TTA CAC GAA ATC GCT **NNK** GGC GGT GCA C
	Phem S314X R	G TGC ACC GCC **MNN** AGC GAT TTC GTG TAA ATT ACT AA
G315	Phem G315X F	C GAA ATC GCT TCG **NNK** GGT GCA CCA CTT G
	Phem G315X R	C AAG TGG TGC ACC **MNN** CGAA GCG ATT TCG
G316	Phem G316X F	TC GCT TCG GGC **NNK** GCA CCA CTT GC
	Phem G316X R	GC AAG TGG TGC **MNN** GCC CGA AGC GA
A317	Phem A317X F	C GAA ATC GCT TCG GGC GGT **NNK** CCA CTT GCA AAG GAA GT
	Phem A317X R	AC TTC CTT TGC AAG TGG **MNN** ACC GCC CGA AGC GAT TTC G
P318	Phem P318X F	CT TCG GGC GGT GCA **NNK** CTT GCA AAG GAA
	Phem P318X R	TTC CTT TGC AAG **MNN** TGC ACC GCC CGA AG
L319	Phem L319X F	GCT TCG GGC GGT GCA CCA **NNK** GCA AAG GAA GTG
	Phem L319X R	CAC TTC CTT TGC **MNN** TGG TGC ACC GCC CGA AGC
R337	Phem R337X F	C AAC CTT CGC GGC ATT **NNK** CAA GGG TAC GGG
	Phem R337X R	CCC GTA CCC TTG **MNN** AAT GCC GCG AAG GTT G

Q338	Phem Q338X F	CTT CGC GGC ATT CGC **NNK** GGG TAC GGG TTG AC
	Phem Q338X R	GT CAA CCC GTA CCC **MNN** GCG AAT GCC GCG AAG
G339	Phem G339X F	C GGC ATT CGC CAA **NNK** TAC GGG TTG ACT G
	Phem G339X R	C AGT CAA CCC GTA **MNN** TTG GCG AAT GCC G
Y340	Phem Y340X F	GGC ATT CGC CAA GGG **NNK** GGG TTG ACT GAG AC
	Phem Y340X R	GT CTC AGT CAA CCC **MNN** CCC TTG GCG AAT GCC
G341	Phem G341X F	TT CGC CAA GGG TAC **NNK** TTG ACT GAG ACT AC
	Phem G341X R	GT AGT CTC AGT CAA **MNN** GTA CCC TTG GCG AA
L342	Phem L342X F	CGC CAA GGG TAC GGG **NNK** ACT GAG ACT ACG TC
	Phem L342X R	GA CGT AGT CTC AGT **MNN** CCC GTA CCC TTG GCG
T343	Phem T343X F	CAA GGG TAC GGG TTG **NNK** GAG ACT ACG TCT G
	Phem T343X R	C AGA CGT AGT CTC **MNN** CAA CCC GTA CCC TTG
E344	Phem E344X F	GG TAC GGG TTG ACT **NNK** ACT ACG TCT GCA GT
	Phem E344X R	AC TGC AGA CGT AGT **MNN** AGT CAA CCC GTA CC
T346	Phem T346X F	GGG TAC GGG TTG ACT GAG ACT **NNK** TCT GCA GTT

	Phem T346X R	AAC TGC AGA **MNN** AGT CTC AGT CAA CCC GTA CCC
S347	Phem S347X F	GGG TTG ACT GAG ACT ACG **NNK** GCA GTT ATT ATT ACA C
	Phem S347X R	G TGT AAT AAT AAC TGC **MNN** CGT AGT CTC AGT CAA CCC
A348	Phem A348X F	C GGG TTG ACT GAG ACT ACG TCT **NNK** GTT ATT ATT ACA CCT GAA GGA G
	Phem A348X R	C TCC TTC AGG TGT AAT AAT AAC **MNN** AGA CGT AGT CTC AGT CAA CCC G
V362	Phem V362X F	AT AAG CCT GGC GCT **NNK** GGA AAA GTT GTG CC
	Phem V362X R	GG CAC AAC TTT TCC **MNN** AGC GCC AGG CTT AT
S420	Phem S420X F	T AAA GAT GGA TGG TTG CAC **NNK** GGC GAT ATT AGC TAC TGG
	Phem S420X R	CCA GTA GCT AAT ATC GCC **MNN** GTG CAA CCA TCC ATC TTT A
D422	Phem D422X F	A TGG TTG CAC AGT GGC **NNK** ATT AGC TAC TGG GAT G
	Phem D422X R	C ATC CCA GTA GCT AAT **MNN** GCC ACT GTG CAA CCA T
I434	Phem I434X F	C TGG GAT GAG GAC GGA CAT TTT TTT **NNK** GTC GAT CGT CTT
	Phem I434X R	AAG ACG ATC GAC **MNN** AAA AAA ATG TCC GTC CTC ATC CCA G
R437	Phem R437X F	GAG GAC GGA CAT TTT TTT ATC GTC GAT **NNK** CTT AAG TCC TTA ATC AAA T
	Phem R437X R	A TTT GAT TAA GGA CTT AAG **MNN** ATC GAC GAT AAA AAA ATG TCC GTC CTC

L526	Phem L526X F	AC GAG GTT CCG AAA GGA **NNK** ACG GGC AAA CTT GAC G
	Phem L526X R	C GTC AAG TTT GCC CGT **MNN** TCC TTT CGG AAC CTC GT
T527	Phem T527X F	C GAG GTT CCG AAA GGA TTG **NNK** GGC AAA CTT G
	Phem T527X R	C AAG TTT GCC **MNN** CAA TCC TTT CGG AAC CTC G
K529	Phem K529X F	G AAA GGA TTG ACG GGC **NNK** CTT GAC GCC CGC AAG
	Phem K529X R	CTT GCG GGC GTC AAG **MNN** GCC CGT CAA TCC TTT C

Sequences of site-directed mutagenesis primers. Primer names and amino acid targets are listed alongside the corresponding sequence. 'F' and 'R' in the primer names denotes forward or reverse orientation. The randomised substitution codons NNK and MNN are shown in bold for each primer.

Table 2.5.

Name	**Primer Sequence (5’-3’)**
Lamp Primers	
FIP	TTCCGAAGAACGCTGAAGCGGAACTGATTACAAACATTGGCC
BIP	CGCATTGGCATGGAAGTCACAATTTGATGGCACCTGTGTA
Loop Primers	
LF	GGGGGCAAATTGTGCAATTTG
LB	CTTCGGGAACGTGGTTGACC
Displacement Primers	
F3	ACCAGGAACTAATCAGACAAG
B3	GACTTGATCTTTGAAATTTGGATCT

Sequences of LAMP-BART primers. Primer names and the corresponding sequences listed under the primer role in the LAMP-BART reaction. Primers target a subsection of the N2 gene region in the SARS-CoV-2 RNA sequence (Zhang *et al.* 2020b).

2.1.7. Double-stranded oligonucleotide fragments

As required, whole gene sequences were ordered as double stranded DNA (gBlocks) from Integrated DNA Technologies (IA, USA) as lyophilised stocks to be reconstituted in dH_2O to 50 ng/µl working concentration. Where appropriate gene sequences were codon optimised for *E. coli* expression, and unwanted restriction site removed by codon usage substitution.

2.1.8. Dry preserved Coleoptera

All fireflies discussed were on loan with permission for the outlined work from either Amgueddfa Cymru – National Museum Wales or California Academy of Sciences in San Francisco (CA, USA), via the National Museum Wales. An exception to this was a sample of *Lampyris noctiluca*, which was collected in Cambridge, UK in 2006 by Dr Amit Jathoul, and cold stored in dimethylsulfoxide (DMSO) at -80 ˚C since that time. Additionally, a complete luciferase gene DNA sequence for the lesser British Glow-worm *Phosphaenus hemipterus* was kindly provided by Dr John Day from the UK Centre for Ecology & Hydrology.

2.2. General Molecular Biology and Recombinant DNA Methods

2.2.1. Quantification of DNA concentration

2.2.1.1 NanoDrop

Purified DNA in dH_2O was quantified using the NanoDrop® ND-1000 UV-Vis spectrophotometer (Thermo Fisher Scientific, MA, USA), which quantifies Nucleic acid concentration by UV absorbance. The Nanodrop was 'blanked' using 1 µl of dH_2O prior to analysis of 1 µl volumes of DNA samples. Determined DNA concentrations were reported in ng/µl, alongside an absorbance trace reading taken to indicate sample purity. The absorbance of DNA is optimal at λ_{max} 260 nm, whilst common contaminants of salts and protein absorb at around 230 nm and 280 nm. Readings below 15 ng/µl were not considered reliable due to the detection limits of UV absorbance-based quantification. Guidance from Thermo Scientific suggest that 260/280 nm ratios of ~1.8 are generally accepted as 'pure' for DNA, and ratios appreciably lower are indicative of contaminants such as protein and phenol which absorb strongly ~280 nm. Expected 260/230 nm ratios for pure DNA are commonly in the range of 2.0-2.2, and ratios appreciably lower are indicative of contaminants including carbohydrates, salts, and phenol, which absorb at ~230 nm.

2.2.1.2 Qubit

High sensitivity DNA quantification was performed by Qubit dsDNA high-sensitivity assay on the Qubit 4 Fluorometer (Thermo Fisher Scientific, MA, USA), following manufactures instructions. The Qubit system uses dyes selective to sample types such as dsDNA. Subsequent fluorescence based quantification against high and low concentration calibration standards offers more sensitivity than UV absorbance-based quantification.

2.2.2. DNA sequencing

Sanger sequencing of inserts within the pET16b vector was performed by submission of 15 µl template DNA (50 – 100 ng/µl) to the Eurofins Genomics TubeSeq service (Luxembourg City, Luxembourg), using T7 primers available from the service. Sequencing of inserts was performed in both 5' and 3' directions, to enable full sequencing coverage. The Eurofins service returns multiple file formats, including 'clipped' files which remove lower confidence sequencing data from the 3' and 5' leaving only the high-confidence core which is typically up to ≈900 bp in length. As firefly luciferase genes are ≈1650 bp in length, this necessitated sequencing in both directions to recover sequencing data across the full insert length. Only 'clipped' files were used for analysis, and where clipping had removed significant portions of the sequence length sequencing was reattempted. Purified PCR products were typically sent for sequencing with 15 µl (10 pmol/µl) of their respective amplification primers. Required concentration for PCR products was dependent on products length such that 150 – 300 bp required 1 ng/µl, 300 – 1000 bp required 5 ng/µl, and 1000 – 3000 bp required 10 ng/µl.

2.2.3. End-point PCR

End-Point PCR was typically performed in the Eppendorf Mastercycler (Eppendorf, Hamburg, Germany) using *Taq* PCR Master Mix. Typically, for a 50 µl reaction volume <50 ng template was combined with 200 nM of each primer and 25 µl of *Taq* PCR Master Mix. PCR conditions would include an initial denaturation of 94 ˚C for 3-minutes followed by 30 cycles of 94 ˚C denaturation for 40-seconds, 50-68 ˚C (typical 5 ˚C lower than primer Tm) primer annealing for 40-seconds, and 72 ˚C extension for 1-minute/kb DNA, followed by a final extension at 72 ˚C for 5-minutes.

2.2.4. Agarose gel electrophoresis

Linear DNA such as the products of PCR or restriction digest was analysed by electrophoretic separation based on molecule size. Agarose gels were prepared to a concentration of 1% (w/v). Solid agarose powder (Bioline, London, UK) was dissolved in TAE buffer (1x) heated by microwave. SafeView Nucleic Acid Stain (NBS biologicals, Cambridgeshire, UK) was added to cooled liquid gel solution following manufacturer's concentration recommendations. DNA samples loaded into wells were mixed with an appropriate volume of 6x gel loading dye purple (NEB. MA, USA). Gels were submerged in 1x TAE and ran for ≈30-minutes at 100 V. Gels were analysed and images produced using the Vilber smart imaging E-Box gel documentation system (Marne-la-Vallée, France).

2.2.5. Growth and maintenance of *E. coli* strains

Glycerol stocks stored at -80 ˚C of *E. coli* containing pET16b constructs were spread onto LB agar plates prepared with 100 µg/ml carbenicillin and grown at 37 ˚C overnight. Single colonies picked from these plates were grown overnight at 37 ˚C with shaking (200 rpm) in 5 ml of LB broth prepared with 100 µg/ml carbenicillin. Liquid cultures of strains were prepared as 1 ml 20% glycerol stocks for long term storage at −80 ˚C. Aseptic technique was practised during all handling of *E. coli*.

2.2.6. Preparation of glycerol stocks

Liquid cultures of *E. coli* were mixed at a ratio of 1:1 with a 40% Glycerol solution in dH_2O. 1 ml aliquots in screw-top Cryogenic tubes (Sigma-Aldrich, MA, USA) were stored for future use at −80 ˚C.

2.2.7. Transformation of chemical competent cells

Plasmid DNA was transformed into *E. coli* DH5α and BL21 by heat shock, according to respective manufacturer’s instructions. Transformations were spread onto LB plates prepared with 100 µg/ml carbenicillin. DH5α were used primarily to acquire mass product from low concentration plasmid such as ligations. Single transformant colonies were picked into 5 ml of LB broth for overnight growth and subsequent bulk plasmid purification. Sequence confirmation was performed by Eurofins Genomics TubeSeq service (Luxembourg, Luxembourg) where required. BL21 were used only for the transformation of high concentration sequence confirmed plasmid DNA for protein expression, with the exception of the transformation of mutagenized libraries.

2.2.8. Purification of plasmid DNA

Purification of plasmid DNA from 5 ml of *E. coli* DH5α and BL21 was performed using the QIAprep Spin Miniprep Kit (Qiagen, Venlo, Netherlands) following manufacturer's instructions. QIAprep utilises alkaline lysis of pelleted cells and purification using a chaotropic salt and silica membrane spin column method which isolates up to 20 µg of plasmid DNA. Purified DNA was eluted into 65 ˚C dH_2O as opposed to the kit elution buffer to enable downstream molecular work. DNA concentration and degree of purity were estimated by applying 1 µl volumes to the NanoDrop® ND-1000 UV-Vis spectrophotometer (Thermo Fisher Scientific, MA, USA).

2.2.9. Purification of linear DNA

Linear DNA such as the products of PCR were purified using the QIAquick PCR Purification Kit (Qiagen, Venlo, Netherlands) following manufacturer's instructions. QIAquick utilises a spin column method which isolates DNA between 100 – 10,000 bp in length. In the instance where more than one linear DNA product was present in a sample such as the products of restriction digest, gel extraction was performed using the Zymoclean Gel DNA Recovery Kit (Zymo Research, CA, USA) following manufactures instructions. To perform gel extraction, DNA products were initially run on a standard electrophoresis gel, and the desired band of expected size excised. These gel slices were weighed to calculate the appropriate volume of kit buffers required, and purified following manufacturer's instructions. Zymoclean utilises a spin column method which isolates DNA between 50 – 23,000 bp. Purified DNA was eluted into 65 ˚C dH_2O as opposed to kit elution buffers in both PCR purification and gel extraction to enable downstream molecular work.

2.2.10. Restriction digestion

Restriction digests were most commonly performed using NEB Type II restriction enzymes *NdeI*, *BamH1*, and the methylation sensitive *DpnI* in 1 x CutSmart buffer (50 mM potassium acetate, 20 mM Tris-acetate, 10 mM magnesium acetate, 100 µg/ml BSA, pH 7.9). Reactions were performed in 50 µl reaction volumes, and incubated at 37 ˚C for 1-hour per µg of DNA used. Type II restriction enzyme were heat inactivated at 65 ˚C for 20-minutes.

2.2.11. Ligation

DNA ligation was performed using 1 µl of T4 DNA ligase (NEB, MA, USA) and T4 DNA Ligase Reaction Buffer (50 mM Tris-HCl, 10 mM MgCl2, 1 mM ATP, 10 mM DTT, pH 7.5) with insert and vector DNA in a 20 µl total reaction volume. 50 ng of vector DNA was used with a 3-fold molar excess of insert. Ligation reactions were performed at 16 ˚C for 30-minute or overnight at 4 ˚C, followed by heat inactivation at 65 ˚C for 10-minutes. Ligation reactions could then be immediately transformed into DH5α competent cells.

2.3. Bioprospecting Methodologies

2.3.1. Non-destructive DNA extraction

DNA was extracted from whole Coleoptera using the Non-Destructive DNA Extraction Buffer previously described. Each specimen was submerged in the minimal amount of buffer required to fully submerge each specimen, rounded to the nearest 50 µl (300-800 µl typically). Each of these samples was incubated at 55 ˚C in buffer for 20-hrs with gentle agitation (300 rpm). Specimens were subsequently transferred to 100% ethanol for 4-hours before return to collections. DNA from the extraction buffer was purified by phenol: chloroform: isoamyl alcohol extraction and subsequently ethanol precipitated overnight at -20 ˚C (Sambrook and Russell 2001). Purified DNA extract concentration was assessed by Qubit (Invitrogen. MA, USA) and DNA quality by TapeStation 4200 (Agilent. CA, USA), following manufacturers protocols.

2.3.2. TapeStation analysis

TapeStation analysis was performed following manufacturer’s instructions on the TapeStation 4200 System (Agilent, CA, USA) in the Cardiff University BIOSI Genomics Research Hub to conduct high-resolution electrophoresis analysis of the fragment size distribution within a DNA sample. TapeStation reagents and ScreenTapes were obtained from Agilent (CA, USA).

2.3.3. COI ‘mini-barcode’ quantitative PCR

Mini-barcode PCR was typically performed in the Eppendorf Mastercycler using *Taq* PCR Master Mix. For a 50 µl reaction volume <10 ng template was combined with 500 nM of each primer and 25 µl of *Taq* PCR Master Mix. Cycling conditions followed guidance from Zeale *et al* (2011). Initial denaturation of 95 ˚C for 3-minutes preceded 16 cycles of 94 ˚C denaturation for 30-seconds, 61 ˚C (-0.5/cycle touchdown) primer annealing for 30-seconds, and 72 ˚C extension for 30-seconds, followed by 24 further cycles with primer annealing temperature reduced to 53 ˚C and a final extension at 72 ˚C for 10-minutes.

2.3.4. CODEHOP quantitative PCR

CODEHOP quantitative PCR was performed in the Corbett Rotor Gene (Qiagen, Venlo, Netherlands) using SYBR Green JumpStart *Taq* ReadyMix. Typically, for a 20 µl reaction volume <10 ng template was combined with 200 nM of each CODEHOP primer and 10 µl SYBR Green JumpStart *Taq* ReadyMix. Following optimization, reactions were supplemented with 4% DMSO. Initial denaturation of 94 ˚C for 3-minutes preceded 35 cycles of 94 ˚C denaturation for 40-seconds, 50 ˚C primer annealing for 40-seconds, and 72 ˚C extension and fluorescence reading, followed by a final extension at 72 ˚C for 3-minutes.

2.3.5. Illumina library preparation

Illumina sequencing libraries were prepared from gDNA extracts using the NEXTFLEX® Rapid DNA-Seq Kit for Illumina® Platforms (PerkinElmer, MA, USA). Cycle numbers in the final amplification of the library prep were increased to give a final library concentration of up to 1 µg. Amplified libraries were purified using Agencourt AMPure XP SPRI beads (Beckman Coulter, CA, USA) following manufactures instructions.

2.3.6. Biotin probe construction

Cross-species capture probes were produced by performing a biotin-incorporating PCR using a gBlock of the mRNA from the *Photinus pyralis* luciferase gene designed and ordered from Integrated DNA Technologies (IA, USA) using terminal primers *Ppy*_F and *Ppy*_R. The PCR conditions were as follows: An initial denaturation of 94 ˚C for 3-minutes was followed by 30 cycles of 94 ˚C for 40-seconds, 60 ˚C for 40-seconds, and 72 ˚C for 2-minutes followed by a final extension at 72 ˚C for 5-minutes. <50 ng gBlock was used as template with 200 nM of each terminal primer, 25 µl of Taq PCR Master Mix, supplemented with 25 µM Biotin-11-dUTP Solution (Thermo Fisher Scientific, MA, USA) made up to 50 µl final volume with dH_2O. This gave a final concentration of 25 µM Biotin-11-dUTP vs a final concentration of 250 µM dTTP, resulting in an approximate 10% inclusion rate of biotin-11-dUTP in the place of dTTP. This gave the final probe products an approximate biotinylated nucleotide inclusion rate of 2.5%. The resulting biotinylated amplicon of ≈1650 bp was analysed on a 1% agarose gel stained with SafeView Nucleic Acid Stain (NBS Biologicals, Cambridgeshire, UK) ran at 100V for 30-minutes. Following confirmation of an appropriately sized amplicon, reactions were pooled and purified using the QIAquick PCR Purification Kit (Qiagen, Venlo, Netherlands) following manufactures instructions.

2.3.7. Cross-species affinity hybridization

Illumina libraries of up to 1 µg were combined with ≈50 ng of biotinylated probe. This mix was made up to 7.5 µl with the addition of dH_2O or reduced down using an Eppendorf concentrator plus (Eppendorf, Hamburg, Germany), where appropriate. The library and biotinylated probe mix was combined with an equal volume of 2x Hybridization solution (see buffers) and overlaid with 50 µl of mineral oil (Sigma-Aldrich, MA, USA). Hybridization reactions were then heated to 99 ˚C for 5-minutes to denature all double stranded DNA molecules before dropping to the annealing range starting at 65 ˚C. A touch down approach was taken in an attempt to recover more divergent library targets. The hybridization temperature starting at 65 ˚C reduced by 0.2 ˚C per hour over a 50-hour period to a final temperature of 55 ˚C. Theoretically, this allowed increased recovery of more specific targets preferentially over the divergent sequences which would be more likely to be captured toward the lower end of the temperature range.

2.3.8. Elution of captured targets

Dynabeads™ M-280 Streptavidin (Invitrogen, CA, USA) were used to exploit the high binding affinity of the biotin incorporated into the probes to streptavidin. The beads were thoroughly resuspended by vortex and 100 µl (1 mg of bead by dry weight) of the suspension taken. Beads were pulled out of solution using a magnetic particle concentrator holder and washed three times in 100 µl TEN buffer (see buffers) by serial magnetic separation and buffer resuspension. The beads were resuspended in a final volume of 100 µl TEN buffer to which the 15 µl hybridization reactions were added and mixed by vortex. The bead and hybridization mix was slowly rotated or agitated at room temperature over a 20-minute period to prevent the beads from settling.

Following this binding period the beads were separated and washed twice in Low stringency wash buffer (see buffers), each wash was slowly rotated or agitated at room temperature over a 15-minute period. These low stringency washes were followed by three consecutive High stringency (see buffers) washes performed at 65 ˚C for 15-minutes each. Each wash was mixed occasionally by gentle vortexing to prevent the beads from settling.

Captured library DNA targets were eluted by resuspending the beads in 50 µl of freshly prepared 0.1 M NaOH and slowly rotating or agitating at room temperature over a 20-minute period to prevent the beads from settling. The beads were finally magnetically separated and the supernatant removed, now containing the eluted DNA targets. The supernatant was then neutralised by the addition of an equal volume of 1 M Tris-HCl pH 7.5. The neutralised supernatant was finally passed through a G50 spin column (Epoch Life Science, Inc., TX, USA) following manufacturer's instructions.

2.3.9. Illumina library sequencing

Illumina libraries were sequenced as 150 bp paired-end reads on the Illumina NextSeq 500 Sequencer (Illumina, Inc., CA, USA) in the Cardiff University BIOSI Genomics Research Hub.

2.3.10. NGS bioinformatic analysis

Bioinformatic analysis was conducted using the Cardiff University School of Biosciences Biocomputing Hub's High Throughput Computing cluster (YSGO), which facilitates the processing and storage of information generated by data-intensive research within the School of Biosciences and makes available pre-prepared modules of common bioinformatics tools/packages. Details and discussion of the bioinformatics process and availability of tools used can be found in Chapter 3. Annotated bash scripts are available in the Appendices.

2.4. Infraluciferin Engineering Methodologies

2.4.1. Protein homology modelling

Homology models of *Phem*Luc were constructed using SWISS-MODEL which is freely available at https://swissmodel.expasy.org/. A model representing *Phem*Luc with native *D*-luciferin bound within the active site was generated using the crystal structure of *Ppy* Fluc in the adenylate-forming conformation bound to DLSA, resolved to 2.62 Å, from Protein Data Bank (PDB) file 4G36.pdb (Sundlov *et al.* 2012). A second model was constructed representing *Phem*Luc bound to infraluciferin, using the template of a second adenylate-forming conformation of *Ppy* Luc bound to iDLSA and resolved to 3.10 Å, from PDB file 6HPS.pdb (Stowe *et al.* 2019). The two *Phem*Luc models, along with the *Ppy* Fluc crystal structures from which they were derived, were then analysed within The PyMOL Molecular Graphics System, Version 2.1.1 Schrödinger, LLC. Amino acid residues measured to be within 4 Å of the respective bound ligand were identified for site-directed mutagenesis.

2.4.2. Site-directed mutagenesis and mutagenic library construction

The QuikChange® II Primer Design Program (Agilent, CA, USA) was used to design mutagenic primers targeted against positions in *Phem*Luc identified through homology modelling. The *Phem*Luc template was uploaded to the program, target positions highlighted and an arbitrary amino acid substitution selected. The mutagenic primers output by the program would then have their mutagenic codon changed to the desired NNK in the 5' forward primer, and MNN in the 3' reverse primer. The use of such primer pairs would result in a sequence library containing all possible coding codons in this position, enabling screening of the effect from all amino acids. Thirty-two primer sets were designed in total, detailed in Table 2.4.

Site-directed mutagenesis was performed using the QuikChange II Site-Directed Mutagenesis Kit (Agilent, CA, USA), according to manufacturer's instructions. Mutant strands were synthesized by PCR in the Eppendorf Mastercycler with 2.5 U of *Pfu* Ultra HF DNA polymerase, 50 ng of *Phem*Luc in pET16b, and 125 ng of each mutagenic primer. Thermocycling conditions included an initial denaturation of 95 ˚C for 30-seconds, followed by 16 cycles of 95 ˚C denaturation for 30-seconds, 55 ˚C primer annealing for 1-minute, and 68 ˚C extension for 8-minutes. Following amplification, SDM reactions were digested with the addition of 1ul *DpnI* and incubation for 1-hour at 37 ˚C. *DpnI* selectively digests adenomethylated plasmids from Dam+ *E. coli* such as DH5α, but leaves the non-methylated, mutagenized products of amplification undigested.

The digested mix was then used to transform XL1-Blue supercompetent cells for overnight growth at 37 ˚C. Plated transformant colonies were then resuspended in LB broth and plasmid DNA purified using the QIAprep Spin Miniprep Kit (Qiagen, Venlo, Netherlands). The extracted plasmid DNA served as libraries containing each possible codon for the target mutagenized position for subsequent transformation into BL21 cells for protein expression and screening.

2.5. Thermostability Engineering Methodologies

2.5.1. DNA shuffling

All DNA shuffling thermocycling processes were performed in the Eppendorf Mastercycler.

2.5.1.1 Template amplification

To generate template for shuffling, 10 ng of *Phem*Luc and *Phem*Luc x15 in pET16b were amplified in 50 µl reaction volumes using 500 nm each of the *Phem*Luc terminal primers and 200 µM of mixed dNTPs with 1 U of Q5® High-Fidelity DNA Polymerase and Q5 Reaction Buffer (NEB, MA, USA). Thermocycling conditions included an initial denaturation of 98 ˚C for 30-seconds, followed by 30 cycles of 98 ˚C denaturation for 10-seconds, 55 ˚C primer annealing for 30-seconds, and 72 ˚C extension for 50-seconds, with a final extension at 72 ˚C for 2-minutes. PCR products were purified using the QIAquick PCR Purification Kit (Qiagen, Venlo, Netherlands).

2.5.1.2 DNase1 digestion

2 µg of amplified template DNA was combined with 1x DNase1 buffer (see buffers) and made up to a final reaction volume of 50 µl. Reactions were incubated at 15 ˚C for 5-minutes prior to the addition of 0.3 U or 0.5 U of DNase1 (NEB, MA, USA). Following the addition of DNase1, reactions were held at 15 ˚C for 3-minutes followed by heat inactivation of DNase1 at 80 ˚C for 10-minutes. Digest products were purified using the QIAquick PCR Purification Kit (Qiagen, Venlo, Netherlands).

2.5.1.3 *Pfu* reassembly

100 ng of purified digest DNA from each template (200 ng total for two genes) was reassembled in 20 µl reaction volume using 400 µM of mixed dNTPs with 2 U of *Pfu* DNA Polymerase (Promega, WI. USA) in the supplied reaction buffer (200 mM Tris-HCl pH 8.8, 100 mM KCl, 100 mM $(NH_4)_2SO_4$, 20 mM $MgSO_4$, 1.0% Triton® X-100 and 1 mg/ml nuclease-free BSA), in the absence of primers. Thermocycling conditions included an initial denaturation at 96 ˚C for 3-minutes, followed by 40 cycles of 94 ˚C denaturation for 1-minute, 55 ˚C overlap annealing for 1-minute, and 72 ˚C extension for 1-minute, with a final extension at 72 ˚C for 7-minutes.

2.5.1.4 Terminal primer amplification

1 µl of the raw *Pfu* reassembly reaction was amplified in 50 µl reaction volumes using 500 nm each of the *Phem*Luc terminal primers and 200 µM of mixed dNTPs with 1 U of Q5® High-Fidelity DNA Polymerase and Q5 Reaction Buffer (NEB, MA, USA). Thermocycling conditions included an initial denaturation of 98 ˚C for 30-seconds, followed by 25 cycles of 98 ˚C denaturation for 10-seconds, 55 ˚C primer annealing for 30-seconds, and 72 ˚C extension for 50-seconds, with a final extension at 72 ˚C for 2-minutes. PCR products were purified by gel extraction using the Zymoclean Gel DNA Recovery Kit (Zymo Research, CA, USA).

2.5.2. Error-prone polymerase chain reaction mutagenesis with $MnCl_2$ and D_2O

Error-prone polymerase chain reaction (epPCR) was performed in the Eppendorf Mastercycler using $MnCl_2$ and deuterium oxide heavy water (D_2O) (Sigma-Aldrich, MA, USA) to increase the error-rate of *Taq* polymerase. 50 ng of primer DNA was used as template and combined with 250 µM of each *Phem*Luc terminal primer, 200 µM mixed dNTPs, and QIAGEN PCR buffer (Qiagen, Venlo, Netherlands). Four different concentrations were set up using 0, 0.1, 0.2, and 0.3 mM $MnCl_2$. Each partial set-up reaction was reduced to <1 µl using the Eppendorf concentrator plus, prior to the addition of 2.5 U *Taq* polymerase (Qiagen) and resuspension to a final volume of 50 µl in D_2O. Thermocycling conditions included an initial denaturation at 94 ˚C for 3-minutes, followed by 30 cycles of 94 ˚C denaturation for 40-seconds, 60 ˚C primer annealing for 40-seconds, and 72 ˚C extension for 2-minute, with a final extension at 72 ˚C for 10-minutes.

2.6. Methods for Bioluminescence Screening

2.6.1. Primary screening with LH_2

Plated *E. coli* BL21 colonies were transferred using nylon Hybond-N membrane and placed face up on LB agar plates prepared with 100 µg/ml carbenicillin and previously spread with 200 µl LB broth containing 12.5 µl of 1 M IPTG. The colonies were then left to induce at room temperature for 3-4 hours. Following induction, the colonies were screened by spraying each plate with 500 µM *D*-luciferin (10 ml 0.1 M sodium citrate (pH 5) with 159 µl *D*-luciferin [Registech, IL, USA]). Bioluminescence emissions were measured in a small-animal imaging device, PhotonIMAGER Optima (Biospace Labs, Paris, France) at room temperature. Bioluminescence was typically recorded over a 20-second integration period without any filters, and the data analysed in the M3 Vision software package, available under license from https://biospacelab.com.

2.6.2. Primary screening with iLH_2

Plated *E. coli* BL21 colonies were transferred using nylon Hybond-N membrane and placed face up on LB agar plates prepared with 100 µg/ml carbenicillin and previously spread with

200 µl LB broth containing 12.5 µl of 1 M IPTG. The colonies were then left to induce at room temperature for 3-4 hours. Following induction, the colonies were screened by spraying each plate with 500 µM infraluciferin in 0.1 M sodium citrate (pH 5). Bioluminescence emissions were measured in a small-animal imaging device, PhotonIMAGER Optima at room temperature. Bioluminescence was typically recorded over a 1-minute integration period without any filters, and the data analysed in the M3 Vision software package, available under license from https://biospacelab.com.

2.6.3. Primary screening of thermal resistance

Plated *E. coli* BL21 colonies were transferred using nylon Hybond-N membrane and placed face up on LB agar plates prepared with 100 µg/ml carbenicillin and previously spread with 200 µl LB broth containing 12.5 µl of 1 M IPTG. The colonies were then left to induce at room temperature for 3-4 hours. Following induction, the plated colonies were incubated at 50 ˚C for 1-hour prior to screening by spraying each plate with 500 µM *D*-luciferin (10 ml 0.1 M sodium citrate (pH 5) with 159 µl *D*-luciferin). Bioluminescence emissions were measured in a small-animal imaging device, PhotonIMAGER Optima at room temperature. Bioluminescence was typically recorded over a 20-second integration period without any filters, and the data analysed in the M3 Vision software package.

2.6.4. Secondary screening

Colonies of interest from the primary screens were identified on their original growth plates from before membrane transfer and picked for replication in triplicate on fresh LB agar plates prepared with 100 µg/ml carbenicillin. Colonies were replicated in a known grid layout to enable better visualisation, and the screening process performed identically to the primary screen, in order to confirm observations.

2.7. Overexpression and Purification of Luciferases

Details of all buffers referenced can be found in 2.1.2 Buffers.

2.7.1. Overexpression of luciferases

Single colonies of *E. coli* BL21 previously transformed with a pET16b Fluc construct were picked and used to inoculate a 5 ml LB broth medium supplemented with 100 µg/ml carbenicillin. The liquid culture was grown overnight at 37 ˚C with shaking (200 rpm) and used to further inoculate 300 ml of fresh LB broth supplemented with 100 µg/ml carbenicillin, under the previous incubation conditions. The microbial growth of the culture was monitored by measuring the optical density at 600 nm (OD_{600}) of 1 ml aliquots using a spectrophotometer (Pharmacia Biotech, Sweden) previously blanked with fresh LB broth supplemented with carbenicillin. Once the OD_{600} was measured between 0.6 – 0.7 AU, which indicated the log phase of growth, IPTG was added to the liquid culture at a final concentration of 1 mM in order to induce protein expression. The induced culture was then incubated overnight at 18 ˚C with shaking (200 rpm). The induced culture was then placed to cool on ice for 5 – 10 minutes, and the cells pelleted by centrifugation at 4000 x g for 40-minutes at 4 ˚C and the supernatant discarded.

2.7.2. Cell lysis and purification

Induced bacterial pellets of overexpressed Flucs were resuspended in 5 ml of chilled lysis buffer per gram of pellet on ice. The final constituents of Triton X-100, lysozyme, and Benzonase were added once the pellet was fully resuspended, and the complete mix incubated on ice for 30-minutes. The cell lysates were centrifuged at 40,000 x g for 90-minutes at 4 ˚C, whilst empty PD-10 columns were prepared with 5 ml of HisPur Ni-NTA Resin in a cold-room facility. The Ni-NTA columns were pre-calibrated with 5 ml of 20 mM IMD solution and the supernatant from centrifugation of the lysate applied. Flow-through supernatant was reapplied twice before 2.5 ml of 50 mM IMD solution was applied to elute non-specifically bound proteins which lack the 10x His-tag present in the Flucs. His-tagged Flucs were then eluted with increasing concentrations of the IMD solutions in 2.5 ml applications in the order of one application of 200 mM IMD, three applications of 300 mM IMD, and three applications of 500 mM IMD. 20 µl of the eluted IMD fractions were kept on ice and assayed by luminometry (see 2.7.3). Fractions containing the highest bioluminescence activity were desalted by applying them to disposable PD-10 desalting columns pre-calibrated in 25 ml of storage buffer. Purified protein was eluted from columns in 3.5 ml of storage buffer and samples of the same protein from multiple desalting columns combined and divided across 500 µl aliquots for storage at -80 ˚C.

2.7.3. Luminometric quantification during protein purification

The bioluminescence activity of 1 µl from each fraction was assayed in triplicate in the PhotonIMAGER Optima by the addition of 500 mM LH_2 and 1 mM ATP in TEM (pH 7.8).

2.8. Assessment of Purification

2.8.1. Quantification of protein concentration by Bradford assay

The protein-dye binding Bradford assay (Bradford 1976) to quantify total protein content was performed using the Bio-Rad Quick Start Bradford Protein Assay kit (Bio-Rad Laboratories, CA, USA), according to manufacturer's instructions. 5 µl of each bovine serum albumin (BSA) concentration standard and protein sample were combined with 250 µl of Dye Reagent and incubated at room temperature for 5-minutes. Absorbance was measured at 595 nm using the CLARIOstar Plus Microplate reader (BMG LABTECH, Ortenberg, Germany). Protein concentration was determined by linear regression of the BSA standard plot.

2.8.2. SDS-PAGE of purified protein

Purified proteins were diluted to the lowest concentration sample as identified by Bradford assay and 10 µl samples were prepared with 10 µl Laemmli loading buffer and heating at 95 ˚C for 10-minutes to denature protein. 10 µl of prepared samples were loaded onto a sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) comprising a 4% stacking layer and 10% separating gel. 5 µl of Fisher BioReagents™ EZ-Run™ Prestained Rec Protein Ladder (MA, USA) was loaded into the outer well, and the gel was ran at 100 V for 30-minutes and 150 V for 5-hours. The constituents of each gel are detailed below.

	10% separating gel	**4% stacking gel**
30% Acrylamide/Bis Solution, 29:1*	13.9 ml	2.8 ml
1.5 M Tris-HCl, pH 8.8	8.1 ml	-
0.5 M Tris-HCl, pH 6.8	-	1.25 ml
20% SDS solution	200 µl	65 µl
dH_2O	7.9 ml	5.725 ml
TEMED*	10 µl	10 µl
20% Ammonium persulfate	225 µl	150 µl

*(Bio-Rad Laboratories, CA, USA)

2.8.3. SDS-PAGE gel staining

The 4% stack layer was trimmed and the 10% gels were fixed in methanol fixing solution for 30-minutes, then immersed overnight in Coomassie staining solution at room temperature with gentle agitation. Gels were destained by immersing in destaining solution at room temperature with gentle agitation for ≈4-hours until gel background was uniformly transparent and protein bands clearly visible. Used destain solution was decanted off and replaced with fresh solution 2-3 times during the destaining process.

2.8.4. SDS-PAGE gel analysis

Destained gels were imaged on an office document scanner. The image of the gel was analysed in ImageJ (Schneider *et al.* 2012) by outlining the protein bands in rectangular boxes and plotting the box areas as histograms of band intensity. The protein bands were represented as peaks on the histogram plot trace, that could be isolated with the straight line tool, and the area of the enclosed peaks selected using the wand tool. Area of the peak indicated the intensity of the band relative to the background and was used to adjust the

Bradford determined protein concentration such that each protein could better be normalised through dilution.

2.9. Firefly Luciferase Characterization Methodologies

2.9.1. Luminometric methods

2.9.1.1 General principles

Final concentration of each Fluc in all assays was consistent to 0.167 µM, whilst substrates ATP and LH_2 were typically 1 mM and 500 µM, unless varied for the determination of Michaelis-Menten kinetic parameters. All reaction constituents were diluted in TEM buffer typically at pH 7.8 (±0.05), with the exception of pH-dependence assays where the pH of TEM was varied. In all experiments where saturation conditions of LH_2 and ATP substrates were used, concentrations were used which have previously been shown to be approximately 10x the K_M for the respective substrate in *Ppy* Fluc and x11 (Jathoul 2008; Halliwell 2015). All assay constituents were prepared and maintained on ice prior to use, and all measurements obtained in triplicate. All assays were performed in 96-well microtitre plates.

2.9.1.2 PhotonIMAGER Optima

Bioluminescence emissions from *E. coli* BL21 and protein samples were measured in a small-animal imaging device, PhotonIMAGER Optima (Biospace Labs, Paris, France) at room temperature. The PhotonIMAGER is equipped with a height adjustable baseplate which was used to position samples to be imaged at an appropriate distance from the camera lens. The PhotonIMAGER possesses a Photonmultiplier tube (PMT) which enables extremely sensitive detection of light from the ultra-violet, visible light, and near-infrared ranges of the electromagnetic spectrum, and is capable of detecting at a single photon level of resolution.

To measure bioluminescence a bright field image was captured which would later serve as a background image to be overlaid with the bioluminescence acquisitions. A non-filtered PMT image was acquired over a duration inversely correlated with the intensity of the bioluminescence signal from each sample in order to moderate the output data file size.

Typically, 20-second acquisition durations were used whilst imaging with LH_2, and 60-second durations with iLH_2. If spectra were to be obtained following the acquisition of the initial non-filtered PMT acquisition, bioluminescence would be recorded through successive PMT acquisitions through the available bandpass filters with a midpoint stepwidth of 25 nm. Immediately following the filtered acquisitions, a second non-filtered PMT image was acquired to enable the calculation of compensation required to account for the bioluminescence signal reduction over the duration of the filtered acquisitions. All acquisitions were performed at room temperature. Data was analysed in the M3 Vision software package, available under license from https://biospacelab.com, and exported to Excel for later analysis.

2.9.1.3 BMG LABTECH CLARIOstar Plus

Bioluminescence emission from purified protein samples were measured in the CLARIOstar Plus Microplate reader (BMG LABTECH, Ortenberg, Germany). Whilst the CLARIOstar lacks the sensitivity of the PhotonIMAGER and can only accept samples in a microtitre plate, it possesses several advantages which make it more amenable to luciferase bioluminescence characterisation. The CLARIOstar is equipped with a monochromator which enables the acquisition of spectra with a resolution of up to 1 nm. It is additionally equipped with a pump and injection needle to facilitate automatic substrate dispensing and immediate capture of bioluminescence signal.

The CLARIOstar was preconfigured with the Firefly preset Optic setting, and gain set to 2500 across all assays. The focal height was set to 11 mm, and the substrate mix pump volume was 100 µl, with a pump speed of 430 µl/s. All acquisitions were performed at room temperature. Data was opened in the included MARS analysis software and exported to Excel for later analysis.

2.9.1.4 LUCY

The 'LUCY' is composed of a programmable TRobot thermal cycler from Biometra (Gottingen, Germany) within a Syngene Chemi Genius Bio Imaging System from Synoptics (Cambridge, UK). A CCD camera above the thermal cycler is used to record light output from each well, using specially designed React IVD software developed by Synoptics (Hardinge 2014).

2.9.2. Measurement of bioluminescence spectra

High resolution bioluminescence spectra were obtained using the CLARIOstar Plus Microplate reader by injection of 100 µl substrate mix onto 50 µl of Fluc such that final assay concentrations were equal to 1 mM ATP, 500 µM LH_2, and 0.167 µM protein. Each reaction constituent previously diluted in chilled TEM buffer of pH 7.8 (±0.05) and total reaction volume equal to 150 µl. Following substrate injection, each reaction was held at RT for 30-seconds prior to acquisition of spectra by integrating light emissions over 2-seconds for 221 wavelength scanpoints between 490 nm and 710 nm, with a stepwidth of 1 nm. All measurements were made in triplicate for each Fluc.

2.9.3. pH dependence of bioluminescence spectra

The bioluminescence spectra for all Flucs were recorded across a range of pH conditions using the CLARIOstar Plus Microplate reader by injection of 100 µl substrate mix onto 50 µl of Fluc such that final assay concentrations were equal to 1 mM ATP, 500 µM LH_2, 0.167 µM protein, and total reaction volume equal to 150 µl. Each reaction constituent was previously diluted in chilled TEM buffer of varied pH (pH 6.3, 6.8, 7.3, 7.8, 8.3, 8.8 [±0.05]), adjusted using acetic acid or sodium hydroxide. Following substrate injection, each reaction was held at RT for 30-seconds prior to acquisition of spectra by integrating light emissions over 2-seconds for 36 wavelength scanpoints between 450 nm and 800 nm, with a stepwidth of 10 nm. All measurements were made in triplicate for each Fluc.

2.9.4. pH dependence of flash kinetics

The flash kinetics for all Flucs were measured across a range of pH conditions using the CLARIOstar Plus Microplate reader by injection of 100 µl substrate mix onto 50 µl of Fluc such that final assay concentrations were equal to 1 mM ATP, 500 µM LH_2, 0.167 µM protein, and total reaction volume equal to 150 µl. Each reaction constituent was previously diluted in chilled TEM buffer of varied pH (pH 6.3, 6.8, 7.3, 7.8, 8.3, 8.8 [±0.05]), adjusted using acetic acid or sodium hydroxide. Following substrate injection light emission were immediately integrated over 20 ms for 1000 consecutive measurements. All measurements were made in triplicate for each Fluc.

2.9.5. Determination of kinetic constants

It is understood that kinetic parameters can be derived for Flucs by interpreting the peak intensity (I_{max}) as a proxy for the pre steady-state of maximal light intensity, which can be used to extrapolate kinetic parameter values using the Michaelis-Menten (MM) equation (Ugarova 1989; Brovko *et al.* 1994). The point of I_{max} for all Flucs were measured across a range of substrate concentrations using the CLARIOstar Plus Microplate reader by injection of 100 µl substrate mix onto 50 µl of 0.5 µM Fluc, and total reaction volume was equal to 150 µl. Measurements to determine kinetic constants with regards to ATP were performed such that final assay substrate concentrations for ATP were varied to include 0.1, 0.5, 10, 25, 50, 100, 200, 400, 800, and 1000 µM, whilst LH_2 was maintained at 500 µM. Measurements to determine kinetic constants with regards to LH_2 were performed such that final assay substrate concentrations for LH_2 were varied to include 0.1, 0.5, 1, 5, 10, 20, 35, 70, 140, and 200 µM, whilst ATP was maintained at 1 mM. Each reaction constituent was previously diluted in chilled TEM buffer of pH 7.8 (±0.05). Following substrate injection light emissions were integrated over 20 ms for 1000 consecutive measurements. All measurements were made in triplicate for each Fluc. Data was analysed by plotting I_{max} values against each substrate concentration and subsequently kinetic constants were derived by implementing a linearized rearrangement of the Michaelis-Menten plot, commonly referred to as a Hanes-Woolf plot (Hanes 1932; Hofstee 1952).

2.9.6. Measurement of bioluminescent spectra with iLH_2

Bioluminescence spectra with iLH_2 were obtained in the PhotonIMAGER Optima by manual pipetting of 100 µl substrate mix onto each Fluc such that final assay concentrations were equal to 1 mM ATP, 500 µM iLH_2, and 0.167 µM protein. Each reaction constituent was previously diluted in chilled TEM buffer at pH 7.8 (±0.05) and total reaction volume was equal to 150 µl. Following substrate injection, each reaction was held at RT for 60-seconds prior to a non-filtered PMT acquisition of total bioluminescence yield, immediately followed by PMT acquisition across 14 band pass filters, with a midpoint range between 472 nm and 800 nm and a midpoint stepwidth of 25 nm. Following acquisition of spectra, a second non-filtered PMT acquisition was recorded to enable the calculation of compensation required to account for the bioluminescence signal reduction over the duration of the filtered acquisitions. Light emissions were integrated over 60-seconds for all acquisitions and all measurements were made in triplicate for each Fluc.

2.9.7. Determination of specific activity with iLH$_2$

Specific activities with iLH$_2$ were determined from the non-filtered PMT acquisitions obtained before and after spectra acquisition in the PhotonIMAGER Optima (see 2.9.6).

2.9.8. Determination of thermal inactivation

Four 200 µl aliquots of 0.5 µM Flucs in TEM buffer of pH 7.8 (±0.05) were incubated in a circulating digital water bath set at 20 ˚C, 30 ˚C, 35 ˚C, 40 ˚C, 45 ˚C, 50 ˚C, 55 ˚C, and 60 ˚C. Every 15-minutes over a 1-hour period an aliquot was removed onto ice, and the bioluminescence immediately recorded in triplicate by integrating light emission over 20 ms for 1000 consecutive measurements using the CLARIOstar Plus Microplate reader by injection of 100 µl substrate mix onto 50 µl of Fluc such that final assay concentrations were equal to 1 mM ATP, 500 µM LH$_2$, and 0.167 µM protein. Each reaction constituent was previously diluted in chilled TEM buffer of pH 7.8 (±0.05) and total reaction volume equal to 150 µl. A 0-minute incubation measurement was obtained in triplicate, and used as the comparative 0-minute incubation point for all incubation temperatures.

2.9.9. Thermal degradation analysis between 50-60 ˚C

Bioluminescence activity degradation for every 2 ˚C between 50-60 ˚C was measured using the LUCY imager by manual pipetting of 100 µl substrate mix onto 50 µl Fluc such that final concentrations were equal to 1 mM ATP, 500 µM LH$_2$, and 0.167 µM protein. Each reaction constituent was previously diluted in chilled TEM buffer at pH 7.8 (±0.05) and total reaction volume equal to 150 µl. Following substrate injection, each reaction was overlaid with mineral oil before transfer to the heat block and acquisition of the bioluminescence signal. Bioluminescence activity was recorded over a 30-minute duration by integrated 10 seconds of bioluminescence signal every 20 seconds. All measurements were made in triplicate for each Fluc.

2.10. LAMP-BART

Trial LAMP-BART assays were performed at 50 ˚C and set up in 20 µl reaction volumes such that the final concentration of all reagents were: 1x Isothermal Amplification Buffer (NEB, MA, USA), 5 ng/µl salmon sperm carrier DNA (SSDNA) (Invitrogen, CA, USA), 10 mM DTT, 0.4 mg/ml polyvinylpyrrolidone, 60 mM potassium chloride, 300 µM each of dNTPs, 100 µg/ml LH_2, 250 µM adenosine-5'-O-phophosulphate (Biolog, Bremen, Germany), 375 mU/ml ATP sulphurylase (NEB, MA USA), 0.32 U/µl *Bst* 2.0 WarmStart DNA Polymerase (NEB, MA, USA), 0.3 U/µl WarmStart RTx Reverse Transcriptase (NEB, MA, USA), 0.8 µM each of LAMP primers (FIP and BIP), 0.4 µM each of Loop primers (LF and LB), 0.2 µM each of displacement primers (F3 and B3), and 500 copies of the SARS-CoV-2 RNA per reaction as template. 5.5 µg/ml of Ultra-Glo™ recombinant luciferase (Promega, WI, USA) was used in the control reactions, and 5 µg/ml of speculative BART Flucs x11 and x16 in their respective reactions. A 20x concentration reaction with 100 µg/ml of x16 was additionally performed. The RNA template was supplied by Dr Patrick Hardinge and all ethical consents required for its use were received. Reactions were set up in quadruplicate in a DNA clean area with single use batched aliquots alongside appropriate controls and made up to a final volume of 20 µl in dH_2O and overlaid in mineral oil to prevent evaporation during the assay. The reaction plates were sealed under a clear adhesive film cover and assayed on the 'LUCY' imager at 50 ˚C. The bioluminescence emissions from 1-minute time integrals were collected for 60 scanpoints over a 1-hour duration, and saved to Excel for later analysis.

2.11. General Bioinformatics Tools

The European Nucleotide Archive (ENA) (available at https://www.ebi.ac.uk/ena/home) was used to search for and access nucleotide sequences using GenBank accession numbers for known sequence data, and keyword search terms relating to the availability of unknown sequence data. The Basic Local Alignment Search Tool (BLAST) (available at https://blast.ncbi.nlm.nih.gov) was used to identify nucleotide and protein sequences of high similarity to input queries. Clustal Omega (available at https://www.ebi.ac.uk/Tools/msa/clustalo) was used for the alignment of two or more nucleotide or protein sequences. The Expasy translate tool (available at https://web.expasy.org/translate) was used to translate nucleotide sequences to protein sequences. QIAGEN CLC sequence viewer v8 was used for the construction of nucleotide and protein alignment images. CLC sequence viewer is no longer available, but updated software package are available at https://digitalinsights.qiagen.com/downloads/product-downloads.

2.12. Statistical analysis

All data are presented as the mean ± standard error of the mean (SEM). P-values were determined by one-way analysis of variance (ANOVA) with Tukey's post-hoc analysis for multiple comparisons using GraphPad Prism version 7 for Windows (GraphPad Software, CA, USA). Differences were considered significant when $P<0.05$. Where appropriate graphs are annotated with lettering to indicate significant differences between measurements. For all measurements with the same letter, the difference between the means is not statistically significant. Where two measurements have different letters, they are significantly different. The omission of any letter indicates the mean of the respective measurement is statistically different to all other measurements.

Chapter 3

Bioprospecting for Coleopteran Luciferase

3.1. Chapter Summary

In this chapter DNA was non-destructively extracted from unidentified dry-preserved Lampyridae from the National Museum Cardiff. Identification was attempted through amplification of ‘mini-barcode’ sequences, and analysis of complementarity with available barcode sequences. The potential to amplify a short fragment luciferase gene target from varied firefly species was explored using a CODEHOP primer set and a method of luciferase gene cross-species affinity enrichment was developed in Illumina sequencing libraries. Enriched libraries were sequenced and a bioinformatic process undertaken to identify complete luciferase gene sequences. The bioinformatic strategy enabled the isolation of a novel luciferase gene from an unidentified Costa Rican firefly, and the resulting enzyme was demonstrated to produce a bioluminescence signal.

3.2. Introduction

A vast repository of unexplored genetic information exists in the diverse collections of dry-preserved insect specimens around the world. Whilst these collections can serve as important records of species diversity, their use in genetic analyses has been limited due to the destructive nature of common DNA extraction practises and the degradation of genetic material over years of inadequate storage. Genetic analyses of any species, whether they be *Insecta* or otherwise have commonly been performed on specifically collected fresh samples. However, recent changes introduced through the implementation of the Nagoya Protocol on 12th October 2014 restricts access to samples collected from this date onwards unless prior informed consent has been obtained to access the genetic resource, along with mutually agreed terms for undertaking research and development. The temporal scope of the Nagoya Protocol is an ambiguous issue, as some countries consider utilization the trigger for benefit sharing obligation, whilst others consider it as physical access in the country of origin. The Nagoya Protocol has left it up to each member State to clarify this ambiguity through their implementing legislation. The EU regulation considers physical access in the country of origin the point at which access and benefits sharing obligations are triggered. Therefore, compliance measures are only required when using resources physically accessed after the Nagoya Protocol has been ratified by both the EU and the country of origin (Drews *et al.* 2016). Therefore, as the Nagoya Protocol does not retrospectively cover samples collected earlier than the date of its implementation in the EU, samples collected prior to this can be freely accessed for genetic analyses, such as those that exist in the National Museum Cardiff.

There are believed to be in excess of 2000 species of fireflies worldwide (https://www.firefly.org/), each of which presumably possesses a unique luciferase enzyme used to generate the characteristic bioluminescence signal. However, the luciferase gene sequence has been isolated from only ≈30 firefly species (Oba *et al.* 2020). Firefly luciferases are an essential tool in research and industry, which exploit their bioluminescence properties in a range of applications. Whilst many of these applications depend upon enhanced enzyme characteristics through engineering, natural variation remains an underexplored reservoir of novel enzyme properties.

To account for the continued need for novel luciferase gene sequences and the legislative restrictions of the Nagoya Protocol, this study sought to explore the feasibility of extracting DNA from dry-preserved Lampyridae with limited damage conferred, and to subsequently use these genomic extracts to bioprospect for novel luciferase gene sequences. With the use of a cross-species affinity enrichment strategy and Next Generation Sequencing (NGS), a novel luciferase gene sequence was isolated from an unidentified Costa Rican firefly and subsequently demonstrated to encode a functioning bioluminescent enzyme.

3.3. Results and Discussion

3.3.1. Trial non-destructive DNA extractions

3.3.1.1 DNA extraction from insects

Most methods of DNA extraction from insects depend upon the use of ground insect tissue (Asghar *et al.* 2014), or limited destructive approaches where only small parts of the specimens are used, such as individual legs (Watts *et al.* 2007). Whilst the drawback to pulverizing a museum-archived specimen is obvious, the issue with leg extraction is that template DNA yields are extremely low quantity and often insufficient for applications beyond genotyping. Fortunately, a method exist which enables the non-destructive extraction of DNA from preserved Coleoptera that has been demonstrated to provide PCR-amplifiable mitochondrial and nuclear DNA on beetles collected up to 50 years previously, without conferring external morphological damage (Gilbert *et al.* 2007). This same method has been used successfully on museum samples collected as far back as 1820 AD (Thomsen *et al.* 2009).

Figure 3.1. provides an overview of the non-destructive DNA extraction method of Gilbert *et al* (2007). Whole specimens are fully immersed in a proteinase K based extraction buffer and incubated for 16 – 20 hours at 55 ˚C, with gentle agitation. Specimens can then be transferred from the extraction buffer to 100% ethanol for 2 – 4 hours to prevent further digestion and returned to their collections once air-dried. Nucleic acids can be purified from the retained extraction buffer using phenol-chloroform extraction (see Chapter 2 for further details).

Non-destructive DNA extraction was trialled on three specimens of Lampyridae. Two dry-preserved samples were provided by the National Museum Cardiff, which were the North American Firefly *Photinus pyralis* (*Ppy*) collected in 1996, and an unidentified Bornean firefly collected in 1987. The third specimen was the British Glow-worm *Lampyris noctiluca* (*Lnoc*), which had been collected by Dr Amit Jathoul in 2006, and stored at -80 ˚C. Figure 3.2. shows images of each sample before and after the DNA extraction process and demonstrates the limited morphological damage that has been conferred. The greatest difference can be observed in the *Lnoc* sample, which is shown to lose pigment from the light organ. However, as this sample was collected fresh in 2006 and then stored at -80 ˚C, this colour change can likely be attributed to leaching of the heavily pigmented luciferin substrate of the bioluminescence reaction into the extraction buffer. The presence of comparable levels of luciferin in any dry-preserved museum specimen of Lampyridae is unlikely due to degradation under less optimal storage conditions.

3.3.1.2 Quality of trial DNA extracts

The quality of DNA extracted in Figure 3.2. was assessed for concentration and the average size of the DNA fragments. DNA concentration was measured by fluorometric quantification on the Qubit 4 Fluorometer (Thermo Fisher Scientific, MA, USA) and average fragment size of DNA extracts measured by analysis on the 4200 TapeStation System (Agilent, CA, USA). The results of these analyses are reported in Table 3.1. High concentration DNA extracts were obtained from both *Ppy* and *Lnoc*, but no DNA could be detected in the extract from the Bornean firefly. Whilst an absolute cause cannot be attributed to, DNA is known to degrade as a function of time and heat (Lindahl 1993). Additionally, many insect specimens are dispatched with chemical asphyxiates including ethyl acetate and formalin, which are known to cause extensive double-stranded breaks throughout the genome, accelerating the degradation process independent of the specimen age (Dillon *et al.* 1996). The use of such chemicals is common practise, and information on their use is rarely recorded. The role of storage condition is further highlighted by the difference in average fragment size between the dry-preserved specimen *Ppy* recorded at 217 bp, and *Lnoc* stored fresh at -80 ˚C yielding an average fragment size of 17273 bp.

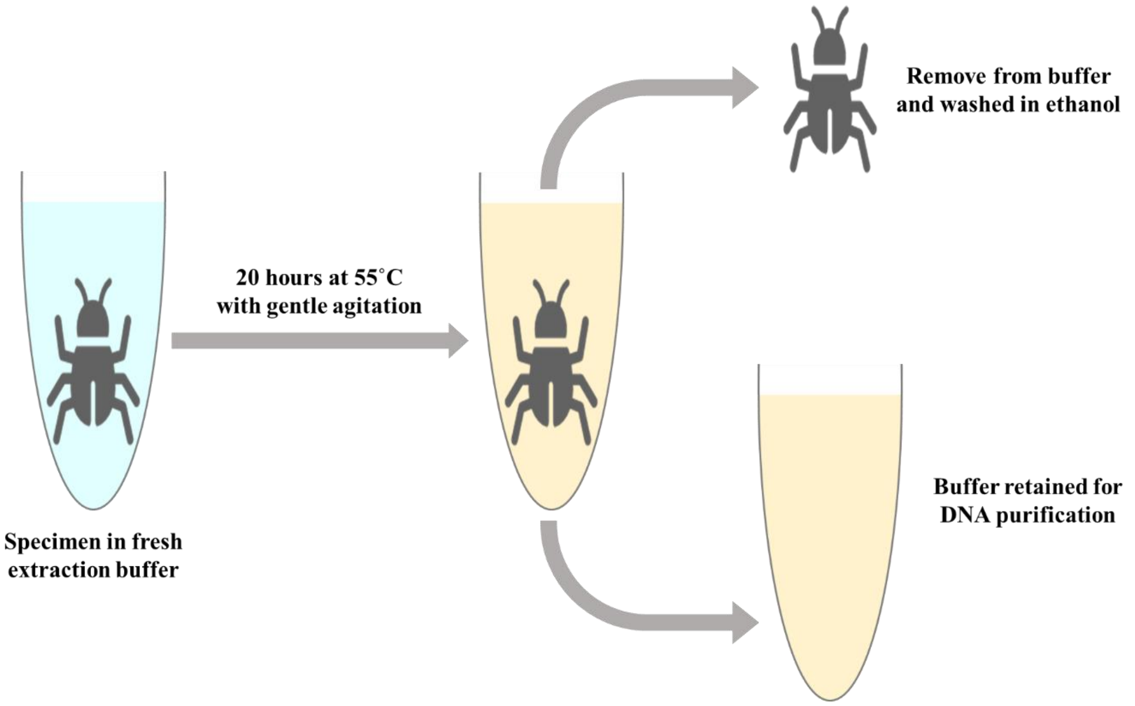
As the results observed in the *Ppy* specimen are indicative of the quality of DNA that can likely be expected from further extraction of dry-preserved Coleoptera, the feasibility of identifying a complete 1650 bp luciferase gene (>2000 bp with intron sequences) by direct approaches such as PCR is extremely low. Furthermore, this problem would be confounded in an unidentified firefly specimen where the target sequence of the luciferase is unknown for primer design. However, PCR of short DNA regions was not ruled out as an approach to establish whether an unidentified firefly DNA extraction could potentially contain a novel luciferase gene sequence.

3.3.1.3 Amplification of 'mini-barcodes'

To investigate whether PCR was possible using the DNA extracts, amplification of a short fragment of the mitochondrial cytochrome *c* oxidase I (COI) gene was attempted using a taxon-specific primer set for the universal amplification of arthropod COI 'mini-barcodes', originally designed to amplify a 157 bp region of digestion degraded DNA extracts from insectivorous bat faecal samples (Zeale *et al.* 2011). The target fragment of this primer set lies within the ≈650 bp 'DNA Barcoding' region of the COI, and therefore sequencing of the PCR product would enable identification of the species, or closest relative where a match was unavailable on the Barcode of Life Database (BOLD), available at www.boldsystems.org (Ratnasingham and Hebert 2007). Mini-barcodes lack the equivalent accuracy of the full ≈650 bp region, which provides >97% species-level specificity in arthropods (Hajibabaei *et al.* 2006a), but have previously been demonstrated to provided >90% species-level resolution in degraded DNA samples (Hajibabaei *et al.* 2006b; Meusnier *et al.* 2008).

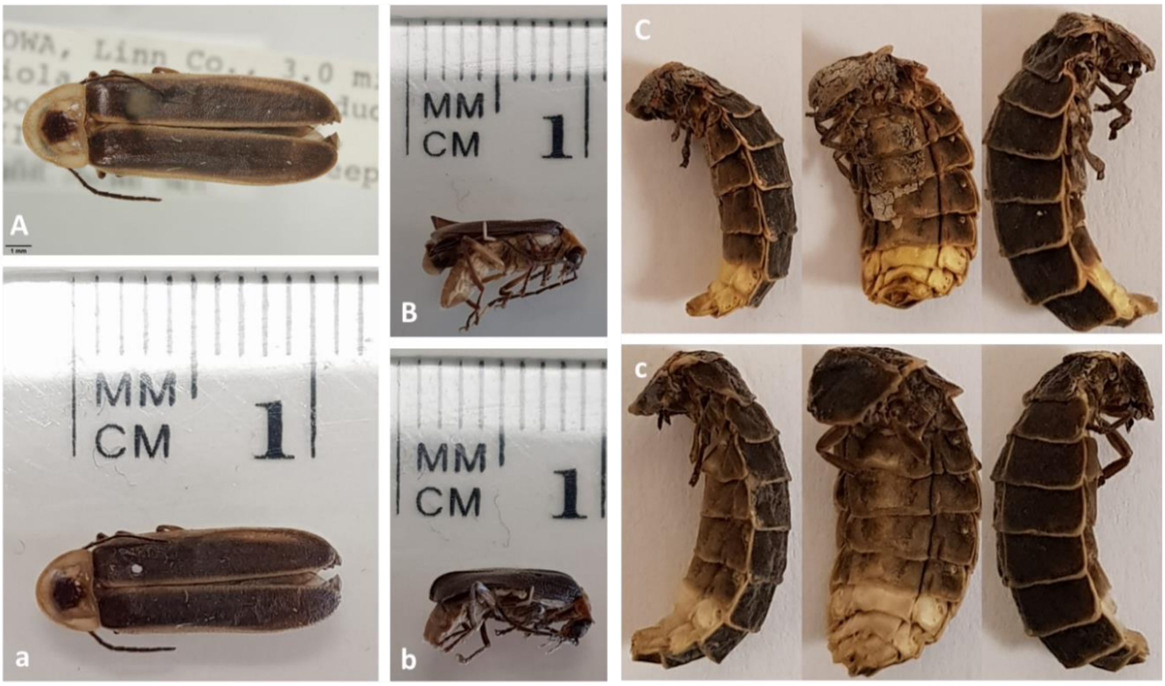
Amplification of the mini-barcode region was successful in the *Lnoc* gDNA extract, matching 98.11% with the *Lnoc* mitochondrion, partial genome (Appendices Table 9.1.). Amplification was not successful with the *Ppy* gDNA extract, which may be attributable to sequence divergence at the primer binding site and not due to the quality of the DNA extract itself.

Figure 3.1.


Remove from buffer and washed in ethanol
20 hours at 55˚C with gentle agitation
Specimen in fresh extraction buffer
Buffer retained for DNA purification

Overview of non-destructive DNA extraction. Simplified overview of the non-destructive DNA extraction method from Gilbert *et al.* (2007). Dried insect specimens are incubated at 55 ˚C overnight in a proteinase K based extraction buffer, with gentle agitation. Insect samples are then removed, and DNA purified from the used buffer by phenol chloroform DNA extraction. See Chapter 2 for further details.

Figure 3.2.



Coleoptera pre and post non-destructive DNA extraction. Photgraphs of *Photinus pyralis* – 1996 (A/a), an unknown Bornean Firefly – 1987 (B/b), and *Lampyris noctiluca* – 2006 (C/c). The top row uppercase images are prior to non-destructive DNA extraction, and the bottom row lowercase images are following non-detructive DNA extraction.

Table 3.1.

Sample	DNA Concentration	Average Fragment Size
Photinus pyralis, 1996	99.2 ng/µl	217 bp
Unknown, Borneo, 1987	ND	ND
Lampyris noctiluca, 2006	164 ng/µl	17273 bp

Quality of trial DNA extractions. DNA concentration and average fragment size of the genomic DNA extracts from collected samples of *Photinus pyralis,* an unknown Bornean Firefly, and *Lampyris noctiluca.* DNA concentration as measured by fluorometric quantification on the Qubit 4 Fluorometer (Thermo Fisher Scientific, MA, USA) and average fragment size of DNA extracts measured by analysis on the 4200 TapeStation System (Agilent, CA, USA). ND: not detected.

3.3.2. Detecting luciferase genes with CODEHOPs

3.3.2.1 CODEHOP primer design

As reported in Table 3.1., the average DNA fragment size of the *Ppy* specimen was 217 bp, suggesting this is representative of the quality of DNA that can be expected from further dry-preserved specimen DNA extractions. Similarly, as the mini-barcode amplification was unsuccessful, amplification might not occur in other samples. To mitigate this, a second more direct approach was sought that might be able to indicate whether an unidentified museum Lampyrid possessed a novel luciferase gene sequence. To enable this, a universal primer set was required which was capable of amplifying reliably a short fragment of the luciferase gene in a diverse range of firefly species.

Common approaches to designing universal primer sets include the use of consensus primers which target the consensus sequence from multi-gene alignment in a conserved region as primer binding sites, and degenerate primer pools that similarly target conserved regions but have several possible bases at positions of variation instead of the majority consensus. However, both approaches have problems regarding sensitivity and specificity. A third option is provided by a primer design approach which takes advantage of the strengths from both consensus and degenerate primers, and combines the two concepts to create Consensus-Degenerate Hybrid Oligonucleotide Primers (CODEHOPs) (Rose *et al.* 2003; Boyce *et al.* 2009). CODEHOP primes contain a short 3' degenerate core and a 5' consensus 'clamp', demonstrated in Figure 3.3.

The 3' degenerate core consists of a pool of sequences containing all possible codons for a 3 – 4 amino acid motif that is highly conserved in multiple sequence alignments of known members of a protein family. The 5' consensus clamp is a nondegenerate nucleotide sequence derived from a codon consensus across the aligned amino acid sequences flanking the conserved motif of the degenerate core (Staheli *et al.* 2011). In the initial rounds of amplification, a small proportion of the primer population will have sufficient complementarity in the 3' degenerate core for template annealing, to prime amplification and the production of amplicons incorporating the 5' consensus clamp sequence. In the subsequent rounds of PCR, full complementarity between the 5' consensus clamp in primers and amplicons allows the participation of all primers in the pool, and drives the exponential amplification of product, regardless of mismatches in the 3' degenerate core.

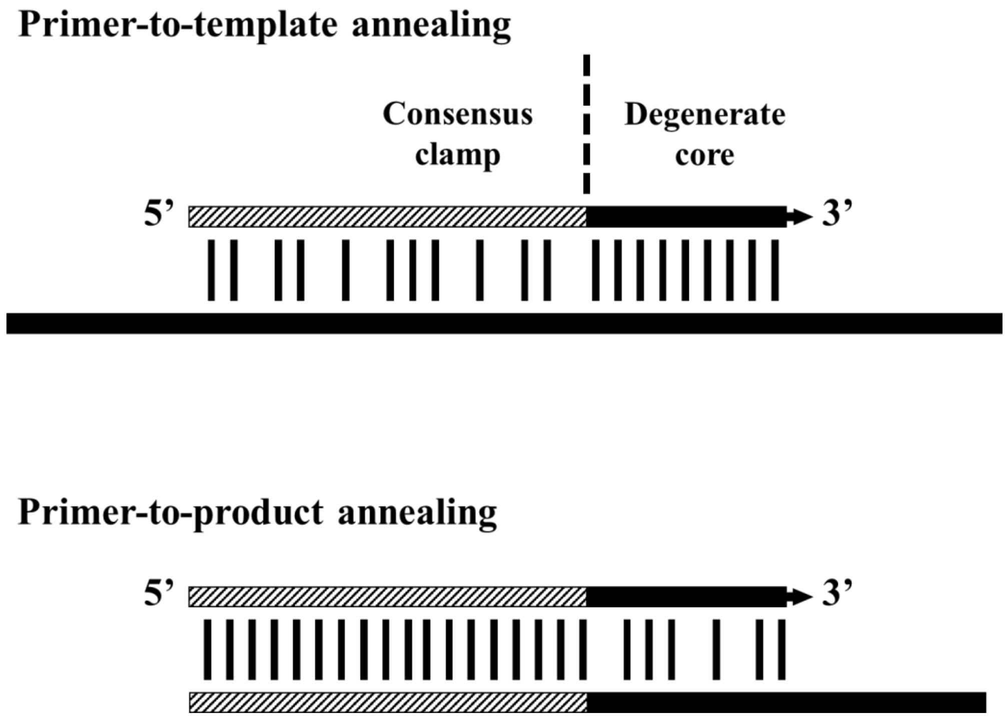
Using the CODEHOP primer design strategy eleven coleopteran luciferase protein sequences (Appendices Table 9.2.) were used to design a CODEHOP primer pair capable of targeting a short fragment of the luciferase gene. The resulting primer pair DKYD-F 8x > GYG-R 32x is detailed in Table 3.2. and targets a firefly luciferase fragment of ≈ 218 bp in length.

3.3.2.2 CODEHOP optimization

The CODEHOP primer set was initially tested by PCR amplification of the gDNA extracts previously obtained from *Lnoc* and *Ppy*. Electrophoretic analysis was conducted (Figure 3.4.) which verified that amplification had been successful for both samples. The amplification targets were gel extracted and Sanger sequenced (Sanger sequencing clipped results available in Appendices), and a BLASTn (available at https://blast.ncbi.nlm.nih.gov/Blast.cgi (Altschul *et al.* 1990)) search performed. The sequencing results of both samples matched their respective luciferase genes, with 99.28% percent identity match for *Lnoc,* and 94.67% for *Ppy* (Appendices Table 9.3.).

Although amplification had been successful for both samples, product band resolution in the electrophoretic analysis (Figure 3.4.) impeded the ease with which bands could be gel extracted for sequencing. In an attempt to optimize primer performance and improve band resolution, the effect of dimethylsulfoxide (DMSO) supplementation on *Lnoc* amplification was investigated by quantitative PCR and subsequent electrophoretic analysis of products (Figure 3.5.). The qPCR data in Figure 3.5A. indicated that each 1% increase in DMSO concentration correlated to an inhibition of amplification, which presented as a delay to the commencement of the exponential phase of amplification in each reaction. However, electrophoretic analysis of the qPCR products in Figure 3.5B. indicated that significantly improved resolution of the product bands could be achieved with the supplementation of 4% DMSO, without significant inhibition to the amplification in qPCR. From this result, all future use of the CODEHOP primer set was supplemented with 4% DMSO.

Figure 3.3.



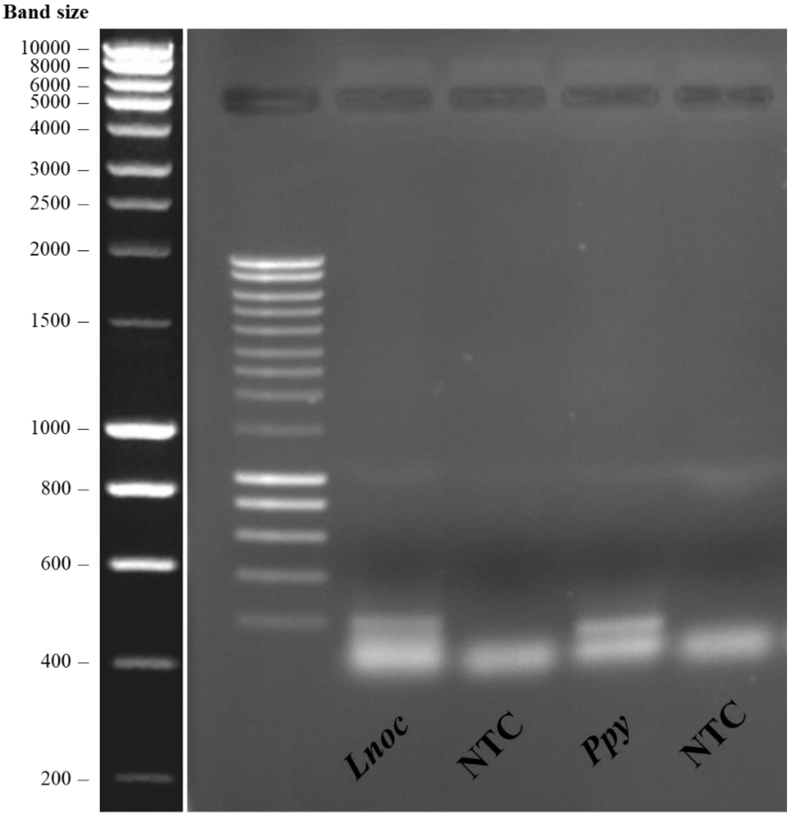
Overview of CODEHOP priming. In the initial rounds of amplification, a small proportion of the primer population will have sufficient complementarity in the degenerate core for template annealing, to prime amplification and the production of amplicons incorporating the consensus clamp sequence. In the subsequent rounds of PCR, full complementarity between the consensus clamp in primers and amplicons drives the exponential amplification of product, regardless of mismatches in the degenerate core. Graphic adapted from similar work by Rose *et al* (2003).

Table 3.2.

Name	**Orientation**	**Sequence**	**Product length** **≈ 218 bp**
DKYD-F 8x	Forward	5’-CTTCTTCGCCAAGTCCACGCTGGTCGAYAARTAYGA-3’	
GYG-R 32x	Reverse	5’-TGATAGCGGAGGTGGTCTCGGTCAGNCCRTANCC-3’	

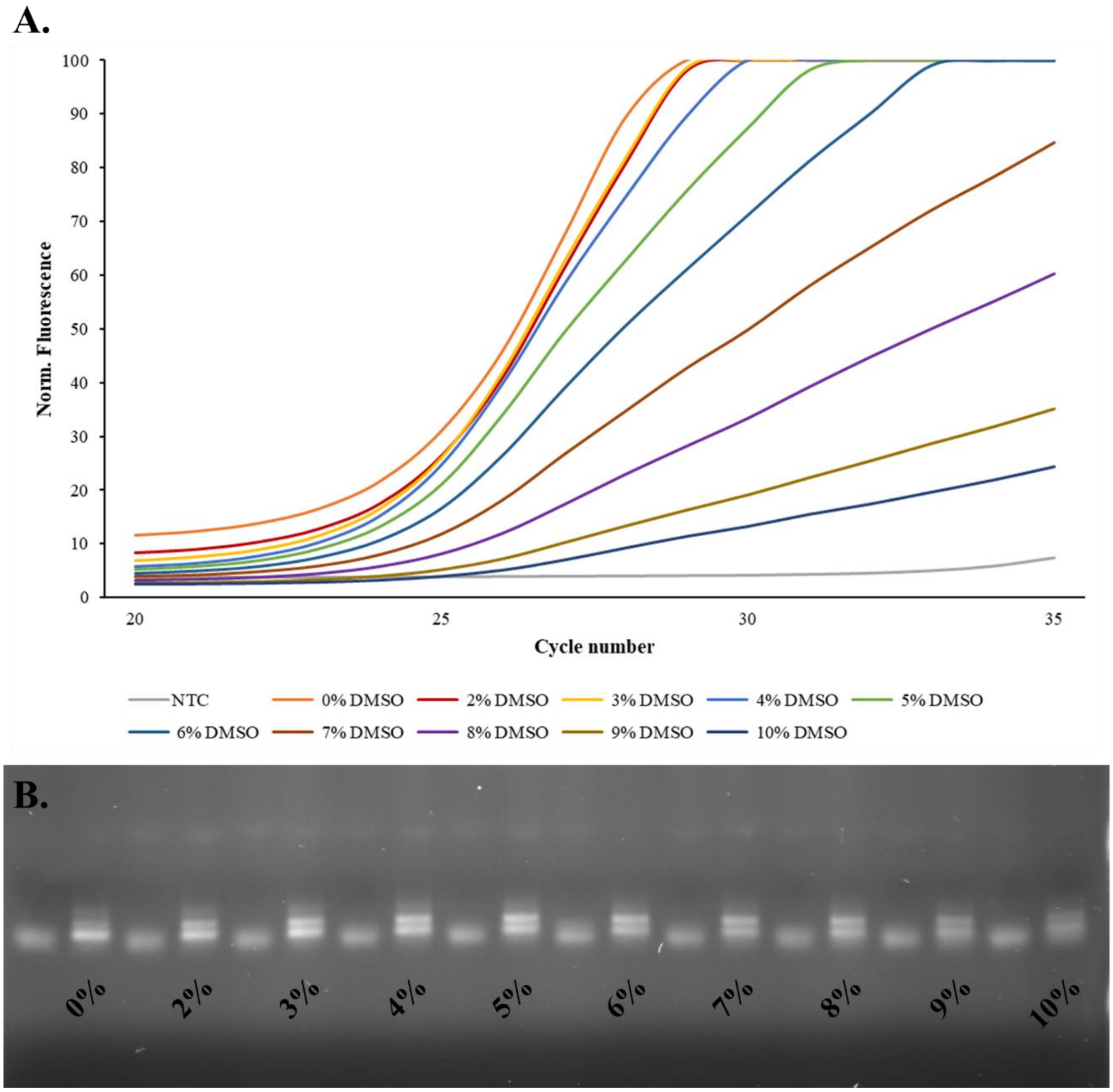
CODEHOP primer pairing. The forward (DKYD-F) and reverse (GYG-R) CODEHOP primers designed for this study. 8x and 32x refer to the degeneracy of the degenerate core sequence, shown in red. Product length is approximately 218 bp.

Figure 3.4.



Electrophoretic analysis of DKYD-F > GYG-R PCR products. Electrophoretic analysis of the amplicons from *Taq* DNA polymerase endpoint PCR of *Lnoc* and *Ppy* gDNA with CODEHOP primers DKYD and GYG**.** NTC's for each respective sample are included in the lane on the right. Ladder band sized in base pairs are displayed on the left.

Figure 3.5.



DMSO optimization of DKYD-F > GYG-R CODEHOP primers. (**A**) Quantitative PCR over 35 cycles displaying the effect of DMSO supplementation between 0 – 10% on amplification of *Lnoc* gDNA with CODEOP primers DKYD-F and GYG-R. Results have been truncated to begin at cycle 20 for improved visualisation. Separate NTC's were run for each DMSO concentration, and the average displayed. All reaction performed in triplicate, and averaged data presented. (**B**) Electrophoretic analysis of the amplicons from **A.** NTC's for each respective DMSO concentration are included in the lane on the left of each sample.

3.3.3. Cross-species affinity enrichment

3.3.3.1 Rational for enrichment

Whilst COI mini-barcoding and CODEHOP amplification are not directly related to the identification of the complete coding DNA sequence (CDS) for novel luciferase genes, their testing and optimization was pursued with the intent of developing tools capable of identifying whether the genomic DNA extracted from an unidentified firefly of the National Museum Cardiff collection possessed a novel luciferase gene sequence. This information would confirm whether investigation of a sample would be continued through the use of Next Generation Sequence (NGS) with its considerable associated expense.

Most NGS technologies do not rely on large DNA fragments and are designed to operate with short nucleotide sequences of 100 – 400 bp as sequencing template. To meet this requirement, most input samples are processed by sonication to produce fragment sizes within this range, as part of sequencing library preparation. As demonstrated by *Ppy* gDNA extraction (Table 3.1.), DNA extracted from historic specimens often provides fragment sizes within this range without the need for sonication during library preparation, making NGS technologies highly applicable to historic specimens. However, whilst many factors affect the final quality of NGS datasets, larger genomes can present additional issues as sequencing coverage correlates with total reads and the length of a given genome, i.e. 1,000,000 sequencing reads provides greater coverage on a small genome than a large genome. Using flow-cytometry, North American firefly genomes have previously been shown to range between 433 – 2572 Mb (Lower *et al.* 2017). Even when considering the smallest size of this range, a ≈ 2000 bp luciferase gene with intron sequences would account for only $4.6x10^{-4}$ of the total genome and would need to be completely mapped with sufficient coverage in order to identify the gene. With the input template originating as degraded gDNA extracts from dry-preserved Coleoptera, it is unlikely that *de novo* sequencing would be possible, and for the purposes of this study would be superfluous to requirements. Therefore, in order to increase the possibility of NGS sequencing recovering the luciferase gene region of interest, nucleotide sequence fragments which related to this region would first need to be enriched relative to all other alternative sequences present in the gDNA population.

3.3.3.2 Affinity enrichment design

Affinity enrichment of luciferase gene sequences was attempted using the gDNA extracts of *Ppy* and *Lnoc*, based on the principles of cross-species capture hybridisation outlined by Mason *et al* (2011). Libraries from both samples were initially prepared with the NEXTFLEX® Rapid DNA-Seq Kit for Illumina® Platforms (PerkinElmer, MA, USA) following manufacturer's instructions, with the omission of sonication for *Ppy* due to the short fragment size average revealed by TapeStation analysis (Table 3.1.). Performing the library preparation prior to enrichment was done to later enable the sequencing adapters to be targeted in PCR, and hence amplify the enriched targets to the concentration required for downstream Illumina 2x150 bp sequencing by the Cardiff University BIOSI Genomics Research Hub. As detailed in Chapter 2, biotinylated enrichment probes were constructed by performing PCR to incorporate biotin on the *Ppy* Fluc cDNA sequence, using Biotin-11-dUTP Solution (Thermo Fisher Scientific, MA, USA). A concentration of 25 µM biotin-11-dUTP was used together with 250 µM dTTP, resulting in ≈10% inclusion rate of biotin-11-dUTP in the place of dTTP. This gave the final *Ppy* probe products an estimated approximate biotinylated nucleotide inclusion rate of ≈2.5%.

With Illumina sequencing libraries prepared and biotinylated probes constructed, an enrichment strategy was attempted as outlined in Figure 3.6. With this method, Illumina libraries were first combined with the biotinylated probes of the *Ppy* Fluc gene and heated to denature all double stranded DNA molecules. The temperature was then reduced to enable the probes to selectively hybridise to library sequences with sufficient complementarity. As detailed in Chapter 2, a touch down approach to temperature reduction was taken in an attempt to initially capture targets of higher specificity, followed by the recovery of library targets with greater sequence divergence. Dynabeads™ M-280 Streptavidin (Invitrogen, CA, USA) were then used to capture the biotinylated probes together with their hybridised library targets, as streptavidin has a high natural binding affinity for biotin. Magnetic force was then used to pull the beads with bound probe-target constructs out of solution so that several round of buffer exchange could be performed to discard non-target sequences. The captured targets were then eluted from the streptavidin beads and biotin probes by NaOH alkaline denaturation. Amplification of the captured target with the NEXTFLEX® Rapid DNA-Seq Kit PCR reagents was then performed and purified using Agencourt AMPure XP SPRI beads (Beckman Coulter, CA, USA), to increase the concentration of targets to levels sufficient for Illumina sequencing and to remove any contaminating biotinylated *Ppy* probe sequences through the size selection process of SPRI beads purification (see Chapter 2 for full details of the enrichment process).

3.3.3.3 Validating enrichment

To validate the success of the enrichment process, quantitative PCR was performed using the CODEHOP primer set on the enriched and non-enriched *Lnoc* and *Ppy* Illumina libraries (Figure 3.7.). For both samples, the relative abundance of target luciferase gene sequence in the enriched libraries was significantly increased in comparison to their non-enriched equivalent libraries, as indicated by the number of cycles required for the exponential phase of amplification to commence. Sanger sequencing of the amplification products confirmed the correct sequence of *Lnoc* and *Ppy* luciferase, as earlier recorded in gDNA amplifications (Figure 3.4.).

Figure 3.6.

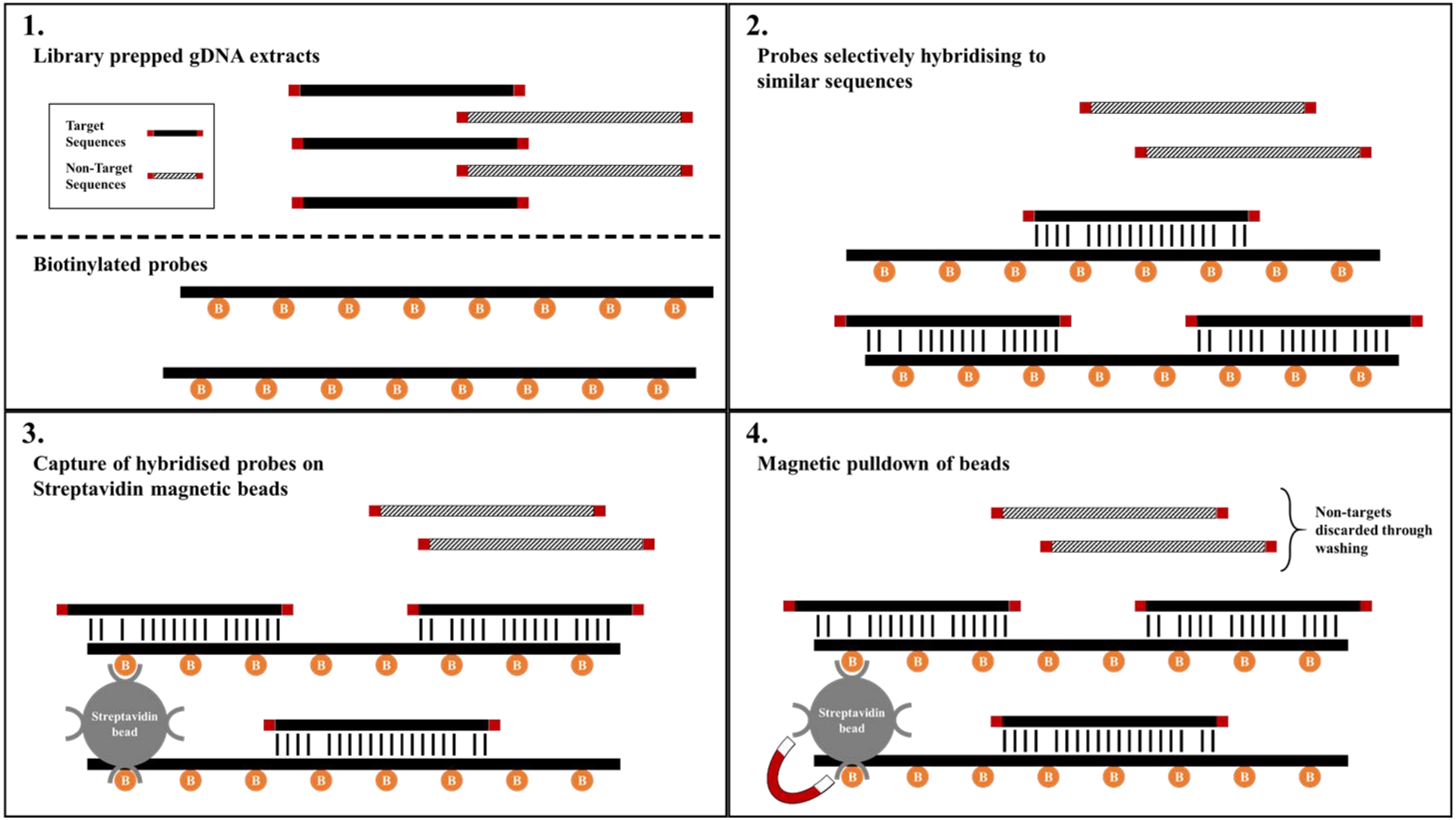
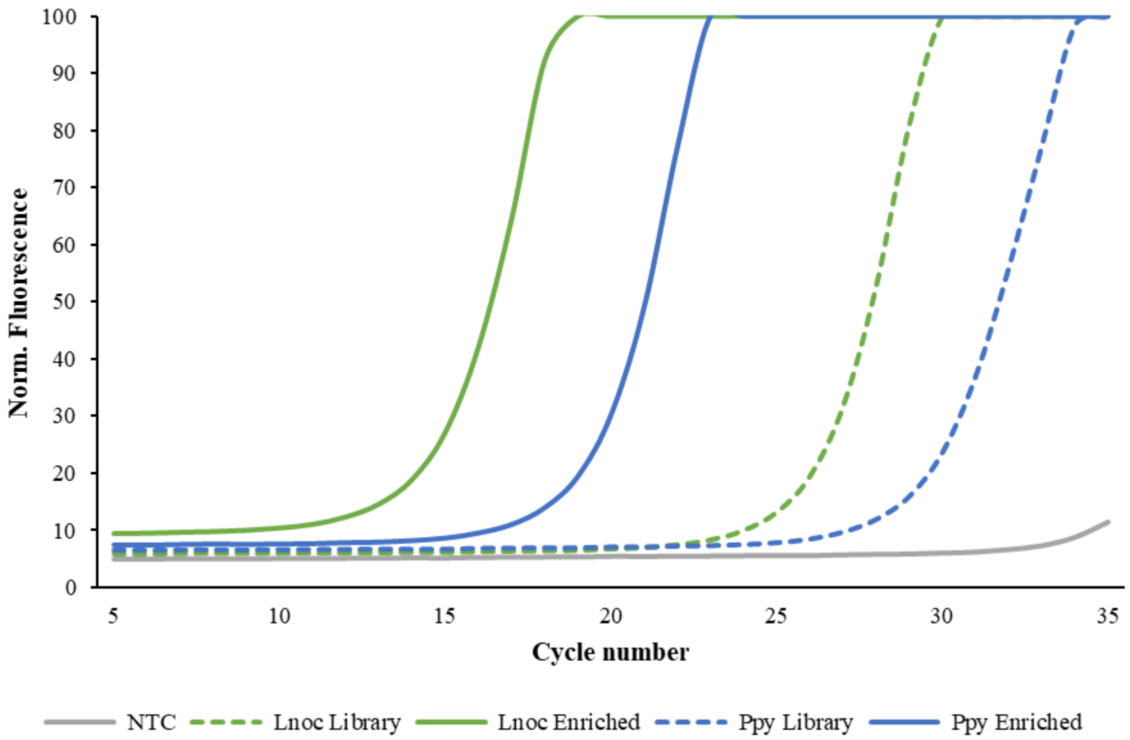

1.
Library prepped gDNA extracts
Target Sequences
Non-Target Sequences
Biotinylated probes
B
2.
Probes selectively hybridising to similar sequences
3.
Capture of hybridised probes on Streptavidin magnetic beads
Streptavidin bead
4.
Magnetic pulldown of beads
Non-targets discarded through washing

Diagram of affinity enrichment process. A method for the enrichment of DNA sequences homologous to the *Ppy* Fluc gene in Illumina libraries of bioluminescent Coleoptera. (**1.**) Library prepared from genomic DNA is incubated with biotinylated probes of the *Ppy* Fluc gene. Positioning of biotin molecules in the figure is for illustrative purposes only and does not accurately reflect biotin inclusion rate. (**2.**) The *Ppy* biotin probes selectively hybridise to sequences with sufficient sequence complementarity. Non-target sequences from the library remain unbound in solution. (**3.**) Magnetic streptavidin coated beads are introduced which have high affinity for biotin. Biotin probes with hybridised library sequences bind to the streptavidin bead coating. (**4.**) Magnetic force is used to pull the streptavidin beads with bound probes and target out of solution, to the bottom/side of the tube. Several round of buffer exchange are then used to discard non-target sequences. Enriched target sequences can then be eluted from probe bound beads. See Chapter 2 for further details.

Figure 3.7.



Verifying enrichment in *Lnoc* and *Ppy*. Quantitative PCR over 35 cycles assessing the relative abundance of luciferase gene sequences in the *Lnoc* and *Ppy* NEXTFLEX libraries before and after affinity enrichment, as indicated by amplification with CODEHOP primers DKYD-F and GYG-R targeting the luciferase gene. Reactions performed in the presence of 4% DMSO. Results have been truncated to begin at cycle 5 for improved visualisation. All reactions performed in triplicate, and averaged data presented.

3.3.4. DNA extraction and enrichment from unidentified Lampyrids

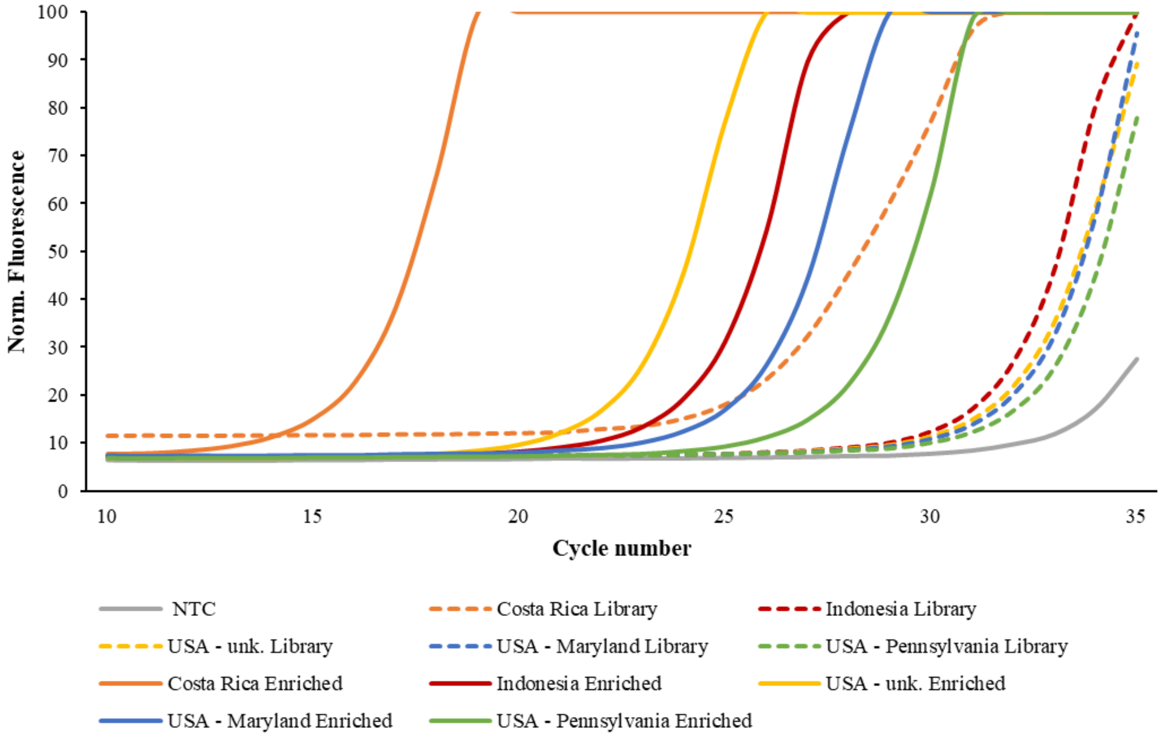
From the previous work carried out using *Ppy* and *Lnoc* samples, a pipeline has been established which enables the non-destructive extraction of genomic DNA from dry-preserved Lampyrids, mini-barcoding to attempt species identification, detection of luciferase gene content with CODEHOP primers, and enrichment of luciferase gene fragments with cross-species biotinylated probes. Five unidentified fireflies from the National Museum Wales Coleoptera collection were selected and processed through the entire pipeline. The origins of these five fireflies were Costa Rica collected 2012, Indonesia collected 1985, USA – unknown State collected 2013, USA – Maryland collected 2015, and USA – Pennsylvania collected 2015. The collection date of the specimens from Costa Rica and Indonesia precedes the implementation of the Nagoya Protocol, whilst the other three samples originated in the USA, which is not a signatory to the Nagoya Protocol.

DNA extracts from the five specimens were prepared for Illumina sequencing with the NEXTFLEX® Rapid DNA-Seq Kit for Illumina® Platforms (PerkinElmer, MA, USA) by Cardiff University BIOSI Genomics Research Hub, with the omission of fragment sonication. The average fragment size of the DNA extracts was measured before and after library preparation by analysis on the 4200 TapeStation System (Agilent, CA, USA) (Appendices Table 9.4.). The Costa Rica firefly had the largest average fragment size before library preparation of 881 bp, whilst the remaining four samples were all measured at <200 bp. Following library preparation, average fragment size recorded for Costa Rica was reduced to 323 bp, and increased to between 250 – 300 bp for all other libraries. Although sonication was omitted, size selection by SPRI beads purification is likely to be the source for both the increases and decreases observed in library fragment sizes relative to the input gDNA.

Amplification of the COI mini-barcode was attempted for all five firefly libraries, and was successful for all but the Indonesia specimen (Appendices Table 9.1.). Sanger sequencing of the mini-barcodes and subsequent enquiry by BLAST indicated that the Costa Rica firefly shared 92.31% sequence identity to the *Photinus australis* COI gene. The USA – Maryland specimen was identified as a 95.24% match to *Photinus interdius*, whilst both USA – unk. (unk. denotes ‘unknown State’) and USA – Pennsylvania shared >98% identity to species of *Lucidota* fireflies. The top BLAST match of USA – Pennsylvania was specified as *Lucidota atra*, whilst USA – unk. had matched to an unspecified species of *Lucidota*. Further analysis of the BLAST identified *Lucidota* sequences revealed they shared 99.66% sequence identity (one mismatch), suggesting the unspecified *Lucidota* sequence was likely to also be *Lucidota atra*. Inspection of the mini-barcodes from USA – Pennsylvania and USA – unk. indicated that they shared 97.80% sequence identity.

Enrichment was performed using the method outlined in Figure 3.6., using the biotinylated amplification of *Ppy* Fluc as probe, and validation of enrichment was subsequently performed using the -CODEHOP primer set targeting the luciferase fragment using quantitative PCR (Figure 3.8.). The qPCR data suggests an increase in the relative abundance of luciferase sequence fragments in all enriched libraries relative to the non-enriched equivalents, with the most significant enrichment being observed in the Costa Rica library. However, Sanger sequencing of the CODEHOP amplification products was less successful than in *Ppy* and *Lnoc* (Appendices Table 9.3.). BLAST analysis indicated that all five sequences were best matched with *Photinus pyralis*, with the least percent identity being present in the Costa Rica sample at 86.83%, whilst the four remaining samples ranged between 91.30 – 93.16%. These findings would not alone be unusual, but further inspection of the 5 sequences revealed a conserved 49 bp region which by BLAST was found to match with *Ppy* Fluc at 87.76%. The variation in percent identity from *Ppy* Fluc suggests that probe contamination as a source of contamination can be ruled out. Additionally, where Sanger sequencing was successful in the non-enriched libraries and genomic DNA extracts, the sequences were conserved compared to the enriched sequences (disregarding terminal variation). Furthermore, all NTCs performed in qPCR were clear of amplification products, indicating that contamination did not occur during qPCR set up. It remains unclear what was the original source of the conserved region observed in the CODEHOP amplification products. Regardless, the mini-barcode amplifications were sufficient to verify that the five fireflies possessed novel luciferase gene sequences, as no luciferase gene sequences are available for any of the highest percent identity matches identified by BLAST, and therefore all five enriched libraries were taken forward to Illumina sequencing.

Figure 3.8.



Verifying enrichment in museum Coleoptera sequencing libraries. Quantitative PCR over 35 cycles assessing the relative abundance of luciferase gene sequences in the Museum Coleoptera NEXTFLEX libraries before and after affinity enrichment, as indicated by amplification with CODEHOP primers DKYD-F and GYG-R targeting the luciferase gene. Reactions performed in the presence of 4% DMSO. Results have been truncated to begin at cycle 10 for improved visualisation. All reaction performed in triplicate, and averaged data presented.

3.3.5. Identification of a novel luciferase

3.3.5.1 Sequencing and bioinformatic processing

The five enriched Illumina libraries from unidentified fireflies were sequenced as 150 bp paired-end reads on the Illumina NextSeq 500 Sequencer (Illumina, Inc., CA, USA) in the Cardiff University BIOSI Genomics Research Hub. Bioinformatic analysis was conducted using the Cardiff University School of Biosciences Biocomputing hubs High Throughput Computing cluster (YSGO), which facilitates the processing and storage of information generated by data-intensive research within the School of Biosciences and makes available pre-prepared modules of common bioinformatics tools/packages. Illumina sequencing data was submitted to the NCBI Sequence Read Archive (SRA), and can be accessed using the SRA Run Selector (available at https://www.ncbi.nlm.nih.gov/Traces/study/) under the BioProject accession PRJNA802557. Direct accessions to individual libraries are available in Appendices Table 9.7.

Using this computing environment, all five sequenced libraries were initially processed with Trimmomatic (available at https://github.com/usadellab/Trimmomatic) to perform quality trimming and adapter clipping (Bolger *et al.* 2014). Reports on the quality of the trimmed sequenced data were subsequently generated using FastQC (available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), an overview of which can be found in Appendices Table 9.5. For all paired trimmed datasets, no sequences were flagged as poor quality and read lengths were as expected (≈150 bp). Of the enriched libraries, the Costa Rica paired trimmed dataset possessed the greatest total paired sequences of 114752, which was ≈2x the total paired sequences of any of the other enriched library paired trimmed datasets. Annotated example scripts for the execution of Trimmomatic and FastQC are available in Appendices Figures 9.1 – 9.2.

With the sequencing datasets trimmed and assessed for quality, the bioinformatic search for novel luciferase genes could be commenced. For this purpose, a bash master script was designed capable of receiving the trimmed dataset FASTQ format file as input, and outputting assembled contigs of reads comprising luciferase genes (annotated bash master script is available in Appendices Figure 9.3.). This script made use of three common bioinformatics packages: Bowtie2, SAMtools, and SPAdes.

Bowtie2 is a fast and sensitive tool designed for the alignment of short sequencing reads to larger reference sequences, or genomes. Bowtie2 outputs alignments in SAM (Sequence

Alignment/Map) format, which is interoperable with many other downstream tools, which has enabled Bowtie2 to become a common first step in many bioinformatics pipelines (Langmead and Salzberg 2012). Bowtie2 is available at https://sourceforge.net/projects/bowtie-bio/.

SAMtools is a toolkit of utilities for post-processing alignments in SAM format, with functions including the ability to convert SAM files to BAM (its binary counterpart with which computers work better), and sort, index, and filter alignments. Alignments in SAM/BAM occur in the order that the sequences are present in the original FASTQ input files. These basic SAMtools functions allow BAM alignments to be sorted instead into "genome order", indexed by genomic coordinates, and region of interest extracted which relate to known coordinates (Li *et al.* 2009). SAMtools is available at http://samtools.sourceforge.net.

SPAdes is a De Bruijn graph assembler used to perform *de novo* assembly of overlapping sequence reads into larger consensus segments known as contigs, which SPAdes refers to as Nodes. The output Nodes are numbered in order of largest to smallest and are reported alongside a coverage score which indicates the mean number of times sequenced nucleotide bases were mapped across the consensus region that comprises an individual Node (Bankevich *et al.* 2012). SPAdes is available at https://github.com/ablab/spades.

The trimmed FASTQ datasets from enriched firefly sequencing were input into the bioinformatics pipeline of the master script (Appendices Figure 9.3.). Using Bowtie2, the reads from each data set were individually aligned to two firefly reference genomes, *Photinus pyralis* and *Aquatica lateralis,* and one click beetle genome, *Ignelater luminosus.* Reference genomes Alat1.4, Ppyr1.4, and Ilumi1.3 are available at http://fireflybase.org/ (Fallon *et al.* 2018). The alignment SAM files were then converted by SAMtools into BAM files with the alignments sorted by the genome order. The sorted BAM files were then indexed, and reads which aligned to the genomic coordinates of the luciferase gene region extracted into a separate file. These regions were identified by using the Firefly BLAST server (http://blast.fireflybase.org/) with the *Photinus pyralis* luciferase gene complete cds as input query, for each genome. Reads mapped to the extracted regions of the three reference genome were then converted back into FASTQ formatted files of forward reads (R1), reverse reads (R2) and unpaired singletons. The sets of output files from each reference genome were then grouped such that each enriched firefly dataset had a single combined extract file for each of R1, R2, and singletons. Finally, these extracted FASTQ files were assembled into contigs by SPAdes.

3.3.5.2 Output analysis

The bioinformatic pipeline was successful for only one of the enriched sequenced datasets, Costa Rica, 2012 which had three Nodes output by SPAdes (Table 3.3.). To investigate what had prevented SPAdes from assembling contigs from the remaining four datasets, the extracted FASTQ output files from SAMtools were reviewed for each dataset, a summary of which can be found in Appendices Table 9.6. Immediately apparent for all enriched library datasets was that reads had only successful mapped to the luciferase gene region of the Ppyr1.4 reference genome, and none to the extract regions of Alat1.4 or Ilumi1.3. Costa Rica, 2012 had 182 paired reads and 889 singletons, with the second most mapped reads being only 7 paired and 4 singletons for USA – unk., 2013. The reads for the four unsuccessful datasets were investigated with BLAST, and all proved to be high similarity matches with *Photinus pyralis* luciferase of varied percent identity. Whether the enrichment process or the Bowtie2 alignment failed is unclear. An attempt to assemble the trimmed FASTQ datasets in SPAdes without the Bowtie2 and SAMtools processing was also unsuccessful for all libraries, including Costa Rica, 2012. Additionally, modifications to the master script to substitute the mapping to reference genomes with mapping to a collection of firefly luciferase gene DNA sequences proved to be incapable of mapping any reads for all libraries, including Costa Rica, 2012.

Additionally, a non-enriched Costa Rica, 2012 library was sequenced and processed with the same bioinformatic pipeline. Although this non-enriched library had ≈4x the total paired reads of its enriched equivalent as input for the master script (Appendices Table 9.5.), only 2 paired sequences successfully mapped to the extracted region of reference genome Ppyr1.4 (Appendices Table 9.6.), indicating that although the other libraries failed, the enrichment process was critical in recovering luciferase mapping reads and subsequently contigs from Costa Rica, 2012.

3.3.5.3 Costa Rica Fluc reconstruction

In total, three Nodes were assembled by SPAdes for the enriched Costa Rica, 2012 dataset. The length and coverage of these Nodes are detailed in Table 3.3. along with the accession and shared identity percentage of the top BLASTn match. The three nodes ranged in length between 213 – 1265 bp, and all possessed ≈90% shared identity with voucher specimens of *Ppy* luciferase complete CDS. Individual alignments of all three Nodes to the *Ppy* luciferase complete CDS were attempted in CLC sequence viewer (Qiagen, Venlo, Netherlands), but was only successful for Node_2 (Figure 3.9.), which aligned from base 1152 of the *Ppy* Fluc CDS onward, extending beyond the final base of 2092.

As Node_1 and Node_3 had failed to align to the *Ppy* Fluc CDS despite a high degree of sequence similarity reported by BLAST, reverse complements of both Nodes were constructed (Node_1RC and Node_3RC) and alignment reattempted. Alignment of the reverse complements was successful for both Node_1RC (Figure 3.10.) which aligned from the start of the reference *Ppy* Fluc CDS up to base 1228, and Node_3RC (Figure 3.11.) which aligned between bases 1152 and 1364.

Node_3RC presents as a region of overlap between the three Node sequences. This common region was aligned in Figure 3.12., which found that although Node_1RC and Node_3RC were identical, three mismatches with Node_2 were present. The role of these mismatched positions were investigated in the *Ppy* Fluc CDS which revealed that the first two mismatches occurred in the 4th intron, which would have no influence on the protein sequence. The location of the third mismatch was identified as within the 5th exon and would substitute the 337th codon from CGA to CGC, which are both conserved for arginine, meaning that either of the mismatched bases at this position would not change the final protein sequence. With this information, a consensus sequence was constructed from Node_2 and Node_1RC using the sequence of Node_2 in the region of overlap due to the higher coverage score of 74 reported by SPAdes in Table 3.3, even though Node_1RC and Node_3RC were in agreement. The 2304 bp Node consensus sequence can be viewed in alignment to the *Ppy* Fluc complete CDS in Figure 3.13.

With the construction of a Node consensus complete, an alignment with the cDNA sequence of *Ppy* Fluc was performed in Figure 3.14. to identify the corresponding regions in the Node consensus which are predicted to comprise the seven exons of the Costa Rica Fluc. These seven predicted exon sequences were extracted and realigned to the Node consensus in

Figure 3.15. to visualise the layout of predicted exon and intron sequences in the complete CDS of Costa Rica Fluc. Finally, a translation of the full predicted exon sequence was aligned with the amino acid sequence of *Ppy* Fluc (Figure 3.16.), suggesting that the proposed Costa Rica Fluc would be a 550AA protein which shares 93.45% sequence identity with *Ppy* Fluc.

3.3.5.4 Bioluminescence from a museum Coleopteran

With a Costa Rica Fluc protein sequence proposed from the SPAdes output Nodes, functional verification was needed to establish whether the speculative enzyme could catalyse the luciferase reaction and produce a bioluminescence signal. An *E. coli* codon optimized gene was synthesized and incorporated into the pET16b plasmid to enable transformation of *E. coli* BL21 (DE3) competent cells. The colonies which arose from the transformation were induced for protein production with IPTG prior to screening with *D*-luciferin citrate spray in the PhotonIMAGER Optima.

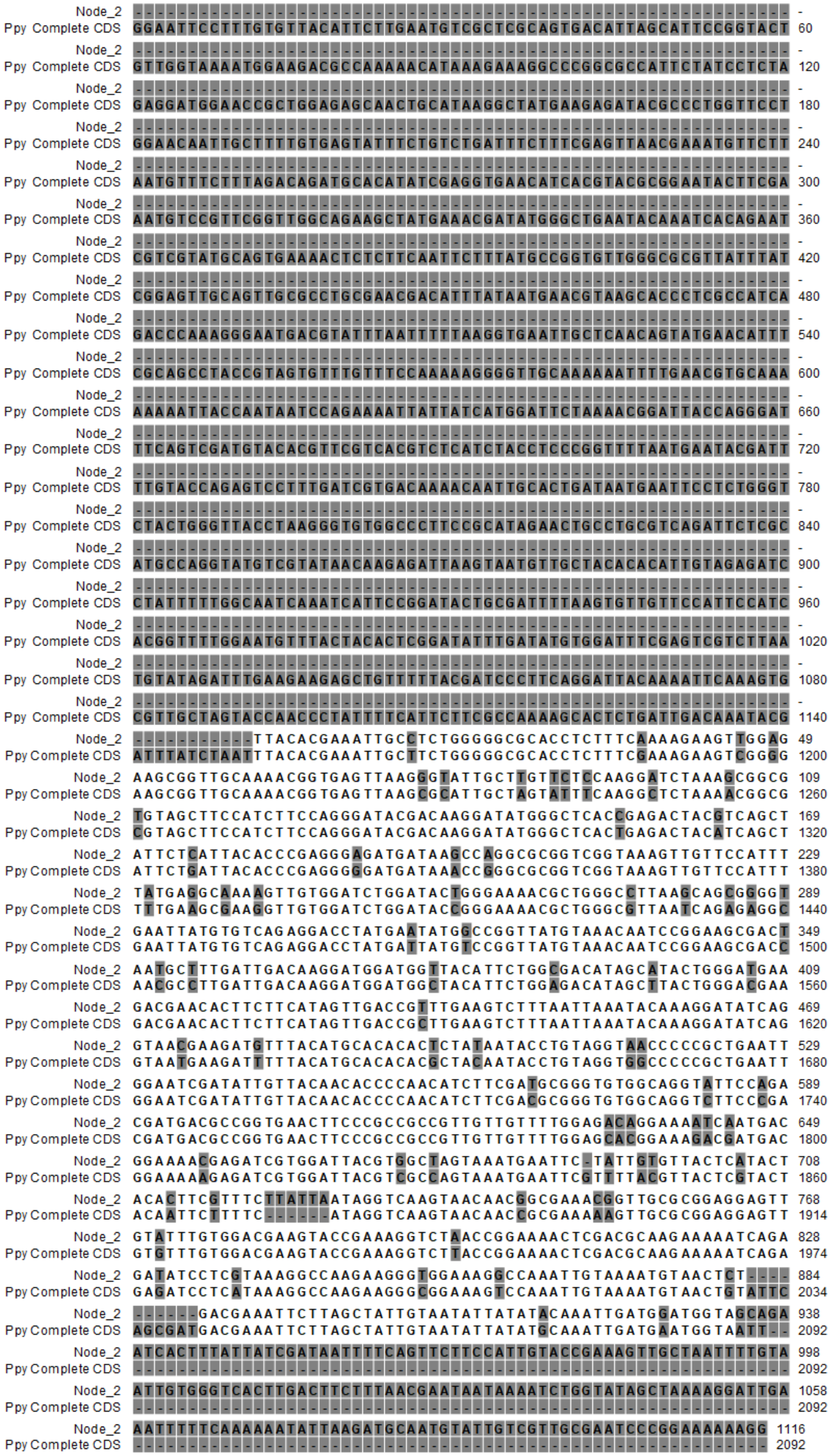
The bioluminescence data acquired during the screens was analysed in the M3 Vision software package, available under license from https://biospacelab.com. The bioluminescence signal of the primary screen as interpreted by M3 Vision is displayed in Figure 3.17A. A secondary screen was conducted using colonies picked from this plate and is shown in Figure 3.17B. Both the primary and secondary screen confirmed that the proposed Costa Rica Fluc was a functioning coleopteran luciferase, capable of producing a strong bioluminescence signal.

Table 3.3.

	Contig Length	**Contig Coverage**	**Top BLASTn match Accession**	**Shared Identity Percentage**
Node_1	1265 bp	5.790404	MH759023.1	89.96%
Node_2	1116 bp	74.004812	MH759210.1	90.25%
Node_3	213 bp	34.544118	MH759210.1	91.55%

Costa Rica firefly sequencing assembled Nodes. Details of the three contigs assembled from the reads extracted through bioinformatics interrogation of the enriched Costa Rica 2012 firefly NEXTFLEX library sequencing data. Contig length and coverage scores are shown, along with the highest percent identity sequence match identified via BLASTn (available at https://blast.ncbi.nlm.nih.gov/Blast.cgi). The top BLASTn match for all contigs are the complete cds of luciferase genes from voucher specimens of *Photinus pyralis.*

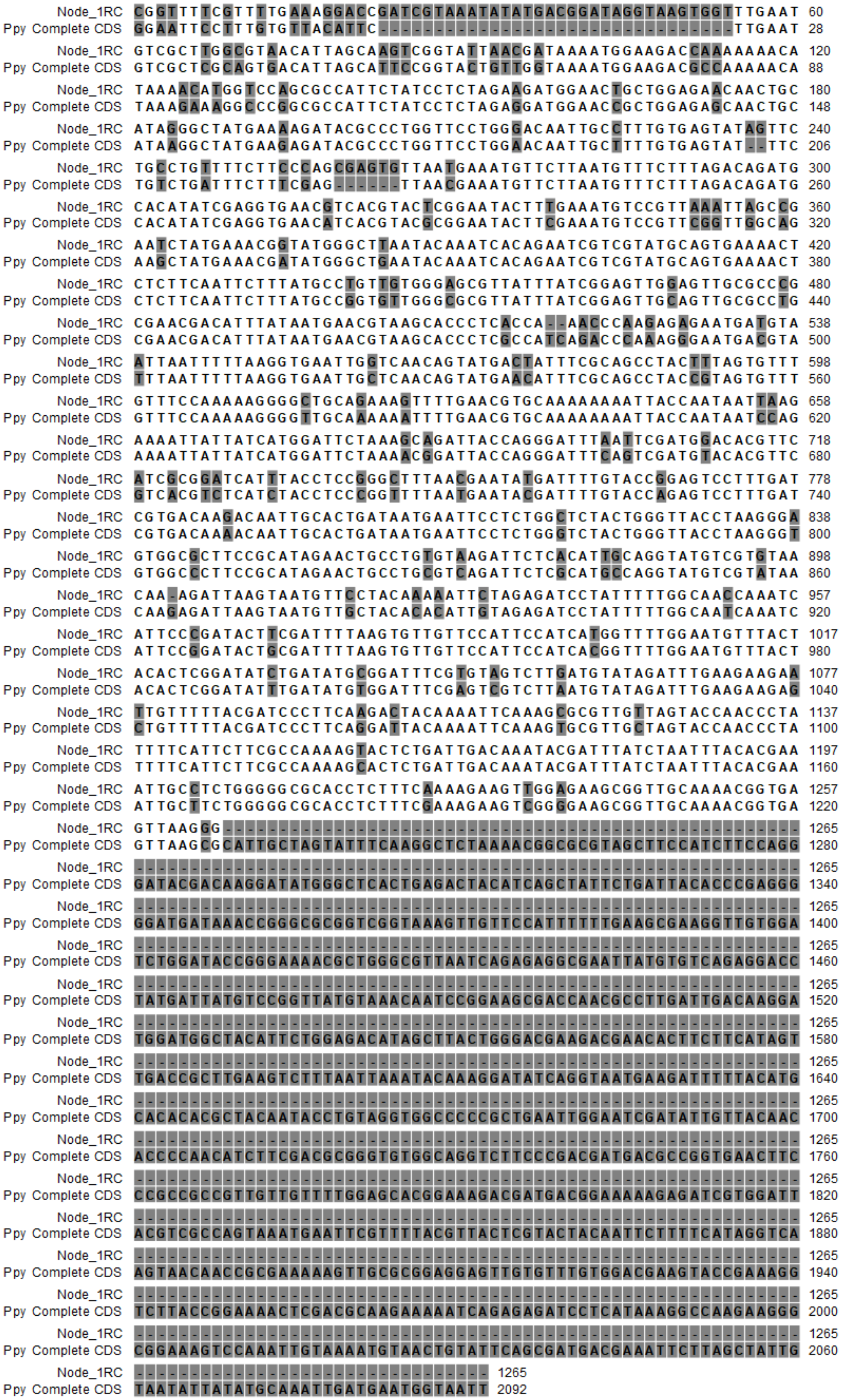
Figure 3.9.



```
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  GGAATTCCTTTGTGTTACATTCTTGAATGTCGCTCGCAGTGACATTAGCATTCCGGTACT 60
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  GTTGGTAAAATGGAAGACGCCAAAAACATAAAGAAAGGCCCGGCGCCATTCTATCCTCTA 120
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  GAGGATGGAACCGCTGGAGAGCAACTGCATAAGGCTATGAAGAGATACGCCCTGGTTCCT 180
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  GGAACAATTGCTTTTGTGAGTATTTCTGTCTGATTTCTTTCGAGTTAACGAAATGTTCTT 240
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  AATGTTTCTTTAGACAGATGCACATATCGAGGTGAACATCACGTACGCGGAATACTTCGA 300
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  AATGTCCGTTCGGTTGGCAGAAGCTATGAAACGATATGGGCTGAATACAAATCACAGAAT 360
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  CGTCGTATGCAGTGAAAACTCTCTTCAATTCTTTATGCCGGTGTTGGGCGCGTTATTTAT 420
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  CGGAGTTGCAGTTGCGCCTGCGAACGACATTTATAATGAACGTAAGCACCCTCGCCATCA 480
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  GACCCAAAGGGAATGACGTATTTAATTTTTAAGGTGAATTGCTCAACAGTATGAACATTT 540
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  CGCAGCCTACCGTAGTGTTTGTTTCCAAAAAGGGGTTGCAAAAAATTTTGAACGTGCAAA 600
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  AAAAATTACCAATAATCCAGAAAATTATTATCATGGATTCTAAAACGGATTACCAGGGAT 660
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  TTCAGTCGATGTACACGTTCGTCACGTCTCATCTACCTCCCGGTTTTAATGAATACGATT 720
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  TTGTACCAGAGTCCTTTGATCGTGACAAAACAATTGCACTGATAATGAATTCCTCTGGGT 780
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  CTACTGGGTTACCTAAGGGTGTGGCCCTTCCGCATAGAACTGCCTGCGTCAGATTCTCGC 840
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  ATGCCAGGTATGTCGTATAACAAGAGATTAAGTAATGTTGCTACACACATTGTAGAGATC 900
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  CTATTTTTGGCAATCAAATCATTCCGGATACTGCGATTTTAAGTGTTGTTCCATTCCATC 960
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  ACGGTTTTGGAATGTTTACTACACTCGGATATTTGATATGTGGATTTCGAGTCGTCTTAA 1020
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  TGTATAGATTTGAAGAAGAGCTGTTTTTACGATCCCTTCAGGATTACAAAATTCAAAGTG 1080
          Node_2  ------------------------------------------------------------ -
Ppy Complete CDS  CGTTGCTAGTACCAACCCTATTTTCATTCTTCGCCAAAAGCACTCTGATTGACAAATACG 1140
          Node_2  -----------TTACACGAAATTGCCTCTGGGGGCGCACCTCTTTCAAAAGAAGTTGGAG 49
Ppy Complete CDS  ATTTATCTAATTTACACGAAATTGCTTCTGGGGGCGCACCTCTTTCGAAAGAAGTCGGGG 1200
          Node_2  AAGCGGTTGCAAAACGGTGAGTTAAGGGTATTGCTTGTTCTCCAAGGATCTAAAGCGGCG 109
Ppy Complete CDS  AAGCGGTTGCAAAACGGTGAGTTAAGCGCATTGCTAGTATTTCAAGGCTCTAAAACGGCG 1260
          Node_2  TGTAGCTTCCATCTTCCAGGGATACGACAAGGATATGGGCTCACCGAGACTACGTCAGCT 169
Ppy Complete CDS  CGTAGCTTCCATCTTCCAGGGATACGACAAGGATATGGGCTCACTGAGACTACATCAGCT 1320
          Node_2  ATTCTCATTACACCCGAGGGAGATGATAAGCCAGGCGCGGTCGGTAAAGTTGTTCCATTT 229
Ppy Complete CDS  ATTCTGATTACACCCGAGGGGGATGATAAACCGGGCGCGGTCGGTAAAGTTGTTCCATTT 1380
          Node_2  TATGAGGCAAAAGTTGTGGATCTGGATACTGGGAAAACGCTGGGCCTTAAGCAGCGGGGT 289
Ppy Complete CDS  TTTGAAGCGAAGGTTGTGGATCTGGATACCGGGAAAACGCTGGGCGTTAATCAGAGAGGC 1440
          Node_2  GAATTATGTGTCAGAGGACCTATGAATATGGCCGGTTATGTAAACAATCCGGAAGCGACT 349
Ppy Complete CDS  GAATTATGTGTCAGAGGACCTATGATTATGTCCGGTTATGTAAACAATCCGGAAGCGACC 1500
          Node_2  AATGCTTTGATTGACAAGGATGGATGGTTACATTCTGGCGACATAGCATACTGGGATGAA 409
Ppy Complete CDS  AACGCCTTGATTGACAAGGATGGATGGCTACATTCTGGAGACATAGCTTACTGGGACGAA 1560
          Node_2  GACGAACACTTCTTCATAGTTGACCGTTTGAAGTCTTTAATTAAATACAAAGGATATCAG 469
Ppy Complete CDS  GACGAACACTTCTTCATAGTTGACCGCTTGAAGTCTTTAATTAAATACAAAGGATATCAG 1620
          Node_2  GTAACGAAGATGTTTACATGCACACACTCTATAATACCTGTAGGTAACCCCCGCTGAATT 529
Ppy Complete CDS  GTAATGAAGATTTTTACATGCACACACGCTACAATACCTGTAGGTGGCCCCCGCTGAATT 1680
          Node_2  GGAATCGATATTGTTACAACACCCCAACATCTTCGATGCGGGTGTGGCAGGTATTCCAGA 589
Ppy Complete CDS  GGAATCGATATTGTTACAACACCCCAACATCTTCGACGCGGGTGTGGCAGGTCTTCCCGA 1740
          Node_2  CGATGACGCCGGTGAACTTCCCGCCGCCGTTGTTGTTTTGGAGACAGGAAAATCAATGAC 649
Ppy Complete CDS  CGATGACGCCGGTGAACTTCCCGCCGCCGTTGTTGTTTTGGAGCACGGAAAGACGATGAC 1800
          Node_2  GGAAAACGAGATCGTGGATTACGTGGCTAGTAAATGAATTC-TATTGTGTTACTCATACT 708
Ppy Complete CDS  GGAAAAAGAGATCGTGGATTACGTCGCCAGTAAATGAATTCGTTTTACGTTACTCGTACT 1860
          Node_2  ACACTTCGTTTCTTATTAATAGGTCAAGTAACAACGGCGAAACGGTTGCGCGGAGGAGTT 768
Ppy Complete CDS  ACAATTCTTTTC------ATAGGTCAAGTAACAACCGCGAAAAAGTTGCGCGGAGGAGTT 1914
          Node_2  GTATTTGTGGACGAAGTACCGAAAGGTCTAACCGGAAAACTCGACGCAAGAAAAATCAGA 828
Ppy Complete CDS  GTGTTTGTGGACGAAGTACCGAAAGGTCTTACCGGAAAACTCGACGCAAGAAAAATCAGA 1974
          Node_2  GATATCCTCGTAAAGGCCAAGAAGGGTGGAAAGGCCAAATTGTAAAATGTAACTCT---- 884
Ppy Complete CDS  GAGATCCTCATAAAGGCCAAGAAGGGCGGAAAGTCCAAATTGTAAAATGTAACTGTATTC 2034
          Node_2  ------GACGAAATTCTTAGCTATTGTAATATTATATACAAATTGATGGATGGTAGCAGA 938
Ppy Complete CDS  AGCGATGACGAAATTCTTAGCTATTGTAATATTATATGCAAATTGATGAATGGTAATT-- 2092
          Node_2  ATCACTTTATTATCGATAATTTTCAGTTCTTCCATTGTACCGAAAGTTGCTAATTTTGTA 998
Ppy Complete CDS  ------------------------------------------------------------ 2092
          Node_2  ATTGTGGGTCACTTGACTTCTTTAACGAATAATAAAATCTGGTATAGCTAAAAGGATTGA 1058
Ppy Complete CDS  ------------------------------------------------------------ 2092
          Node_2  AATTTTTCAAAAAATATTAAGATGCAATGTATTGTCGTTGCGAATCCCGGAAAAAAGG 1116
Ppy Complete CDS  ---------------------------------------------------------- 2092
```

Alignment of Node_2 and *Ppy* complete CDS. The Node_2 contig from SPAdes assembly, aligned to the complete CDS of the *Ppy* Fluc gene. Highlighting in grey indicates a mismatch between the two sequences. Alignment graphic produced in CLC sequence viewer **(**Qiagen, Venlo, Netherlands).

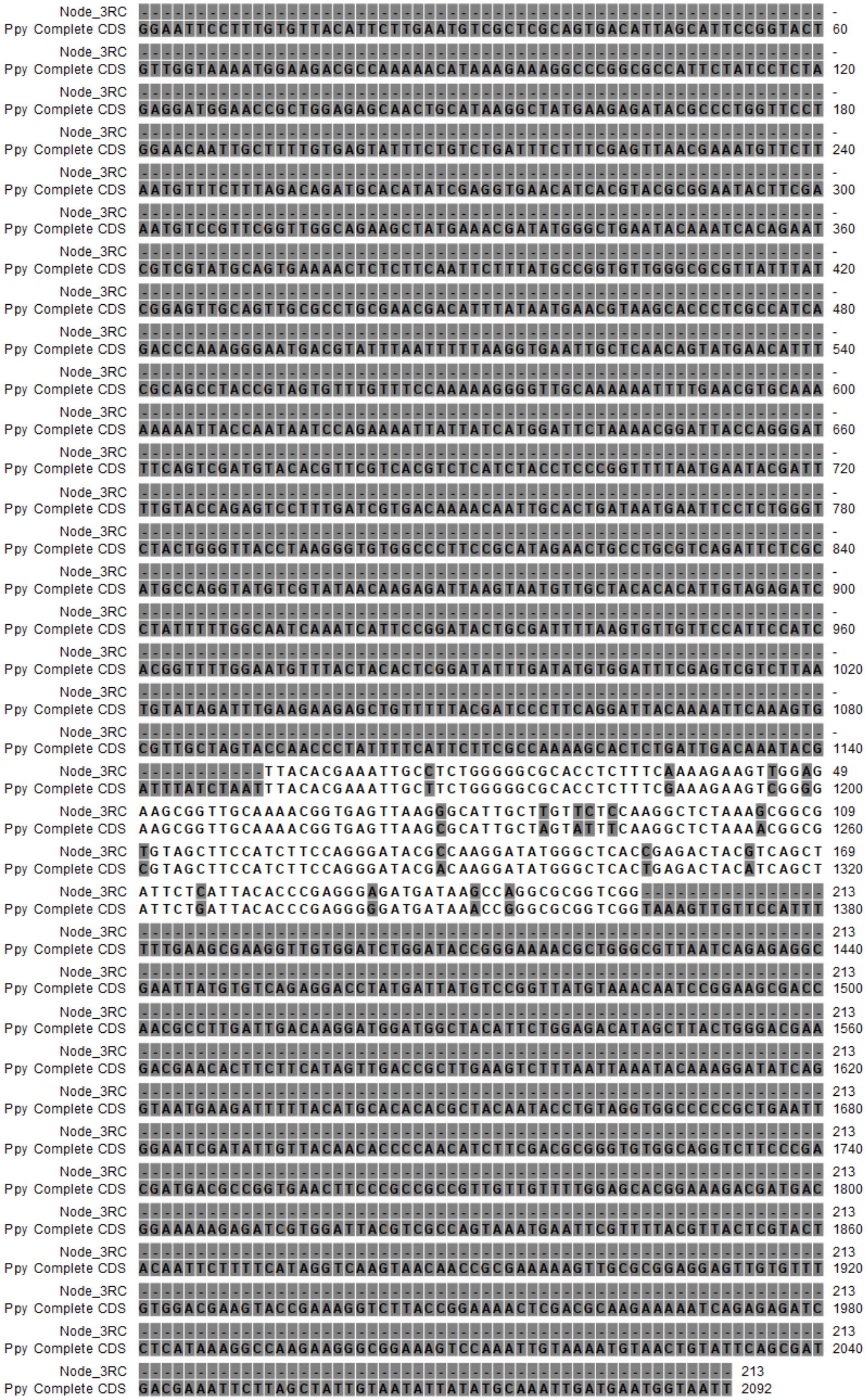
Figure 3.10.



```
         Node_1RC  CGGTTTTCGTTTTGAAAGGACCGATCGTAAATATATGACGGATAGGTAAGTGGTTTGAAT  60
Ppy Complete CDS  GGAATTCCTTTGTGTTACATTC--------------------------------TTGAAT  28

         Node_1RC  GTCGCTTGGCGTAACATTAGCAAGTCGGTATTAACGATAAAATGGAAGACCAAAAAAACA  120
Ppy Complete CDS  GTCGCTCGCAGTGACATTAGCATTCCGGTACTGTTGGTAAAATGGAAGACGCCAAAAACA  88

         Node_1RC  TAAAACATGGTCCAGCGCCATTCTATCCTCTAGAAGATGGAACTGCTGGAGAACAACTGC  180
Ppy Complete CDS  TAAAGAAAGGCCCGGCGCCATTCTATCCTCTAGAGGATGGAACCGCTGGAGAGCAACTGC  148

         Node_1RC  ATAGGGCTATGAAAAGATACGCCCTGGTTCCTGGGACAATTGCCTTTGTGAGTATAGTTC  240
Ppy Complete CDS  ATAAGGCTATGAAGAGATACGCCCTGGTTCCTGGAACAATTGCTTTTGTGAGTAT--TTC  206

         Node_1RC  TGCCTGTTTTCTTCCCAGCGAGTGTTAATGAAATGTTCTTAATGTTTCTTTAGACAGATG  300
Ppy Complete CDS  TGTCTGATTTCTTTCGAG------TTAACGAAATGTTCTTAATGTTTCTTTAGACAGATG  260

         Node_1RC  CACATATCGAGGTGAACGTCACGTACTCGGAATACTTTGAAATGTCCGTTAAATTAGCCG  360
Ppy Complete CDS  CACATATCGAGGTGAACATCACGTACGCGGAATACTTCGAAATGTCCGTTCGGTTGGCAG  320

         Node_1RC  AATCTATGAAACGGTATGGGCTTAATACAAATCACAGAATCGTCGTATGCAGTGAAAACT  420
Ppy Complete CDS  AAGCTATGAAACGATATGGGCTGAATACAAATCACAGAATCGTCGTATGCAGTGAAAACT  380

         Node_1RC  CTCTTCAATTCTTTATGCCTGTTGTGGGAGCGTTATTTATCGGAGTTGGAGTTGCGCCCG  480
Ppy Complete CDS  CTCTTCAATTCTTTATGCCGGTGTTGGGCGCGTTATTTATCGGAGTTGCAGTTGCGCCTG  440

         Node_1RC  CGAACGACATTTATAATGAACGTAAGCACCCTCACCA--AACCCAAGAGAGAATGATGTA  538
Ppy Complete CDS  CGAACGACATTTATAATGAACGTAAGCACCCTCGCCATCAGACCCAAAGGGAATGACGTA  500

         Node_1RC  ATTAATTTTTAAGGTGAATTGGTCAACAGTATGACTATTTCGCAGCCTACTTTAGTGTTT  598
Ppy Complete CDS  TTTAATTTTTAAGGTGAATTGCTCAACAGTATGAACATTTCGCAGCCTACCGTAGTGTTT  560

         Node_1RC  GTTTCCAAAAAGGGGCTGCAGAAAGTTTTGAACGTGCAAAAAAAATTACCAATAATTAAG  658
Ppy Complete CDS  GTTTCCAAAAAGGGGTTGCAAAAAATTTTGAACGTGCAAAAAAAATTACCAATAATCCAG  620

         Node_1RC  AAAATTATTATCATGGATTCTAAAGCAGATTACCAGGGATTTAATTCGATGGACACGTTC  718
Ppy Complete CDS  AAAATTATTATCATGGATTCTAAAACGGATTACCAGGGATTTCAGTCGATGTACACGTTC  680

         Node_1RC  ATCGCGGATCATTTACCTCCGGGCTTTAACGAATATGATTTTGTACCGGAGTCCTTTGAT  778
Ppy Complete CDS  GTCACGTCTCATCTACCTCCCGGTTTTAATGAATACGATTTTGTACCAGAGTCCTTTGAT  740

         Node_1RC  CGTGACAAGACAATTGCACTGATAATGAATTCCTCTGGCTCTACTGGGTTACCTAAGGGA  838
Ppy Complete CDS  CGTGACAAAACAATTGCACTGATAATGAATTCCTCTGGGTCTACTGGGTTACCTAAGGGT  800

         Node_1RC  GTGGCGCTTCCGCATAGAACTGCCTGTGTAAGATTCTCACATTGCAGGTATGTCGTGTAA  898
Ppy Complete CDS  GTGGCCCTTCCGCATAGAACTGCCTGCGTCAGATTCTCGCATGCCAGGTATGTCGTATAA  860

         Node_1RC  CAA-AGATTAAGTAATGTTCCTACAAAAATTCTAGAGATCCTATTTTTGGCAACCAAATC  957
Ppy Complete CDS  CAAGAGATTAAGTAATGTTGCTACACACATTGTAGAGATCCTATTTTTGGCAATCAAATC  920

         Node_1RC  ATTCCCGATACTTCGATTTTAAGTGTTGTTCCATTCCATCATGGTTTTGGAATGTTTACT  1017
Ppy Complete CDS  ATTCCGGATACTGCGATTTTAAGTGTTGTTCCATTCCATCACGGTTTTGGAATGTTTACT  980

         Node_1RC  ACACTCGGATATCTGATATGCGGATTTCGTGTAGTCTTGATGTATAGATTTGAAGAAGAA  1077
Ppy Complete CDS  ACACTCGGATATTTGATATGTGGATTTCGAGTCGTCTTAATGTATAGATTTGAAGAAGAG  1040

         Node_1RC  TTGTTTTTACGATCCCTTCAAGACTACAAAATTCAAAGCGCGTTGTTAGTACCAACCCTA  1137
Ppy Complete CDS  CTGTTTTTACGATCCCTTCAGGATTACAAAATTCAAAGTGCGTTGCTAGTACCAACCCTA  1100

         Node_1RC  TTTTCATTCTTCGCCAAAAGTACTCTGATTGACAAATACGATTTATCTAATTTACACGAA  1197
Ppy Complete CDS  TTTTCATTCTTCGCCAAAAGCACTCTGATTGACAAATACGATTTATCTAATTTACACGAA  1160

         Node_1RC  ATTGCCTCTGGGGGCGCACCTCTTTCAAAAGAAGTTGGAGAAGCGGTTGCAAAACGGTGA  1257
Ppy Complete CDS  ATTGCTTCTGGGGGCGCACCTCTTTCGAAAGAAGTCGGGGAAGCGGTTGCAAAACGGTGA  1220

         Node_1RC  GTTAAGGG----------------------------------------------------  1265
Ppy Complete CDS  GTTAAGCGCATTGCTAGTATTTCAAGGCTCTAAAACGGCGCGTAGCTTCCATCTTCCAGG  1280

         Node_1RC  ------------------------------------------------------------  1265
Ppy Complete CDS  GATACGACAAGGATATGGGCTCACTGAGACTACATCAGCTATTCTGATTACACCCGAGGG  1340

         Node_1RC  ------------------------------------------------------------  1265
Ppy Complete CDS  GGATGATAAACCGGGCGCGGTCGGTAAAGTTGTTCCATTTTTTGAAGCGAAGGTTGTGGA  1400

         Node_1RC  ------------------------------------------------------------  1265
Ppy Complete CDS  TCTGGATACCGGGAAAACGCTGGGCGTTAATCAGAGAGGCGAATTATGTGTCAGAGGACC  1460

         Node_1RC  ------------------------------------------------------------  1265
Ppy Complete CDS  TATGATTATGTCCGGTTATGTAAACAATCCGGAAGCGACCAACGCCTTGATTGACAAGGA  1520

         Node_1RC  ------------------------------------------------------------  1265
Ppy Complete CDS  TGGATGGCTACATTCTGGAGACATAGCTTACTGGGACGAAGACGAACACTTCTTCATAGT  1580

         Node_1RC  ------------------------------------------------------------  1265
Ppy Complete CDS  TGACCGCTTGAAGTCTTTAATTAAATACAAAGGATATCAGGTAATGAAGATTTTTACATG  1640

         Node_1RC  ------------------------------------------------------------  1265
Ppy Complete CDS  CACACACGCTACAATACCTGTAGGTGGCCCCCGCTGAATTGGAATCGATATTGTTACAAC  1700

         Node_1RC  ------------------------------------------------------------  1265
Ppy Complete CDS  ACCCCAACATCTTCGACGCGGGTGTGGCAGGTCTTCCCGACGATGACGCCGGTGAACTTC  1760

         Node_1RC  ------------------------------------------------------------  1265
Ppy Complete CDS  CCGCCGCCGTTGTTGTTTTGGAGCACGGAAAGACGATGACGGAAAAAGAGATCGTGGATT  1820

         Node_1RC  ------------------------------------------------------------  1265
Ppy Complete CDS  ACGTCGCCAGTAAATGAATTCGTTTTACGTTACTCGTACTACAATTCTTTTCATAGGTCA  1880

         Node_1RC  ------------------------------------------------------------  1265
Ppy Complete CDS  AGTAACAACCGCGAAAAAGTTGCGCGGAGGAGTTGTGTTTGTGGACGAAGTACCGAAAGG  1940

         Node_1RC  ------------------------------------------------------------  1265
Ppy Complete CDS  TCTTACCGGAAAACTCGACGCAAGAAAAATCAGAGAGATCCTCATAAAGGCCAAGAAGGG  2000

         Node_1RC  ------------------------------------------------------------  1265
Ppy Complete CDS  CGGAAAGTCCAAATTGTAAAATGTAACTGTATTCAGCGATGACGAAATTCTTAGCTATTG  2060

         Node_1RC  --------------------------------  1265
Ppy Complete CDS  TAATATTATATGCAAATTGATGAATGGTAATT  2092
```

Alignment of Node_1 reverse complement and *Ppy* complete CDS. The reverse complement of the Node_1 contig from SPAdes assembly, aligned to the complete CDS of the *Ppy* Fluc gene. Highlighting in grey indicates a mismatch between the two sequences. Alignment graphic produced in CLC sequence viewer **(**Qiagen, Venlo, Netherlands).

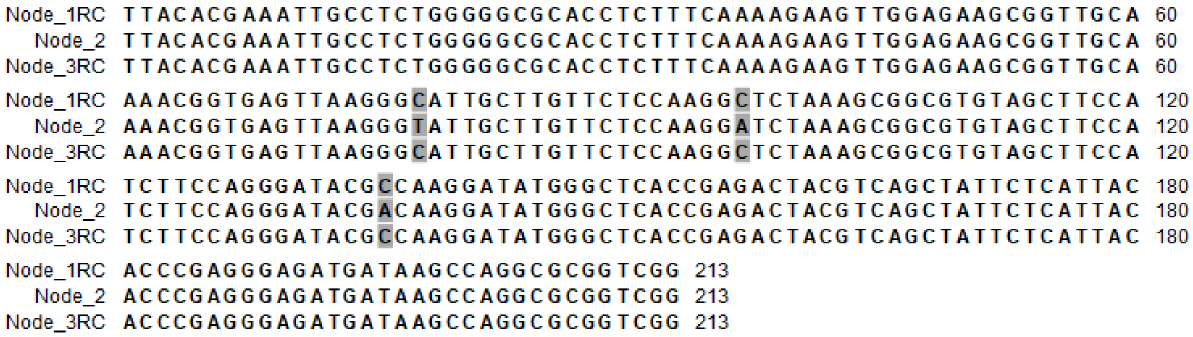
Figure 3.11.



```
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS GGAATTCCTTTGTGTTACATTCTTGAATGTCGCTCGCAGTGACATTAGCATTCCGGTACT 60
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS GTTGGTAAAATGGAAGACGCCAAAAACATAAAGAAAGGCCCGGCGCCATTCTATCCTCTA 120
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS GAGGATGGAACCGCTGGAGAGCAACTGCATAAGGCTATGAAGAGATACGCCCTGGTTCCT 180
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS GGAACAATTGCTTTTGTGAGTATTTCTGTCTGATTTCTTTCGAGTTAACGAAATGTTCTT 240
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS AATGTTTCTTTAGACAGATGCACATATCGAGGTGAACATCACGTACGCGGAATACTTCGA 300
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS AATGTCCGTTCGGTTGGCAGAAGCTATGAAACGATATGGGCTGAATACAAATCACAGAAT 360
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS CGTCGTATGCAGTGAAAACTCTCTTCAATTCTTTATGCCGGTGTTGGGCGCGTTATTTAT 420
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS CGGAGTTGCAGTTGCGCCTGCGAACGACATTTATAATGAACGTAAGCACCCTCGCCATCA 480
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS GACCCAAAGGGAATGACGTATTTAATTTTTAAGGTGAATTGCTCAACAGTATGAACATTT 540
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS CGCAGCCTACCGTAGTGTTTGTTTCCAAAAAGGGGTTGCAAAAAATTTTGAACGTGCAAA 600
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS AAAAATTACCAATAATCCAGAAAATTATTATCATGGATTCTAAAACGGATTACCAGGGAT 660
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS TTCAGTCGATGTACACGTTCGTCACGTCTCATCTACCTCCCGGTTTTAATGAATACGATT 720
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS TTGTACCAGAGTCCTTTGATCGTGACAAAACAATTGCACTGATAATGAATTCCTCTGGGT 780
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS CTACTGGGTTACCTAAGGGTGTGGCCCTTCCGCATAGAACTGCCTGCGTCAGATTCTCGC 840
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS ATGCCAGGTATGTCGTATAACAAGAGATTAAGTAATGTTGCTACACACATTGTAGAGATC 900
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS CTATTTTTGGCAATCAAATCATTCCGGATACTGCGATTTTAAGTGTTGTTCCATTCCATC 960
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS ACGGTTTTGGAATGTTTACTACACTCGGATATTTGATATGTGGATTTCGAGTCGTCTTAA 1020
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS TGTATAGATTTGAAGAAGAGCTGTTTTTACGATCCCTTCAGGATTACAAAATTCAAAGTG 1080
         Node_3RC ------------------------------------------------------------ -
 Ppy Complete CDS CGTTGCTAGTACCAACCCTATTTTCATTCTTCGCCAAAAGCACTCTGATTGACAAATACG 1140
         Node_3RC -----------TTACACGAAATTGCCTCTGGGGGCGCACCTCTTTCAAAAGAAGTTGGAG 49
 Ppy Complete CDS ATTTATCTAATTTACACGAAATTGCTTCTGGGGGCGCACCTCTTTCGAAAGAAGTCGGGG 1200
         Node_3RC AAGCGGTTGCAAAACGGTGAGTTAAGGGCATTGCTTGTTCTCCAAGGCTCTAAAGCGGCG 109
 Ppy Complete CDS AAGCGGTTGCAAAACGGTGAGTTAAGCGCATTGCTAGTATTTCAAGGCTCTAAAACGGCG 1260
         Node_3RC TGTAGCTTCCATCTTCCAGGGATACGCCAAGGATATGGGCTCACCGAGACTACGTCAGCT 169
 Ppy Complete CDS CGTAGCTTCCATCTTCCAGGGATACGACAAGGATATGGGCTCACTGAGACTACATCAGCT 1320
         Node_3RC ATTCTCATTACACCCGAGGGAGATGATAAGCCAGGCGCGGTCGG---------------- 213
 Ppy Complete CDS ATTCTGATTACACCCGAGGGGGATGATAAACCGGGCGCGGTCGGTAAAGTTGTTCCATTT 1380
         Node_3RC ------------------------------------------------------------ 213
 Ppy Complete CDS TTTGAAGCGAAGGTTGTGGATCTGGATACCGGGAAAACGCTGGGCGTTAATCAGAGAGGC 1440
         Node_3RC ------------------------------------------------------------ 213
 Ppy Complete CDS GAATTATGTGTCAGAGGACCTATGATTATGTCCGGTTATGTAAACAATCCGGAAGCGACC 1500
         Node_3RC ------------------------------------------------------------ 213
 Ppy Complete CDS AACGCCTTGATTGACAAGGATGGATGGCTACATTCTGGAGACATAGCTTACTGGGACGAA 1560
         Node_3RC ------------------------------------------------------------ 213
 Ppy Complete CDS GACGAACACTTCTTCATAGTTGACCGCTTGAAGTCTTTAATTAAATACAAAGGATATCAG 1620
         Node_3RC ------------------------------------------------------------ 213
 Ppy Complete CDS GTAATGAAGATTTTTACATGCACACACGCTACAATACCTGTAGGTGGCCCCCGCTGAATT 1680
         Node_3RC ------------------------------------------------------------ 213
 Ppy Complete CDS GGAATCGATATTGTTACAACACCCCAACATCTTCGACGCGGGTGTGGCAGGTCTTCCCGA 1740
         Node_3RC ------------------------------------------------------------ 213
 Ppy Complete CDS CGATGACGCCGGTGAACTTCCCGCCGCCGTTGTTGTTTTGGAGCACGGAAAGACGATGAC 1800
         Node_3RC ------------------------------------------------------------ 213
 Ppy Complete CDS GGAAAAAGAGATCGTGGATTACGTCGCCAGTAAATGAATTCGTTTTACGTTACTCGTACT 1860
         Node_3RC ------------------------------------------------------------ 213
 Ppy Complete CDS ACAATTCTTTTCATAGGTCAAGTAACAACCGCGAAAAAGTTGCGCGGAGGAGTTGTGTTT 1920
         Node_3RC ------------------------------------------------------------ 213
 Ppy Complete CDS GTGGACGAAGTACCGAAAGGTCTTACCGGAAAACTCGACGCAAGAAAAATCAGAGAGATC 1980
         Node_3RC ------------------------------------------------------------ 213
 Ppy Complete CDS CTCATAAAGGCCAAGAAGGGCGGAAAGTCCAAATTGTAAAATGTAACTGTATTCAGCGAT 2040
         Node_3RC ---------------------------------------------------- 213
 Ppy Complete CDS GACGAAATTCTTAGCTATTGTAATATTATATGCAAATTGATGAATGGTAATT 2092
```

Alignment of Node_3 reverse complement and *Ppy* complete CDS. The reverse complement of the Node_3 contig from SPAdes assembly, aligned to the complete CDS of the *Ppy* Fluc gene. Highlighting in grey indicates a mismatch between the two sequences. Alignment graphic produced in CLC sequence viewer **(**Qiagen, Venlo, Netherlands).

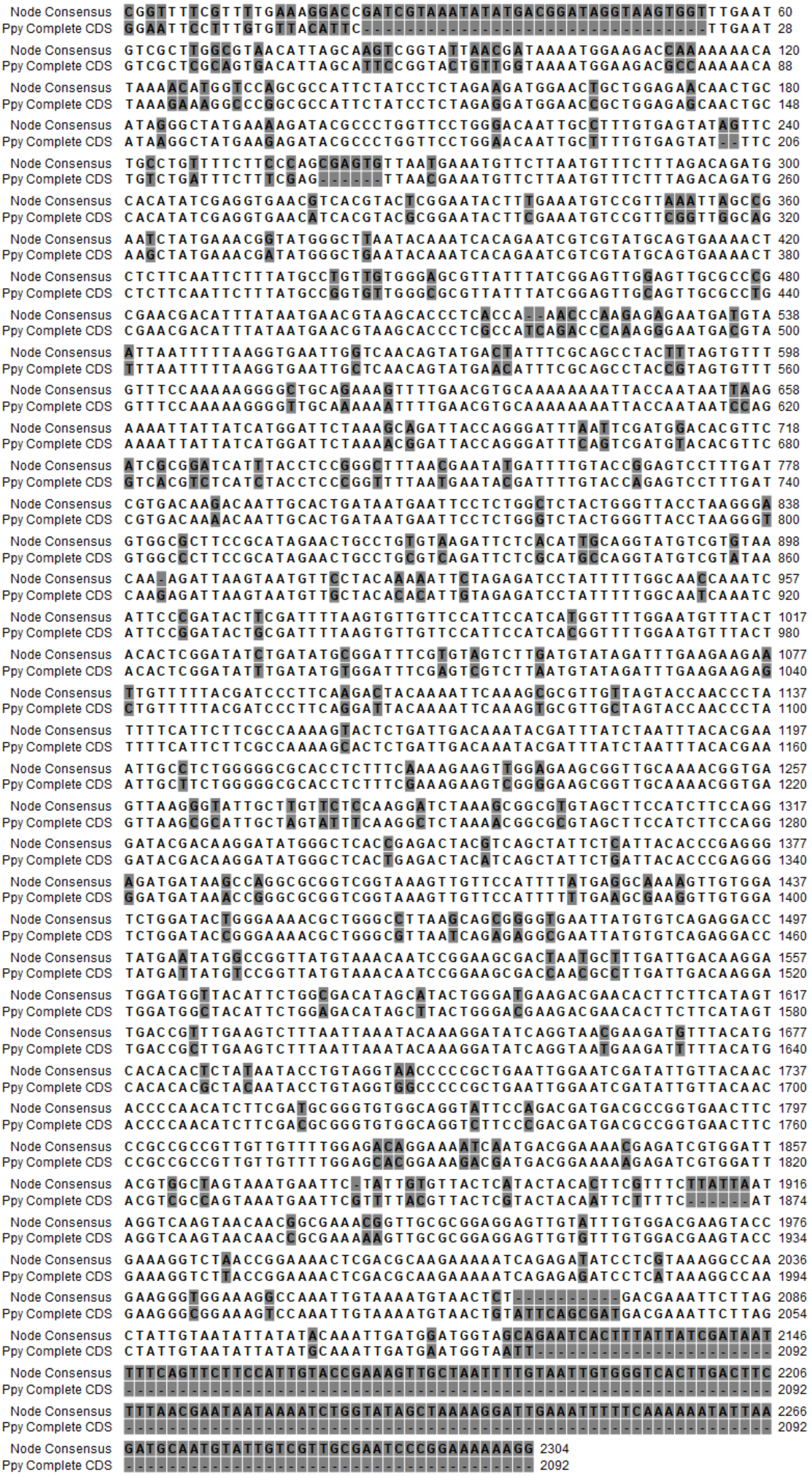
Figure 3.12.



```
Node_1RC TTACACGAAATTGCCTCTGGGGGCGCACCTCTTTCAAAAGAAGTTGGAGAAGCGGTTGCA 60
  Node_2 TTACACGAAATTGCCTCTGGGGGCGCACCTCTTTCAAAAGAAGTTGGAGAAGCGGTTGCA 60
Node_3RC TTACACGAAATTGCCTCTGGGGGCGCACCTCTTTCAAAAGAAGTTGGAGAAGCGGTTGCA 60

Node_1RC AAACGGTGAGTTAAGGGCATTGCTTGTTCTCCAAGGCTCTAAAGCGGCGTGTAGCTTCCA 120
  Node_2 AAACGGTGAGTTAAGGGTATTGCTTGTTCTCCAAGGATCTAAAGCGGCGTGTAGCTTCCA 120
Node_3RC AAACGGTGAGTTAAGGGCATTGCTTGTTCTCCAAGGCTCTAAAGCGGCGTGTAGCTTCCA 120

Node_1RC TCTTCCAGGGATACGCCAAGGATATGGGCTCACCGAGACTACGTCAGCTATTCTCATTAC 180
  Node_2 TCTTCCAGGGATACGACAAGGATATGGGCTCACCGAGACTACGTCAGCTATTCTCATTAC 180
Node_3RC TCTTCCAGGGATACGCCAAGGATATGGGCTCACCGAGACTACGTCAGCTATTCTCATTAC 180

Node_1RC ACCCGAGGGAGATGATAAGCCAGGCGCGGTCGG 213
  Node_2 ACCCGAGGGAGATGATAAGCCAGGCGCGGTCGG 213
Node_3RC ACCCGAGGGAGATGATAAGCCAGGCGCGGTCGG 213
```

Alignment of Node_1 reverse complement and Node_2 to Node_3 reverse complement. All three nodes in the correct orientation cropped to the region of Node_3. Highlighting in grey indicates a mismatch between the two sequences. Alignment graphic produced in CLC sequence viewer **(**Qiagen, Venlo, Netherlands).

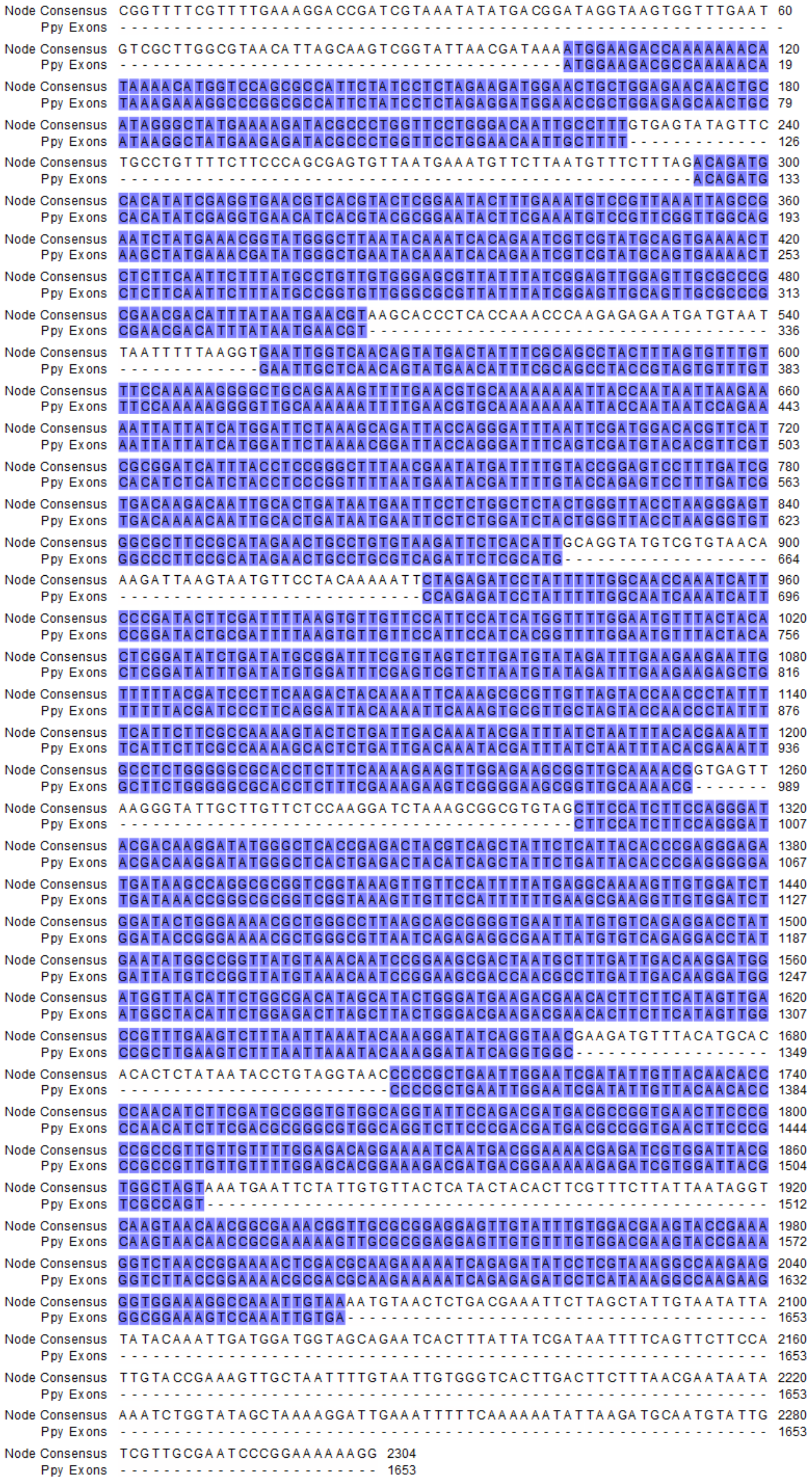
Figure 3.13.



```
  Node Consensus  CGGTTTTCGTTTTGAAAGGACCGATCGTAAATATATGACGGATAGGTAAGTGGTTTGAAT 60
Ppy Complete CDS  GGAATTCCTTTGTGTTACATTC--------------------------------TTGAAT 28

  Node Consensus  GTCGCTTGGCGTAACATTAGCAAGTCGGTATTAACGATAAAATGGAAGACCAAAAAAACA 120
Ppy Complete CDS  GTCGCTCGCAGTGACATTAGCATTCCGGTACTGTTGGTAAAATGGAAGACGCCAAAAACA 88

  Node Consensus  TAAAACATGGTCCAGCGCCATTCTATCCTCTAGAAGATGGAACTGCTGGAGAACAACTGC 180
Ppy Complete CDS  TAAAGAAAGGCCCGGCGCCATTCTATCCTCTAGAGGATGGAACCGCTGGAGAGCAACTGC 148

  Node Consensus  ATAGGGCTATGAAAAGATACGCCCTGGTTCCTGGGACAATTGCCTTTGTGAGTATAGTTC 240
Ppy Complete CDS  ATAAGGCTATGAAGAGATACGCCCTGGTTCCTGGAACAATTGCTTTTGTGAGTAT--TTC 206

  Node Consensus  TGCCTGTTTTCTTCCCAGCGAGTGTTAATGAAATGTTCTTAATGTTTCTTTAGACAGATG 300
Ppy Complete CDS  TGTCTGATTTCTTTCGAG------TTAACGAAATGTTCTTAATGTTTCTTTAGACAGATG 260

  Node Consensus  CACATATCGAGGTGAACGTCACGTACTCGGAATACTTTGAAATGTCCGTTAAATTAGCCG 360
Ppy Complete CDS  CACATATCGAGGTGAACATCACGTACGCGGAATACTTCGAAATGTCCGTTCGGTTGGCAG 320

  Node Consensus  AATCTATGAAACGGTATGGGCTTAATACAAATCACAGAATCGTCGTATGCAGTGAAAACT 420
Ppy Complete CDS  AAGCTATGAAACGATATGGGCTGAATACAAATCACAGAATCGTCGTATGCAGTGAAAACT 380

  Node Consensus  CTCTTCAATTCTTTATGCCTGTTGTGGGAGCGTTATTTATCGGAGTTGGAGTTGCGCCCG 480
Ppy Complete CDS  CTCTTCAATTCTTTATGCCGGTGTTGGGCGCGTTATTTATCGGAGTTGCAGTTGCGCCTG 440

  Node Consensus  CGAACGACATTTATAATGAACGTAAGCACCCTCACCA--AACCCAAGAGAGAATGATGTA 538
Ppy Complete CDS  CGAACGACATTTATAATGAACGTAAGCACCCTCGCCATCAGACCCAAAGGGAATGACGTA 500

  Node Consensus  ATTAATTTTTAAGGTGAATTGGTCAACAGTATGACTATTTCGCAGCCTACTTTAGTGTTT 598
Ppy Complete CDS  TTTAATTTTTAAGGTGAATTGCTCAACAGTATGAACATTTCGCAGCCTACCGTAGTGTTT 560

  Node Consensus  GTTTCCAAAAAGGGGCTGCAGAAAGTTTTGAACGTGCAAAAAAAATTACCAATAATTAAG 658
Ppy Complete CDS  GTTTCCAAAAAGGGGTTGCAAAAAATTTTGAACGTGCAAAAAAAATTACCAATAATCCAG 620

  Node Consensus  AAAATTATTATCATGGATTCTAAAGCAGATTACCAGGGATTTAATTCGATGGACACGTTC 718
Ppy Complete CDS  AAAATTATTATCATGGATTCTAAAACGGATTACCAGGGATTTCAGTCGATGTACACGTTC 680

  Node Consensus  ATCGCGGATCATTTACCTCCGGGCTTTAACGAATATGATTTTGTACCGGAGTCCTTTGAT 778
Ppy Complete CDS  GTCACGTCTCATCTACCTCCCGGTTTTAATGAATACGATTTTGTACCAGAGTCCTTTGAT 740

  Node Consensus  CGTGACAAGACAATTGCACTGATAATGAATTCCTCTGGCTCTACTGGGTTACCTAAGGGA 838
Ppy Complete CDS  CGTGACAAAACAATTGCACTGATAATGAATTCCTCTGGGTCTACTGGGTTACCTAAGGGT 800

  Node Consensus  GTGGCGCTTCCGCATAGAACTGCCTGTGTAAGATTCTCACATTGCAGGTATGTCGTGTAA 898
Ppy Complete CDS  GTGGCCCTTCCGCATAGAACTGCCTGCGTCAGATTCTCGCATGCCAGGTATGTCGTATAA 860

  Node Consensus  CAA-AGATTAAGTAATGTTCCTACAAAAATTCTAGAGATCCTATTTTTGGCAACCAAATC 957
Ppy Complete CDS  CAAGAGATTAAGTAATGTTGCTACACACATTGTAGAGATCCTATTTTTGGCAATCAAATC 920

  Node Consensus  ATTCCCGATACTTCGATTTTAAGTGTTGTTCCATTCCATCATGGTTTTGGAATGTTTACT 1017
Ppy Complete CDS  ATTCCGGATACTGCGATTTTAAGTGTTGTTCCATTCCATCACGGTTTTGGAATGTTTACT 980

  Node Consensus  ACACTCGGATATCTGATATGCGGATTTCGTGTAGTCTTGATGTATAGATTTGAAGAAGAA 1077
Ppy Complete CDS  ACACTCGGATATTTGATATGTGGATTTCGAGTCGTCTTAATGTATAGATTTGAAGAAGAG 1040

  Node Consensus  TTGTTTTTACGATCCCTTCAAGACTACAAAATTCAAAGCGCGTTGTTAGTACCAACCCTA 1137
Ppy Complete CDS  CTGTTTTTACGATCCCTTCAGGATTACAAAATTCAAAGTGCGTTGCTAGTACCAACCCTA 1100

  Node Consensus  TTTTCATTCTTCGCCAAAAGTACTCTGATTGACAAATACGATTTATCTAATTTACACGAA 1197
Ppy Complete CDS  TTTTCATTCTTCGCCAAAAGCACTCTGATTGACAAATACGATTTATCTAATTTACACGAA 1160

  Node Consensus  ATTGCCTCTGGGGGCGCACCTCTTTCAAAAGAAGTTGGAGAAGCGGTTGCAAAACGGTGA 1257
Ppy Complete CDS  ATTGCTTCTGGGGGCGCACCTCTTTCGAAAGAAGTCGGGGAAGCGGTTGCAAAACGGTGA 1220

  Node Consensus  GTTAAGGGTATTGCTTGTTCTCCAAGGATCTAAAGCGGCGTGTAGCTTCCATCTTCCAGG 1317
Ppy Complete CDS  GTTAAGCGCATTGCTAGTATTTCAAGGCTCTAAAACGGCGCGTAGCTTCCATCTTCCAGG 1280

  Node Consensus  GATACGACAAGGATATGGGCTCACCGAGACTACGTCAGCTATTCTCATTACACCCGAGGG 1377
Ppy Complete CDS  GATACGACAAGGATATGGGCTCACTGAGACTACATCAGCTATTCTGATTACACCCGAGGG 1340

  Node Consensus  AGATGATAAGCCAGGCGCGGTCGGTAAAGTTGTTCCATTTTATGAGGCAAAAGTTGTGGA 1437
Ppy Complete CDS  GGATGATAAACCGGGCGCGGTCGGTAAAGTTGTTCCATTTTTTGAAGCGAAGGTTGTGGA 1400

  Node Consensus  TCTGGATACTGGGAAAACGCTGGGCCTTAAGCAGCGGGGTGAATTATGTGTCAGAGGACC 1497
Ppy Complete CDS  TCTGGATACCGGGAAAACGCTGGGCGTTAATCAGAGAGGCGAATTATGTGTCAGAGGACC 1460

  Node Consensus  TATGAATATGGCCGGTTATGTAAACAATCCGGAAGCGACTAATGCTTTGATTGACAAGGA 1557
Ppy Complete CDS  TATGATTATGTCCGGTTATGTAAACAATCCGGAAGCGACCAACGCCTTGATTGACAAGGA 1520

  Node Consensus  TGGATGGTTACATTCTGGCGACATAGCATACTGGGATGAAGACGAACACTTCTTCATAGT 1617
Ppy Complete CDS  TGGATGGCTACATTCTGGAGACATAGCTTACTGGGACGAAGACGAACACTTCTTCATAGT 1580

  Node Consensus  TGACCGTTTGAAGTCTTTAATTAAATACAAAGGATATCAGGTAACGAAGATGTTTACATG 1677
Ppy Complete CDS  TGACCGCTTGAAGTCTTTAATTAAATACAAAGGATATCAGGTAATGAAGATTTTTACATG 1640

  Node Consensus  CACACACTCTATAATACCTGTAGGTAACCCCCGCTGAATTGGAATCGATATTGTTACAAC 1737
Ppy Complete CDS  CACACACGCTACAATACCTGTAGGTGGCCCCCGCTGAATTGGAATCGATATTGTTACAAC 1700

  Node Consensus  ACCCCAACATCTTCGATGCGGGTGTGGCAGGTATTCCAGACGATGACGCCGGTGAACTTC 1797
Ppy Complete CDS  ACCCCAACATCTTCGACGCGGGTGTGGCAGGTCTTCCCGACGATGACGCCGGTGAACTTC 1760

  Node Consensus  CCGCCGCCGTTGTTGTTTTGGAGACAGGAAAATCAATGACGGAAAACGAGATCGTGGATT 1857
Ppy Complete CDS  CCGCCGCCGTTGTTGTTTTGGAGCACGGAAAGACGATGACGGAAAAAGAGATCGTGGATT 1820

  Node Consensus  ACGTGGCTAGTAAATGAATTC-TATTGTGTTACTCATACTACACTTCGTTTCTTATTAAT 1916
Ppy Complete CDS  ACGTCGCCAGTAAATGAATTCGTTTTACGTTACTCGTACTACAATTCTTTTC------AT 1874

  Node Consensus  AGGTCAAGTAACAACGGCGAAACGGTTGCGCGGAGGAGTTGTATTTGTGGACGAAGTACC 1976
Ppy Complete CDS  AGGTCAAGTAACAACCGCGAAAAAGTTGCGCGGAGGAGTTGTGTTTGTGGACGAAGTACC 1934

  Node Consensus  GAAAGGTCTAACCGGAAAACTCGACGCAAGAAAAATCAGAGATATCCTCGTAAAGGCCAA 2036
Ppy Complete CDS  GAAAGGTCTTACCGGAAAACTCGACGCAAGAAAAATCAGAGAGATCCTCATAAAGGCCAA 1994

  Node Consensus  GAAGGGTGGAAAGGCCAAATTGTAAAATGTAACTCT----------GACGAAATTCTTAG 2086
Ppy Complete CDS  GAAGGGCGGAAAGTCCAAATTGTAAAATGTAACTGTATTCAGCGATGACGAAATTCTTAG 2054

  Node Consensus  CTATTGTAATATTATATACAAATTGATGGATGGTAGCAGAATCACTTTATTATCGATAAT 2146
Ppy Complete CDS  CTATTGTAATATTATATGCAAATTGATGAATGGTAATT---------------------- 2092

  Node Consensus  TTTCAGTTCTTCCATTGTACCGAAAGTTGCTAATTTTGTAATTGTGGGTCACTTGACTTC 2206
Ppy Complete CDS  ------------------------------------------------------------ 2092

  Node Consensus  TTTAACGAATAATAAAATCTGGTATAGCTAAAAGGATTGAAATTTTTCAAAAAATATTAA 2266
Ppy Complete CDS  ------------------------------------------------------------ 2092

  Node Consensus  GATGCAATGTATTGTCGTTGCGAATCCCGGAAAAAAGG 2304
Ppy Complete CDS  -------------------------------------- 2092
```

Alignment of Node consensus and *Ppy* complete CDS. The consensus sequence of the Node contigs from SPAdes assembly, aligned to the complete CDS of the *Ppy* Fluc gene. Highlighting in grey indicates a mismatch between the two sequences. Percent identity between the two sequences is 88.79%. Alignment graphic produced in CLC sequence viewer **(**Qiagen, Venlo, Netherlands).

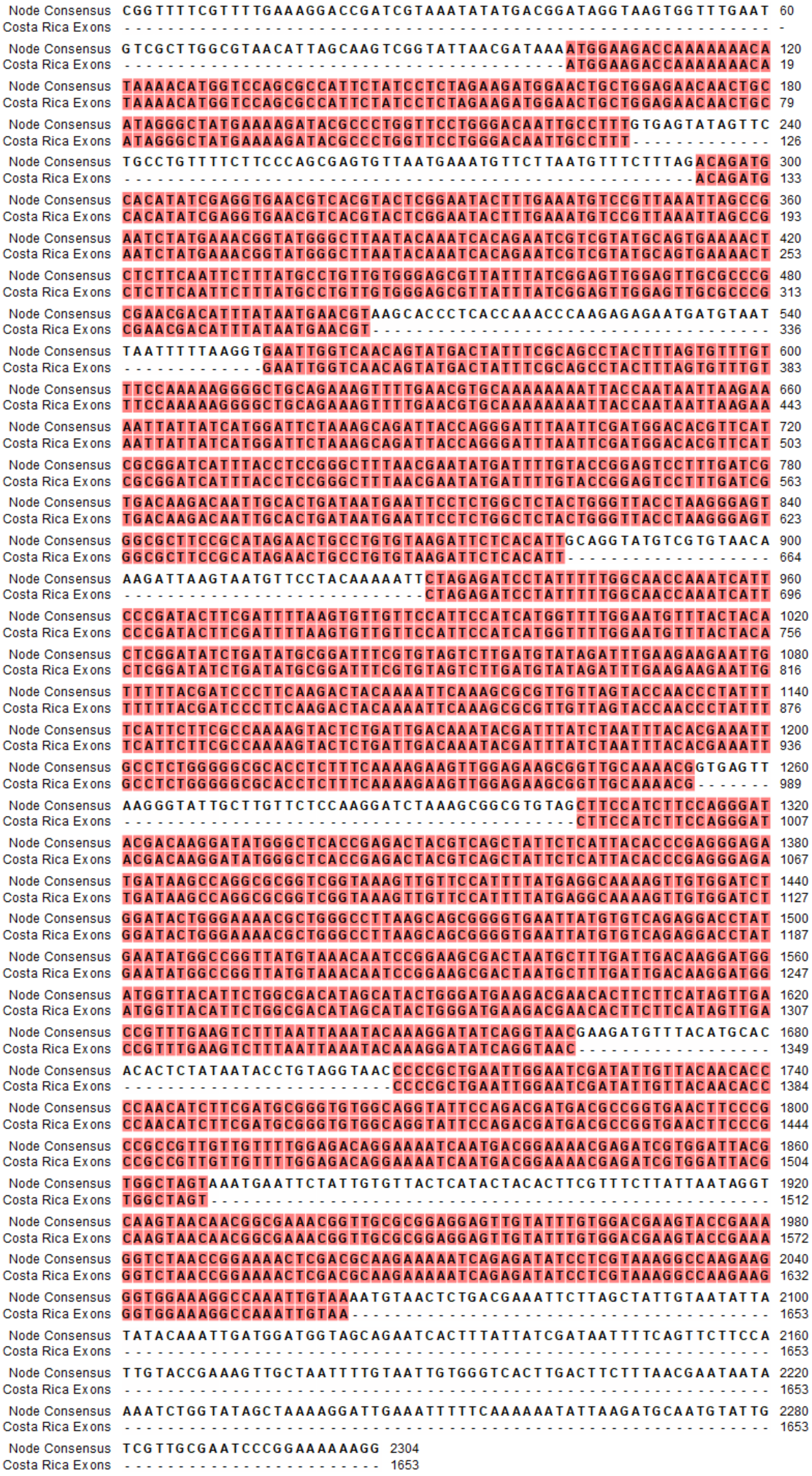
Figure 3.14.



```
Node Consensus CGGTTTTCGTTTTGAAAGGACCGATCGTAAATATATGACGGATAGGTAAGTGGTTTGAAT 60
     Ppy Exons ------------------------------------------------------------ -

Node Consensus GTCGCTTGGCGTAACATTAGCAAGTCGGTATTAACGATAAAATGGAAGACCAAAAAAACA 120
     Ppy Exons -----------------------------------------ATGGAAGACGCCAAAAACA 19

Node Consensus TAAAACATGGTCCAGCGCCATTCTATCCTCTAGAAGATGGAACTGCTGGAGAACAACTGC 180
     Ppy Exons TAAAGAAAGGCCCGGCGCCATTCTATCCTCTAGAGGATGGAACCGCTGGAGAGCAACTGC 79

Node Consensus ATAGGGCTATGAAAAGATACGCCCTGGTTCCTGGGACAATTGCCTTTGTGAGTATAGTTC 240
     Ppy Exons ATAAGGCTATGAAGAGATACGCCCTGGTTCCTGGAACAATTGCTTTT------------- 126

Node Consensus TGCCTGTTTTCTTCCCAGCGAGTGTTAATGAAATGTTCTTAATGTTTCTTTAGACAGATG 300
     Ppy Exons -----------------------------------------------------ACAGATG 133

Node Consensus CACATATCGAGGTGAACGTCACGTACTCGGAATACTTTGAAATGTCCGTTAAATTAGCCG 360
     Ppy Exons CACATATCGAGGTGAACATCACGTACGCGGAATACTTCGAAATGTCCGTTCGGTTGGCAG 193

Node Consensus AATCTATGAAACGGTATGGGCTTAATACAAATCACAGAATCGTCGTATGCAGTGAAAACT 420
     Ppy Exons AAGCTATGAAACGATATGGGCTGAATACAAATCACAGAATCGTCGTATGCAGTGAAAACT 253

Node Consensus CTCTTCAATTCTTTATGCCTGTTGTGGGAGCGTTATTTATCGGAGTTGGAGTTGCGCCCG 480
     Ppy Exons CTCTTCAATTCTTTATGCCGGTGTTGGGCGCGTTATTTATCGGAGTTGCAGTTGCGCCCG 313

Node Consensus CGAACGACATTTATAATGAACGTAAGCACCCTCACCAAACCCAAGAGAGAATGATGTAAT 540
     Ppy Exons CGAACGACATTTATAATGAACGT------------------------------------- 336

Node Consensus TAATTTTTAAGGTGAATTGGTCAACAGTATGACTATTTCGCAGCCTACTTTAGTGTTTGT 600
     Ppy Exons -------------GAATTGCTCAACAGTATGAACATTTCGCAGCCTACCGTAGTGTTTGT 383

Node Consensus TTCCAAAAAGGGGCTGCAGAAAGTTTTGAACGTGCAAAAAAAATTACCAATAATTAAGAA 660
     Ppy Exons TTCCAAAAAGGGGTTGCAAAAAATTTTGAACGTGCAAAAAAAATTACCAATAATCCAGAA 443

Node Consensus AATTATTATCATGGATTCTAAAGCAGATTACCAGGGATTTAATTCGATGGACACGTTCAT 720
     Ppy Exons AATTATTATCATGGATTCTAAAACGGATTACCAGGGATTTCAGTCGATGTACACGTTCGT 503

Node Consensus CGCGGATCATTTACCTCCGGGCTTTAACGAATATGATTTTGTACCGGAGTCCTTTGATCG 780
     Ppy Exons CACATCTCATCTACCTCCCGGTTTTAATGAATACGATTTTGTACCAGAGTCCTTTGATCG 563

Node Consensus TGACAAGACAATTGCACTGATAATGAATTCCTCTGGCTCTACTGGGTTACCTAAGGGAGT 840
     Ppy Exons TGACAAAACAATTGCACTGATAATGAATTCCTCTGGATCTACTGGGTTACCTAAGGGTGT 623

Node Consensus GGCGCTTCCGCATAGAACTGCCTGTGTAAGATTCTCACATTGCAGGTATGTCGTGTAACA 900
     Ppy Exons GGCCCTTCCGCATAGAACTGCCTGCGTCAGATTCTCGCATG------------------- 664

Node Consensus AAGATTAAGTAATGTTCCTACAAAAATTCTAGAGATCCTATTTTTGGCAACCAAATCATT 960
     Ppy Exons ----------------------------CCAGAGATCCTATTTTTGGCAATCAAATCATT 696

Node Consensus CCCGATACTTCGATTTTAAGTGTTGTTCCATTCCATCATGGTTTTGGAATGTTTACTACA 1020
     Ppy Exons CCGGATACTGCGATTTTAAGTGTTGTTCCATTCCATCACGGTTTTGGAATGTTTACTACA 756

Node Consensus CTCGGATATCTGATATGCGGATTTCGTGTAGTCTTGATGTATAGATTTGAAGAAGAATTG 1080
     Ppy Exons CTCGGATATTTGATATGTGGATTTCGAGTCGTCTTAATGTATAGATTTGAAGAAGAGCTG 816

Node Consensus TTTTTACGATCCCTTCAAGACTACAAAATTCAAAGCGCGTTGTTAGTACCAACCCTATTT 1140
     Ppy Exons TTTTTACGATCCCTTCAGGATTACAAAATTCAAAGTGCGTTGCTAGTACCAACCCTATTT 876

Node Consensus TCATTCTTCGCCAAAAGTACTCTGATTGACAAATACGATTTATCTAATTTACACGAAATT 1200
     Ppy Exons TCATTCTTCGCCAAAAGCACTCTGATTGACAAATACGATTTATCTAATTTACACGAAATT 936

Node Consensus GCCTCTGGGGGCGCACCTCTTTCAAAAGAAGTTGGAGAAGCGGTTGCAAAACGGTGAGTT 1260
     Ppy Exons GCTTCTGGGGGCGCACCTCTTTCGAAAGAAGTCGGGGAAGCGGTTGCAAAACG------- 989

Node Consensus AAGGGTATTGCTTGTTCTCCAAGGATCTAAAGCGGCGTGTAGCTTCCATCTTCCAGGGAT 1320
     Ppy Exons ------------------------------------------CTTCCATCTTCCAGGGAT 1007

Node Consensus ACGACAAGGATATGGGCTCACCGAGACTACGTCAGCTATTCTCATTACACCCGAGGGAGA 1380
     Ppy Exons ACGACAAGGATATGGGCTCACTGAGACTACATCAGCTATTCTGATTACACCCGAGGGGGA 1067

Node Consensus TGATAAGCCAGGCGCGGTCGGTAAAGTTGTTCCATTTTATGAGGCAAAAGTTGTGGATCT 1440
     Ppy Exons TGATAAACCGGGCGCGGTCGGTAAAGTTGTTCCATTTTTTGAAGCGAAGGTTGTGGATCT 1127

Node Consensus GGATACTGGGAAAACGCTGGGCCTTAAGCAGCGGGGTGAATTATGTGTCAGAGGACCTAT 1500
     Ppy Exons GGATACCGGGAAAACGCTGGGCGTTAATCAGAGAGGCGAATTATGTGTCAGAGGACCTAT 1187

Node Consensus GAATATGGCCGGTTATGTAAACAATCCGGAAGCGACTAATGCTTTGATTGACAAGGATGG 1560
     Ppy Exons GATTATGTCCGGTTATGTAAACAATCCGGAAGCGACCAACGCCTTGATTGACAAGGATGG 1247

Node Consensus ATGGTTACATTCTGGCGACATAGCATACTGGGATGAAGACGAACACTTCTTCATAGTTGA 1620
     Ppy Exons ATGGCTACATTCTGGAGACTTAGCTTACTGGGACGAAGACGAACACTTCTTCATAGTTGG 1307

Node Consensus CCGTTTGAAGTCTTTAATTAAATACAAAGGATATCAGGTAACGAAGATGTTTACATGCAC 1680
     Ppy Exons CCGCTTGAAGTCTTTAATTAAATACAAAGGATATCAGGTGGC------------------ 1349

Node Consensus ACACTCTATAATACCTGTAGGTAACCCCCGCTGAATTGGAATCGATATTGTTACAACACC 1740
     Ppy Exons -------------------------CCCCGCTGAATTGGAATCGATATTGTTACAACACC 1384

Node Consensus CCAACATCTTCGATGCGGGTGTGGCAGGTATTCCAGACGATGACGCCGGTGAACTTCCCG 1800
     Ppy Exons CCAACATCTTCGACGCGGGCGTGGCAGGTCTTCCCGACGATGACGCCGGTGAACTTCCCG 1444

Node Consensus CCGCCGTTGTTGTTTTGGAGACAGGAAAATCAATGACGGAAAACGAGATCGTGGATTACG 1860
     Ppy Exons CCGCCGTTGTTGTTTTGGAGCACGGAAAGACGATGACGGAAAAAGAGATCGTGGATTACG 1504

Node Consensus TGGCTAGTAAATGAATTCTATTGTGTTACTCATACTACACTTCGTTTCTTATTAATAGGT 1920
     Ppy Exons TCGCCAGT---------------------------------------------------- 1512

Node Consensus CAAGTAACAACGGCGAAACGGTTGCGCGGAGGAGTTGTATTTGTGGACGAAGTACCGAAA 1980
     Ppy Exons CAAGTAACAACCGCGAAAAAGTTGCGCGGAGGAGTTGTGTTTGTGGACGAAGTACCGAAA 1572

Node Consensus GGTCTAACCGGAAAACTCGACGCAAGAAAAATCAGAGATATCCTCGTAAAGGCCAAGAAG 2040
     Ppy Exons GGTCTTACCGGAAAACGCGACGCAAGAAAAATCAGAGAGATCCTCATAAAGGCCAAGAAG 1632

Node Consensus GGTGGAAAGGCCAAATTGTAAAATGTAACTCTGACGAAATTCTTAGCTATTGTAATATTA 2100
     Ppy Exons GGCGGAAAGTCCAAATTGTGA--------------------------------------- 1653

Node Consensus TATACAAATTGATGGATGGTAGCAGAATCACTTTATTATCGATAATTTTCAGTTCTTCCA 2160
     Ppy Exons ------------------------------------------------------------ 1653

Node Consensus TTGTACCGAAAGTTGCTAATTTTGTAATTGTGGGTCACTTGACTTCTTTAACGAATAATA 2220
     Ppy Exons ------------------------------------------------------------ 1653

Node Consensus AAATCTGGTATAGCTAAAAGGATTGAAATTTTTCAAAAAATATTAAGATGCAATGTATTG 2280
     Ppy Exons ------------------------------------------------------------ 1653

Node Consensus TCGTTGCGAATCCCGGAAAAAAGG 2304
     Ppy Exons ------------------------ 1653
```

Alignment of Node consensus and *Ppy* cDNA. The consensus sequence of the Node contigs from SPAdes assembly, aligned to only the cDNA of the *Ppy* Fluc gene. Highlighting in blue indicates the seven exon regions of the two sequences. Alignment graphic produced in CLC sequence viewer **(**Qiagen, Venlo, Netherlands).

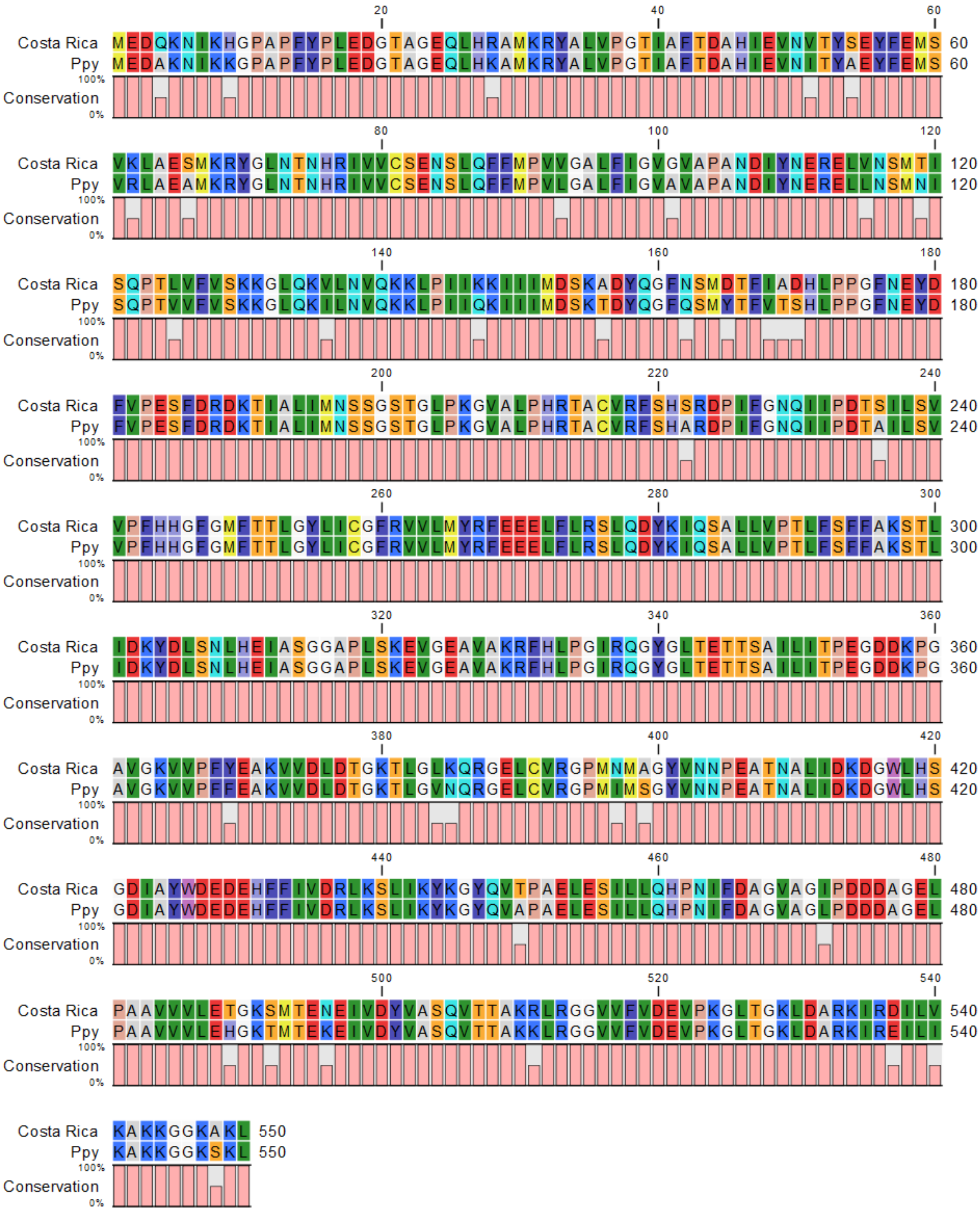
Figure 3.15.



```
Node Consensus    CGGTTTTCGTTTTGAAAGGACCGATCGTAAATATATGACGGATAGGTAAGTGGTTTGAAT 60
Costa Rica Exons  ------------------------------------------------------------ -
Node Consensus    GTCGCTTGGCGTAACATTAGCAAGTCGGTATTAACGATAAAATGGAAGACCAAAAAAACA 120
Costa Rica Exons  -----------------------------------------ATGGAAGACCAAAAAAACA 19
Node Consensus    TAAAACATGGTCCAGCGCCATTCTATCCTCTAGAAGATGGAACTGCTGGAGAACAACTGC 180
Costa Rica Exons  TAAAACATGGTCCAGCGCCATTCTATCCTCTAGAAGATGGAACTGCTGGAGAACAACTGC 79
Node Consensus    ATAGGGCTATGAAAAGATACGCCCTGGTTCCTGGGACAATTGCCTTTGTGAGTATAGTTC 240
Costa Rica Exons  ATAGGGCTATGAAAAGATACGCCCTGGTTCCTGGGACAATTGCCTTT------------- 126
Node Consensus    TGCCTGTTTTCTTCCCAGCGAGTGTTAATGAAATGTTCTTAATGTTTCTTTAGACAGATG 300
Costa Rica Exons  -----------------------------------------------------ACAGATG 133
Node Consensus    CACATATCGAGGTGAACGTCACGTACTCGGAATACTTTGAAATGTCCGTTAAATTAGCCG 360
Costa Rica Exons  CACATATCGAGGTGAACGTCACGTACTCGGAATACTTTGAAATGTCCGTTAAATTAGCCG 193
Node Consensus    AATCTATGAAACGGTATGGGCTTAATACAAATCACAGAATCGTCGTATGCAGTGAAAACT 420
Costa Rica Exons  AATCTATGAAACGGTATGGGCTTAATACAAATCACAGAATCGTCGTATGCAGTGAAAACT 253
Node Consensus    CTCTTCAATTCTTTATGCCTGTTGTGGGAGCGTTATTTATCGGAGTTGGAGTTGCGCCCG 480
Costa Rica Exons  CTCTTCAATTCTTTATGCCTGTTGTGGGAGCGTTATTTATCGGAGTTGGAGTTGCGCCCG 313
Node Consensus    CGAACGACATTTATAATGAACGTAAGCACCCTCACCAAACCCAAGAGAGAATGATGTAAT 540
Costa Rica Exons  CGAACGACATTTATAATGAACGT------------------------------------- 336
Node Consensus    TAATTTTTAAGGTGAATTGGTCAACAGTATGACTATTTCGCAGCCTACTTTAGTGTTTGT 600
Costa Rica Exons  -------------GAATTGGTCAACAGTATGACTATTTCGCAGCCTACTTTAGTGTTTGT 383
Node Consensus    TTCCAAAAAGGGGCTGCAGAAAGTTTTGAACGTGCAAAAAAAATTACCAATAATTAAGAA 660
Costa Rica Exons  TTCCAAAAAGGGGCTGCAGAAAGTTTTGAACGTGCAAAAAAAATTACCAATAATTAAGAA 443
Node Consensus    AATTATTATCATGGATTCTAAAGCAGATTACCAGGGATTTAATTCGATGGACACGTTCAT 720
Costa Rica Exons  AATTATTATCATGGATTCTAAAGCAGATTACCAGGGATTTAATTCGATGGACACGTTCAT 503
Node Consensus    CGCGGATCATTTACCTCCGGGCTTTAACGAATATGATTTTGTACCGGAGTCCTTTGATCG 780
Costa Rica Exons  CGCGGATCATTTACCTCCGGGCTTTAACGAATATGATTTTGTACCGGAGTCCTTTGATCG 563
Node Consensus    TGACAAGACAATTGCACTGATAATGAATTCCTCTGGCTCTACTGGGTTACCTAAGGGAGT 840
Costa Rica Exons  TGACAAGACAATTGCACTGATAATGAATTCCTCTGGCTCTACTGGGTTACCTAAGGGAGT 623
Node Consensus    GGCGCTTCCGCATAGAACTGCCTGTGTAAGATTCTCACATTGCAGGTATGTCGTGTAACA 900
Costa Rica Exons  GGCGCTTCCGCATAGAACTGCCTGTGTAAGATTCTCACATT------------------- 664
Node Consensus    AAGATTAAGTAATGTTCCTACAAAAATTCTAGAGATCCTATTTTTGGCAACCAAATCATT 960
Costa Rica Exons  ----------------------------CTAGAGATCCTATTTTTGGCAACCAAATCATT 696
Node Consensus    CCCGATACTTCGATTTTAAGTGTTGTTCCATTCCATCATGGTTTTGGAATGTTTACTACA 1020
Costa Rica Exons  CCCGATACTTCGATTTTAAGTGTTGTTCCATTCCATCATGGTTTTGGAATGTTTACTACA 756
Node Consensus    CTCGGATATCTGATATGCGGATTTCGTGTAGTCTTGATGTATAGATTTGAAGAAGAATTG 1080
Costa Rica Exons  CTCGGATATCTGATATGCGGATTTCGTGTAGTCTTGATGTATAGATTTGAAGAAGAATTG 816
Node Consensus    TTTTTACGATCCCTTCAAGACTACAAAATTCAAAGCGCGTTGTTAGTACCAACCCTATTT 1140
Costa Rica Exons  TTTTTACGATCCCTTCAAGACTACAAAATTCAAAGCGCGTTGTTAGTACCAACCCTATTT 876
Node Consensus    TCATTCTTCGCCAAAAGTACTCTGATTGACAAATACGATTTATCTAATTTACACGAAATT 1200
Costa Rica Exons  TCATTCTTCGCCAAAAGTACTCTGATTGACAAATACGATTTATCTAATTTACACGAAATT 936
Node Consensus    GCCTCTGGGGGCGCACCTCTTTCAAAAGAAGTTGGAGAAGCGGTTGCAAAACGGTGAGTT 1260
Costa Rica Exons  GCCTCTGGGGGCGCACCTCTTTCAAAAGAAGTTGGAGAAGCGGTTGCAAAACG------- 989
Node Consensus    AAGGGTATTGCTTGTTCTCCAAGGATCTAAAGCGGCGTGTAGCTTCCATCTTCCAGGGAT 1320
Costa Rica Exons  ------------------------------------------CTTCCATCTTCCAGGGAT 1007
Node Consensus    ACGACAAGGATATGGGCTCACCGAGACTACGTCAGCTATTCTCATTACACCCGAGGGAGA 1380
Costa Rica Exons  ACGACAAGGATATGGGCTCACCGAGACTACGTCAGCTATTCTCATTACACCCGAGGGAGA 1067
Node Consensus    TGATAAGCCAGGCGCGGTCGGTAAAGTTGTTCCATTTTATGAGGCAAAAGTTGTGGATCT 1440
Costa Rica Exons  TGATAAGCCAGGCGCGGTCGGTAAAGTTGTTCCATTTTATGAGGCAAAAGTTGTGGATCT 1127
Node Consensus    GGATACTGGGAAAACGCTGGGCCTTAAGCAGCGGGGTGAATTATGTGTCAGAGGACCTAT 1500
Costa Rica Exons  GGATACTGGGAAAACGCTGGGCCTTAAGCAGCGGGGTGAATTATGTGTCAGAGGACCTAT 1187
Node Consensus    GAATATGGCCGGTTATGTAAACAATCCGGAAGCGACTAATGCTTTGATTGACAAGGATGG 1560
Costa Rica Exons  GAATATGGCCGGTTATGTAAACAATCCGGAAGCGACTAATGCTTTGATTGACAAGGATGG 1247
Node Consensus    ATGGTTACATTCTGGCGACATAGCATACTGGGATGAAGACGAACACTTCTTCATAGTTGA 1620
Costa Rica Exons  ATGGTTACATTCTGGCGACATAGCATACTGGGATGAAGACGAACACTTCTTCATAGTTGA 1307
Node Consensus    CCGTTTGAAGTCTTTAATTAAATACAAAGGATATCAGGTAACGAAGATGTTTACATGCAC 1680
Costa Rica Exons  CCGTTTGAAGTCTTTAATTAAATACAAAGGATATCAGGTAAC------------------ 1349
Node Consensus    ACACTCTATAATACCTGTAGGTAACCCCCGCTGAATTGGAATCGATATTGTTACAACACC 1740
Costa Rica Exons  -------------------------CCCCGCTGAATTGGAATCGATATTGTTACAACACC 1384
Node Consensus    CCAACATCTTCGATGCGGGTGTGGCAGGTATTCCAGACGATGACGCCGGTGAACTTCCCG 1800
Costa Rica Exons  CCAACATCTTCGATGCGGGTGTGGCAGGTATTCCAGACGATGACGCCGGTGAACTTCCCG 1444
Node Consensus    CCGCCGTTGTTGTTTTGGAGACAGGAAAATCAATGACGGAAAACGAGATCGTGGATTACG 1860
Costa Rica Exons  CCGCCGTTGTTGTTTTGGAGACAGGAAAATCAATGACGGAAAACGAGATCGTGGATTACG 1504
Node Consensus    TGGCTAGTAAATGAATTCTATTGTGTTACTCATACTACACTTCGTTTCTTATTAATAGGT 1920
Costa Rica Exons  TGGCTAGT---------------------------------------------------- 1512
Node Consensus    CAAGTAACAACGGCGAAACGGTTGCGCGGAGGAGTTGTATTTGTGGACGAAGTACCGAAA 1980
Costa Rica Exons  CAAGTAACAACGGCGAAACGGTTGCGCGGAGGAGTTGTATTTGTGGACGAAGTACCGAAA 1572
Node Consensus    GGTCTAACCGGAAAACTCGACGCAAGAAAAATCAGAGATATCCTCGTAAAGGCCAAGAAG 2040
Costa Rica Exons  GGTCTAACCGGAAAACTCGACGCAAGAAAAATCAGAGATATCCTCGTAAAGGCCAAGAAG 1632
Node Consensus    GGTGGAAAGGCCAAATTGTAAAATGTAACTCTGACGAAATTCTTAGCTATTGTAATATTA 2100
Costa Rica Exons  GGTGGAAAGGCCAAATTGTAA--------------------------------------- 1653
Node Consensus    TATACAAATTGATGGATGGTAGCAGAATCACTTTATTATCGATAATTTTCAGTTCTTCCA 2160
Costa Rica Exons  ------------------------------------------------------------ 1653
Node Consensus    TTGTACCGAAAGTTGCTAATTTTGTAATTGTGGGTCACTTGACTTCTTTAACGAATAATA 2220
Costa Rica Exons  ------------------------------------------------------------ 1653
Node Consensus    AAATCTGGTATAGCTAAAAGGATTGAAATTTTTCAAAAAATATTAAGATGCAATGTATTG 2280
Costa Rica Exons  ------------------------------------------------------------ 1653
Node Consensus    TCGTTGCGAATCCCGGAAAAAAGG 2304
Costa Rica Exons  ------------------------ 1653
```

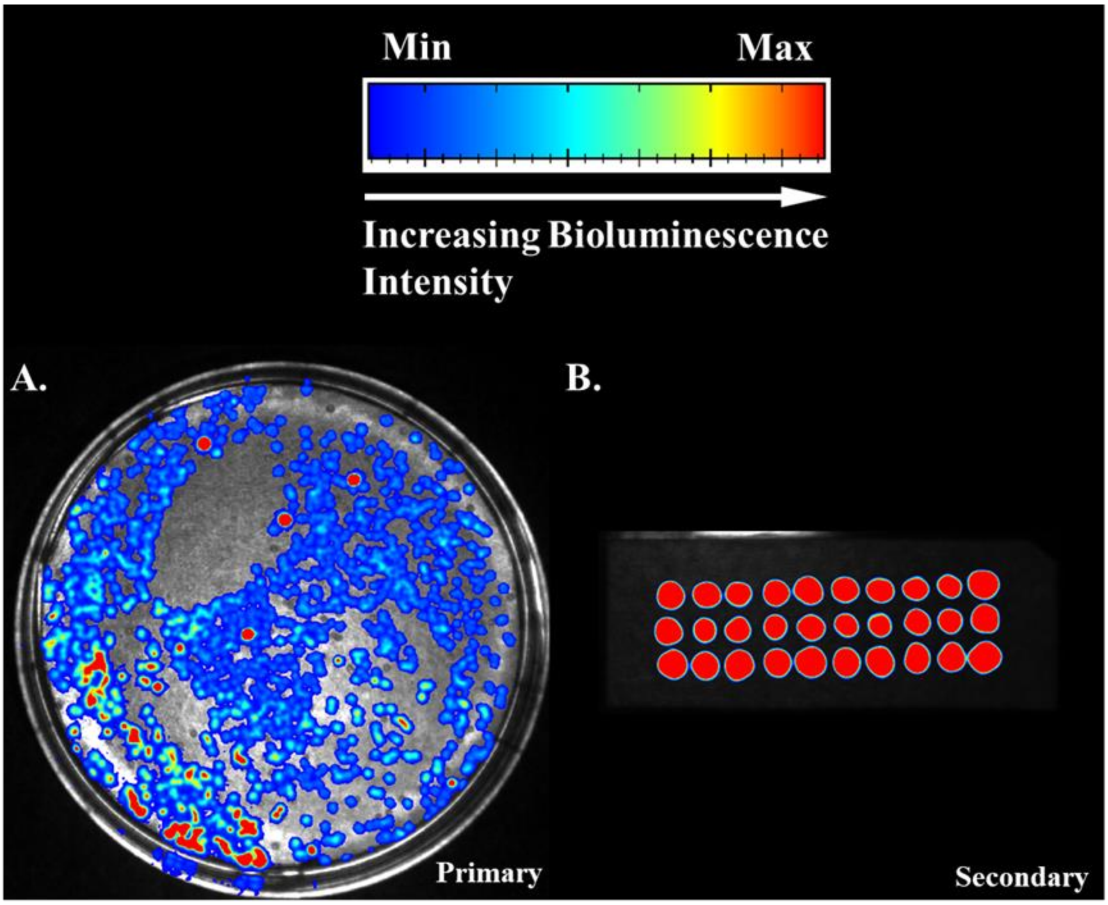
Alignment of the Node consensus and Costa Rica Fluc predicted exons. The consensus sequence of the Node contigs from SPAdes assembly, aligned to the predicted exons of the Costa Rica Fluc. Highlighting in red indicates the seven exon regions. Alignment graphic produced in CLC sequence viewer **(**Qiagen, Venlo, Netherlands).

Figure 3.16.


20
40
60
Costa Rica MEDQKNIKHGPAPFYPLEDGTAGEQLHRAMKRYALVPGTIAFTDAHIEVNVTYSEYFEMS 60
Ppy MEDAKNIKKGPAPFYPLEDGTAGEQLHKAMKRYALVPGTIAFTDAHIEVNITYAEYFEMS 60
100%
Conservation
0%
80
100
120
Costa Rica VKLAESMKRYGLNTNHRIVVCSENSLQFFMPVVGALFIGVGVAPANDIYNERELVNSMTI 120
Ppy VRLAEAMKRYGLNTNHRIVVCSENSLQFFMPVLGALFIGVAVAPANDIYNERELLNSMNI 120
100%
Conservation
0%
140
160
180
Costa Rica SQPTLVFVSKKGLQKVLNVQKKLPIIKKIIIMDSKADYQGFNSMDTFIADHLPPGFNEYD 180
Ppy SQPTVVFVSKKGLQKILNVQKKLPIIQKIIIMDSKTDYQGFQSMYTFVTSHLPPGFNEYD 180
100%
Conservation
0%
200
220
240
Costa Rica FVPESFDRDKTIALIMNSSGSTGLPKGVALPHRTACVRFSHSRDPIFGNQIIPDTSILSV 240
Ppy FVPESFDRDKTIALIMNSSGSTGLPKGVALPHRTACVRFSHARDPIFGNQIIPDTAILSV 240
100%
Conservation
0%
260
280
300
Costa Rica VPFHHGFGMFTTLGYLICGFRVVLMYRFEEELFLRSLQDYKIQSALLVPTLFSFFAKSTL 300
Ppy VPFHHGFGMFTTLGYLICGFRVVLMYRFEEELFLRSLQDYKIQSALLVPTLFSFFAKSTL 300
100%
Conservation
0%
320
340
360
Costa Rica IDKYDLSNLHEIASGGAPLSKEVGEAVAKRFHLPGIRQGYGLTETTSAILITPEGDDKPG 360
Ppy IDKYDLSNLHEIASGGAPLSKEVGEAVAKRFHLPGIRQGYGLTETTSAILITPEGDDKPG 360
100%
Conservation
0%
380
400
420
Costa Rica AVGKVVPFYEAKVVDLDTGKTLGLKQRGELCVRGPMNMAGYVNNPEATNALIDKDGWLHS 420
Ppy AVGKVVPFFEAKVVDLDTGKTLGVNQRGELCVRGPMIMSGYVNNPEATNALIDKDGWLHS 420
100%
Conservation
0%
440
460
480
Costa Rica GDIAYWDEDEHFFIVDRLKSLIKYKGYQVTPAELESILLQHPNIFDAGVAGIPDDDAGEL 480
Ppy GDIAYWDEDEHFFIVDRLKSLIKYKGYQVAPAELESILLQHPNIFDAGVAGLPDDDAGEL 480
100%
Conservation
0%
500
520
540
Costa Rica PAAVVVLETGKSMTENEIVDYVASQVTTAKRLRGGVVFVDEVPKGLTGKLDARKIRDILV 540
Ppy PAAVVVLEHGKTMTEKEIVDYVASQVTTAKKLRGGVVFVDEVPKGLTGKLDARKIREILI 540
100%
Conservation
0%
Costa Rica KAKKGGKAKL 550
Ppy KAKKGGKSKL 550
100%
Conservation
0%

Alignment of Costa Rica Fluc and *Ppy* Fluc. The proposed translated amino acid sequence of Costa Rica Fluc aligned to the amino acid sequence of *Ppy* Fluc. A conservation bar plot indicates the mismatches between the two protein sequences. Percent identity was recorded at 93.45%. Alignment graphic produced in CLC sequence viewer **(**Qiagen, Venlo, Netherlands).

Figure 3.17.



Bioluminescence activity screening of Costa Rica Fluc. The colony formations from *E. coli* BL21 (DE3) transformed with pET16b plasmid containing Costa Rica Fluc, here transferred onto nitrocellulose membranes (Chapter 2). Colony transferred membranes were induced for 4 hours at RT with IPTG (1 mM). All plates were subsequently screened with 500 µM LH_2 and imaged in the PhotonIMAGER Optima (Biospace Labs, Paris, France). The upper colour spectrum illustrates how bioluminescence signal intensity is represented relative to the maximum signal detected, with minimum intensity indicated in blue and maximum intensity indicated in red. (**A**) Bioluminescence activity from the primary screen of plated transformants over an integrating period of 20 seconds. (**B**) Bioluminescence activity from the secondary screen of randomly picked colonies from the primary screen, over an integrating period of 60 seconds. Images produced in the M3 Vision software package. Intensity scaling between **A** and **B** is not directly comparable.

3.4. Further Discussion

The initial investigation of the non-destructive DNA extraction process from Gilbert *et al* (2007) using three samples of varied age and storage condition provided equally varied results in regards to the quality of the extracted DNA. An obvious correlation was observed between specimen age and storage condition, as indicated by the 217 bp average fragment size recorded in the dry-preserved *Ppy* collected in 1996, relative to average size of 17273 bp recorded in the 2006 collected *Lnoc* stored fresh at -80 ˚C. The failure to recover a quantifiable level of DNA from the Bornean firefly collected in 1987 brought awareness to the possibility that DNA from museum specimens may have experienced extensive degradation from the use of chemical asphyxiates including ethyl acetate, which is commonly used to dispatch collected insects as future genetic analyses are not a common consideration (Dillon *et al.* 1996). Importantly, the double stranded break mechanism of accelerated degradation from the use of chemical asphyxiates occurs independently of the specimen age, and therefore without this information the age of a museum specimen could not be taken as a direct proxy for the quality of DNA available for extraction. With this in consideration, no correlation was observed in the DNA extracts from the five fireflies of interest between age and the average fragment size. The Costa Rica firefly collected in 2012 had an average fragment size of 881 bp, whilst the average size of the three USA fireflies collected between 2013 – 2015 was 146 – 178 bp, which was in line with the 146 bp average recorded for the Indonesia firefly collected ≈ 30 years previously in 1985. Although information on asphyxiate usage was not recorded for any of the five fireflies, an explanation for the recorded average fragment sizes may be that the USA samples could have been exposed to such chemical asphyxiates, in contrast with the samples from Costa Rica and Indonesia which have likely experienced degradation as a product of their age and general storage conditions such as temperature and humidity.

Mini-barcoding was initially performed to explore the possibility of PCR amplification of short fragment targets, as mini-barcodes have previously been demonstrated to provide >90% species-level resolution in degraded DNA samples (Hajibabaei *et al.* 2006b; Meusnier *et al.* 2008). The sequence information from successful amplifications of the mini-barcodes enabled the identification of the highest identity match through BLAST. This analysis was used to determine whether the five unidentified museum fireflies could be matched to an identified species with a known luciferase gene sequence. If an unidentified museum firefly had been matched with an identified species which had a known luciferase gene sequence associated, it could have been dismissed from further analysis.

The mini-barcodes from USA –Pennsylvania and USA – unk. shared a sequence identity of 97.80%, and individually matched >98% to their BLAST hits of the COI from *Lucidota atra*. Although this suggested a high possibility that both fireflies were *Lucidota atra* or a closely related species of *Lucidota*, no luciferase gene sequence information is available for any species of *Lucidota*, and for this reason investigation of both samples was continued, regardless of the high possibility for redundancy. The mini-barcode from the USA – Maryland specimen was discovered to share 95.24% sequence identity with *Photinus interdius*, a fully diurnal species from Panama (Vencl *et al.* 2017). Although the previously discussed species-level resolution of mini-barcodes in degraded DNA of >90% implies the possibility that USA – Maryland could be *Photinus interdius*, the >2000-mile separation and entirely different environmental conditions between Maryland – USA and Panama suggest the two to be distinct, but closely related species. Similarly, although the mini-barcode of the Costa Rica specimen shared 92.31% sequence identity with the North American firefly *Photnius australis* COI region, this result most likely indicates that the Costa Rica specimen is an unidentified firefly that currently is most closely associated with *Photnius australis*, in absence of further Lampyridae barcode availability.

Tropical America is presumed to be the origin and region of greatest lampyrid diversity, yet a large proportion of this diversity remain unknown (Stanger-Hall *et al.* 2007). *Photinus* are the largest genus of Lampyridae in the Americas, with over 235 members identified (Vencl *et al.* 2017). Currently, the BOLD database (www.boldsystems.org) contains 486 *Photinus* specimen barcodes, from only 28 species, and only 177 species with barcodes in the family Lampyridae. Without a considerable expansion to global understanding of Lampyridae biological diversity, the use of barcoding as a strategy to identify phylogenetic relationships will remain extremely limited.

Development of a CODEHOP primer set was explored to provide a direct approach to detecting the presence of novel luciferase gene sequences in the DNA extract libraries from an unidentified museum firefly. It would additionally demonstrate whether nuclear DNA targets could be amplified similarly to the mitochondrial DNA which contains the mini-barcode COI target. CODEHOP amplification and product sequencing was successful in both trial libraries of *Ppy* and *Lnoc.* Although amplification initially appeared successful in the Museum firefly libraries, the identification of a 49 bp conserved region in all of the amplification sequences undermines the validity of the amplifications seen in qPCR. As the sequence of the conserved region could not be matched directly with the *Ppy* luciferase gene, probe contamination could be ruled out. Attempts to sequence the amplification products from the pre-enriched libraries were less successful and produced sequence reads with significant clipping, possibly owing to the poor performance of the DKYD-F CODEHOP primer in Sanger sequencing. Only the amplification product from the pre-enriched Costa Rica library could be successfully sequenced, which disregarding terminal variation and a single mismatch was identical to the sequence of the CODEHOP amplification in the enriched Costa Rica library.

In a retrospective attempt to understand the CODEHOP results, the full Costa Rica CODEHOP amplification and the 49 bp conserved region were aligned to the completed CDS assembled from the bioinformatic analysis of the enriched Costa Rica library. These alignments identified only 79.65% shared identity of the full CODEHOP amplification and 89.90% in the conserved region, indicating that these amplifications had not been of the Costa Rica luciferase gene sequence.

Prior to the unexpected sequencing results observed in CODEHOP amplification, enrichment of luciferase gene sequences was performed using cross-species affinity enrichment probes constructed from the *Ppy* luciferase gene sequence. Although qPCR verification of enrichment initially appeared successful, due to the issues with the CODEHOP sequencing results, the verification of enrichment with the CODEHOP primer set is undermined, and therefore failure to successfully enrich the four libraries other than Costa Rica could explain their ultimate failure in the bioinformatic analyses.

Failure to recover luciferase gene contigs from the four libraries other than Costa Rica might also be attributed to failures in the bioinformatic process. Initial analysis of the trimmed sequencing reads by FastQC was good for all enriched libraries. A high total volume of paired reads was produced for each library, although Costa Rica had approximately double the paired reads of any other enriched library. In the failed libraries, likely explanations for the low total of reads successfully mapped are either insufficient sequence complementarity to the probes in enrichment, or to the reference genomes in the Bowtie2 alignment. From the successful mapping of reads in the enriched Costa Rica library to the *Ppy* reference, a gene sequence was derived which shared 88.79% sequence identity with *Ppy* Fluc complete CDS.

However, no reads from the enriched Costa Rica library were able to map to the luciferase gene region in the reference genomes for *Aquatica lateralis* and *Ignelater luminosus*. To understand this, the Costa Rica Fluc complete CDS was compared with a luciferase gene sequence available for *A. lateralis* (GenBank: Z49891.1), revealing only 60.38% sequence identity (no luciferase gene sequence is available for *I. luminosus*). This suggests that the inability of Costa Rica reads to map to the reference of *A. lateralis* and *I. luminosus* is due to the excessive divergence in sequence identity, which could also explain the reduced ability of the unsuccessful enriched libraries to map to the *Ppy* reference genome. The few reads that were successfully mapped from these libraries were discovered to share >88% sequence identity with the *Ppy* Fluc gene (Appendices Table 9.6.), and therefore possessed sufficient complementarity to exceed the unknown threshold. Attempts to assemble the trimmed reads with SPAdes without mapping by Bowtie2 were entirely unsuccessful for all libraries, including the enriched Costa Rica library. This indicates that mapping and subsequent assembly of only the reads associated with the luciferase gene region, was necessary to the success of the bioinformatic process in the enriched Costa Rica library.

To understand whether the enrichment process had been inconsequential to the ultimate success of the Costa Rica library, a sample of the non-enriched library was sequenced and the data processed using the same scripts. This non-enriched library failed to reproduce the results of the enriched library, confirming that the enrichment strategy was necessary for the success in the Costa Rica library, regardless of the issues verifying enrichment with the CODEHOP primer set. The principles of this method for cross-species affinity enrichment have previously been demonstrated to be capable of enriching sequences 10-13% divergent from the biotin probe identity (Mason *et al.* 2011). It may be that the four unsuccessful libraries possessed sequence divergences relative to *Ppy* that exceeded this threshold, and therefore failed to hybridize to enable enrichment. If the failure had been with the enrichment and not the read mapping by Bowtie2, the use of multiple firefly luciferase genes to generate a pool of varied probe sequences may have enabled a greater opportunity for successful enrichment.

From the successfully mapped reads of the enriched Costa Rica library, SPAdes was able to assemble three Nodes, which could be overlapped to produce a single contig of 2304 bp in length. The 213 bp region of overlap between the three Nodes revealed three mismatched positions. Ultimately, these mismatches would be inconsequential on the associated Costa Rica Fluc, as two were identified as located within the 4th intron, and the third mismatch conferred a synonymous substitution at the 337th codon, such that both variants would contain codons for arginine at this position. However, three mismatches at this short region of overlap suggest that further erroneous bases could be present throughout the derived Costa Rica gene sequence, which may not be as inconsequential as the three mismatches that were observed. With no way to verify the validity of the luciferase gene sequence from the Costa Rica firefly, it can currently only be considered an approximation of the true gene as opposed to directly representative.

Regardless of the uncertainty in the sequence accuracy, extraction of the information which related only to the exons through comparative analysis of the exon-intron structure in the *Ppy* Fluc enabled the derivation of a speculative protein sequence for the Costa Rica Fluc which shared 93.45% amino acid sequence identity with the *Ppy* Fluc. This speculative enzyme was subsequently screened in transformed *E. coli*, which verified its bioluminescence capability. This result confirmed that dry-preserved museum specimens of Lampyridae could be investigated to provide novel luciferase gene sequences, although further work would be required to characterise the bioluminescence properties of the Costa Rica Fluc (see Chapter 6).

3.5. Conclusions

Dry-preserved museum Coleoptera collections were demonstrated as a valuable genetic repository from which simple analyses can be readily conducted such as the targeted amplification of mini-barcodes. However, the limited availability of Lampyridae species with barcode sequences inhibits their current use as a method of identification. Cross-species affinity enrichment was validated as a strategy to improve the read mapping of a targeted gene in Illumina sequencing analysis, without which the bioinformatic processing was subject to failure. Although, this enrichment process was dependent on a high degree of shared sequence identity between the probe and target sequences, and could likely be improved with the use of a probe pool of divergent gene sequences. Ultimately, a novel luciferase gene sequence was isolated for the Costa Rican firefly and was demonstrated to encode a luciferase enzyme capable of bioluminescent functionality. Whilst this enzyme cannot be definitively claimed as an accurate representation of the *wild-type* enzyme in nature, it serves as a novel source of luciferase gene variation, which is yet to be meaningfully characterised.

Chapter 4

Engineering Infraluciferin Compatibility by Rational Design

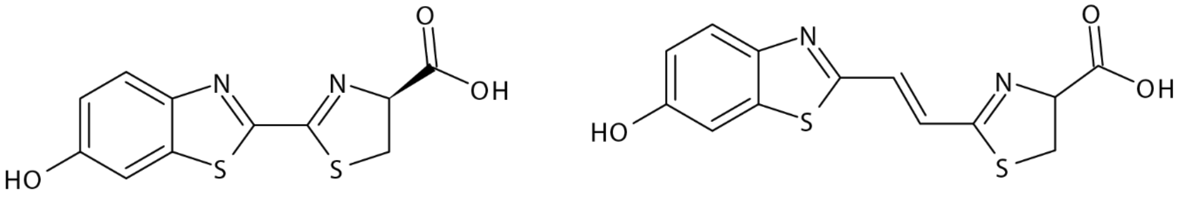
4.1. Chapter Summary

In this chapter the primary and secondary structure of a novel luciferase from the lesser British Glow-worm *Phosphaenus hemipterus* (*Phem*Luc) was analysed, and compatibility with the *Photinus pyralis* (*Ppy*) firefly luciferase (Fluc) assessed for the purpose of homology modelling. Available crystal structures of *Ppy* Fluc bound to structural analogues of luciferin (LH_2) and Infraluciferin (iLH_2) were used to construct homology models of *Phem*Luc catalysing each substrate. These models were used to identify 32 amino acids targets to mutagenize by SDM in order to identify improved bioluminescence activity variants. Two potential mutations were identified and combined to make a dual-mutant termed “x2 Infra” which was purified for further analysis. Analysis of the x2 Infra enzyme verified an increase in the bioluminescence yield relative to the *wild-type Phem*Luc. An attempt was made to explain the improvement through further homology modelling of x2 Infra, but no clear explanation could be found.

4.2. Introduction

Bioluminescence imaging (BLI) is a non-invasive imaging application which allows the longitudinal tracking of disease models in living animals through the incorporation of luciferase reporter genes into the tissue of interest (see Chapter 1) (Adams and Miller 2014). A limitation for BLI is presented by the presence of haemoglobin and myoglobin in mammalian tissues which scatter and attenuate visible light shorter than 600 nm in wavelength (Rice *et al.* 2002; Rice and Contag 2009). Therefore modern approaches to BLI systems exploit luciferase bioluminescence which is emitted in the ideal wavelength range of 600-800 nm, otherwise known as the bio-optical window for imaging *in vivo* (Iwano *et al.* 2013). Whilst red-shifted bioluminescence emissions can be achieved though luciferase engineering alone, current efforts seek to develop synthetic substrate analogues capable of producing significant redshifts to emission independent of the paired luciferase. One such analogue of Coleopteran beetle luciferin (LH_2) has been described termed Infraluciferin (iLH_2), which functions as the first dual-colour, far red to near-infrared synthetic substrate analogue for multiparametric imaging, with a peak emission wavelength (λ_{max}) of 706 nm (further detail in Chapter 1) (Jathoul *et al.* 2014).

Figure 4.1.



Chemical structures of native *D*-Luciferin (*D*-LH_2) and analogue infraluciferin (*DL*-iLH_2). Luciferin and analogue illustrations demonstrating the extended carbon linker of *DL*-Infraluciferin relative to *D*-Luciferin. Infraluciferin utilised within this work was comprised of a racemic mix of (*DL*-). Molecule structures produced using ChemSketch. ACD/ChemSketch, version 2021.1.1, Advanced Chemistry Development, Inc., Toronto, ON, Canada, www.acdlabs.com, 2021.

The structure of iLH_2 incorporates an extended carbon linker relative to the native LH_2, demonstrated in Figure 4.1. Whilst this addition confers the desired ability to produce near-infrared emissions, the total bioluminescence emission from the paired luciferase enzyme is significantly diminished in comparison to catalysis of LH_2. In order to produce bioluminescence yields sufficient for the purposes of BLI, engineered variants of firefly luciferases are required to improve activity. Therefore, this chapter sought to investigate whether mutagenesis of amino acids positions in greatest proximity to bound LH_2 or iLH_2 could improve the bioluminescence output of a novel luciferase from the lesser British Glow-worm *Phosphaenus hemipterus* (*Phem*Luc) with iLH_2. As previously discussed in the main introduction, whilst *Phem*Luc remains uncharacterised it presents the opportunity to discover novel mutagenic functionality which may be specific to the *Phem*Luc protein, or ultimately contribute to the pool of known conserved mutagenic targets across related enzyme variants. In order to identify mutagenesis targets in *Phem*Luc, protein models were required to identify amino acid residues that may be involved in the active site and subsequent reaction catalysis, due to their proximity with the bound substrate. Through this modelling work and mutagenic strategy a dual mutant termed x2 Infra was constructed which displayed improved bioluminescence activity with iLH_2 over *Phem*Luc. Further investigation of how these two mutations augment iLH_2 catalysis is required to understand their function and whether this action is conserved for mutations at corresponding positions in homologous luciferases.

4.3. Results and Discussion

4.3.1. Primary structural analysis of *Phem*Luc

4.3.1.1 Aligning *Phem*Luc to *Ppy*

As the aim of this work is to engineer the previously uncharacterised luciferase *Phosphaenus hemipteris* luciferase (*Phem*Luc) for improved activity with the LH_2 analogue iLH_2, we first sought to 3D model the novel enzyme to aid *in silico* analyses. To begin, an initial assessment of the shared amino acid sequence identity was required to understand whether there would be sufficient compatibility with the well-studied *Photinus pyralis* (*Ppy*) firefly luciferase (Fluc), for which several crystal structures exist. Although the *Ppy* Fluc could be engineered directly for its own improved activity with iLH_2 it has previously been extensively studied through mutagenesis, whereas *Phem*Luc presents the opportunity to explore a related novel scaffold. The computational approach of homology modelling is a long-established and reliable method for generating protein models of an unknown structure, if a sufficiently similar protein crystal structure is available. A high degree of shared amino acid sequence identity can increase the reliability of any structures produced (Waterhouse *et al.* 2018). For this purpose, a homology alignment was performed of *Phem*Luc and *Ppy* Fluc (GenBank: AAA29795.1 (De Wet *et al.* 1987)) using CLC Sequence Viewer (CLC Sequence Viewer, version 8.0, QIAGEN Aarhus, Denmark, www.digitalinsights.qiagen.com). From the alignment (Figure 4.2.) it was shown that *Phem*Luc possesses a high degree of sequence identity to *Ppy* Fluc at the amino acid level (87.07%), suggesting that *Phem*Luc would be a good candidate for homology modelling with *Ppy* Fluc.

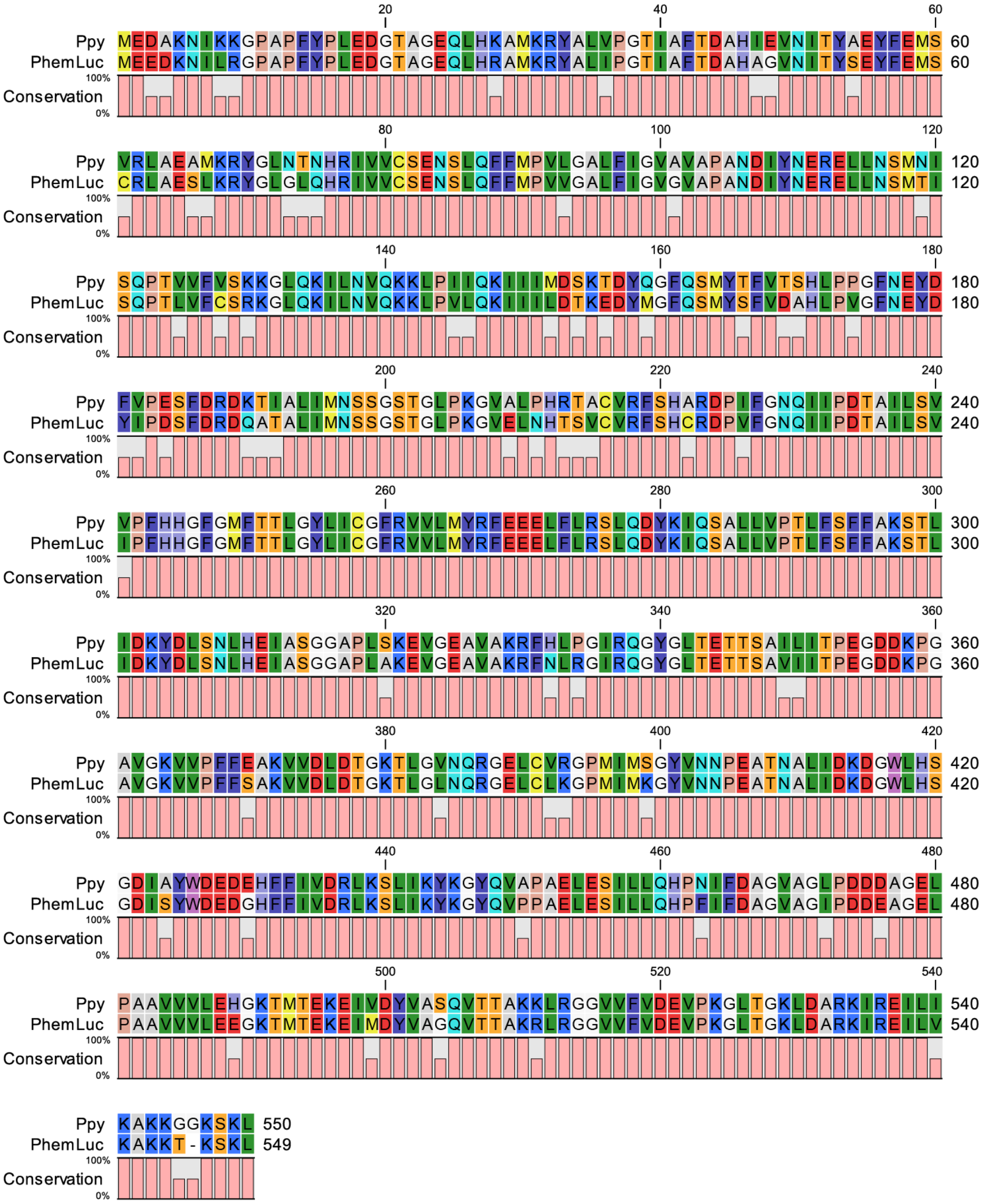
4.3.1.2 Analysis of residues unique to *Phem*Luc

The *Phem*Luc protein sequence was further aligned against a total of 13 firefly luciferase sequences identified from the NCBI database. This multiple sequence alignment (Figure 4.3.) was used to assess the conservation of *Phem*Luc amongst a homologous population of firefly luciferases, and more specifically to assess any amino acid positions unique to *Phem*Luc, shown in Table 4.1. The polarity of amino acids unique to *Phem*Luc were additionally assessed relative to the population diversity at each respective position within the alignment.

Polarity of amino acids is determined by the side chains (R groups), i.e. whether they are hydrophobic, (non-polar amino acids,) or hydrophilic, resulting from polar amino acids. The order of amino acids within a protein sequence dictates the emergent structure and catalytic activity of the corresponding protein. This is largely determined by the hydrophilic or hydrophobic status of each amino acid, as the hydrophobic positions are more likely to be located towards the core of the protein, and hydrophilic positions will typically be solvent exposed, where the polarity will allow the formation of hydrogen bonds with polar molecules in the aqueous environment, including water (Khan *et al.* 2017).

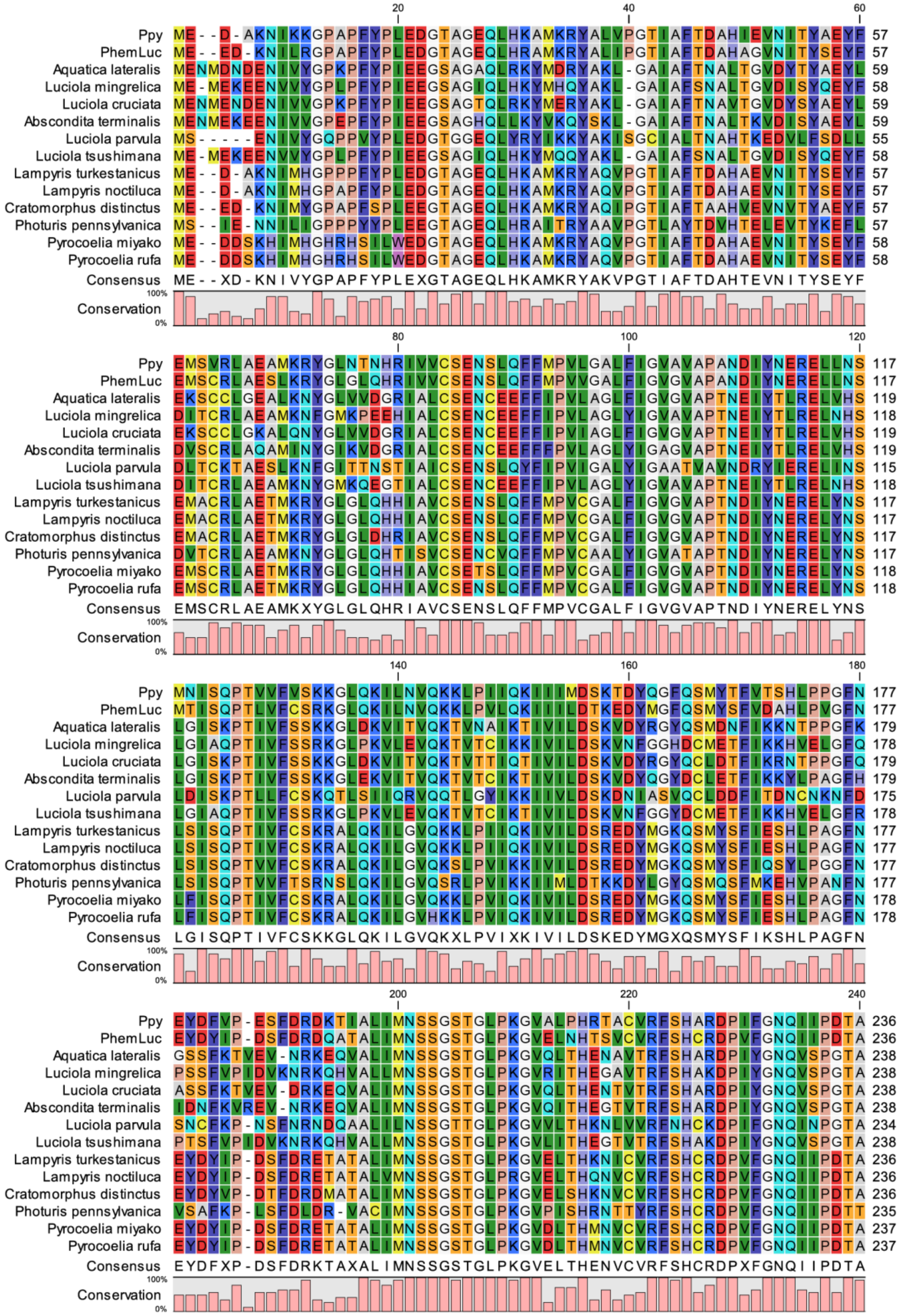
Firefly luciferases are typically well conserved, as can be visualised in Figure 4.3., and the significant majority of amino acids at any given position are common to 2 or more of the 14 sequences aligned in total. In *Phem*Luc, only 12 positions could be identified which contained amino acids unique only to *Phem*Luc (Table 4.1.). Of these unique amino acids, the majority occur as substitutions for highly similar amino acids, such as at position 146. Amongst all other sequences in the alignment, position 146 is conserved for isoleucine in contrast with *Phem*Luc which contains the structural isomer leucine, meaning that the molecular formula is identical, and the position is conserved for amino acid properties. More significant substitutions can only be found at 2 positions. Firstly, position 170 which is occupied by alanine in *Phem*Luc, making it uniquely non-polar relative to the population of luciferase sequences in the alignment which are all polar. The second is position 424, which is conserved for non-polar sequences, with the exception of serine in *Phem*Luc, making it uniquely polar at this site.

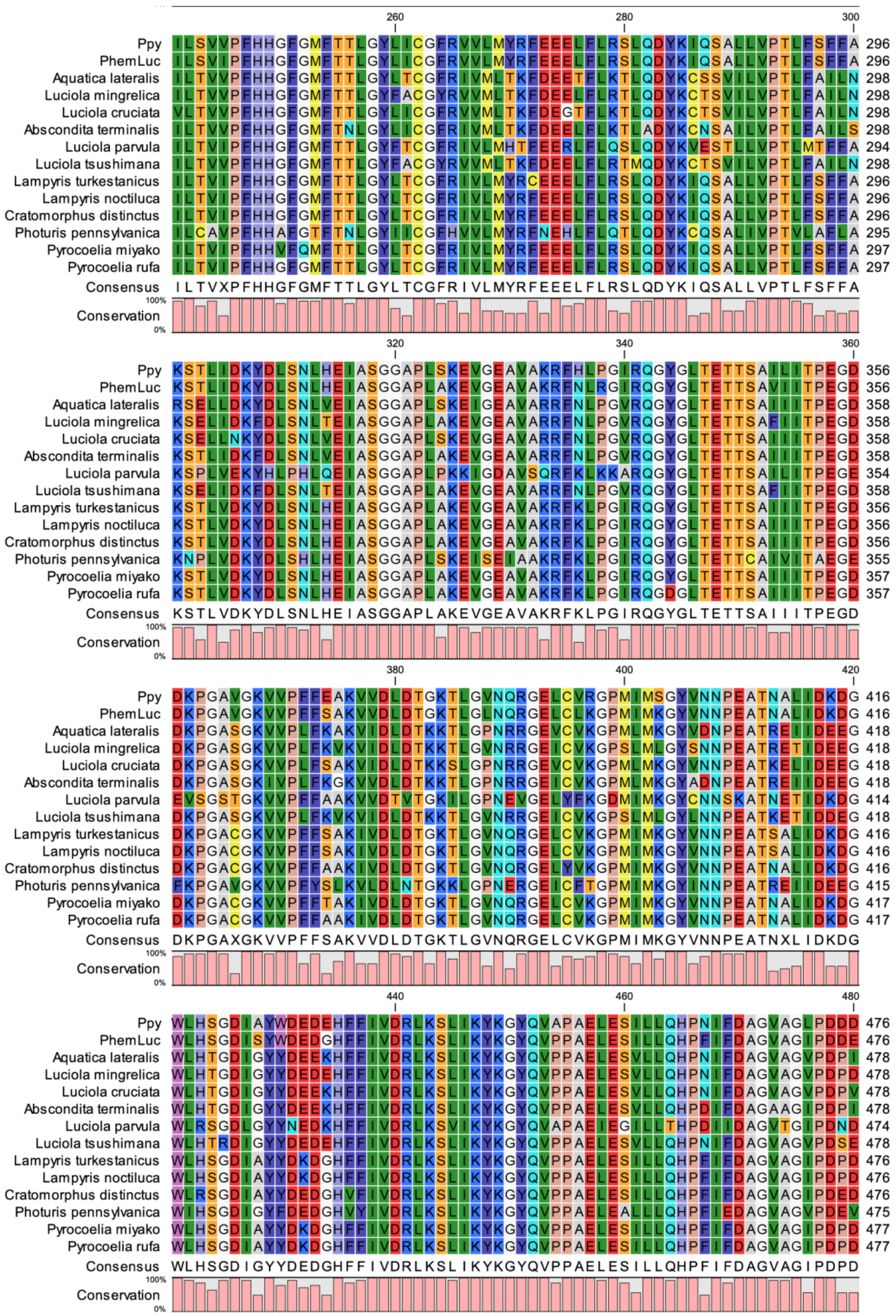
Figure 4.2.

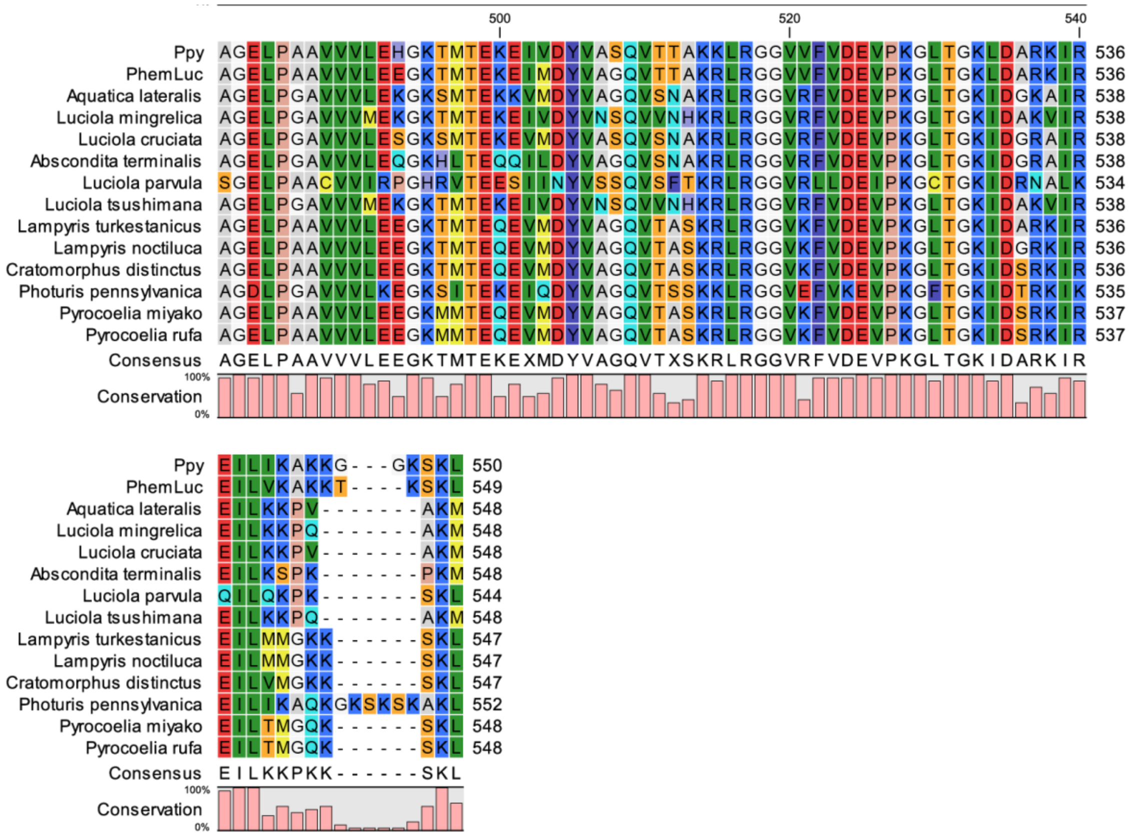

20
40
60
Ppy MEDAKNIKKGPAPFYPLEDGTAGEQLHKAMKRYALVPGTIAFTDAHIEVNITYAEYFEMS 60
PhemLuc MEEDKNILRGPAPFYPLEDGTAGEQLHRAMKRYALIPGTIAFTDAHAGVNITYSEYFEMS 60
100%
Conservation
0%
80
100
120
Ppy VRLAEAMKRYGLNTNHRIVVCSENSLQFFMPVLGALFIGVAVAPANDIYNERELLNSMNI 120
PhemLuc CRLAESLKRYGLGLQHRIVVCSENSLQFFMPVVGALFIGVGVAPANDIYNERELLNSMTI 120
100%
Conservation
0%
140
160
180
Ppy SQPTVVFVSKKGLQKILNVQKKLPIIQKIIIMDSKTDYQGFQSMYTFVTSHLPPGFNEYD 180
PhemLuc SQPTLVFCSRKGLQKILNVQKKLPVLQKIIILDTKEDYMGFQSMYSFVDAHLPVGFNEYD 180
100%
Conservation
0%
200
220
240
Ppy FVPESFDRDKTIALIMNSSGSTGLPKGVALPHRTACVRFSHARDPIFGNQIIPDTAILSV 240
PhemLuc YIPDSFDRDQATALIMNSSGSTGLPKGVELNHTSVCVRFSHCRDPVFGNQIIPDTAILSV 240
100%
Conservation
0%
260
280
300
Ppy VPFHHGFGMFTTLGYLICGFRVVLMYRFEEELFLRSLQDYKIQSALLVPTLFSFFAKSTL 300
PhemLuc IPFHHGFGMFTTLGYLICGFRVVLMYRFEEELFLRSLQDYKIQSALLVPTLFSFFAKSTL 300
100%
Conservation
0%
320
340
360
Ppy IDKYDLSNLHEIASGGAPLSKEVGEAVAKRFHLPGIRQGYGLTETTSAILITPEGDDKPG 360
PhemLuc IDKYDLSNLHEIASGGAPLAKEVGEAVAKRFNLRGIRQGYGLTETTSAVIITPEGDDKPG 360
100%
Conservation
0%
380
400
420
Ppy AVGKVVPFFEAKVVDLDTGKTLGVNQRGELCVRGPMIMSGYVNNPEATNALIDKDGWLHS 420
PhemLuc AVGKVVPFFSAKVVDLDTGKTLGLNQRGELCLKGPMIMKGYVNNPEATNALIDKDGWLHS 420
100%
Conservation
0%
440
460
480
Ppy GDIAYWDEDEHFFIVDRLKSLIKYKGYQVAPAELESILLQHPNIFDAGVAGLPDDDAGEL 480
PhemLuc GDISYWDEDGHFFIVDRLKSLIKYKGYQVPPAELESILLQHPFIFDAGVAGIPDDEAGEL 480
100%
Conservation
0%
500
520
540
Ppy PAAVVVLEHGKTMTEKEIVDYVASQVTTAKKLRGGVVFVDEVPKGLTGKLDARKIREILI 540
PhemLuc PAAVVVLEEGKTMTEKEIMDYVAGQVTTAKRLRGGVVFVDEVPKGLTGKLDARKIREILV 540
100%
Conservation
0%
Ppy KAKKGGKSKL 550
PhemLuc KAKKT-KSKL 549
100%
Conservation
0%

Homology alignment of *Ppy* Fluc and *Phem*Luc amino acid sequences. Shown are the amino acid sequences of *Photinus pyralis* luciferase, *Ppy* Fluc, and *Phosphaenus hemipterus* luciferase, *Phem*Luc. Degree of conservation is represented below each amino acid as a bar plot. Alignment performed using CLC Sequence Viewer.

Figure 4.3.


20
40
60
Ppy ME--D-AKNIKKGPAPFYPLEDGTAGEQLHKAMKRYALVPGTIAFTDAHIEVNITYAEYF 57
PhemLuc ME--ED-KNILRGPAPFYPLEDGTAGEQLHRAMKRYALIPGTIAFTDAHAGVNITYSEYF 57
Aquatica lateralis MENMDNDENIVYGPKPFYPIEEGSAGAQLRKYMDRYAKL-GAIAFTNALTGVDYTYAEYL 59
Luciola mingrelica ME-MEKEENVVYGPLPFYPIEEGSAGIQLHKYMHQYAKL-GAIAFSNALTGVDISYQEYF 58
Luciola cruciata MENMENDENIVVGPKPFYPIEEGSAGTQLRKYMERYAKL-GAIAFTNAVTGVDYSYAEYL 59
Abscondita terminalis MENMEKEENIVVGPEPFYPIEEGSAGHQLLKYVKQYSKL-GAIAFTNALTKVDISYAEYL 59
Luciola parvula MS-----ENIVYGQPPVYPLEDGTGGEQLYRYIKKYAKISGCIALTNAHTKEDVLFSDLL 55
Luciola tsushimana ME-MEKEENVVYGPLPFYPIEEGSAGIQLHKYMQQYAKL-GAIAFSNALTGVDISYQEYF 58
Lampyris turkestanicus ME--D-AKNIMHGPPPFYPLEDGTAGEQLHKAMKRYAQVPGTIAFTDAHAEVNITYSEYF 57
Lampyris noctiluca ME--D-AKNIMHGPAPFYPLEDGTAGEQLHKAMKRYAQVPGTIAFTDAHAEVNITYSEYF 57
Cratomorphus distinctus ME--ED-KNIMYGPAPFSPLEEGTAGEQLHKAMKRYAQIPGTIAFTAAHVEVNVTYAEYF 57
Photuris pennsylvanica MS--IE-NNILIGPPPYYPLEEGTAGEQLHRAITRYAAVPGTLAYTDVHTELEVTYKEFL 57
Pyrocoelia miyako ME--DDSKHIMHGHRHSILWEDGTAGEQLHKAMKRYAQVPGTIAFTDAHAEVNITYSEYF 58
Pyrocoelia rufa ME--DDSKHIMHGHRHSILWEDGTAGEQLHKAMKRYAQVPGTIAFTDAHAEVNITYSEYF 58
Consensus ME--XD-KNIVYGPAPFYPLEXGTAGEQLHKAMKRYAKVPGTIAFTDAHTEVNITYSEYF
100%
Conservation
0%
80
100
120
Ppy EMSVRLAEAMKRYGLNTNHRIVVCSENSLQFFMPVLGALFIGVAVAPANDIYNERELLNS 117
PhemLuc EMSCRLAESLKRYGLGLQHRIVVCSENSLQFFMPVVGALFIGVGVAPANDIYNERELLNS 117
Aquatica lateralis EKSCCLGEALKNYGLVVDGRIALCSENCEEFFIPVLAGLFIGVGVAPTNEIYTLRELVHS 119
Luciola mingrelica DITCRLAEAMKNFGMKPEEHIALCSENCEEFFIPVLAGLYIGVAVAPTNEIYTLRELNHS 118
Luciola cruciata EKSCCLGKALQNYGLVVDGRIALCSENCEEFFIPVIAGLFIGVGVAPTNEIYTLRELVHS 119
Abscondita terminalis DVSCRLAQAMINYGIKVDGRIALCSENCEEFFFPVLAGLYIGAGVAPTNEIYTLRELVHS 119
Luciola parvula DLTCKTAESLKNFGITTNSTIAICSENSLQYFIPVIGALYIGAATVAVNDRYIERELINS 115
Luciola tsushimana DITCRLAEAMKNYGMKQEGTIALCSENCEEFFIPVLAGLYIGVAVAPTNEIYTLRELNHS 118
Lampyris turkestanicus EMACRLAETMKRYGLGLQHHIAVCSENSLQFFMPVCGALFIGVGVAPTNDIYNERELYNS 117
Lampyris noctiluca EMACRLAETMKRYGLGLQHHIAVCSENSLQFFMPVCGALFIGVGVAPTNDIYNERELYNS 117
Cratomorphus distinctus EMACRLAETMKRYGLGLDHRIAVCSENSLQFFMPVCGALFIGVGVAPTNDIYNERELYNS 117
Photuris pennsylvanica DVTCRLAEAMKNYGLGLQHTISVCSENCVQFFMPVCAALYIGVATAPTNDIYNERELYNS 117
Pyrocoelia miyako EMSCRLAETMKRYGLGLQHHIAVCSETSLQFFMPVCGALFIGVGVAPTNDIYNERELYNS 118
Pyrocoelia rufa EMSCRLAETMKRYGLGLQHHIAVCSENSLQFFMPVCGALFIGVGVAPTNDIYNERELYNS 118
Consensus EMSCRLAEAMKXYGLGLQHRIAVCSENSLQFFMPVCGALFIGVGVAPTNDIYNERELYNS
100%
Conservation
0%
140
160
180
Ppy MNISQPTVVFVSKKGLQKILNVQKKLPIIQKIIIMDSKTDYQGFQSMYTFVTSHLPPGFN 177
PhemLuc MTISQPTLVFCSRKGLQKILNVQKKLPVLQKIIILDTKEDYMGFQSMYSFVDAHLPVGFN 177
Aquatica lateralis LGISKPTIVFSSKKGLDKVITVQKTVNAIKTIVILDSKVDYRGYQSMDNFIKKNTPPGFK 179
Luciola mingrelica LGIAQPTIVFSSRKGLPKVLEVQKTVTCIKKIVILDSKVNFGGHDCMETFIKKHVELGFQ 178
Luciola cruciata LGISKPTIVFSSKKGLDKVITVQKTVTTIQTIVILDSKVDYRGYQCLDTFIKRNTPPGFQ 179
Abscondita terminalis LGISKPTIVFSSKKGLEKVITVQKTVTCIKTIVILDSKVDYQGYDCLETFIKKYLPAGFH 179
Luciola parvula LDISKPTLLFCSKQTLSIIQRVQQTLGYIKKIIVLDSKDNIASVQCLDDFITDNCNKNFD 175
Luciola tsushimana LGIAQPTIVFSSRKGLPKVLEVQKTVTCIKTIVILDSKVNFGGYDCMETFIKKHVELGFR 178
Lampyris turkestanicus LSISQPTIVFCSKRALQKILGVQKKLPIIQKIVILDSREDYMGKQSMYSFIESHLPAGFN 177
Lampyris noctiluca LSISQPTIVFCSKRALQKILGVQKKLPIIQKIVILDSREDYMGKQSMYSFIESHLPAGFN 177
Cratomorphus distinctus LSISQPTVVFCSKRALQKILGVQKSLPVIKKIVILDSREDYMGKQSMYSFIQSYLPGGFN 177
Photuris pennsylvanica LSISQPTVVFTSRNSLQKILGVQSRLPVIKKIIMLDTKKDYLGYQSMQSFMKEHVPANFN 177
Pyrocoelia miyako LFISQPTIVFCSKRALQKILGVQKKLPVIQKIVILDSREDYMGKQSMYSFIESHLPAGFN 178
Pyrocoelia rufa LFISQPTIVFCSKRALQKILGVHKKLPVIQKIVILDSREDYMGKQSMYSFIESHLPAGFN 178
Consensus LGISQPTIVFCSKKGLQKILGVQKXLPVIXKIVILDSKEDYMGXQSMYSFIKSHLPAGFN
100%
Conservation
0%
200
220
240
Ppy EYDFVP-ESFDRDKTIALIMNSSGSTGLPKGVALPHRTACVRFSHARDPIFGNQIIPDTA 236
PhemLuc EYDYIP-DSFDRDQATALIMNSSGSTGLPKGVELNHTSVCVRFSHCRDPVFGNQIIPDTA 236
Aquatica lateralis GSSFKTVEV-NRKEQVALIMNSSGSTGLPKGVQLTHENAVTRFSHARDPIYGNQVSPGTA 238
Luciola mingrelica PSSFVPIDVKNRKQHVALLMNSSGSTGLPKGVRITHEGAVTRFSHAKDPIYGNQVSPGTA 238
Luciola cruciata ASSFKTVEV-DRKEQVALIMNSSGSTGLPKGVQLTHENTVTRFSHARDPIYGNQVSPGTA 238
Abscondita terminalis IDNFKVREV-NRKEQVALIMNSSGSTGLPKGVQITHEGTVTRFSHARDPIYGNQVSPGTA 238
Luciola parvula SNCFKP-NSFNRNDQAALILNSSGTTGLPKGVVLTHKNLVVRFNHCKDPIFGNQINPGTA 234
Luciola tsushimana PTSFVPIDVKNRKQHVALLMNSSGSTGLPKGVLITHEGTVTRFSHAKDPIYGNQVSPGTA 238
Lampyris turkestanicus EYDYIP-DSFDRETATALIMNSSGSTGLPKGVELTHKNICVRFSHCRDPVFGNQIIPDTA 236
Lampyris noctiluca EYDYIP-DSFDRETATALVMNSSGSTGLPRGVELTHQNVCVRFSHCRDPVFGNQIIPDTA 236
Cratomorphus distinctus EYDYVP-DTFDRDMATALIMNSSGSTGLPKGVELSHKNVCVRFSHCRDPVFGNQIIPDTA 236
Photuris pennsylvanica VSAFKP-LSFDLDR-VACIMNSSGSTGLPKGVPISHRNTTYRFSHCRDPVFGNQIIPDTT 235
Pyrocoelia miyako EYDYIP-DSFDRETATALIMNSSGSTGLPKGVDLTHMNVCVRFSHCRDPVFGNQIIPDTA 237
Pyrocoelia rufa EYDYIP-DSFDRETATALIMNSSGSTGLPKGVDLTHMNVCVRFSHCRDPVFGNQIIPDTA 237
Consensus EYDFXP-DSFDRKTAXALIMNSSGSTGLPKGVELTHENVCVRFSHCRDPXFGNQIIPDTA
100%
Conservation
0%

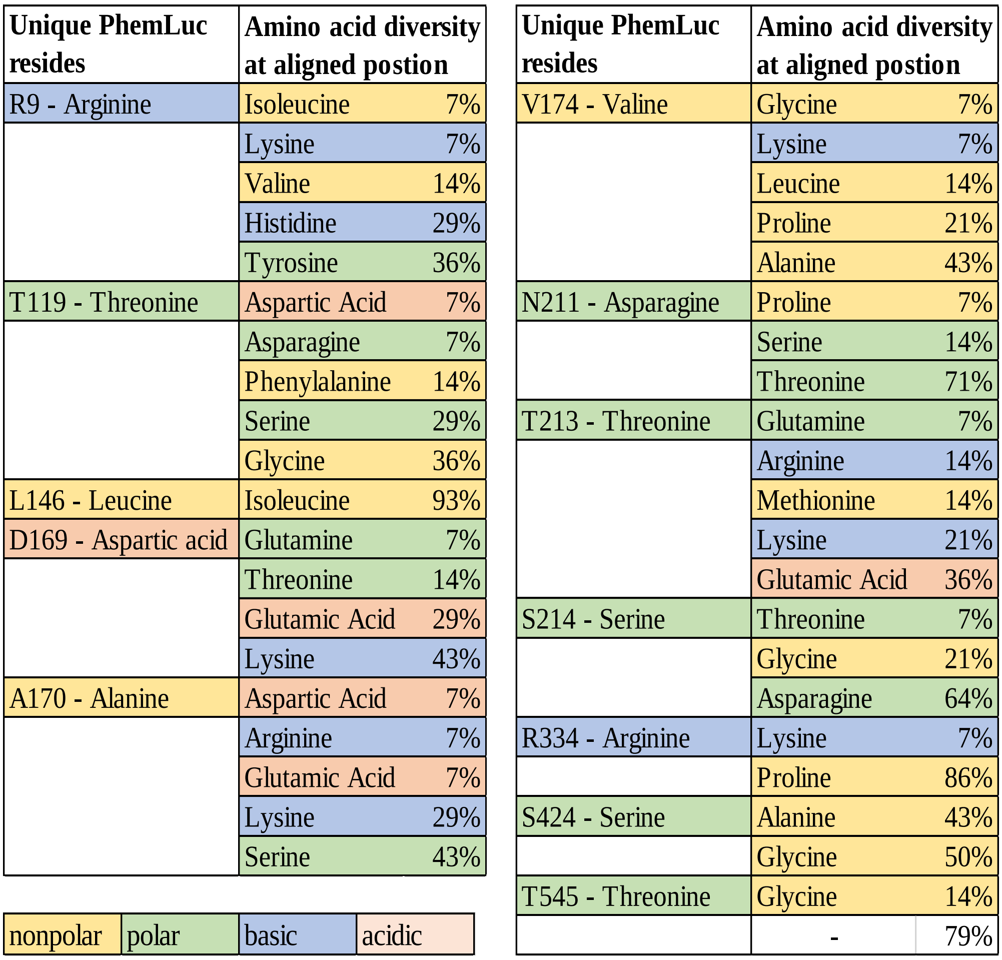

260
280
300
Ppy ILSVVPFHHGFGMFTTLGYLICGFRVVLMYRFEEELFLRSLQDYKIQSALLVPTLFSFFA 296
PhemLuc ILSVIPFHHGFGMFTTLGYLICGFRVVLMYRFEEELFLRSLQDYKIQSALLVPTLFSFFA 296
Aquatica lateralis ILTVVPFHHGFGMFTTLGYLTCGFRIVMLTKFDEETFLKTLQDYKCSSVILVPTLFAILN 298
Luciola mingrelica ILTVVPFHHGFGMFTTLGYFACGYRVVMLTKFDEELFLRTLQDYKCTSVILVPTLFAILN 298
Luciola cruciata VLTVVPFHHGFGMFTTLGYLICGFRVVMLTKFDEGTFLKTLQDYKCTSVILVPTLFAILN 298
Abscondita terminalis ILTVVPFHHGFGMFTNLGYLICGFRIVMLTKFDEELFLKTLADYKCNSAILVPTLFAILS 298
Luciola parvula ILTVIPFHHGFGMFTTLGYFTCGFRIVLMHTFEERLFLQSLQDYKVESTLLVPTLMTFFA 294
Luciola tsushimana ILTVVPFHHGFGMFTTLGYFACGYRVVMLTKFDEELFLRTMQDYKCTSVILVPTLFAILN 298
Lampyris turkestanicus ILTVIPFHHGFGMFTTLGYLTCGFRIVLMYRCEEELFLRSLQDYKIQSALLVPTLFSFFA 296
Lampyris noctiluca ILTVIPFHHGFGMFTTLGYLTCGFRIVLMYRFEEELFLRSLQDYKIQSALLVPTLFSFFA 296
Cratomorphus distinctus ILTVIPFHHGFGMFTTLGYLTCGFRIVLMYRFEEELFLRSLQDYKIQSALLVPTLFSFFA 296
Photuris pennsylvanica ILCAVPFHHAFGTFTNLGYIICGFHVVLMYRFNEHLFLQTLQDYKCQSALIVPTVLAFLA 295
Pyrocoelia miyako ILTVIPFHHVFQMFTTLGYLTCGFRIVLMYRFEEELFLRSLQDYKIQSALLVPTLFSFFA 297
Pyrocoelia rufa ILTVIPFHHGFGMFTTLGYLTCGFRIVLMYRFEEELFLRSLQDYKIQSALLVPTLFSFFA 297
Consensus ILTVXPFHHGFGMFTTLGYLTCGFRIVLMYRFEEELFLRSLQDYKIQSALLVPTLFSFFA
100%
Conservation
0%
320
340
360
Ppy KSTLIDKYDLSNLHEIASGGAPLSKEVGEAVAKRFHLPGIRQGYGLTETTSAILITPEGD 356
PhemLuc KSTLIDKYDLSNLHEIASGGAPLAKEVGEAVAKRFNLRGIRQGYGLTETTSAVIITPEGD 356
Aquatica lateralis RSELLDKYDLSNLVEIASGGAPLSKEIGEAVARRFNLPGVRQGYGLTETTSAIIITPEGD 358
Luciola mingrelica KSELIDKFDLSNLTEIASGGAPLAKEVGEAVARRFNLPGVRQGYGLTETTSAFIITPEGD 358
Luciola cruciata KSELLNKYDLSNLVEIASGGAPLSKEVGEAVARRFNLPGVRQGYGLTETTSAIIITPEGD 358
Abscondita terminalis KSTLIDKFDLSNLVEIASGGAPLAKEVGEAVARRFNLPGVRQGYGLTETTSAIIITPEGD 358
Luciola parvula KSPLVEKYHLPHLQEIASGGAPLPKKIGDAVSQRFKLKKARQGYGLTETTSAILITPEGE 354
Luciola tsushimana KSELIDKFDLSNLTEIASGGAPLAKEVGEAVARRFNLPGVRQGYGLTETTSAFIITPEGD 358
Lampyris turkestanicus KSTLVDKYDLSNLHEIASGGAPLAKEVGEAVAKRFKLPGIRQGYGLTETTSAIIITPEGD 356
Lampyris noctiluca KSTLVDKYDLSNLHEIASGGAPLAKEVGEAVAKRFKLPGIRQGYGLTETTSAIIITPEGD 356
Cratomorphus distinctus KSTLVDKYDLSNLHEIASGGAPLAKEVGEAVAKRFKLPGIRQGYGLTETTSAIIITPEGD 356
Photuris pennsylvanica KNPLVDKYDLSHLHEIASGGAPLSKEISEIAAKRFKLPGIRQGYGLTETTCAIVITAEGE 355
Pyrocoelia miyako KSTLVDKYDLSNLHEIASGGAPLAKEVGEAVAKRFKLPGIRQGYGLTETTSAIIITPEGD 357
Pyrocoelia rufa KSTLVDKYDLSNLHEIASGGAPLAKEVGEAVAKRFKLPGIRQGDGLTETTSAIIITPEGD 357
Consensus KSTLVDKYDLSNLHEIASGGAPLAKEVGEAVAKRFKLPGIRQGYGLTETTSAIIITPEGD
100%
Conservation
0%
380
400
420
Ppy DKPGAVGKVVPFFEAKVVDLDTGKTLGVNQRGELCVRGPMIMSGYVNNPEATNALIDKDG 416
PhemLuc DKPGAVGKVVPFFSAKVVDLDTGKTLGLNQRGELCLKGPMIMKGYVNNPEATNALIDKDG 416
Aquatica lateralis DKPGASGKVVPLFKAKVIDLDTKKTLGPNRRGEVCVKGPMLMKGYVDNPEATREIIDEEG 418
Luciola mingrelica DKPGASGKVVPLFKVKVIDLDTKKTLGVNRRGEICVKGPSLMLGYSNNPEATRETIDEEG 418
Luciola cruciata DKPGASGKVVPLFSAKVIDLDTKKSLGPNRRGEVCVKGPMLMKGYVNNPEATKELIDEEG 418
Abscondita terminalis DKPGASGKIVPLFKGKVVDLDTKKTLGPNRRGEICVKGPMLMKGYADNPEATREIIDEEG 418
Luciola parvula EVSGSTGKVVPFFAAKVVDTVTGKILGPNEVGELYFKGDMIMKGYCNNSKATNETIDKDG 414
Luciola tsushimana DKPGASGKVVPLFKVKVIDLDTKKTLGVNRRGEICVKGPSLMLGYLNNPEATKETIDDEG 418
Lampyris turkestanicus DKPGACGKVVPFFSAKIVDLDTGKTLGVNQRGELCVKGPMIMKGYVNNPEATSALIDKDG 416
Lampyris noctiluca DKPGACGKVVPFFSAKIVDLDTGKTLGVNQRGELCVKGPMIMKGYVNNPEATSALIDKDG 416
Cratomorphus distinctus DKPGACGKVVPFFAAKIVDLDTGKTLGVNQRGELYVKGPMIMKGYVNNPEATNALIDKDG 416
Photuris pennsylvanica FKPGAVGKVVPFYSLKVLDLNTGKKLGPNERGEICFTGPMIMKGYINNPEATREIIDEEG 415
Pyrocoelia miyako DKPGACGKVVPFFTAKIVDLDTGKTLGVNQRGELCVKGPMIMKGYVNNPEATNALIDKDG 417
Pyrocoelia rufa DKPGACGKVVPFFAAKIVDLDTGKTLGVNQRGELCVKGPMIMKGYVNNPEATNALIDKDG 417
Consensus DKPGAXGKVVPFFSAKVVDLDTGKTLGVNQRGELCVKGPMIMKGYVNNPEATNXLIDKDG
100%
Conservation
0%
440
460
480
Ppy WLHSGDIAYWDEDEHFFIVDRLKSLIKYKGYQVAPAELESILLQHPNIFDAGVAGLPDDD 476
PhemLuc WLHSGDISYWDEDGHFFIVDRLKSLIKYKGYQVPPAELESILLQHPFIFDAGVAGIPDDE 476
Aquatica lateralis WLHTGDIGYYDEEKHFFIVDRLKSLIKYKGYQVPPAELESVLLQHPNIFDAGVAGVPDPI 478
Luciola mingrelica WLHTGDIGYYDEDEHFFIVDRLKSLIKYKGYQVPPAELESVLLQHPNIFDAGVAGVPDPD 478
Luciola cruciata WLHTGDIGYYDEEKHFFIVDRLKSLIKYKGYQVPPAELESVLLQHPNIFDAGVAGVPDPV 478
Abscondita terminalis WLHTGDIGYYDEEKHFFIVDRLKSLIKYKGYQVPPAELESVLLQHPDIFDAGAAGIPDPI 478
Luciola parvula WLRSGDLGYYNEDKHFFIVDRLKSVIKYKGYQVAPAEIEGILLTHPDIIDAGVTGIPDND 474
Luciola tsushimana WLHTRDIGYYDEDEHFFIVDRLKSLIKYKGYQVPPAELESVLLQHPNIFDAGVAGVPDSE 478
Lampyris turkestanicus WLHSGDIAYYDKDGHFFIVDRLKSLIKYKGYQVPPAELESILLQHPFIFDAGVAGIPDPD 476
Lampyris noctiluca WLHSGDIAYYDKDGHFFIVDRLKSLIKYKGYQVPPAELESILLQHPFIFDAGVAGIPDPD 476
Cratomorphus distinctus WLRSGDIAYYDEDGHVFIVDRLKSLIKYKGYQVPPAELESILLQHPFIFDAGVAGIPDED 476
Photuris pennsylvanica WIHSGDIGYFDEDGHVYIVDRLKSLIKYKGYQVPPAELEALLLQHPFIEDAGVAGVPDEV 475
Pyrocoelia miyako WLHSGDIAYYDKDGHFFIVDRLKSLIKYKGYQVPPAELESILLQHPFIFDAGVAGIPDPD 477
Pyrocoelia rufa WLHSGDIAYYDKDGHFFIVDRLKSLIKYKGYQVPPAELESILLQHPFIFDAGVAGIPDPD 477
Consensus WLHSGDIGYYDEDGHFFIVDRLKSLIKYKGYQVPPAELESILLQHPFIFDAGVAGIPDPD
100%
Conservation
0%



```
                                              500                 520                 540
                       Ppy  AGELPAAVVVLEHGKTMTEKEIVDYVASQVTTAKKLRGGVVFVDEVPKGLTGKLDARKIR 536
                   PhemLuc  AGELPAAVVVLEEGKTMTEKEIMDYVAGQVTTAKRLRGGVVFVDEVPKGLTGKLDARKIR 536
         Aquatica lateralis AGELPGAVVVLEKGKSMTEKKVMDYVAGQVSNAKRLRGGVRFVDEVPKGLTGKIDGKAIR 538
         Luciola mingrelica AGELPGAVVVMEKGKTMTEKEIVDYVNSQVVNHKRLRGGVRFVDEVPKGLTGKIDAKVIR 538
           Luciola cruciata AGELPGAVVVLESGKSMTEKEVMDYVASQVSNAKRLRGGVRFVDEVPKGLTGKIDGRAIR 538
      Abscondita terminalis AGELPGAVVVLEQGKHLTEQQILDYVAGQVSNAKRLRGGVRFVDEVPKGLTGKIDGRAIR 538
            Luciola parvula SGELPAACVVIRPGHRVTEESIINYVSSQVSFTKRLRGGVRLLDEIPKGCTGKIDRNALK 534
         Luciola tsushimana AGELPGAVVVMEKGKTMTEKEIVDYVNSQVVNHKRLRGGVRFVDEVPKGLTGKIDAKVIR 538
     Lampyris turkestanicus AGELPAAVVVLEEGKTMTEQEVMDYVAGQVTASKRLRGGVKFVDEVPKGLTGKIDARKIR 536
        Lampyris noctiluca  AGELPAAVVVLEEGKTMTEQEVMDYVAGQVTASKRLRGGVKFVDEVPKGLTGKIDGRKIR 536
     Cratomorphus distinctus AGELPAAVVVLEEGKTMTEQEVMDYVAGQVTASKRLRGGVKFVDEVPKGLTGKIDSRKIR 536
     Photuris pennsylvanica AGDLPGAVVVLKEGKSITEKEIQDYVAGQVTSSKKLRGGVEFVKEVPKGFTGKIDTRKIK 535
          Pyrocoelia miyako AGELPAAVVVLEEGKMMTEQEVMDYVAGQVTASKRLRGGVKFVDEVPKGLTGKIDSRKIR 537
            Pyrocoelia rufa AGELPAAVVVLEEGKMMTEQEVMDYVAGQVTASKRLRGGVKFVDEVPKGLTGKIDSRKIR 537
                  Consensus AGELPAAVVVLEEGKTMTEKEXMDYVAGQVTXSKRLRGGVRFVDEVPKGLTGKIDARKIR
                       100%
               Conservation
                         0%

                       Ppy  EILIKAKKG---GKSKL 550
                   PhemLuc  EILVKAKKT----KSKL 549
         Aquatica lateralis EILKKPV-------AKM 548
         Luciola mingrelica EILKKPQ-------AKM 548
           Luciola cruciata EILKKPV-------AKM 548
      Abscondita terminalis EILKSPK-------PKM 548
            Luciola parvula QILQKPK-------SKL 544
         Luciola tsushimana EILKKPQ-------AKM 548
     Lampyris turkestanicus EILMMGKK------SKL 547
        Lampyris noctiluca  EILMMGKK------SKL 547
     Cratomorphus distinctus EILVMGKK------SKL 547
     Photuris pennsylvanica EILIKAQKGKSKSKAKL 552
          Pyrocoelia miyako EILTMGQK------SKL 548
            Pyrocoelia rufa EILTMGQK------SKL 548
                  Consensus EILKKPKK------SKL
                       100%
               Conservation
                         0%
```

Homology alignment of beetle luciferase amino acid sequences. Shown are the amino acid sequences of *Photinus pyralis* luciferase, *Ppy* Fluc, and *Phosphaenus hemipterus* luciferase, *Phem*Luc, aligned to 12 additional firefly luciferases of varied geographic origin. A consensus sequence derived from the majority amino acid recorded in each position is indicated below. Degree of conservation is represented below the consensus for each amino acid as a bar plot. Alignment performed using CLC Sequence Viewer.

Table 4.1.



Unique PhemLuc resides	Amino acid diversity at aligned postion	
R9 - Arginine	Isoleucine	7%
	Lysine	7%
	Valine	14%
	Histidine	29%
	Tyrosine	36%
T119 - Threonine	Aspartic Acid	7%
	Asparagine	7%
	Phenylalanine	14%
	Serine	29%
	Glycine	36%
L146 - Leucine	Isoleucine	93%
D169 - Aspartic acid	Glutamine	7%
	Threonine	14%
	Glutamic Acid	29%
	Lysine	43%
A170 - Alanine	Aspartic Acid	7%
	Arginine	7%
	Glutamic Acid	7%
	Lysine	29%
	Serine	43%

nonpolar	polar	basic	acidic

Unique PhemLuc resides	Amino acid diversity at aligned postion	
V174 - Valine	Glycine	7%
	Lysine	7%
	Leucine	14%
	Proline	21%
	Alanine	43%
N211 - Asparagine	Proline	7%
	Serine	14%
	Threonine	71%
T213 - Threonine	Glutamine	7%
	Arginine	14%
	Methionine	14%
	Lysine	21%
	Glutamic Acid	36%
S214 - Serine	Threonine	7%
	Glycine	21%
	Asparagine	64%
R334 - Arginine	Lysine	7%
	Proline	86%
S424 - Serine	Alanine	43%
	Glycine	50%
T545 - Threonine	Glycine	14%
	-	79%

Unique *Phem*Luc residues derived from multiple sequence alignment. Amino acids identified as unique to *Phem*Luc relative to 13 homologous firefly luciferase sequences. Colour shading indicates the polarity of the respective amino acid. Yellow: nonpolar, Green: polar, Blue: basic, Pink: acidic. The basic and acidic amino acids can also be categorised as polar. "-" Indicates the presence of a gap at the respective aligned position. Table assembled using information from Figure 4.3.

4.3.2. Secondary structure analysis

4.3.2.1 Predicting three-state secondary structure

The compatibility of *Phem*Luc and *Ppy* Fluc was further assessed on the basis of secondary structure. The secondary structure of a protein is an emergent property of the sequence of the amino acids which make up the primary structure, and describes the local conformation of the segments of polypeptide backbone and amino acid side-chain interactions. However, whilst accurate secondary structure data is available for *Ppy* Fluc, there exists no similar information for *Phem*Luc. Computational approaches for the prediction of protein secondary structure are well established and commonly used where structural data is not available. Multiple secondary structure prediction programs exist, but can vary significantly in their accuracy to predict the location of α-helix and β-sheet. The efficacy of a group of the most commonly utilised programs available online found that the best results for secondary structure prediction could be produced by PSIPRED (Koswatta *et al.* 2012).

PSIPRED is an online protein analysis work bench which has remained amongst the most reliable programs for protein prediction analysis, including secondary structure for over 20 years (PSIPRED web server available from http://bioinf.cs.ucl.ac.uk/psipred/ (Jones 1999; Buchan and Jones 2019)). Whilst PSIPRED is well-established and continues to be utilised over two decades after its inception, newer programs continue to emerge built on more modern computational approaches, such as deep learning, to provide improved overall accuracies. One such program, RaptorX, offers a similar protein analysis service as PSIPRED, but has been shown to improve the accuracy of protein secondary structure analysis (RaptorX web server available from http://raptorx.uchicago.edu/ (Wang *et al.* 2011; Källberg *et al.* 2012)).

For the purpose of analysing *Phem*Luc against *Ppy* Fluc, both PSIPRED and RaptorX were used to produce three-state secondary structure predictions – α-helix, β-sheet and, coil. The three-state predictions made by both programs were compared with the crystal structure of *Ppy* Fluc stored in the Protein Data Bank (PDB files from RCSB) as file 4G36 (Sundlov *et al.* 2012). Secondary structure information was accessed from 4G36.pdb using DSSP (Touw *et al.* 2015; Kabsch and Sander 1983). The comparison of predictions against the known secondary structure of *Ppy* Fluc is shown in Table 4.3. From this comparison, both PSIPRED and RaptorX were shown to produce secondary structure of good accuracies by comparing the predictions made for *Ppy* Fluc against the correct structure from 4G36.pdb. However, both PSIPRED and RaptorX were shown to misattribute several short regions of α-helix and β-sheet to coil environments. This failure to correctly identify α-helix and β-sheet occurred at a lower frequency in RaptorX, resulting in more accurate distribution of α-helix, β-sheet, and coil ratios, and overall greater agreement with the true secondary structure.

4.3.2.2 Analysing RaptorX predicted secondary structures

Due to RaptorX proving to generate improved accuracy predictions for *Ppy* Fluc relative to PSIPRED, the RaptorX *Phem*Luc structure was used for comparison with *Ppy* Fluc. As previously discussed, the *Ppy* Fluc RaptorX structure was shown to vary from the 4G36.pdb structure, including several misattributed regions shown in Table 4.2. Interestingly, these misattributed regions were also present in the prediction generated by RaptorX for *Phem*Luc. Therefore, it was decided that alignment of the RaptorX predictions for both enzymes would allow for better comparative analysis than the alignment of *Phem*Luc to the *Ppy* Fluc secondary structure from 4G36.pdb.

To visualise the comparison of secondary structure, the RaptorX predictions were aligned in CLC Sequence Viewer (Figure 4.4.). From this alignment of the three-state structure it was shown that *Phem*Luc and *Ppy* Fluc are predicted to share 97.63% of secondary structure, and no segments of α-helix or β-sheet elements are unique to either enzyme. The only variation between the two sequences arises from disagreement made in predictions for the exact boundary of α-helix and β-sheet elements in a minority of regions. This broad agreement in predicted secondary structure is a further indicator, in addition to the primary structure analysis, of compatibility between the two enzymes for the purpose of homology modelling.

Table 4.2.

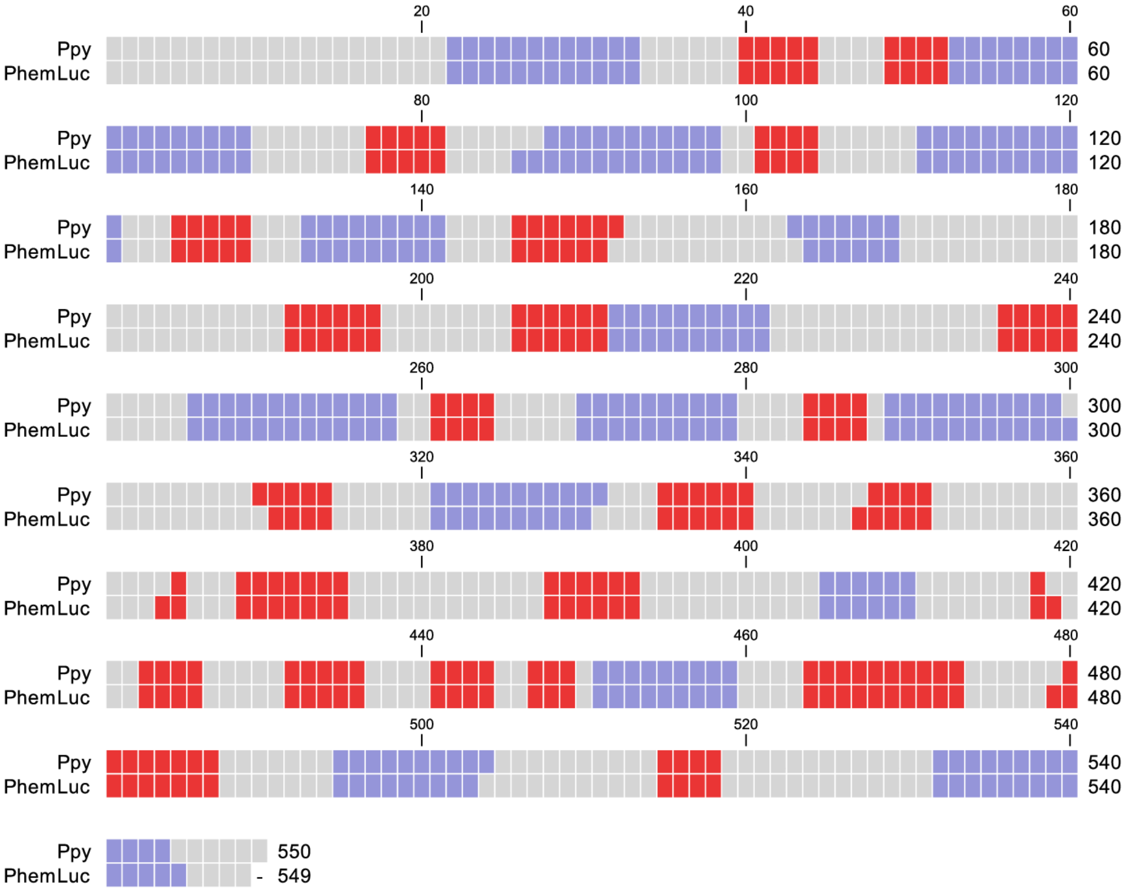
Reported Secondary structure of Ppy	**Ppy RaptorX structure prediction**	**Ppy PSIPRED structure prediction**	**PhemLuc RaptorX structure prediction**	**PhemLuc PSIPRED structure prediction**

H = 31%
E = 20%
C = 50%
	H = 29%
	E = 23%
	C = 48%
		H = 27%
		E = 22%
		C = 51%
			H = 29%
			E = 23%
			C = 48%
				H = 27%
				E = 22%
				C = 51%

H 22 – 34	H 22 – 33	H 22 – 33	H 22 – 33	H 22 – 33
E 40 – 44	E 40 – 44	E 40 – 43	E 40 – 44	E 40 – 43
E 49 – 52	E 49 – 52	E 49 – 52	E 49 – 52	E 49 – 52
H 53 – 70	H 53 – 69	H 53 – 69	H 53 – 69	H 53 – 69
E 77 – 81	E 77 –81	E 77 –81	E 77 –81	E 77 –81
H 89 – 97	H 88 – 98	H 86 – 98	H 86 – 98	H 86 – 98
E 101 – 104	E 101 – 104	E 101 – 104	E 101 – 104	E 101 – 104
H 111 – 121	H 111 – 121	H 111 – 121	H 111 – 121	H 111 – 121
E 125 –128	E 125 – 129	E 125 – 128	E 125 – 129	E 125 – 128
H 130 – 140	H 133 – 141	H 132 – 141	H 133 – 141	H 132 – 141
E 148 – 151	E 146 – 152	E 147 – 152	E 146 – 151	E 147 – 152
H 164 – 171	H 163 – 169	–	H 164 – 169	–
E 192 – 196	E 192 – 197	E 193 – 197	E 192 – 197	E 193 – 197
E 208 – 211	E 206 – 211	E 207 – 210	E 206 – 211	E 207 – 210
H 212 – 223	H 212 – 221	H 212 – 222	H 212 – 221	H 212 – 222
E 236 – 239	E 236 – 240	E 234 – 237	E 236 – 240	E 234 – 237
H 246 – 258	H 246 – 258	H 243 – 256	H 246 – 258	H 243 – 256
E 261 – 264	E 261 – 264	E 257 – 262	E 261 – 264	E 257 – 262
H 270 – 279	H 270 – 279	H 270 – 279	H 270 – 279	H 270 – 279
E 284 – 286	E 284 – 287	E 283 – 287	E 284 – 287	E 283 – 287
H 289 – 295	H 289 – 299	H 289 – 297	H 289 – 300	H 289 – 297
H 300 – 302	–	–	–	–
E 311 – 315	E 310 – 314	E 311 – 315	E 311 – 314	E 311 – 315
H 321 – 330	H 321 – 331	H 321 – 330	H 321 – 330	H 321 – 330
E 337 – 340	E 335 – 340	E 335 – 341	E 335 – 340	E 335 – 341
H 343 – 354	–	–	–	–
E 350 – 351	E 348 – 351	E 346 – 351	E 347 – 351	E 346 – 351
E 364 – 365	E 365	–	E 364 – 365	–
E 370 – 374	E 369 – 375	E 370 – 375	E 369 – 375	E 370 – 375
E 388 – 393	E 388 – 393	E 389 – 393	E 388 – 393	E 389 – 393
E 400 – 401	–	–	–	–
H 405 – 409	H 405 – 410	–	H 405 – 410	–
E 418 – 426	E 418	E 418 – 419	E 418 – 419	E 418 – 419
–	E 423 – 426	E 423 – 426	E 423 – 426	E 423 – 426
E 432 – 437	E 432 – 436	E 431 – 436	E 432 – 436	E 431 – 436
H 438 – 440	–	–	–	–
E 442 – 444	E 441 – 444	E 442 – 444	E 441 – 444	E 442 – 444
E 447 – 449	E 447 – 449	E 447 – 449	E 447 – 449	E 447 – 449
H 451 – 458	H 451 – 459	H 451 – 460	H 451 – 459	H 451 – 460
E 467 – 472	E 464 – 473	E 464 – 473	E 464 – 473	E 464 – 473
E 480 – 485	E 480 – 487	E 480 – 487	E 479 – 487	E 480 – 487
H 495 – 502	H 495 – 504	H 495 – 505	H 495 – 503	H 495 – 505
H 508 – 510	–	–	–	–
E 516 – 518	E 515 – 518	E 515 – 518	E 515 – 518	E 515 – 518
H 535 – 539	H 532 – 544	H 532 – 541	H 532 – 545	H 532 – 541

Comparison of known *Ppy* Fluc secondary structure against computational predictions using RaptorX and PSIPRED. The recorded secondary structure positons of α-helix and β-sheet for *Ppy* Fluc from PDB file 4G36, accessed through DSSP, compared with the secondary structure predictions made for both *Ppy* Fluc and *Phem*Luc by RaptorX and PSIPRED. H: α-helix, E: β-sheet.

Figure 4.4.



Alignment of RaptorX predicted secondary structures for *Ppy* Fluc and *Phem*Luc. Secondary structure as detailed in Table 3.1 for RaptorX predictions of *Ppy* Fluc and *Phem*Luc. Alignment performed in CLC Sequence Viewer by processing the secondary structure sequence as an amino acid sequence for the purposes of alignment. Blue: α-helix, Red: β-sheet, Grey: Coil environments.

4.3.3. Protein and substrate interaction modelling

4.3.3.1 Substrate specificity of the luciferase active site

All members of the Lampyridae family share a common bioluminescence reaction dependent on homologous luciferases (from 98.91% to 55.06% max/min shared sequence identity amongst the sequences of the multiple protein alignment in Figure 4.3), and an identical substrate, *D*-luciferin (Day *et al.* 2004). The variability in the luciferase enzyme gives rise to the broad variability in the properties of bioluminescence signals seen across enzymes, in terms of intensity, spectra, kinetics and pH dependencies. However, as *D*-luciferin is common to all firefly luciferases, the active site has evolved to accommodate this structure such that many residues within proximity of the binding pocket are well conserved (Branchini *et al.* 1998; Branchini *et al.* 2003; Viviani *et al.* 2007). This evolutionary selection for *D*-luciferin-specific compatibility is likely to correlate with diminished bioluminescence intensity when pairing luciferases with synthetic substrate analogues (see 4.2. Introduction).

As detailed in the Introduction, many synthetic analogues of *D*-luciferin have emerged in recent years for *in vitro* and *in vivo* applications in biotechnology and/ or biomedicine, aiming to use engineered synthetic substrate variants to modulate the bioluminescence signal to a greater extent than can be achieved by protein engineering alone (Kaskova *et al.* 2016). In the case of the development of infraluciferin, a luciferin analogue was sought which could reliably redshift the bioluminescence signal to the bio-optical window (ca. 600≤800 nm) for *in vivo* bioluminescence imaging (BLI) (Iwano *et al.* 2013). However, an issue common to luciferin substrate analogues is that the structure can vary such that it is no longer compatible with the enzymes active site and may fail to bind or interact appropriately with the luciferin, ATP or intermediates for the different steps of the bioluminescence reaction (Adams and Miller 2014; Jones *et al.* 2017; Amadeo *et al.* 2021). As described, the beetle luciferase bioluminescence reaction occurs in two steps (adenylation and oxidation), which depend on two distinct conformations of the enzyme, derived from a 140˚ rotation of the C-domain to allow for adenylation by K529 and subsequent oxidation by K443, lysine residues juxtaposed on the opposing sides of the C-terminal domain (Sundlov *et al.* 2012). Consequently, the incompatibility of a synthetic analogue such as infraluciferin can occur at either the adenylation or oxidative stage of bioluminescence catalysis, or both (Berraud-Pache and Navizet 2016).

Whilst efforts have been made to engineer *Ppy* Fluc for improved activity with infraluciferin (Stowe *et al.* 2019), *Phem*Luc remains uncharacterised, and without protein models which can be utilised for the purposes of rational mutagenesis. With no further information than the protein sequence available, work was undertaken to alter the substrate specificity of *Phem*Luc with *D*-luciferin, to allow for improved activity with the synthetic substrate analogue, infraluciferin.

4.3.3.2 Construction of homology models

As well as being highly conserved (Sections 4.3.1-2), *Ppy* Fluc was additionally selected as crystal structures are available with bound substrate analogues of *D*-luciferin and infraluciferin, DLSA and iDLSA, respectively. DLSA (5’-O-[N-(dehydroluciferyl)-sulfamoyl]-adenosine) is an analogue of the adenylated form of luciferin, luciferyl-AMP, and acts as a reversible inhibitor of firefly luciferase activity (Branchini *et al.* 2005a). DLSA can be bound by the first conformation of the luciferase enzyme, but is unable to undergo catalysis, and thus prevents the binding enzyme from proceeding further to the oxidative stage of the reaction. The DLSA-luciferase complex is stable, and resistant to hydrolysis, oxidation, and proteolysis, allowing for the production of a protein sample of uniform conformation for the purpose of crystal structure studies. More recently, an infraluciferin high-energy intermediate analogue known as iDLSA (5'-O-[N-(dehydroinfraluciferyl)-sulfamoyl]-adenosine) has been developed based on DLSA for crystal structure studies (Stowe *et al.* 2019).

Homology models of *Phem*Luc were constructed using SWISS-MODEL (Waterhouse *et al.* 2018). To generate a model indicative of *Phem*Luc containing native *D*-luciferin bound within the active site, the crystal structure of *Ppy* Fluc in the adenylate-forming conformation bound to DLSA, resolved to 2.62 Å, was selected from Protein Data Bank (PDB) file 4G36.pdb to serve as template (Sundlov *et al.* 2012). A second model was constructed to represent the structure of *Phem*Luc bound to infraluciferin, using the template of a second adenylate-forming conformation of *Photinus pyralis* bound to iDLSA and resolved to 3.10 Å, from PDB file 6HPS.pdb (Stowe *et al.* 2019). The two *Phem*Luc models, along with the *Ppy* Fluc crystal structures from which they were derived, were then analysed within The PyMOL Molecular Graphics System, Version 2.1.1 Schrödinger, LLC. A total of 4 models were processed to display their three-state secondary structure, and identify amino acid residues measured to be within 4 Å of their respective bound ligand, DLSA or iDLSA (Figures 4.5. – 4.8.). These models were however limited to identifying residues within 4 Å during the adenylation conformation, as no models currently exist of a firefly luciferase bound to an infraluciferin analogue in the oxidation conformation. Protein model quality was assessed using Molprobity (available at http://molprobity.biochem.duke.edu/) (Williams et al. 2018). Clashscores and Molprobity scores are detailed in the appendices Table 9.8 and suggest that all models are of good quality, where good indicates ≥ 66th percentile, as determined by Molprobity.

4.3.3.3 Identification of target for mutagenesis

A total of 32 amino acid residues were identified as being within 4 Å of bound ligand for one or more of the models, detailed in Table 4.3. Of these 32 sites, 18 were found to be common across all 4 models. Further reinforcing that residues directly involved in, or proximal to the active site are highly conserved, 26 of the positions identified were fully conserved for the identified amino acid across the 14 firefly luciferases aligned in Figure 4.3. All 32 positions were conserved for the respective amino acid between *Ppy* Fluc and *Phem*Luc, regardless of whether it was identified within a model.

Additionally, the polar contacts in each model with DLSA or iDLSA were identified and are detailed in Table 4.4. This analysis indicated considerable differences between *Ppy* Fluc and *Phem*Luc in the amount of polar contacts in each model, the distance of interactions, and which residues were found to be interacting. Unsurprisingly, all interacting residues were present within 4 Å of the bound ligand and therefore included in Table 4.3 for mutagenesis targeting.

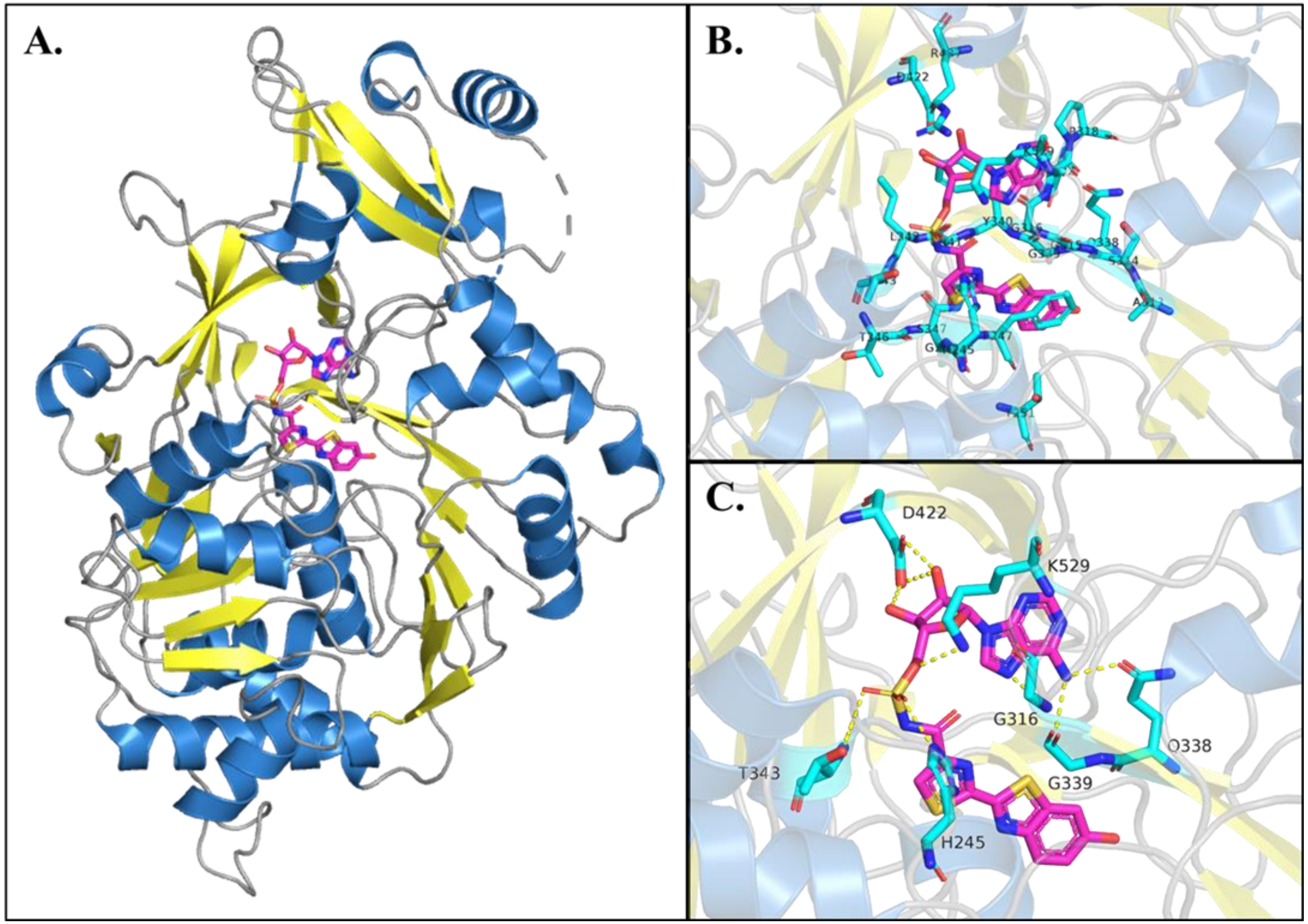
From the work to crystallise *Ppy* Fluc with DLSA by Sundlov *et al* (2012), 7 residues are known to have a role in the active site LH_2-AMP complex in the enzymes first conformation, which increases to 8 residues during the second conformation which follows the 140˚ domain rotation. Of these sites, 6 are common to both conformations – H245, F247, A317, Y340, T343, and S347. The adenylation conformation of the active site contains the unique reside K529, whereas K443 and Q448 are only involved in the active site of the second conformation. All 7 residues of the primary conformation were identified in both *Ppy* Fluc models with DLSA and iDLSA. However, only 5 of these positions were identified in both

*Phem*Luc models, and in both the DLSA and iDLSA model of *Phem*Luc, H245 and K529 were not present. The residues unique to the second conformation, K443 and Q448, were not identified in any of the models as they have no involvement in the adenylation catalysis and are only brought into proximity with the bound ligand following the C-domain rotation.

Due to the proximity of the 32 sites identified through modelling to either DLSA or iDLSA, it is posited that systematic mutagenesis of these sites in *Phem*Luc may uncover mutations which allow for improved compatibility with infraluciferin, which would correlate with a greater bioluminescence signal. Any mutations which are found to confer an increase to the observed bioluminescence may be doing so either directly by interaction with the bound ligand, or indirectly by influencing how the residues which directly interact with the substrate are positioned within the active site conformation. However, due to the high conservation scores of many of these positions detailed in Table 4.3., mutations made at these positions are likely to bring about significant deleterious effects on bioluminescence activity, regardless of whether the substitutions are of amino acids sharing similar biochemical properties, or significant differences such as charge and polarity. If the active site were more resilient to mutagenesis, it would likely correlate with a greater variation of amino acids found across these positions in the firefly luciferase sequences in Figure 4.3.

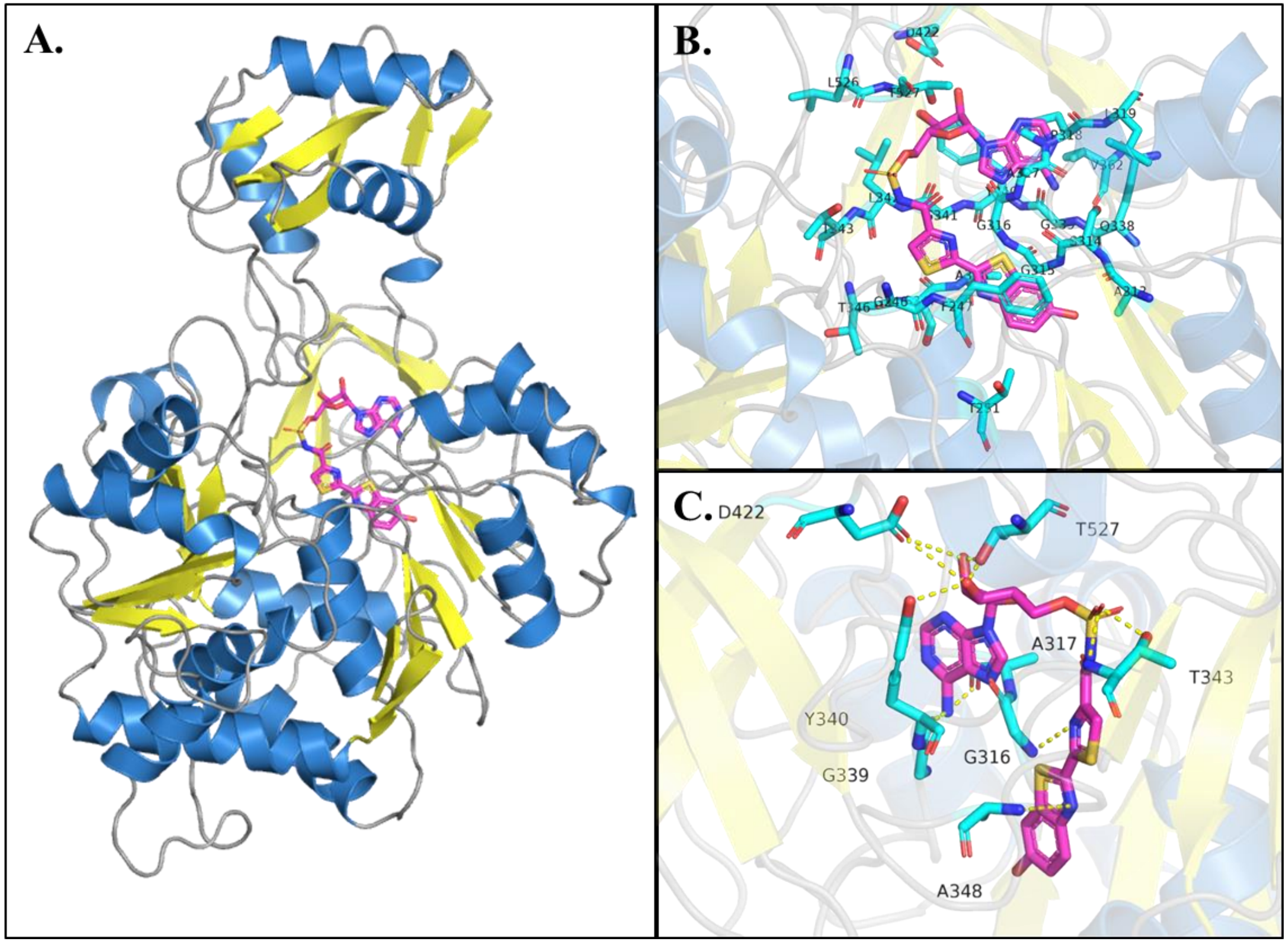
To investigate whether *Phem*Luc compatibility with infraluciferin could be improved, site-directed mutagenesis (SDM) was performed at each position individually to create libraries containing *Phem*Luc sequences incorporating all 20 possible natural amino acids for each of the 32 positions identified. The SDM libraries were created using primers containing mutagenic codon for the target sequences of NNK in the 5’ forward primer, and MNN in the 3’ reverse primer. In the IUPAC nucleotide code N represents any base, K would be G or T, and M is A or C. The use of such primer pairs reduces redundancy and disallows stop codons, resulting in a sequence library containing all possible coding codons in this position, allowing screening the effect of all 20 natural amino acids for the 32 SDM libraries. This would allow the screening of 640 unique sequences, from which mutations conferring improved compatibility with infraluciferin could be identified.

Figure 4.5.


A.
B.
D422
P318
Y340
L342
Q338
T346
S347
C.
D422
K529
G316
Q338
T343
G339
H245

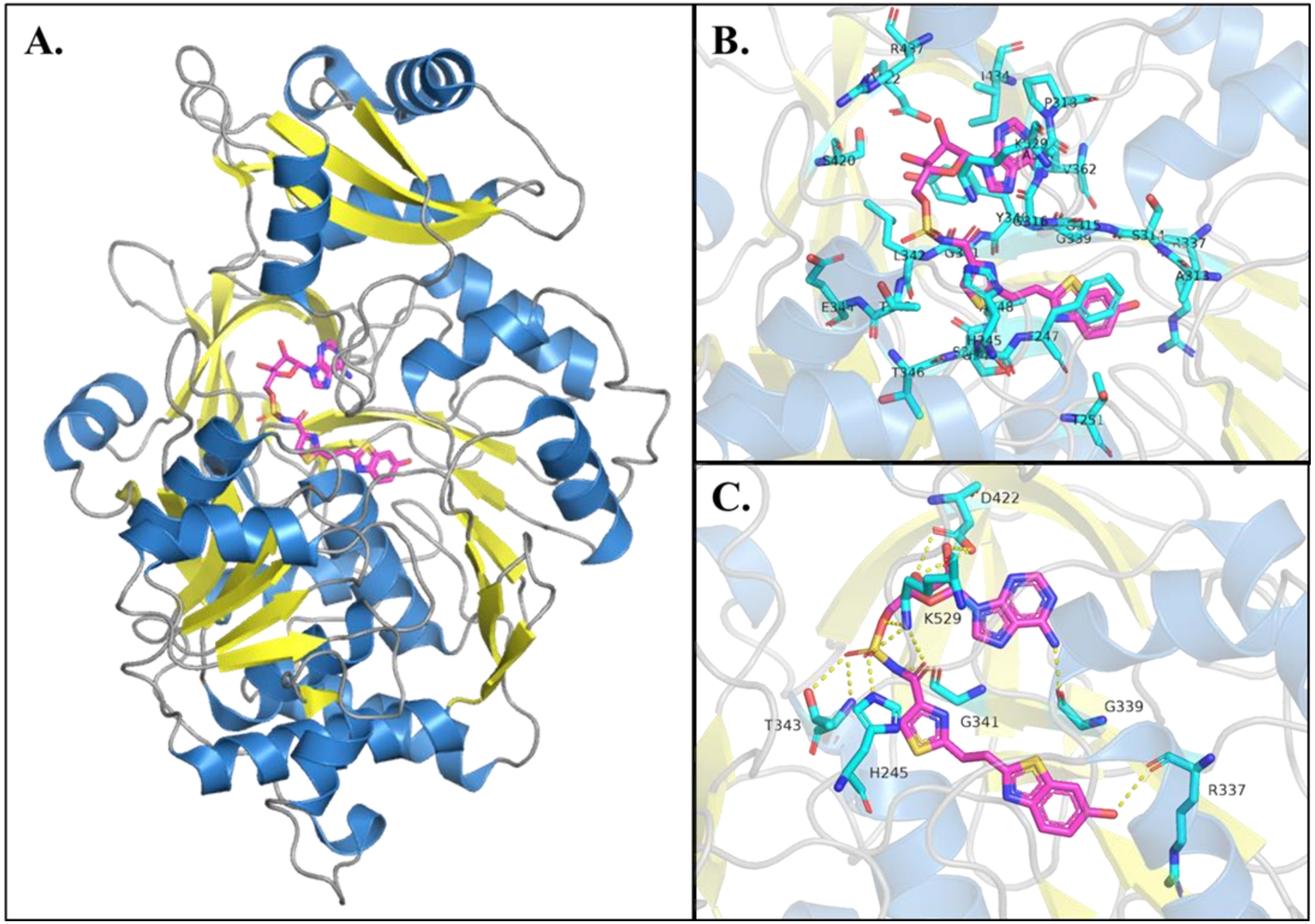
Modelling of *Ppy* Fluc and residues within 4 Å of bound LH_2 analogue. (**A**) Macro view of *Ppy* Fluc modelled in complex with the luciferyl-adenylate analogue 5’-O-[(N-dehydroluciferyl)-sulfamoyl]-adenosine (DLSA). Adapted from available crystal structure (PDB ID: 4G36). Blue: α-helix, Yellow: β-sheet, Grey: Coil environments. The bound DLSA ligand is contrasted with the colouring Magenta: Carbon, Dark blue: Nitrogen, Red: Oxygen, Yellow: Sulphur. (**B**) Focused view of bound DLSA with residues within 4 Å displayed as sticks with the Carbon backbone contrasted in Cyan and side chain elements Nitrogen and Oxygen in Dark Blue and Red, respectively. Orientation had been preserved relative to **A**. (**C**) Polar interactions between DLSA and *Ppy* Fluc indicated by yellow dashed lines and the interacting residues. Orientation has been adjusted to improve visualisation. Interactions are detailed in Table 4.4. Model analysis and imaging performed in PyMOL.

Figure 4.6.


A.
B.
D422
L526
T527
L319
P318
V362
A317
L342
G341
T343
G316
G339
Q338
S314
G315
A312
T346
G246
F247
T251
C.
D422
T527
A317
T343
Y340
G316
G339
A348

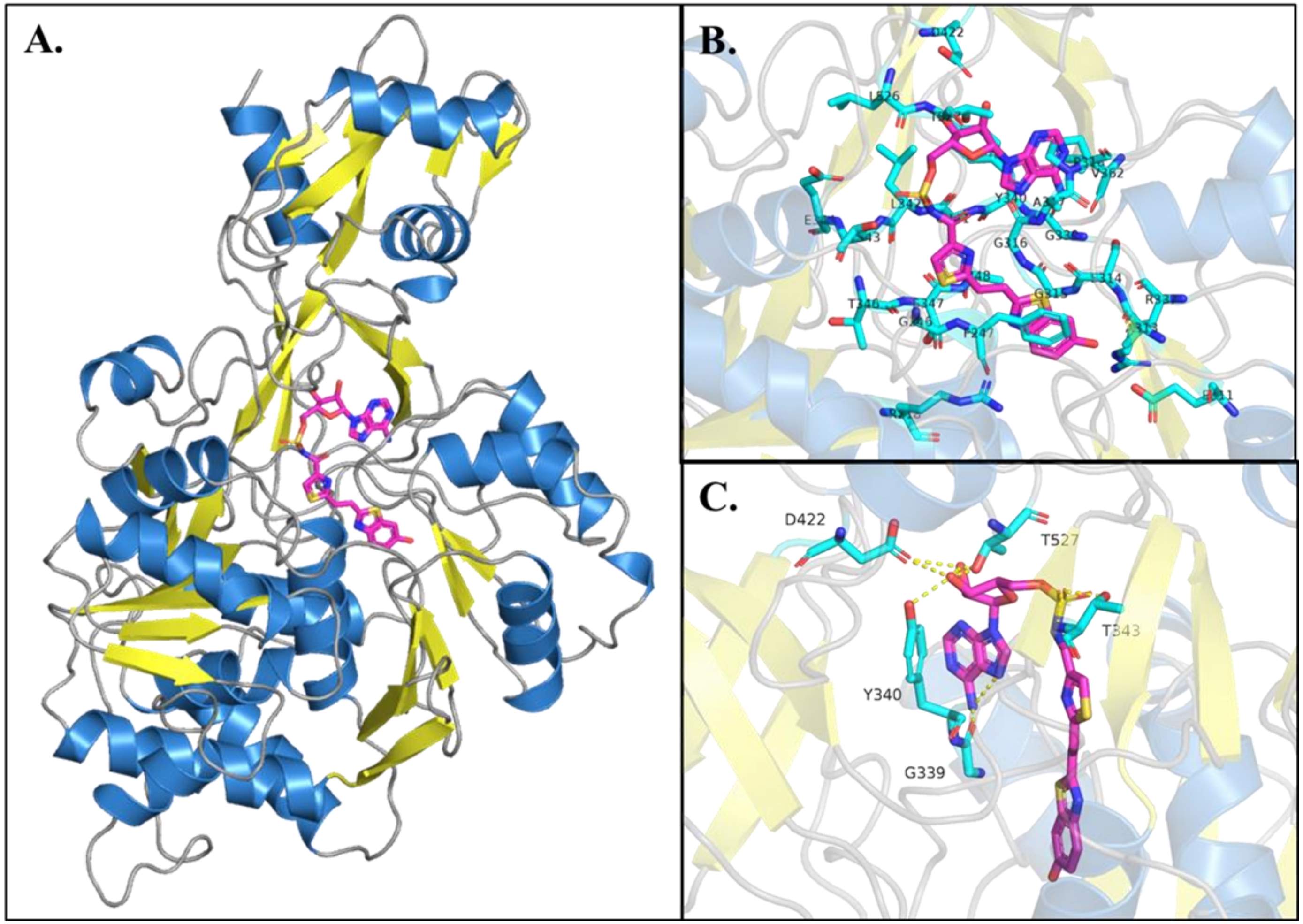
Modelling of *Phem*Luc and residues within 4 Å of bound LH$_2$ analogue. (**A**) Macro view of *Phem*Luc modelled in complex with the luciferyl-adenylate analogue 5'-O-[(N-dehydroluciferyl)-sulfamoyl]-adenosine (DLSA). Constructed from homology modelling of *Phem*Luc amino acid sequence to available *Ppy* Fluc crystal structure (PDB ID: 4G36) using SWISS-MODEL, and superimposition of DLSA by alignment to 4G36 in PyMOL (see Chapter 2). Blue: α-helix, Yellow: β-sheet, Grey: Coil environments. The bound DLSA ligand is contrasted with the colouring Magenta: Carbon, Dark blue: Nitrogen, Red: Oxygen, Yellow: Sulphur. (**B**) Focused view of bound DLSA with residues within 4 Å displayed as sticks with the Carbon backbone contrasted in Cyan and side chain elements Nitrogen and Oxygen in Dark Blue and Red, respectively. Orientation had been preserved relative to **A**. (**C**) Polar interactions between DLSA and *Phem*Luc indicated by yellow dashed lines and the interacting residues. Orientation has been adjusted to improve visualisation. Interactions are detailed in Table 4.4. Model analysis and imaging performed in PyMOL.

Figure 4.7.


A.
B.
R437
I434
P318
K529
S420
V362
Y340
G316
G315
G339
S314
R337
L342
A313
E344
H245
T247
T346
T251
C.
D422
K529
G339
T343
G341
H245
R337

Modelling of *Ppy* Fluc and residues within 4 Å of bound iLH_2 analogue. (**A**) Macro view of *Ppy* Fluc modelled in complex with the infraluciferyl-adenylate analogue 5'-O-[(N-dehydroinfraluciferyl)-sulfamoyl]-adenosine (iDLSA). Adapted from available crystal structure (PDB ID: 6HPS). Blue: α-helix, Yellow: β-sheet, Grey: Coil environments. The bound iDLSA ligand is contrasted with the colouring Magenta: Carbon, Dark blue: Nitrogen, Red: Oxygen, Yellow: Sulphur. (**B**) Focused view of bound iDLSA with residues within 4 Å displayed as sticks with the Carbon backbone contrasted in Cyan and side chain elements Nitrogen and Oxygen in Dark Blue and Red, respectively. Orientation had been preserved relative to **A**. (**C**) Polar interactions between iDLSA and *Ppy* Fluc indicated by yellow dashed lines and the interacting residues. Orientation has been adjusted to improve visualisation. Interactions are detailed in Table 4.4. Model analysis and imaging performed in PyMOL.

Figure 4.8.


A.
B.
D422
L526
V362
L342
Y340
G316
T346
F247
C.
D422
T527
T343
Y340
G339

Modelling of *Phem*Luc and residues within 4 Å of bound iLH_2 analogue. (**A**) Macro view of *Phem*Luc modelled in complex with the infraluciferyl-adenylate analogue 5'-O-[(N-dehydroinfraluciferyl)-sulfamoyl]-adenosine (iDLSA). Constructed from homology modelling of *Phem*Luc amino acid sequence to available *Ppy* Fluc crystal structure (PDB ID: 6HPS) using SWISS-MODEL, and superimposition of iDLSA by alignment to 6HPS in PyMOL (see Chapter 2). Blue: α-helix, Yellow: β-sheet, Grey: Coil environments. The bound iDLSA ligand is contrasted with the colouring Magenta: Carbon, Dark blue: Nitrogen, Red: Oxygen, Yellow: Sulphur. (**B**) Focused view of bound iDLSA with residues within 4 Å displayed as sticks with the Carbon backbone contrasted in Cyan and side chain elements Nitrogen and Oxygen in Dark Blue and Red, respectively. Orientation had been preserved relative to **A**. (**C**) Polar interactions between iDLSA and *Phem*Luc indicated by yellow dashed lines and the interacting residues. Orientation has been adjusted to improve visualisation. Interactions are detailed in Table 4.4. Model analysis and imaging performed in PyMOL.

Table 4.3.

	***Ppy* DLSA contacts**	***Phem*Luc DLSA contacts**	***Ppy* iDLSA contacts**	***Phem*Luc iDLSA contacts**	**Residue Conservation**
R218				✓	100%
H245	✓		✓		100%
G246	✓	✓	✓	✓	86%
F247	✓	✓	✓	✓	100%
T251	✓	✓	✓		100%
E311				✓	100%
A313	✓	✓	✓	✓	100%
S314	✓	✓	✓	✓	100%
G315	✓	✓	✓	✓	100%
G316	✓	✓	✓	✓	100%
A317	✓	✓	✓	✓	100%
P318	✓	✓	✓	✓	100%
L319		✓			100%
R337			✓	✓	100%
Q338	✓	✓			100%
G339	✓	✓	✓	✓	100%
Y340	✓	✓	✓	✓	93%
G341	✓	✓	✓	✓	100%
L342	✓	✓	✓	✓	100%
T343	✓	✓	✓	✓	100%
E344			✓	✓	100%
T346	✓	✓	✓	✓	100%
S347	✓	✓	✓	✓	93%
A348	✓	✓	✓	✓	100%
V362	✓	✓	✓	✓	21%
S420			✓		64%
D422	✓	✓	✓	✓	100%
I434			✓		100%
R437	✓		✓		100%
L526		✓		✓	86%
T527		✓		✓	100%
K529	✓		✓		100%

Residues identified within 4 Å of bound substrates across protein models. All 32 amino acid residues identified from all 4 models are displayed in bold on the left. Tick marks indicate where the given site has been identified within 4 Å for each respective model. Residue conservation indicates the percentage at which the identified amino acid is found in the respective position across the firefly luciferases aligned in Figure 4.3. Underlined residues are known to function in the active site of the first conformation of luciferase. Italicized residues are known to function in the second conformation (Sundlov *et al.* 2012). Residues common to both conformations are both underlined and italicized.

Table 4.4

Residue	***Ppy* DLSA**			***Phem*Luc DLSA**		
	Distance (Å)	**Sidechain Interaction Centre**	**Interacting DLSA Atom**	**Distance (Å)**	**Sidechain Interaction Centre**	**Interacting DLSA Atom**
H245	3	NE2	O18			
G316	3	NE2	N37	3.2	N	N15
A317				2.2	O	N38
Q338	2.7	OE1	N38			
G339	2.7	OE1	N38	1.5	O	N38
Y340				3	OH	O27
T343	3.3	NE2	O19	3.1	N	O19
T343	3.2	OG1	O19	2.7	OG1	O19
A348				3.5	N	N7
D422	2.7	OD1	O28	3.3	OD2	O27
D422	2.8	OD2	O27	3.4	OD2	O28
D422	3.2	OD2	O28			
T527				1.6	OG1	O27
T527				1.6	OG1	O28

Residue	***Ppy* iDLSA**			***Phem*Luc iDLSA**		
	Distance (Å)	**Sidechain Interaction Centre**	**Interacting iDLSA Atom**	**Distance (Å)**	**Sidechain Interaction Centre**	**Interacting iDLSA Atom**
H245	2.7	NE2	OBK			
R337	3.3	O	OAA			
G339	2.5	O	N6	1.2	O	N6
Y340				2.6	OH1	O3'
G341	3	O	NAO			
T343	3.1	NZ	OBJ	2.4	N	OBJ
T343	2.8	OG1	OBJ	2.6	OG1	OBJ
D422	3.1	OD1	O3'	2.9	OD1	O2'
D422	3.4	OD2	O3'	3	OD1	O3'
D422	3.4	OD2	O2'			
T527				1.5	OG1	O2'
T527				2.1	OG1	O3'
K529	3.4	NZ	O5'			
K529	3	NZ	OBK			
K529	3.1	NZ	OBL			

Polar contacts with DLSA and iDLSA identified across protein models. All polar contacts identified across the protein models of Figures 4.5C – 4.8C and the distance in angstroms of the interactions. The sidechain interaction centres as defined by Bahar and Jernigan, 1996 and the interacting ligand atom are detailed. Grey fills are used to indicate the absence of the respective bond at that residue.

4.3.4. Expression and *in vitro* bioluminescence analysis of substrate contact libraries in *E. coli*

4.3.4.1 Estimating SDM library quality and confidence in diversity

The use of NNN codons in primer design allows incorporation of the full degeneracy of the genetic code into the constructed library. This method of randomizing a targeted position has a significant potential to incorporate premature stop codons in a large proportion of the total library (Patrick and Firth 2005). To circumvent this issue, the SDM primers for the target positions identified in Table 4.3. were designed using NNK/MNN primer pairs, which encode 32 equiprobable sequence variants that encompass all 20 natural amino acids. All 32 variants would need to be screened to determine which amino acid conferred the greatest improvements to activity in this position. Equation 4.1. estimates the library size required to have a 95% chance of being 100% complete, where complete here means representing each equiprobable sequence variant at least once (Patrick *et al.* 2003). The implication of this equation is that for a transformation of SDM products to be deemed successful and worth screening, a minimum of 206 colonies would be required to maximise the likelihood of screening all 32 possible sequence variants. Whilst the use of NNK/MNN codons prevents the inclusion of stop codons and therefore reduces the percentage of non-functional mutants, a bias still remains due to the degeneracy of the amino acid code. The amino acids arginine, leucine, and serine are each encoded by 6 unique codons, whilst methionine and tryptophan are each encoded by only a single codon. The implication from this codon redundancy is that methionine and tryptophan mutations would be the rarest to occur in each library and both had the greatest chance of being excluded if the calculated library size of 206 colonies was not achieved. For this reason, any libraries that did not transform efficiently were recreated to achieve libraries which when transformed would produce >206 colonies.

4.3.4.2 Primary colony screening and basis of selection

In any protein engineering study, a method of screening protein activity is required in order to identify any divergences in protein characteristics within the mutant population relative to the original enzyme, regardless of whether these changes are beneficial or deleterious to the desired enzyme activity. For luciferases, their bioluminescence activity enables a powerful screening strategy in transformant colonies of *E. coli* by identifying changes to the bioluminescence signal representing a desired phenotype relative to the original enzyme. This method was first shown to be an effectual luciferase screening strategy by Wood and Deluca (1987), and has here been adapted to identify *Phem*Luc mutants which exhibit improved compatibility with iLH_2, as indicated by a greater yield of bioluminescence.

Transformed SDM libraries of >206 colonies were induced for recombinant protein production with IPTG, prior to screening bioluminescence activity with an iLH_2-citrate spray in the PhotonIMAGER Optima (Biospace Labs, Paris, France), as further explained in Chapter 2. The subsequent bioluminescence data was analysed in the M3 Vision software package, available under license from https://biospacelab.com. Figure 4.9. illustrates how bioluminescence activity is represented in M3 Vision, where lower bioluminescence signals are indicated in blue, and greater signals in red. Total bioluminescent emission are recorded as radiant flux (Ph/s/cm^2/sr) which evaluates the photon emission (Ph) as a product of time (s), area or colony size (cm^2), and steradian (sr), which accounts for the emission of bioluminescence signal in 3D space, as opposed to a direct line of emission between the emitting colony and the imaging camera. The *Phem*Luc transformation control depicts the bioluminescence signal of a uniform unaltered construct, relative to the diversity in bioluminescence signal from the example SDM libraries A313X and V362X. The A313X SDM library produces a diverse range of bioluminescence signals, including several that appear to be high activity, as indicated in red. In contrast, the V362X library contains many low activity or entirely inactive colonies. Sequencing of the highest activity colonies from such a library reveals the original *Phem*Luc sequence, indicating that the substitution of any other amino acid are highly deleterious at these positions.

4.3.4.3 Secondary and tertiary screening against *Phem*Luc control

Regardless of whether SDM libraries contained high activity phenotypes such as A313X, or appeared to be similar to *wild-type Phem*Luc, the highest activity colonies from each plate were replicated in triplicate on secondary screening plates to allow comparison against a sequence verified transformation of *Phem*Luc (Figure 4.10.). In this secondary screening process, many inactive or lower activity colonies were brought through from the primary screens, especially in SDM libraries that appeared to display higher activity phenotypes amongst some of the colonies that were also taken forward to the secondary screen. This was carried out as identifying the colony on the original growth plate which corresponded to the bioluminescence data recorded in M3 Vision for the nitrocellulose membrane transfers (see Chapter 2) could be challenging if the improved activity phenotype colony was located amongst a high density of colonies on the original SDM library transformed plate, as was often the case. To ensure the inclusion of the high activity phenotype in the secondary screen, the entire region from which it originated would be selected.

Many of the ‘highest activity colonies’ brought forward to the secondary screens from SDM libraries without significantly improved activity phenotypes (i.e. colony radiant flux – measured in Ph/S/cm^2/sr) produced bioluminescence signals comparable to the *Phem*Luc control included on each secondary plate. Sequencing these colonies revealed that they were indeed reproductions of the *wild-type Phem*Luc sequence within the SDM library. As previously discussed, even with a different substrate analogue, mutation of conserved residues impacted negatively on function, possibly owing to the similarity in structures between *D*-LH$_2$ and *DL*-iLH$_2$ (Figure 4.1.). It is important to note that the high frequency of *wild-type Phem*Luc amongst the colonies picked was due to reversion mutagenesis to the relatively high activity original *Phem*Luc sequence (with variation in codon usage for the respective position) compared to mutants from the 32 equiprobable sequence variants of each SDM library, and not due to contamination of the final SDM libraries with the *Phem*Luc template which would have been removed by *Dpn*1 digest (see Chapter 2).

From all the colonies assessed in the secondary screen, only 16 were taken on to the tertiary screening in Figure 4.10. Of these 16 colonies, Sanger sequencing revealed that 7 were reconstructions of the *Phem*Luc sequence and a further 3 were replicates. This left only 6 mutations to be carried forward for further analysis – H245W, E311S, A313G, S314V, S347T, and A348V.

4.3.4.4 Comparison of sequence verified isolates and cumulative effects

The final 6 isolated sequences from the SDM libraries in pET16b were transformed into fresh *E.coli* BL21 (DE3) stocks and the screening was again repeated as carried out for all screening with iLH$_2$-citrate spray. The resulting bioluminescence signal images were scaled to identical intensity ranges and compared in M3 Vision (Figure 4.12.). Following the performance of fresh transformations, H245W appeared to produce the greatest yield of bioluminescence relative to the other 5 final isolates. The mutations A313G produced the second highest activity phenotype, followed by A348V. The bioluminescence activity of E311S, S314V, and S347T could not be visualised under this scaling. Increasing the LUT range in M3 vision to indicate their respective intensities saturated the intensity representations of H245W and A313G. Although each plate represents colonies of *E. coli* transformed with a sequence verified homogenous DNA construct, significant variation can be observed between plated colonies. Accurately attributing this to a proper cause is challenging, as it may be that expression levels of the construct varied significantly between colonies, or more likely relates to the method of iLH_2 delivery in citrate spray, which has never been assessed or optimised as a method of luciferase screening in *E. coli.*

The single mutations identified to improve the bioluminescence activity of *Phem*Luc with iLH_2 suggested an adaptation of the active site conformation to better accommodate the increased size of the iLH_2 structure relative to LH_2, so an attempt was made to investigate whether these single mutations could be combined to produce combinatorial mutations with cumulative effects producing a phenotype improved relative to any single mutation alone. For this purpose, two different cumulative mutants of varying complexity were constructed. The first, x6 Infra, combined all 6 of the final SDM library isolates, and the second less complex combination was a double mutant of the two most active SDM isolates, H245W and A313G, called x2 Infra. x6 Infra was constructed using a synthesized gene, whilst x2 Infra was constructed by restriction digest and ligation of H245W and A313G (see Chapter 2 for further details).

Both combinatorial mutants were screened for bioluminescence activity in *E.coli* against a *Phem*Luc control and the 6 SDM isolates from which they were derived, in Figure 4.13. A313G remained the highest activity single mutation observed in this screen, followed by H245W. Interestingly, H245W produced a reduced bioluminescence signal in comparison to *Phem*Luc, yet in combination with A313G in the x2 Infra mutant, the bioluminescence signal exceeded that of any single mutation alone. However, the increased activity of x2 Infra was not found to be significantly different (P=0.9517) to A313G in this screen. The higher complexity combinatorial mutant x6 Infra suffered significant deleterious effects to its bioluminescence activity. Although it is difficult to speculate accurately on the cause of the activity inhibition, one explanation may be the inclusion of multiple mutations in proximity to the active site, which are adjacent to the ligand such as E311S, A313G, and S314V, or S347T and A348V. Whilst the inclusion of these as single mutations may produce only lesser modifications in the conformation of the active site, the addition of so many in such close proximity appear to produce a significant alteration to the active site conformity, which presumably impedes its ability to catalyse the bioluminescence reaction.

To confirm whether x2 Infra possessed a bioluminescent activity consistently greater than the activity of its individual mutations, it was further investigated in *E. coli* against *Phem*Luc, H245W, and A313G (Figure 4.14.). From this screen, a considerable degree of variation can be seen in the bioluminescent activity of *Phem*Luc relative to the replicate colony groups for H245W, A313G and x2 Infra, which produce a more consistent bioluminescence signal. Regardless of the variation inflating the average activity of *Phem*Luc plotted in Figure 4.14B., the A313G phenotype produced a 9.7% greater bioluminescent yield. Whilst H245W appears to produce comparable bioluminescent activity to *Phem*Luc in this screen (99.2%), when combined with A313G in x2 Infra, the bioluminescent yield is recorded as 25% greater than the *wild-type Phem*Luc. So whilst H245W appears to produce minor deleterious effects on the activity of *Phem*Luc alone, it is compensated by the addition of A313G by a yet unknown mechanism. The activity of x2 Infra was found to be significantly different to A313G during this screen (P=0.0202), in contrast to the lack of difference observed in Figure 4.13. From the activity observed in this screen, x2 Infra was the most active mutant with iLH_2, and was selected to be taken forward for analysis of the bioluminescence activity as purified protein.

Equation 4.1.

$$L = -V \ln\left(-\frac{\ln Pc}{V}\right)$$

L = Library containing a number of clones/colonies (Unknown)

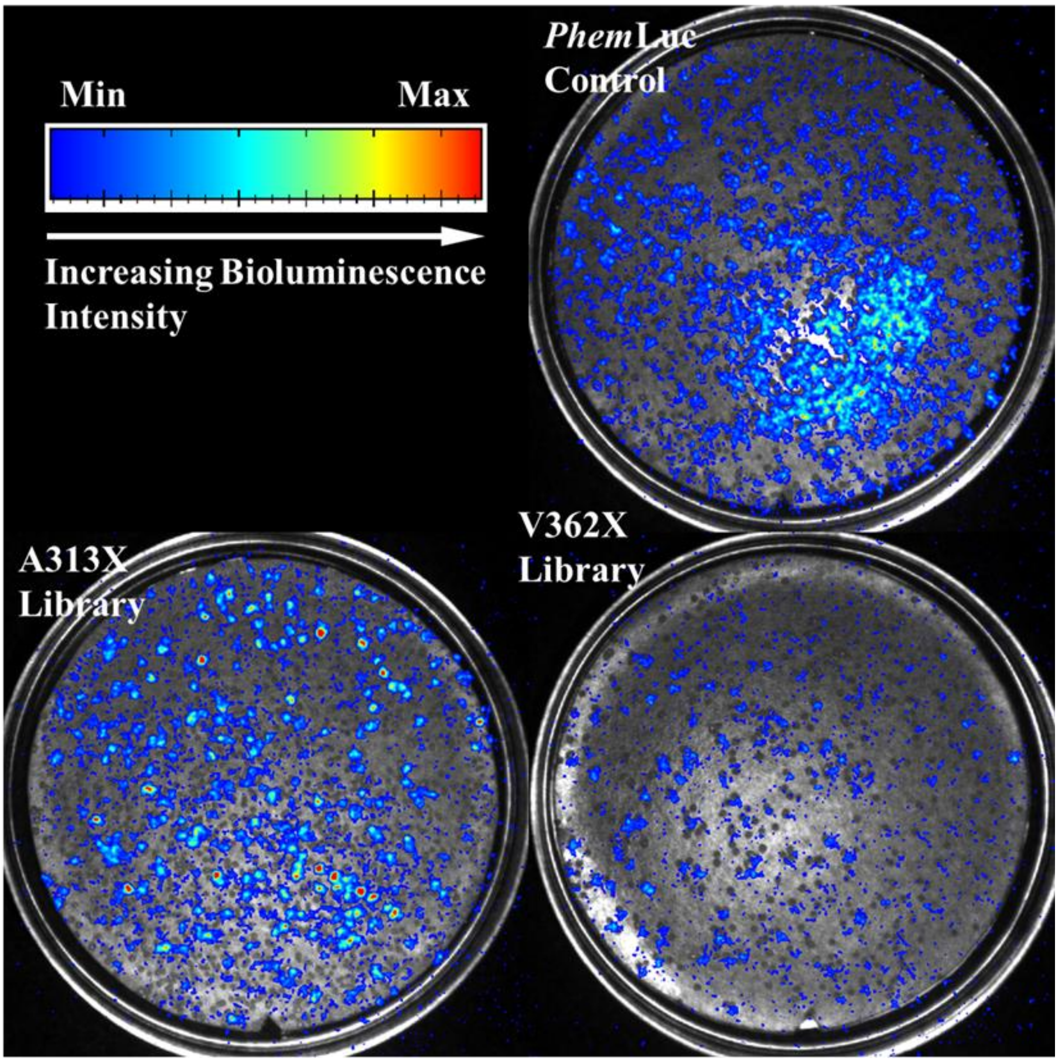
V = Total number of sequence variants (32 possibilities from NNK codon)

Pc = Probability of complete library (95% confidence)

$$L = -32 \ln\left(-\frac{\ln(0.95)}{32}\right) \approx \mathbf{206} \text{ (rounded)}$$

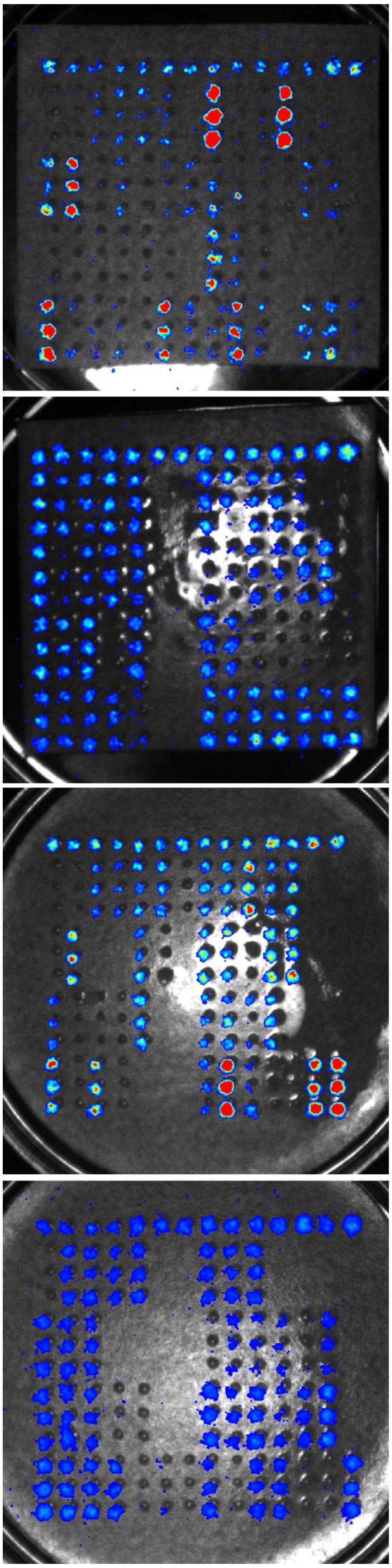
Estimating library size for complete diversity of sequence variants. The above equation from Patrick *et al* (2003) calculates the size of library required to have a 95% chance of containing every possible sequence variant. The degree of over-sampling required positively correlates with the number of sequence variants. NNK/MNN codons give rise to 32 equiprobable sequence variants (unrelated/coincidental to the 32 residues identified in Table 4.3.), meaning to obtain a 95% chance of screening all variants, 206 (rounded) clones/colonies will need to be screened in the given library.

Figure 4.9.



Representation of primary screening of substrate contact libraries in *E.coli* colonies. The colony formations from *E. coli* BL21 (DE3) transformed with substrate contact libraries previously created by SDM, here transferred onto nitrocellulose membranes (Chapter 2). Colony transferred membranes were induced for 3-4 hours at RT with IPTG (1 mM). All plates were subsequently screened with 500 µM iLH$_2$ and imaged in the PhotonIMAGER Optima (Biospace Labs, Paris, France), over an integrating period of 60 seconds. The upper left colour spectrum illustrates how bioluminescence signal intensity is represented relative to the maximum signal detected, with minimum intensity indicated in blue and maximum intensity indicated in red. *Phem*Luc control demonstrates the uniform bioluminescence signal from transforming the unmodified *Phem*Luc construct. A313x library demonstrates a transformed substrate contact library where higher activity mutations can be observed. V362X demonstrates a transformed contact substrate library which is deleterious to activity for all amino acids substituted in the library which differ from the *wild-type Phem*Luc. The three images were acquired independently, but the colour intensity representation has been scaled to be directly comparable in the M3 Vision software package.

Figure 4.10.



*Phem*Luc	
R218X	H245X
G246X	F247X
T251X	E311X
A313X	S314X

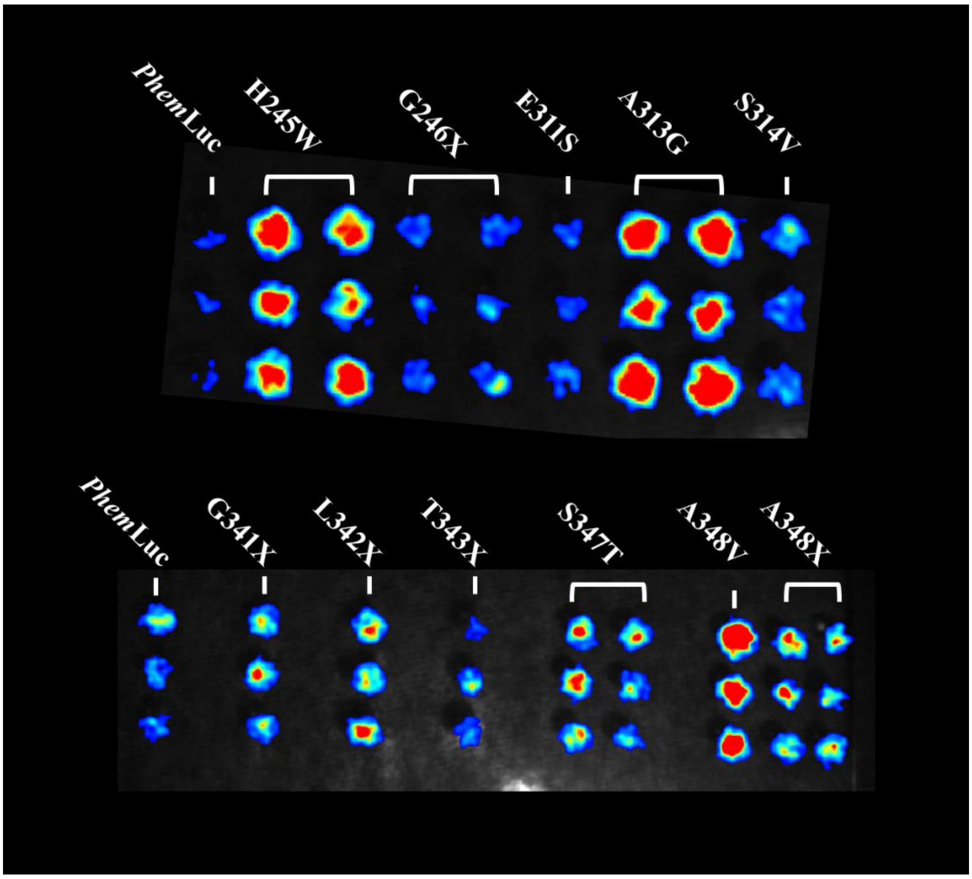
*Phem*Luc	
G315X	G316X
A317X	P318X
L319X	R337X
Q338X	G339X

*Phem*Luc	
Y340X	G341X
L342X	T343X
E344X	T346X
S347X	A348X

*Phem*Luc	
V362X	S420X
D422X	I434X
R437X	L526X
T527X	K529X

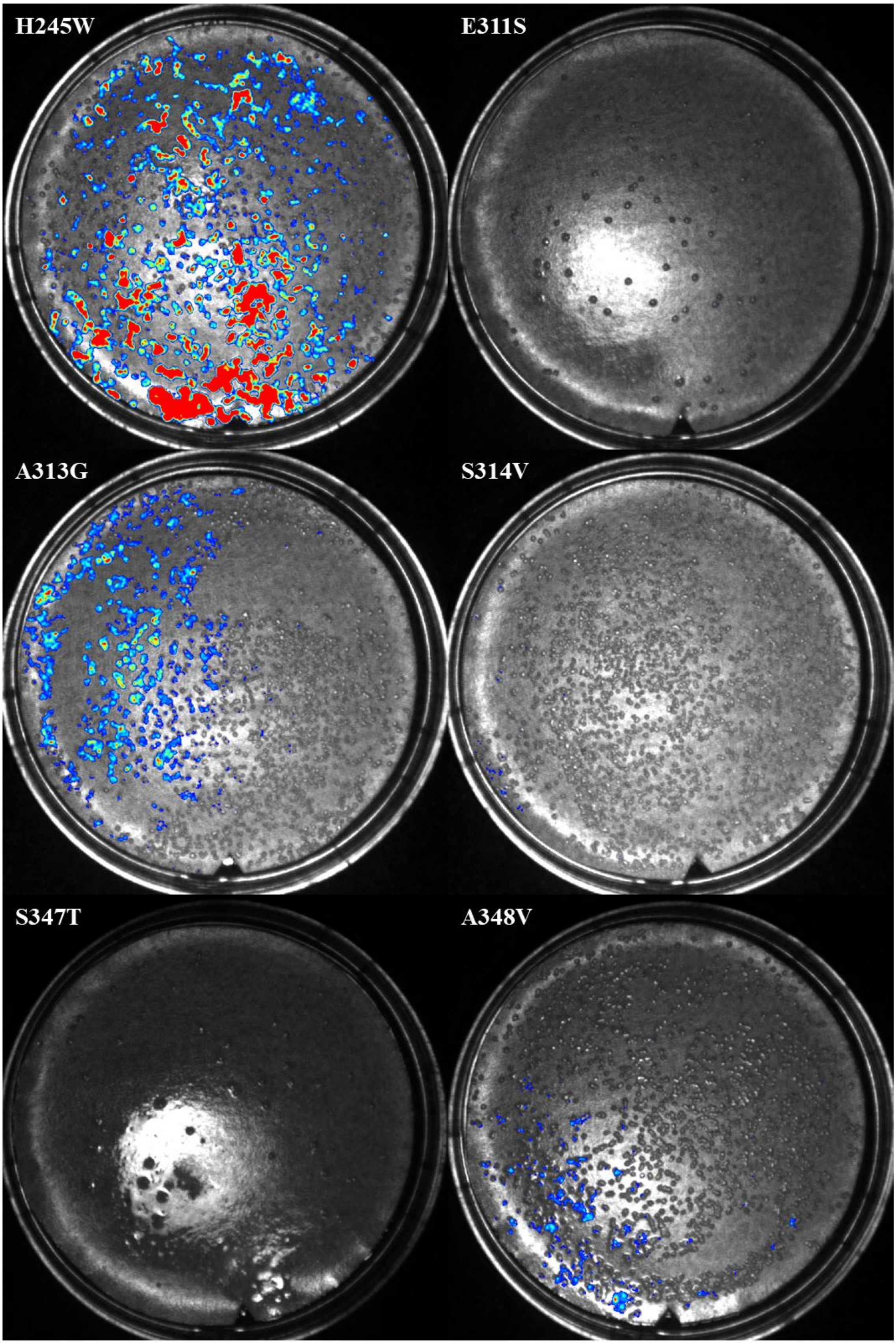
Secondary screening of substrate contact libraries. Nitrocellulose membranes with *E. coli* BL21 (DE3) regrown from colonies exhibiting the highest activities within their respective library. Colony transferred membranes were induced for 3-4 hours at RT with IPTG (1 mM). All plates were subsequently screened with 500 µM iLH_2 and imaged in the PhotonIMAGER Optima (Biospace Labs, Paris, France), over an integrating period of 60 seconds. Tables on the right indicate the library of origin from which the imaged colonies were isolated in the left hand images. All images contain a top row of *Phem*Luc to act as control, and all colonies have been replicated in triplicate. The four images were acquired independently, and the colour intensity representation has not been scaled to be directly comparable in the M3 Vision software package, as was performed in Figure 4.10. Signal intensity is represented relative to the maximum signal detected, with minimum intensity indicated in blue and maximum intensity indicated in red.

Figure 4.11.



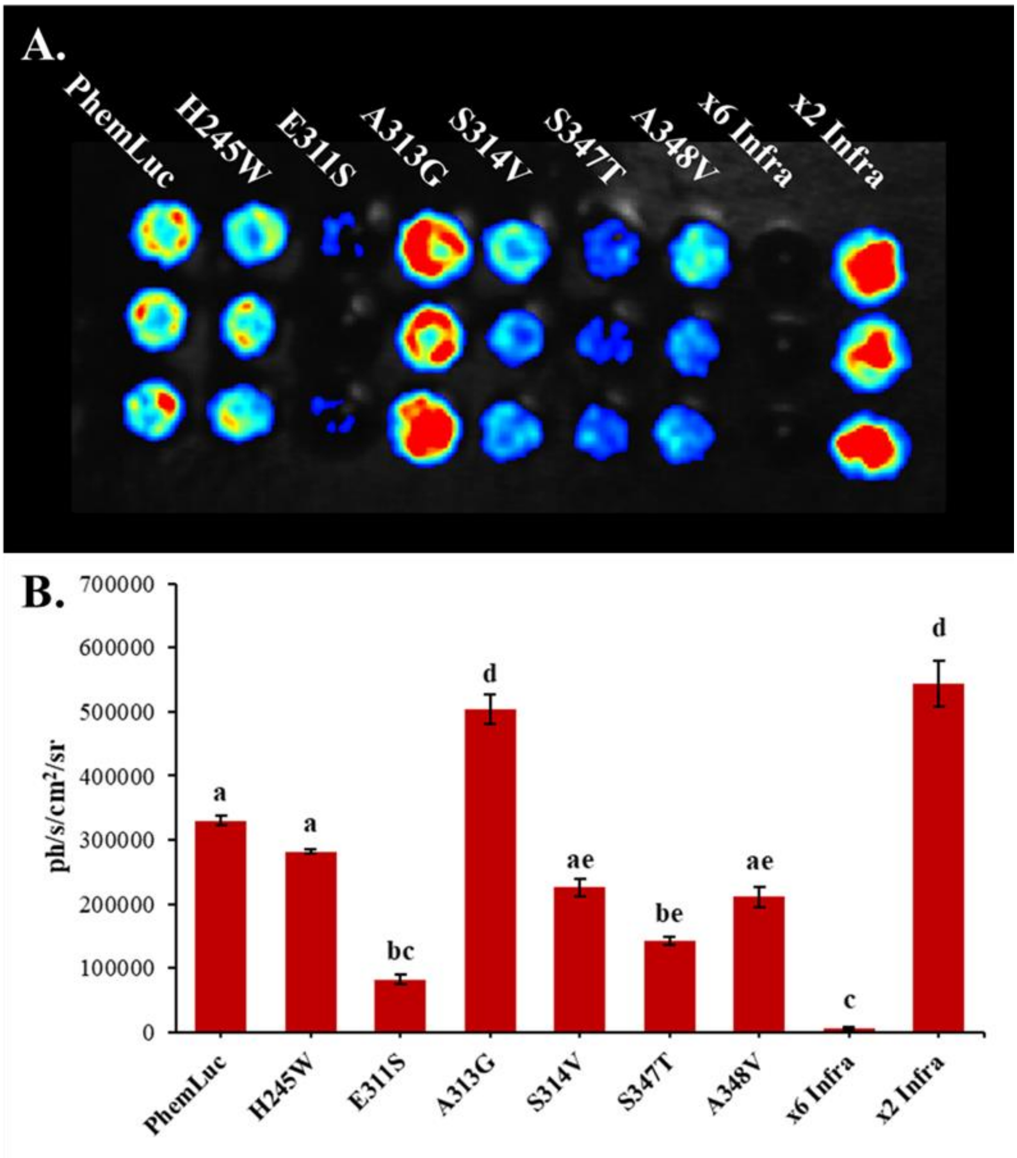
Tertiary screening of substrate contact libraries. Nitrocellulose membranes with *E. coli* BL21 (DE3) regrown from colonies exhibiting the highest activities within their respective library, from the secondary screen. Colony transferred membranes were induced for 3-4 hours at RT with IPTG (1 mM). All plates were subsequently screened with 500 µM iLH_2 and imaged in the PhotonIMAGER Optima (Biospace Labs, Paris, France), over an integrating period of 60 seconds. All colonies are replicated in triplicate, and names indicate the mutations present, as established by Sanger sequencing. All names ending with X represent their library of origin, but were revealed to be *wild-type Phem*Luc by Sanger sequencing. The two images were acquired independently, and the colour intensity representation has not been scaled to be directly comparable in the M3 Vision software package, as was performed in Figure 4.10. Signal intensity is represented relative to the maximum signal detected, with minimum intensity indicated in blue and maximum intensity indicated in red.

Figure 4.12.



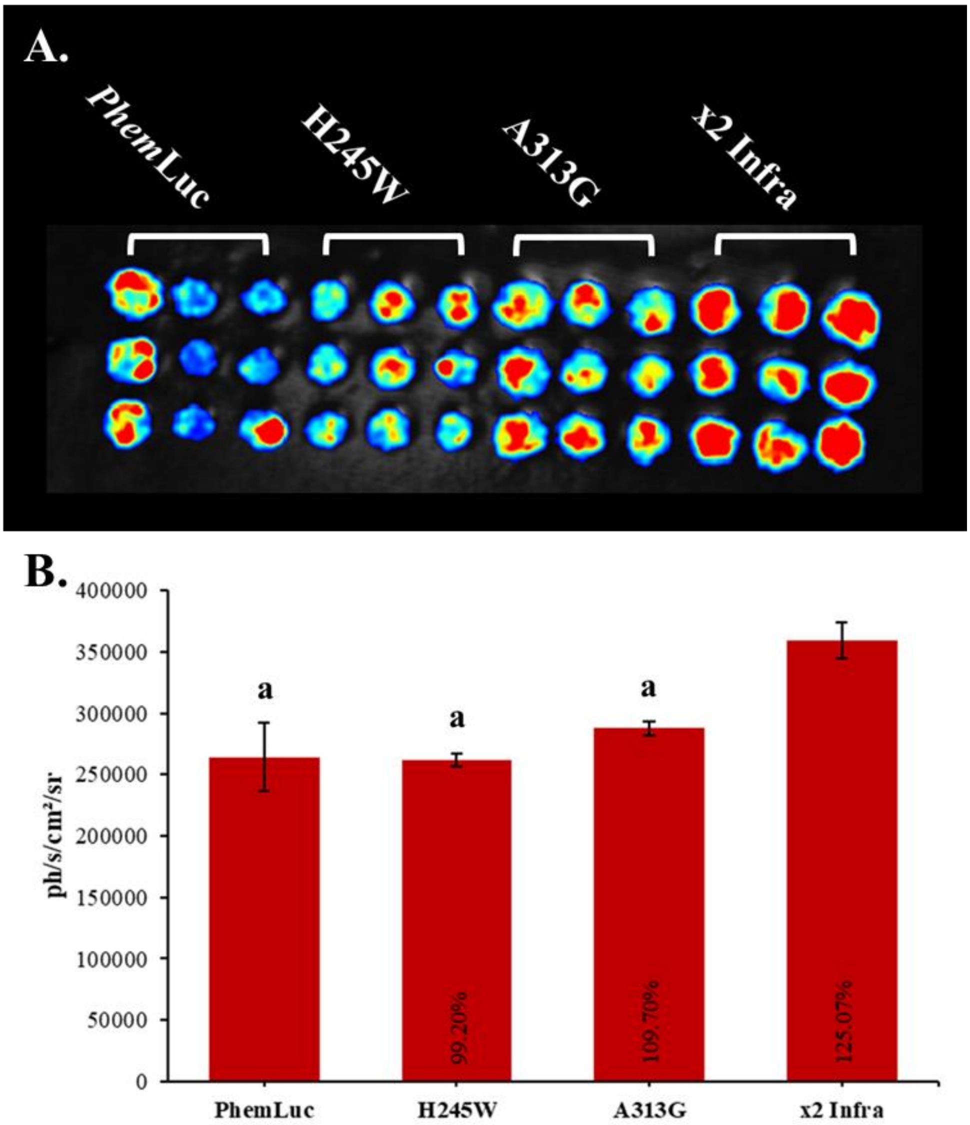
Comparison of final six sequence verified isolates. Nitrocellulose membranes with *E. coli* BL21 (DE3) transformed with the final six sequence verified substrate contact library isolates: H245W, E311S, A313G, S314V, S347T, and A348V. Colony transferred membranes were induced for 3-4 hours at RT with IPTG (1 mM). All plates were subsequently screened with 500 µM iLH_2 and imaged in the PhotonIMAGER Optima (Biospace Labs, Paris, France), over an integrating period of 60 seconds. The six images were acquired independently, and the colour intensity representation has been scaled to be directly comparable in the M3 Vision software package. Signal intensity is represented relative to the maximum signal detected, with minimum intensity indicated in blue and maximum intensity indicated in red.

Figure 4.13.



***In vitro* bioluminescence analysis of combinatorial mutants x6 Infra and x2 Infra.** (**A**) Nitrocellulose membranes with *E. coli* BL21 (DE3) transformed with *wild-type Phem*Luc, the final 6 sequence verified substrate contact library isolates, and 2 combinatorial mutants: *Phem*Luc, H245W, E311S, A313G, S314V, S347T, A348V, x6 Infra, and x2 Infra. Colony transferred membranes were induced for 3-4 hours at RT with IPTG (1 mM). All plates were subsequently screened with 500 µM iLH_2 and imaged in the PhotonIMAGER Optima (Biospace Labs, Paris, France), over an integrating period of 60 seconds. x6 Infra comprises all 6 substrate contact library isolates, and x2 Infra contains only H245W and A313G. Signal intensity is represented relative to the maximum signal detected, with minimum intensity indicated in blue and maximum intensity indicated in red. (**B**) Bar chart displaying the averaged bioluminescence activity measured across the 3 colonies for each luciferase. Error bars represent the standard error of the mean (SEM). Statistical analysis performed as detailed in 2.12.

Figure 4.14.



Comparison of x2 Infra to its single mutations and *Phem*Luc. (**A**) Nitrocellulose membranes with *E. coli* BL21 (DE3) transformed with *wild-type Phem*Luc, H245W, A313G, and x2 Infra. Colony transferred membranes were induced for 3-4 hours at RT with IPTG (1 mM). All plates were subsequently screened with 500 µM iLH_2 and imaged in the PhotonIMAGER Optima (Biospace Labs, Paris, France), over an integrating period of 60 seconds. For each construct, three separate colonies were picked and replicated in triplicate. Signal intensity is represented relative to the maximum signal detected, with minimum intensity indicated in blue and maximum intensity indicated in red. (**B**) Bar chart displaying the averaged bioluminescence activity measured across the 9 colonies for each luciferase. Percentages reflect the activity relative to *wild-type Phem*Luc, and error bars are the standard error of the mean. Statistical analysis performed as detailed in 2.12.

4.3.5. Analysis of purified x2 Infra bioluminescence activity with iLH$_2$

Whilst screens using *E. coli* transformations are broadly indicative of modifications to bioluminescence phenotypes following mutagenesis, bioluminescence assays using purified protein for the measurement of luciferase bioluminescence allow for greater control of variables and provide an enzyme characterization of greater accuracy (see Chapter 2 and 6 for further purification discussion). The bioluminescence properties of x2 Infra were recorded alongside the *wild-type Phem*Luc and the *Ppy* Fluc mutant x11, which has previously been shown to produce moderate bioluminescence activity when catalysing iLH$_2$ (Jathoul *et al.* 2014; Anderson *et al.* 2019). However, the use of iLH$_2$ in the performance of bioluminescence enzyme assays came with limitations not present in the LH$_2$ screens of Chapter 6.

As detailed in Chapter 2, the iLH$_2$ used in this study was synthesized as a racemic mix, meaning it contains equal quantities of dextrorotatory (*D*) and levorotatory (*L*) stereoisomers, whereas LH$_2$ used in Chapters 5-6 contained only the dextrorotatory stereoisomer, as it naturally occurs in nature. The *DL*-iLH$_2$ mix could therefore not be used for enzyme kinetic assays and thus parameters of K_M and K_{cat} could not be derived.

In addition to the racemic quality of the iLH$_2$, supply was extremely limited as it is not commercially available and the difficulty and expense involved in its synthesis (Jathoul *et al.* 2014). Conducting enzyme assays in the BMG Labtech CLARIOstar as was used for all LH$_2$ assays was not possible due to the significant volume of substrate mix required to load the injector pumps. This meant that assays were restricted to those that could be performed in the PhotonIMAGER Optima, where the quantity of iLH$_2$ utilised could be reduced to align with the available supply. The necessary use of the PhotonIMAGER restricted the available enzyme assays to lower resolution bioluminescence spectra than can be recorded in the CLARIOstar, and comparative measurements of enzyme total bioluminescence yields.

4.3.5.1 Bioluminescence spectra of purified x2 Infra with iLH$_2$

The emission wavelengths for luciferases catalysing iLH$_2$ have previously been shown to exhibit a colour-shifting effect, where the peak emission wavelength (λ_{max}) will be red-shifted in excess of 100 nm compared to the λ_{max} values recorded with LH$_2$ (Jathoul *et al.* 2014). To measure the emission colours of the purified enzymes with iLH$_2$, bioluminescence light output was recorded through the successive 30 nm band-pass filters of the PhotonIMAGER. Whilst this does not allow for the 1 nm resolution obtainable for spectra measurements performed in the CLARIOstar, as was done with LH_2 in Chapter 6, the measurements obtained through the PhotonIMAGER band-pass filters are +/-15 nm, which is sufficient resolution to observe a red-shifting effect relative to LH_2 emissions.

Bioluminescence spectra of purified x2 Infra, *Phem*Luc, and *Ppy* Fluc x11 were measured across the PhotonIMAGER band-pass filters in the presence of saturating conditions of iLH_2 and ATP. Whilst true saturating conditions of iLH_2 are unknown, an estimate was made from Jathoul *et al* (2014), where measurements for the K_M of *D*-iLH_2 with *Ppy* Fluc were obtained after saponification of *D*-iLH_2 methyl ester. Substrates and enzymes were mixed by manual pipetting and left at room temperature for 60 seconds. Following this, an initial measurement of total bioluminescence specific activity over 60 seconds was made (Figure 4.16.) prior to acquisition of spectra. To obtain spectra, 60 second measurements were taken across band-pass filters starting at a midpoint of 472 nm up to 797 nm, with a step-width of 25 nm, in respect to the band-pass filter midpoint. The resulting spectral curves produced for each enzyme are shown in Figure 4.15. For each enzyme, λ_{max} was recorded at the 697 nm band-pass filter. The λ_{max} previously reported for *Ppy* Fluc x11 with racemic *DL*-iLH_2 was 685 nm by Anderson *et al.* (2019), which would agree with λ_{max} recording obtained here in the 697 nm band-pass filter. This agreement between the *Ppy* Fluc x11 measurements presented here and reported previously can be considered as an indication of similar reliability of the spectral measurements for *Phem*Luc and x2 Infra.

Full Width Half Maximum (FWHM) is defined as the width of a spectra curve measured between the two opposite points recorded at half maximum intensity. The FWHM cannot be accurately determined from the band-pass filters used here, which are only accurate to +/-15 nm. However, normalization of the spectral curves still allows for comparative visualization of the bandwidths of bioluminescence emission. Normalisation of the spectral measurements as presented in Figure 4.15B. indicates a narrowing of the emission bandwidth in x2 Infra relative to *Phem*Luc, which possesses a larger shoulder to the curve into the shorter wavelength region. This subtle reduction in bandwidth would reduce the proportion of x2 Infra light emission in the green region of the visible light spectrum, allowing for a larger proportion of the total bioluminescence yield to originate from the far-red to near-infrared region of light, which falls better in the bio-optical window of 600-800 nm for *in vivo* imaging of mammalian tissues and would allow for better penetration of signal through haemoglobinised tissues (Rice and Contag 2009; Iwano *et al.* 2013).

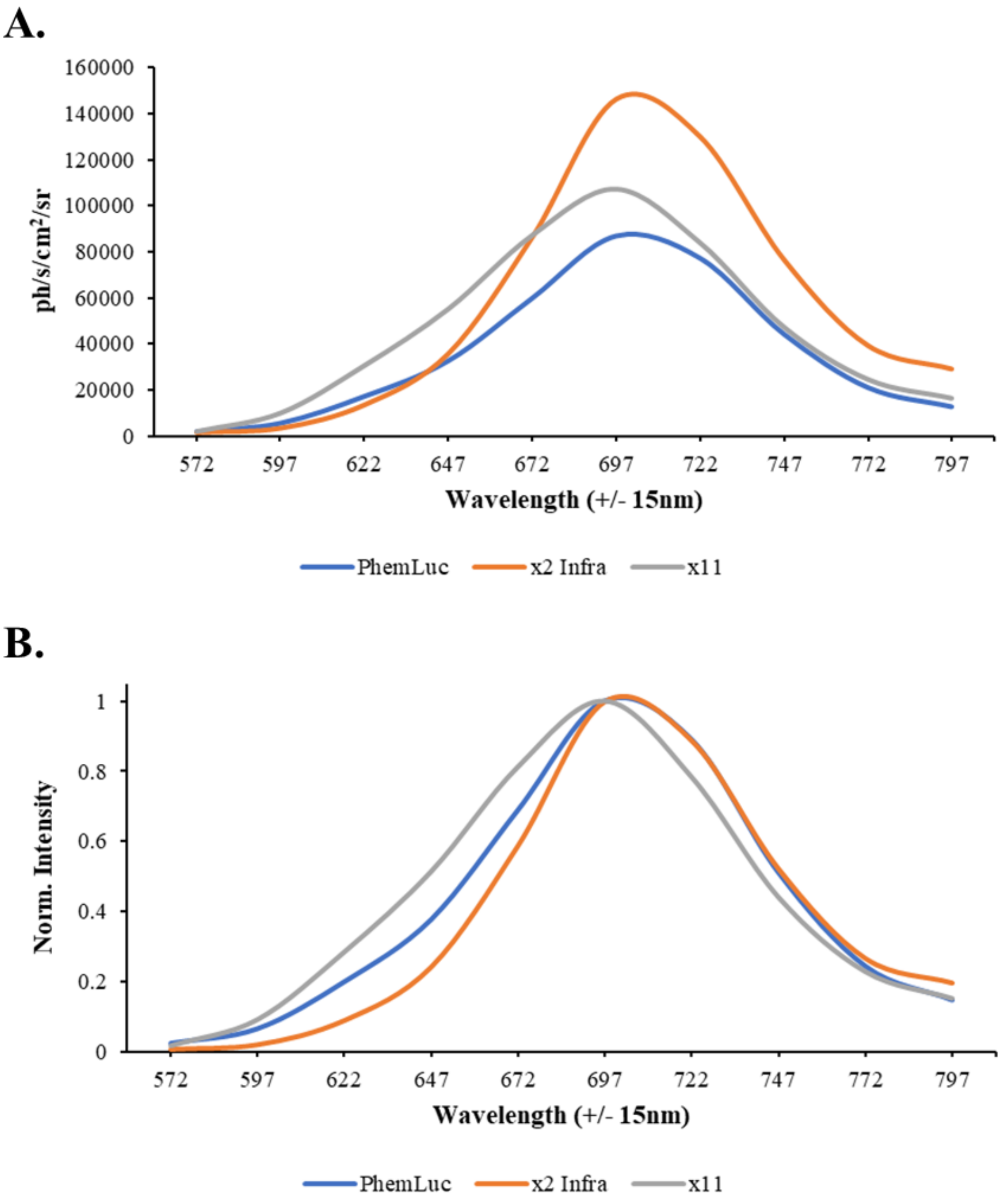
4.3.5.2 Bioluminescence activity of x2 Infra with iLH$_2$

The iLH$_2$ bioluminescence reactions recorded in the PhotonIMAGER were initiated by manual pipetting, and therefore flash kinetics could not be obtained, as was done in Chapter 6 using the automatic injection function of the BMG CLARIOstar. Instead, the total bioluminescence signal recorded across the spectra measurements (equivalent to the area under the spectra curves) were recorded, along with two 60 second measurements of specific activity, the first of which was taken 60 seconds after manual pipetting of the substrate mix onto each enzyme, prior to spectra acquisition, and the second immediately following the acquisition of the last band-pass filter measurement of the spectra.

The total bioluminescence yield from spectra acquisition is shown in Figure 4.16A. Over these consecutive measurements, x2 Infra was observed to produce a significantly increased (P=0.024) bioluminescence yield of 54% over its *wild-type* counterpart *Phem*Luc. Over this same period, it was also shown to produce an emission yield 21% greater than that of *Ppy* Fluc x11, the original enzyme pairing in iLH$_2$ development. However, this measurement was determined to lack statistical significance (P=0.0574).

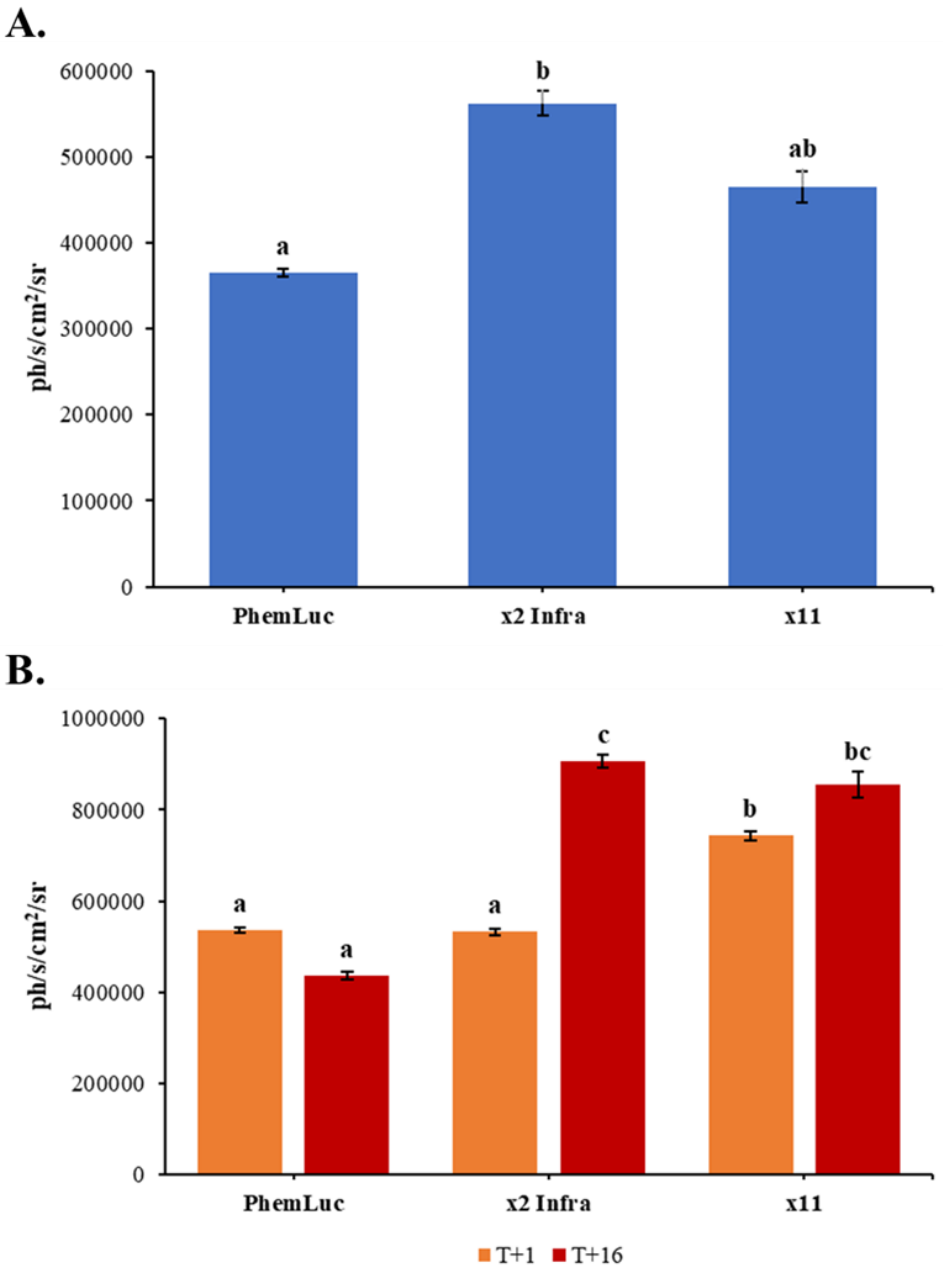
The specific activities recorded before and after spectra acquisition in Figure 4.16B suggest that there is <1% difference in the initial emission of x2 Infra and *Phem*Luc (P>0.9999), but whilst the *Phem*Luc activity decayed by 23% over the acquisition of spectra, x2 Infra bioluminescence specific activity increases significantly, by 70%. At the time these measurements were taken, ca. 16 minutes from reaction initiation, x2 Infra was shown to possess a bioluminescence activity 108% greater than that of its *wild-type Phem*Luc (P<0.0001). The *Ppy* Fluc mutant x11 was shown to exhibit bioluminescence activity 40% greater than x2 Infra prior to the acquisition of spectra (P=0.0005). Similar to x2 Infra, the bioluminescence activity of x11 increased 15% by the second measurement taken after spectra acquisition, but due to its 40% increase in activity, x2 Infra was found to be 6% more active than x11 ca. 16 minutes after reaction initiation (P=0.6734). The measurements taken before and after the spectra acquisition indicate that the bioluminescence emission of *wild-type Phem*Luc begin to diminish soon after the reaction initiation, whereas x2 Infra and x11 continue to increase in bioluminescence activity over this duration, with the gradient of the kinetic profile being greater for x2 Infra. What cannot be seen from the available data is the time at which the peak of bioluminescence emission occurs for each enzyme, or whether a steady state of emission is achieved, and for how long.

Figure 4.15.



Bioluminescence spectra of *Phem*Luc, x2 Infra, and x11 with iLH$_2$. Measurements obtained in the PhotonIMAGER Optima by manual pipetting of substrate mix into each enzyme such that final assay concentrations were equal to 1 mM ATP, 500 µM iLH$_2$, and 0.167 µM protein. Each reaction constituent was previously diluted in chilled TEM buffer at pH 7.8 (±0.05) and total reaction volume was equal to 150 µl. Following substrate injection, each reaction was held at RT for 60 seconds prior to acquisition of total bioluminescence yield (see Figure 4.16.), immediately followed by spectra. Light emissions integrated over 60 seconds for 14 band pass filters, with a midpoint range between 472 nm and 800 nm and a step width of 25 nm (in respect to the midpoint). Assays performed in triplicate, and averaged data are presented. The lower graph represents the same data normalised such that each point is presented as intensity relative to λ_{max}.

Figure 4.16.



Bioluminescence yield of *Phem*Luc, x2 Infra, and x11 with iLH$_2$. Measurements obtained in the PhotonIMAGER Optima by manual pipetting of substrate mix into each enzyme such that final assay concentrations were equal to 1 mM ATP, 500 µM iLH$_2$, and 0.167 µM protein. Each reaction constituent was previously diluted in chilled TEM buffer at pH 7.8 (±0.05) and total reaction volume was equal to 150 µl. Following substrate injection, each reaction held at RT for 60 seconds prior to acquisition of total bioluminescence yield (T+1), immediately followed by spectra (see Figure 4.15.). (**A**) Bar plot representing the total bioluminescence yield for each enzyme recorded over the duration of spectra acquisition. (**B**) Bar plot displaying the total light emissions integrated over 60 seconds before (T+1) and after (T+16) spectra acquisition. T+1 and T+16 indicate the time in minutes that measurements were obtained relative to the initial substrate injection. Assays performed in triplicate, and averaged data are presented. Error bars represent the standard error of the mean. Statistical analysis performed as detailed in 2.12. Significance groupings are discrete between **A** and **B**.

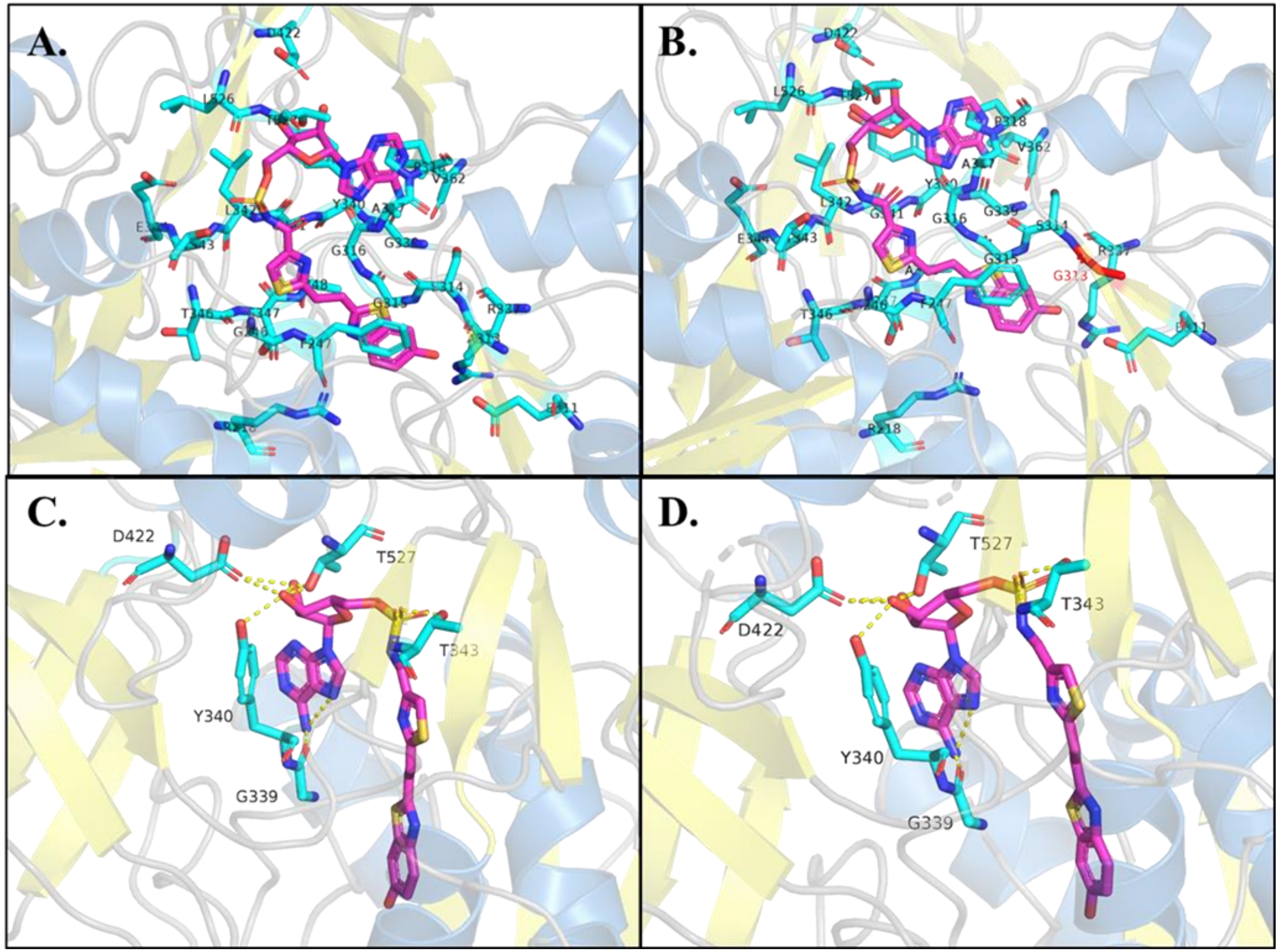
4.3.6. Modelling x2 Infra

A final homology model of x2 Infra was produced following the procedures used to generate the luciferase substrate contact models in Section 4.3.3. The x2 Infra model was identically constructed using the available crystal structure from PDB file 6HPS.pdb of *Ppy* Fluc bound to iDLSA in the adenylate-forming conformation (Stowe *et al.* 2019). The resulting model was then analysed against the original *Phem*Luc iDLSA model from Figure 4.8 in The PyMOL Molecular Graphics System, to identify changes in amino acid residues measured within 4 Å of bound iDLSA (Figure 4.17.). An additional model indicating the positions of x2 Infra mutations in *Phem*Luc is available in Appendices Figure 9.5. In the *Phem*Luc model of Figure 4.17., A313 was originally identified within 4 Å of the bound iDLSA. However, in x2 Infra this position was mutagenized to G313, which is no longer within 4 Å of iDLSA in the x2 Infra model (Figure 4.17A). The other 23 residues identified in PhemLuc were conserved (Table 4.3.). The polar contacts of x2 Infra and iDLSA identified in Figure 4.17 were compared with those identified in *Phem*Luc in Table 4.5. However, this analysis indicated no difference in interactions between the two models, and therefore how the spatial reorganisation of this mutagenized position improves the bioluminescence yield from *Phem*Luc remains unknown. Additionally, it remains unclear how A313G and H245W produce an additive effect on bioluminescence yield which is greater than their individual effects seen in Section 4.3.4.

Although H245W was one of two mutations that were ultimately selected for inclusion in the final x2 Infra, the position H245 was originally identified only in the *Ppy* Fluc models (Table 4.3.). The W245 residue of x2 Infra could not be identified within 4 Å of the bound iDLSA in Figure 4.17A, similarly to the omission of H245 in the *Phem*Luc iDLSA model (Figure 4.17B). It remains unclear how H245W improves the bioluminescence activity of x2 Infra and whether it has a role which augments the primary adenylation of the reaction.

Understanding of the mechanisms through which H245W and A313G improve the bioluminescence activity from *Phem*Luc is further restricted due to the lack of availability of a luciferase crystal structure bound to an iLH_2 analogue in the oxidation-forming conformation of the reaction. The ability to generate such homology models may have provided the insight into the roles of these mutations which is lacking from the adenylation-forming models.

Figure 4.17.


A.
D422
L526
V362
Y340
A317
L342
G316
G339
R337
T346
G246
F247
E311
B.
D422
L526
T527
P318
V362
A317
Y340
L342
G316
G339
S314
E344
S343
G315
R337
G313
T346
F247
E311
R218
C.
D422
T527
T343
Y340
G339
D.
T527
T343
D422
Y340
G339

Modelling of *Phem*Luc and x2 Infra residues within 4 Å of bound iLH$_2$ analogue. x2 Infra (**A**) and *Phem*Luc (**B**) amino acid residues within 4 Å of iDLSA. The position G313 is indicated in red in **B** as its presence is only identified in the *Phem*Luc model **A** (as A313). Polar interactions between *Phem*Luc (**C**) and x2 Infra (**D**) with iDLSA are indicated by yellow dashed lines and the interacting residues. Interactions are detailed in Table 4.5. Model colouring is as described in Figures 4.5-8. Model analysis and imaging performed in PyMOL.

Table 4.5

Residue	**PhemLuc iDLSA**			**x2 Infra iDLSA**		
	Distance (Å)	**Sidechain Interaction Centre**	**Interacting iDLSA Atom**	**Distance (Å)**	**Sidechain Interaction Centre**	**Interacting iDLSA Atom**
G339	1.2	O	N6	1.2	O	N6
Y340	2.6	OH1	O3'	2.6	OH	O3'
T343	2.4	N	OBJ	2.4	N	OBJ
T343	2.6	OG1	OBJ	2.6	OG1	OBJ
D422	2.9	OD1	O2'	2.9	OD1	O2'
D422	3	OD1	O3'	3	OD1	O3'
T527	1.5	OG1	O2'	1.5	OG1	O2'
T527	2.1	OG1	O3'	2.1	OG1	O3'

Polar contacts with iDLSA identified across *Phem*Luc and x2 Infra. All polar contacts identified across the protein models of Figures 4.8C and 4.17C and the distance in angstroms of the interactions. The sidechain interaction centres as defined by Bahar and Jernigan, 1996 and the interacting iDLSA atom are detailed.

4.4. Further Discussion

This work was undertaken to investigate whether the bioluminescence activity of *Phem*Luc with iLH$_2$ could be altered by targeted mutagenesis of amino acid residues hypothesised to have either direct or indirect influence on the conformation of the LH$_2$ specific active site, and introduce improved compatibility with the larger structure of the synthetic substrate analogue, iLH$_2$. For this purpose, homology models of the *Phem*Luc enzyme in adenylated form were constructed utilising available crystal structures of *Ppy* Fluc bound to LH$_2$ and iLH$_2$ structural analogues. These models were then used to identify all amino acid positions within 4 Å of the respective bound ligand. The identified positions were mutagenized using the targeted approach of semi-random site directed mutagenesis (SDM) with NNK/MNN codon primers, which enabled the substitution and sampling of all 20 natural amino acids at every target position. This targeted approach toward residues in proximity to the bound ligand was selected over random mutagenic methods based on the hypothesis that the diminished bioluminescence activity with iLH$_2$ is in part influenced by an inability of either the iLH$_2$ or ATP to position correctly within the luciferase active site due to the extended structure of iLH$_2$ relative to the naturally occurring LH$_2$, for which all beetle luciferases have evolutionarily developed a compatible active site conformation.

Primary structure analysis of *Phem*Luc relative to *Ppy* Fluc indicated 87.07% shared identity at the amino acid level. The protein sequence was further compared with 12 additional firefly luciferases and discovered to contain only 12 positions of unique amino acid identity relative to the aligned majority consensus. Of these 12 residues, only 2 positions were of unique polarity, at positions 170 and 424. Computational predictive secondary structural analysis of *Phem*Luc further confirmed a significant conservation to *Ppy* Fluc. The limited regions of variation from the firefly luciferase consensus along with the considerable shared identity with *Ppy* Fluc were taken as confirmation of sufficient compatibility for the purposes of producing reliable protein homology models.

Two homology models of *Phem*Luc were constructed using available crystal structures for *Ppy* Fluc, in complex with DLSA and iDLSA. These models and the existing *Ppy* Fluc structures were used to identify all amino acids positioned within 4 Å of the bound DLSA or iDLSA. Whilst these models were able to identify the active site residues of luciferases involved in the first adenylation conformation of the LH$_2$-AMP complex detailed by Sundlov *et al* (2012), the active site residues unique to the second oxidation conformation which follows a 140° domain rotation could not be identified. The availability of an iDLSA crystal structure only in the adenylation conformation limits the mutagenic targets to only those involved in the first stage of catalysis. Positions uniquely involved in the secondary catalysis stage of oxidation, following the C-domains rotation, could not be identified and therefore not targeted to improve the iLH_2 compatibility.

SDM libraries were created for the targeted positions using NNK/MNN codon primers to allow for the substitution of 32 equiprobable sequence variants which permitted the inclusion of all 20 natural amino acids, without premature stop codons, and therefore fewer deleterious mutants in each library. However, this approach was not without drawbacks due to the codon redundancy of the genetic code, which would lead to the over-representation of amino acids encoded by multiple codons relative to those with fewer. To mitigate this, the minimum library size required to provide a 95% chance in screening all sequence variants was calculated, and the guidance of these results strictly adhered to.

The luciferase screening method used by Wood and Deluca (1987) was adapted to allow for screening of bioluminescent activity with iLH_2 in the PhotonIMAGER Optima. Earlier studies have shown that LH_2 does not readily pass through prokaryotic or eukaryotic cell membranes at physiological pH, but the efficiency can be significantly increased under the slight acidic pH conditions of sodium citrate buffer (Wood and DeLuca 1987; Jawhara and Mordon 2004). Whilst this adapted method was sufficient for the comparative screening needs of this study, variation between rounds of screening for any given isolated colony was frequent. Whilst this variation could be attributed to a simple cause such as uniformity of the iLH_2 spray delivery method from the plate periphery to centre, more advanced investigation of the iLH_2 delivery and applicability of sodium citrate in the case of iLH_2 should be made, but are outside of the scope of this study due to the limited available supply of $iLH_{2.}$

Of the 32 SDM libraries constructed and screened, many produced a significant proportion of *E. coli* colonies of diminished bioluminescence activity relative to *Phem*Luc. Sequencing of the highest activity colonies from such plates revealed that the original *Phem*Luc sequence had been selected for out of the 20 possible amino acid substitutions. This result was predicted prior to conducting any screens, due to the high conservation scores that many of the residues directly involved in or proximal to the active site exhibit.

A subgroup of the SDM libraries were found to produce beneficial mutations, and 6 tentative advantageous mutants were isolated for further screening. Inconsistency in the recorded bioluminescence amongst the final subgroup was apparent throughout the final screens. The mutant H245W appeared to confer the greatest improvement to bioluminescence activity in the primary screen of fresh *E. coli* transformants, but this effect diminished in subsequent screens, indicating that A313G was the most consistent advantageous mutation across successive screening rounds.

To explore the possibility of cumulative effects amongst the final subgroup, two combinatorial mutants were created, x6 Infra which contained all 6 final mutants, and a dual mutant of H245W and A313G, x2 Infra. The combination of all 6 mutations in x6 Infra was significantly deleterious to the enzyme bioluminescence activity. Whilst this inhibition must relate to an incompatibility of two or more mutations, a likely cause may be the inclusion of several adjacent mutations E311S, A313G, and S314V, which could together be drastically remodelling the conformation of the local region of active site and rendering the enzyme activity with iLH_2 inert. The x2 Infra mutant displayed a bioluminescence activity greater than its individual derivatives, which was consistent across multiple screens. A348V displayed improved activity relative to *Phem*Luc in the primary screens, but this effect diminished in subsequent rounds, similarly to H245W which ultimately went on to enhance the effect of A313G. Based on the synergistic action of x2 Infra, and speculation of the x6 Infra inhibition relating to over inclusion of adjacent mutations, the A348V mutations may be useful to explore in a future x3 Infra, due to its distal position.

The improved activity of x2 Infra with iLH_2 relative to *Phem*Luc was confirmed by protein assays which measured the specific activity and emission spectra with iLH_2. Although no significant changes to spectra could be observed with the resolution available, x2 Infra was shown to produce more than double the bioluminescence signal of *Phem*Luc, 16 minutes after initiation of the reaction. By this stage *Phem*Luc had exhibited a kinetic profile of declining activity from the initial measurement toward the start of the reaction, whereas the activity of x2 Infra displayed a relative increase. Following the kinetic profiles for both enzymes over an extended imaging window would be useful in order to understand the emission kinetics of where peak activity occurs and how long stable emission is maintained, for the purposes of *in vivo* imaging. Unfortunately, such a study could not be performed here with the supply of iLH_2 available.

Revisiting the models used to generate the SDM libraries indicates that substitution of A313G extends the distance of this position to beyond the 4 Å range of the bound iDLSA.

The relocation of this position may be producing greater bioluminescence activity with iLH_2 by creating a larger conformation of the active site which is capable of accommodating the large structure of iLH_2 and therefore allowing its positioning to be improved, along with that of ATP. How H245W is able to enhance this effect in x2 Infra remains unclear from the available models which only capture the primary adenylation conformation with iDLSA. The existence of a crystal structure in complex with an iLH_2 analogue in the secondary catalysis step of oxidation might have revealed how H245W is functioning, and whether there is an additional role of A313G that further benefits catalysis following the domain rotation.

4.5. Conclusions

Homology modelling and targeting of amino acid positions in closest proximity to bound ligand was a useful strategy for improving compatibility of *Phem*Luc with iLH_2 which could further be deployed as a targeted approach for engineering compatibility with additional emerging synthetic substrate variants. The mutations H245W and A313G were both found to improve the bioluminescence activity of *Phem*Luc with iLH_2, with an additive effect that exceeded the activity of either mutation independently. Further investigation of how these two mutations augment iLH_2 catalysis is required to understand their function and whether this action is conserved for mutations at corresponding positions in homologous luciferases, such as *Ppy* Fluc. The future existence of a luciferase crystal structure in complex with an iLH_2 analogue in the secondary catalysis step of oxidation could possibly provide an understanding of the roles of H245W and A313G, whilst additionally presenting further positional targets for mutagenesis.

Chapter 5

Thermostability Engineering by Directed Evolution

5.1. Chapter Summary

In this chapter, work was undertaken to explore whether the novel luciferase enzyme from the lesser British Glow-worm *Phosphaenus hemipterus* (*Phem*Luc) could be engineered to produce variants with higher resistance to thermal inactivation based on known mutations from homologous luciferase enzymes, previously engineered in the lab. It was hypothesised that these could subsequently be refined for a higher activity phenotype using DNA shuffling prior to further enhancement through directed evolution. To achieve this, fifteen mutations known to confer thermostability to the homologous firefly luciferase from *Photinus pyralis* (*Ppy* Fluc) were incorporated into *Phem*Luc. Whilst this initial x15 mutant displayed poor bioluminescence activity under all conditions assessed, its construction enabled the generation of an x14 revertant mutant through DNA shuffling of the x15 mutant with *Phem*Luc. The simplified x14 enzyme presented a greater resistance to thermal inactivation than *wild-type Phem*Luc. A final x16 mutant integrated two independent mutations discovered through mutagenesis and subsequent screening to produce an enzyme which possessed sufficient resistance to thermal inactivation to be deployed in a LAMP-BART assay using SARS-CoV-2 RNA as template.

5.2. Introduction

Thermostability is the degree to which a material or substance can resist irreversible change in structural and functional properties from exposure to excessive conditions of temperature. Exposure of proteins to thermal energy which exceeds their native temperature in the originating organism often produces a cascade of unfolding and denaturation through the disruption of the intramolecular bonds comprising the tertiary structure, which results in the loss of the enzymatic activity. Therefore, the thermostability of a protein can be considered as the resistance to this process of molecular degradation, or more simply the resistance to thermal inactivation. Whilst some proteins naturally possess a high degree of thermostability, beetle luciferases are highly thermolabile, and may even inactivate from exposure to room temperature conditions (Prebble *et al.* 2001; Tisi *et al.* 2002b; Law *et al.* 2006). This core limitation prohibits the use of *wild-type* luciferases for *in vivo* applications of medical imaging where stability at 37 ˚C is required (Baggett *et al.* 2004; Zambito *et al.* 2021), and *in vitro* applications such as the detection of DNA amplification which requires luciferases capable of resisting thermal inactivation at temperature exceeding 60 ˚C (Gandelman *et al.* 2007; Gandelman *et al.* 2010).

Directed evolution and the various methods of mutagenesis it comprises are long established processes for the development of luciferase variants possessing improved thermostability, and have been successfully used to increase the resistance to thermal inactivation of luciferases from varied firefly species (Hall *et al.* 1999; Tisi *et al.* 2002a; Kitayama *et al.* 2003; Koksharov and Ugarova 2011a; Mortazavi and Hosseinkhani 2011; Koksharov and Ugarova 2012). The continued search for novel mutations in multiple luciferases from diverse Coleopteran species has been key to identifying collections of thermostabilising mutations, a number of which are conserved for their thermostabilising effects across enzymes. For example, the single point mutation A217 (or its equivalent position) has been demonstrated to retain its thermostability enhancing properties when transposed between firefly species (Kajiyama and Nakano 1993; Kajiyama and Nakano 1994; Branchini *et al.* 2007). However, not all mutations have been found to conserve their advantageous actions when incorporated within other homologous luciferases (Kitayama *et al.* 2003; Koksharov and Ugarova 2011b).

The most thermostable luciferase developed is Ultra-Glo, which is an engineered variant developed from *Photuris pensylvanica* and is commercially available for a number of assays (Hall *et al.* 1999; Hsiao *et al.* 2016). A key limitation of Ultra-Glo arises due to its patent to protect commercial interests, which prevents its availability as a genetic construct that could be introduced into living organisms for the purposes of bioluminescence imaging (BLI) (Sun *et al.* 2012). Also, Ultra-Glo has relatively low activity with luciferin and amino-luciferin compared to *Ppy*-derivatives (Jathoul *et al.* 2012). Due to these limits, the continued discovery of novel mutations and development of thermostable luciferases is crucial to expand the toolbox of enzymes available for utilization within BLI. To further this purpose, the work described in this chapter sought to develop engineered variants of a novel luciferase from the lesser British Glow-worm *Phosphaenus hemipterus* (*Phem*Luc) which shares

87.07% amino acid identity with the *Ppy* Fluc by first incorporating known thermostabilising mutations from homologous luciferases to use as a foundation from which to discover novel mutations by directed evolution which are capable of conferring further thermostabilising effects.

5.3. Results and Discussion

5.3.1. Thermostabilising *Phem*Luc with mutations from thermostable derivatives of *Ppy* Fluc

5.3.1.1 Identifying *Ppy* Fluc thermostability mutations within *Phem*Luc

It is well established that single mutations altering either amino acids on the protein surface or within the enzyme core of firefly luciferases can improve the overall resistance to thermal inactivation, and that the inclusion of multiple such mutations can have additive effects (Tisi *et al.* 2002b; Law *et al.* 2006). A selection of mutations across the protein surface and core known to confer improved resistance to thermal inactivation in *Ppy* Fluc were identified. An in depth analysis of *Phem*Luc in comparison to *Ppy* Fluc was made in Chapter 4 across primary to tertiary structure. An alignment of the two protein sequences was revisited here in Figure 4.2. for the purpose of identifying the corresponding positions in *Phem*Luc of the mutations shown to confer increased thermal stability to *Ppy* Fluc in previous studies. These mutations are listed in Table 5.1. with their positions in *Ppy* Fluc and the respective *Phem*Luc positions as derived from the alignment. A total of 14 *Ppy* Fluc thermal stability mutations were selected from literature, with the addition of a 15th mutation (S347G) upon the recommendation of Dr. Amit Jathoul, which is an active site mutation hypothesized to interact with the synthetic substrate analogue infraluciferin. S347G was initially included to explore the possibility of generating a dual-function mutant of improved thermostability and activity with infraluciferin, although this approach was discontinued and a separate rational was explored for infraluciferin activity engineering (see Chapter 4).

Unsurprisingly, as *Phem*Luc and *Ppy* Fluc share 87.07% amino acid identity, the majority of the 15 thermal stability mutation positions were conserved for the *wild-type* amino acid in both sequences. However, the 3 positions V182K, T214C, and A215L from *Ppy* Fluc are not fully conserved, and were instead identified as I182K, S214C, and V215L in *Phem*Luc. Notably, all 3 of these positions maintain similar amino acids between *Ppy* Fluc and *Phem*Luc considering amino acid polarity and charge. The positions T/S214C and A/V215L maintain their polarity and charge status following substitution mutagenesis to thermostabilising equivalents from *Ppy* Fluc studies, whilst A/V215L additionally provides a slight reduction in hydrophobicity on the surface. V/I182K substitutes the previously non-polar position with the basic polar group lysine, which confers an increased polarity and positive charge to the previously hydrophobic surface-exposed residue (Law *et al.* 2002). So whilst, these positions may not be directly conserved between the two proteins, with some mutations (A/V215L and V/I182K) the underlying mechanism of increasing surface polarity for the function of the thermostabilising mutations may apply, as for those identified in the development of x5 Fluc (Law *et al.* 2006).

5.3.1.2 Mutations from the x11 enzyme

Many of the mutations listed in Table 5.1 are present in the x11 mutant of *Ppy* Fluc which was constructed by combining a selection of thermal stability mutations both novel and previously reported in available literature. x11 has been demonstrated to possess significant pH tolerance and resistance to thermal inactivation, but still possesses less stability though considerably higher activity than has been shown for the commercially available Ultra-Glo (Jathoul *et al.* 2012). The advantage of a non-commercialised mutant such as x11 is in the published availability of its amino acid composition, which allows for its availability as a genetic construct that can be utilized in application such as a reporter gene for *in vivo* bioluminescence imaging (Jathoul *et al.* 2012; Stowe *et al.* 2019). x11 was originally designed as a x12 mutant, which included the addition of F295L, which can be found in the mutations selected in Table 5.1. It was later reverted for F295 to produce a simplified mutant which retained similar properties of resistance to thermal inactivation. Regardless of the reversion in x11, F295L was still included for assessment in *Phem*Luc. A model indicating the positions of x11 mutations in *Ppy* Fluc is available in Figure 1.7.

5.3.1.3 Generation and screening of an x15 *Phem*Luc mutant

An x15 mutant of *Phem*Luc was designed and synthesized to investigate whether the 15 mutations known to enhance the thermal stability of *Ppy* Fluc listed in Table 5.1 could produce similar improvements to the performance of *Phem*Luc under elevated conditions of temperature. As discussed in Chapter 4 (4.3.4.2), in any work to augment the activity of a given protein through engineering, a method of screening is required in order to identify mutants with advantageous phenotypes of the desired characteristic. For the purpose of identifying variants of *Phem*Luc with greater resistance to thermal inactivation, an adaptation was made to the original luciferase bioluminescence screening strategy in *E. coli* of Wood and Deluca (1987) to incorporate a 1-hour incubation of the colonies at 50 ˚C prior to screening the remaining bioluminescent activity with LH_2.

For an initial assessment of whether x15 retained bioluminescent activity, the synthesized gene was incorporated into the pET16b plasmid and transformed into *E.coli* BL21 (DE3), and the transformed colonies were induced for production of the x15 enzyme with IPTG. Screening of bioluminescence activity was subsequently performed with an LH_2-citrate spray in the PhotonIMAGER Optima (Biospace Labs, Paris, France) at room temperature, as further explained in Chapter 2. An identical screening process was conducted for the *wild-type Phem*Luc. The bioluminescence data from both screens was analysed in the M3 Vision software package, available under license from https://biospacelab.com. The bioluminescence signal as interpreted in M3 Vision is displayed in Figure 5.1, where lower bioluminescent activity is indicated by the look-up-table (LUT) colour of blue, which contrasts to red for the indication of greater bioluminescent signals. The primary screen of *Phem*Luc and x15 are scaled to be directly comparable and are labelled as A. and B., respectively. Room temperature screening of both enzymes revealed that the incorporation of 15 discrete mutations into *Phem*Luc had significant deleterious effects on the bioluminescent activity. This result was verified in a secondary screen (Figure 5.1C.).

Although the bioluminescent activity of x15 was significantly less than the *wild-type Phem*Luc, the activity was further investigated with the proposed method of screening for resistance to thermal inactivation by the inclusion of a 1-hour incubation at 50 ˚C prior to measuring the remaining bioluminescent activity with LH_2. The data from the RT and 50 ˚C screens for both enzymes are displayed in Figure 5.2. At RT x15 is shown to produce a bioluminescence yield equal to 1.42% the activity of *Phem*Luc under the same conditions. Remaining activity of x15 increases to 4.53% of *Phem*Luc remaining activity following incubation of both at 50 ˚C. This suggests that one or more of the mutations incorporated within x15 have a preserving effect on the original activity under conditions of elevated temperature, but is still not sufficient to accomplish the intention of generating an enzyme capable of producing a greater bioluminescent yield than *Phem*Luc under elevated temperature conditions.

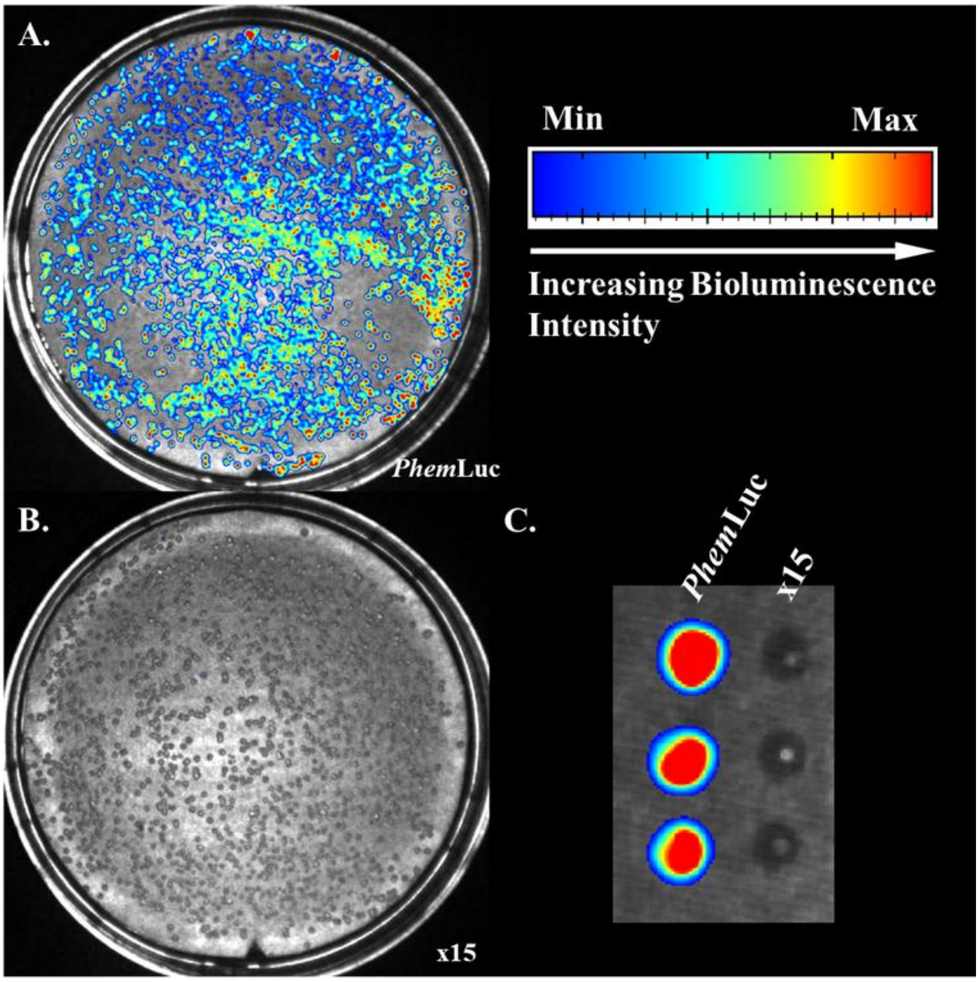
As the activity of x15 is considerably lower than *Phem*Luc at RT, it is likely that one or more of the 15 mutations is significantly deleterious to the enzyme bioluminescent activity under these conditions. In order to restore bioluminescence activity before the continued development of resistance to thermal inactivation, reversion of at least one of the mutations to the *wild-type* sequence may be required, as demonstrated in the development of x11 Fluc (Jathoul *et al.* 2012).

Table 5.1.

***Ppy* mutations**	**Corresponding *Phem*Luc mutations**	**Authors**
F14R	F14R	(Law *et al.* 2006)
L35Q	L35Q	(Law *et al.* 2006)
A105V	A105V	(Jathoul *et al.* 2012)
V182K	I182K	(Law *et al.* 2006)
T214C	S214C	(Prebble *et al.* 2001)
A215L	V215L	(Kitayama *et al.* 2003)
I232K	I232K	(Law *et al.* 2006)
D234G	D234G	(Tisi *et al.* 2002b)
E270K	E270K	(Prebble *et al.* 2001)
F295L	F295L	(Prebble *et al.* 2001)
S347G	S347G	-
E354R	E354R	(White *et al.* 1996)
D357Y	D357Y	(White *et al.* 2002)
S420T	S420T	(Prebble *et al.* 2001)
F465R	F465R	(Law *et al.* 2006)

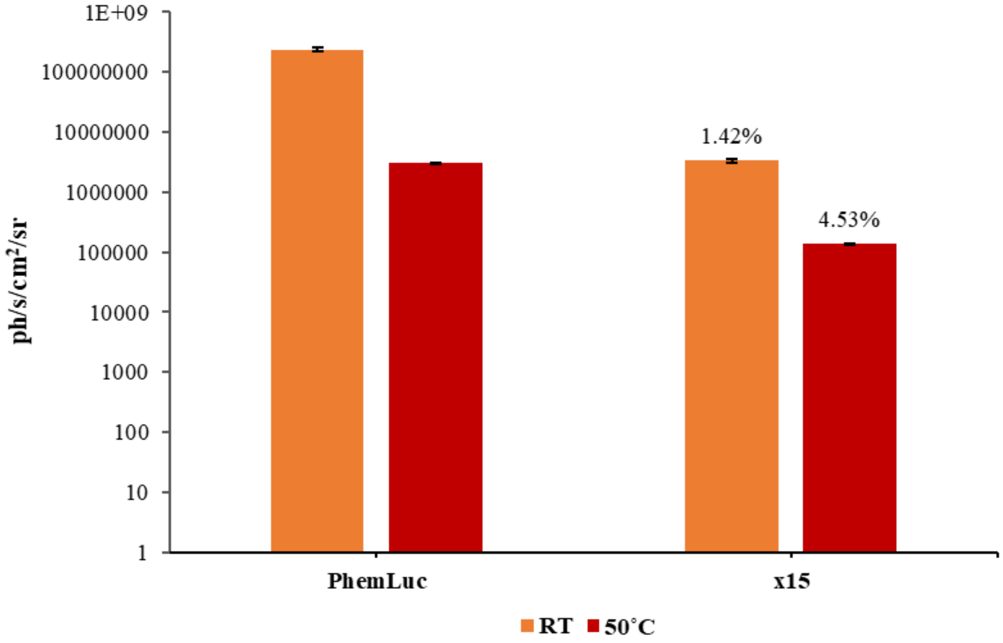
Mutation selected to construct thermostable *Phem*Luc. Primary thermostability mutations incorporated into *Phem*Luc from previous studies in *Ppy*. The inclusion of S347G was suggested from discussions with Dr Amit Jathoul for a discontinued investigation into dual-function thermostabilisation and improved activity with infraluciferin, and has no corresponding literature. Mutations shown in bold for *Ppy* Fluc are present in x11. The three mutations not conserved in *Phem*Luc are shown in red.

Figure 5.1.



Primary screening at RT of x15 against *Phem*Luc. The colony formations from *E. coli* BL21 (DE3) transformed with pET16b plasmid containing x15 or *Phem*Luc, were transferred onto nitrocellulose membranes (Chapter 2). Membranes carrying transferred colonies were induced for 3-4 hours at RT with IPTG (1 mM). All plates were subsequently screened with 500 µM LH_2 and imaged in the PhotonIMAGER Optima (Biospace Labs, Paris, France), over an integrating period of 20 seconds. The upper right colour spectrum illustrates how bioluminescence signal intensity is represented relative to the maximum signal detected, with minimum intensity indicated in blue and maximum intensity indicated in red. (**A**) Bioluminescence activity from control transformation of *wild-type Phem*Luc. (**B**) Bioluminescence activity from transformation of speculative thermostable construct x15. (**C**) Secondary screen of bioluminescence activity from *Phem*Luc against x15. The three images were acquired independently, but the colour intensity representation between **A** and **B** has been scaled to be directly comparable in the M3 Vision software package. Intensity scaling of **C** is not directly comparable.

Figure 5.2.



Log bioluminescence activity of *Phem*Luc and x15 at RT and 50 ˚C. Bar chart displaying the averaged bioluminescence activity measured across 3 colonies of *Phem*Luc and x15 in the secondary screen (Figure 5.1C.). Both plates were induced and screened identically, with the inclusion of a 60 minute incubation at 50 ˚C immediately prior to screening for the 50 ˚C plate. Percentages above the x15 bars indicate the activity as a function of *Phem*Luc activity under the same condition. Error bars represent the standard error of the mean (SEM). Figure produced from a subset of data from Figure 5.9. Statistical analysis can be found in Figure 5.9.

5.3.2. Restoring bioluminescent activity of x15 by reversion

Bioluminescence activity of *Phem*Luc was significantly diminished following the simultaneous incorporation of all mutations from Table 5.1, indicating that at least one of the mutations is incompatible with the *Phem*Luc enzyme or a deleterious interaction occurs between two or more mutations that would not arise as independent mutations. Reverting the enzyme to fifteen discrete x14 enzymes by site directed mutagenesis (SDM) would be an effective approach for the discovery and removal of a single deleterious mutation. However, this strategy would be incapable of restoring the enzyme bioluminescent activity if the deleterious effect was due to two or more mutations.

In order to restore bioluminescent activity, multiple subsets of the mutations would need to be generated and screened in order to identify a novel combinatorial mutant that could be taken forward for further engineering of thermostability. Creating such revertant mutants with a systematic approach such as with SDM would be possible, but the generation of randomized subsets through DNA shuffling was explored as a more efficient strategy for the removal of unknown deleterious mutations.

5.3.2.1 DNA shuffling of *Phem*Luc with x15

DNA shuffling as first demonstrated by Stemmer (1994) is a process of molecular evolution by random fragmentation and reassembly of two or more highly homologous gene sequences (Stemmer 1994). The original principle was inspired by homologous recombination during meiosis and allows for artificially accelerated evolution of target genes towards a desired functionality through the randomised recombination of variant sequences to remove deleterious mutations, and combine the advantageous. The source of the sequence variation can be of natural origin, such as in the shuffling of the collective variation from a gene family, or artificially generated variation from error-prone PCR or most commonly oligonucleotide-directed mutagenesis (Zhao and Arnold 1997; Crameri *et al.* 1998; Meyer *et al.* 2015).

The use of a DNA shuffling strategy would not be without complication, as the shuffling of 15 discrete mutations gives rise to an issue of scale. Equation 5.1. estimates the required shuffle library size of transformed colonies to have a 95% chance of being 100% complete, where complete here means representing each unique combination and subset of the 15 mutations at least once (Patrick *et al.* 2003). The unique combinations of mutants doubles with the inclusion of each additional mutant, so whereas a single mutation only allows for two sequence variants (mutation is either present or absent), the total number of unique sequence variants from 15 mutations would be 32786. This number of colonies alone would be unfeasible for the screening strategy adopted in this work, but in order to have a 95% confidence in screening all sequence variants as described by Equation 5.1, the number of transformed colonies increases an order of magnitude, totalling 438282.

Fortunately, the screening of this number of transformed colonies would only be necessitated by the goal of identifying the highest activity subset of mutations available from the 15 mutations originally included. However, the aim of shuffling x15 back with *Phem*Luc is only to identify at minimum one revertant that displays an improved resistance to thermal inactivation than the *wild-type Phem*Luc. Without the screening of each mutation individually, it is impossible to determine how many preserve their thermostabilising effect from *Ppy* Fluc when incorporated into *Phem*Luc compared with how many are neutral or deleterious. However, thermostabilising mutations from the Japanese firefly *Luciola cruciata* have previously been shown to produce similar thermostabilising effects in *Ppy* Fluc when comparative mutations are made (Tisi *et al.* 2002a). As *Ppy* Fluc and *Phem*Luc share significant sequence identity (87.07% amino acid identity), it is likely that many of the mutations from Table 5.1 are similarly capable of thermostabilising effects in both enzymes. Similar to the exponential growth in the size of library to be screened for completeness, the total number of sequence variants that should display thermostabilising properties doubles with the inclusion of each advantageous mutation. The identification of any one of these through DNA shuffling would yield a mutant sufficient to proceed with further engineering efforts.

The DNA shuffling process utilised was adapted from the original Stemmer (1994) method using the improvement of *Pfu* polymerase as suggested by Zhao and Arnold (1997), and optimised for utilisation with luciferase gene sequences. The stages of shuffling as visualised through gel electrophoresis are shown in Figure 5.3; specific details are available in Chapter 2. The first stage of the DNA shuffling process is the amplification of the genes to be recombined with a high fidelity polymerase, as shown for *Phem*Luc in Figure 5.3A. The purified genes are then randomly fragmented by DNase1 to a desired size to be purified (Figure 5.3B). The fragments from the separate genes of interest are then combined in the absence of terminal primers through cycles of denaturation, annealing, and *Pfu* polymerase extension (Figure 5.3C). Following the *Pfu* reassembly, a final PCR amplification with terminal primers is performed to acquire full length shuffled sequences (Figure 5.3D) which can be cloned into an expression vector to form a complete shuffle library for transformation and screening.

Optimization of the fragmentation process through DNase1 concentration and duration of the digest reaction were critical to obtaining a uniform collection of small fragments, as can be observed in Figure 5.3B for the 0.5U DNase1 example. Smaller fragments allow the advantage of more crossovers to occur, meaning that there would be a greater chance of creating recombinations which separate more proximal mutations. A higher chance of crossover events is also more likely when using high similarity parent sequences (Joern 2003). The *Pfu* reassembly was performed identically to Zhao and Arnold (1997), other than increasing the polymerase extension time to account for the 1650 bp length of the luciferase gene. Optimisation of the primer concentration in the final amplification proved unnecessary.

5.3.2.2 Screening of x15 shuffle products

Once the shuffle of x15 with *Phem*Luc had been completed, the library was cloned into the pET16b vector and the screening process conducted with the inclusion of an incubation at 50 ˚C for 1-hour prior to bioluminescence imaging, as discussed in 5.3.1.3. Whereas a RT screen was performed for the primary screening of x15 in Figure 5.1 to assess the impact of the 15 concurrently introduced mutations on the general bioluminescent activity, primary screening of the shuffle library was only conducted after the 50 ˚C incubation as the primary objective of the shuffling process was to identify a subset mutant with improved bioluminescence signal following heat inactivation, meaning the bioluminescent activity at RT would be considered a secondary characteristic.

The post-incubation primary screen of the x15 shuffle performed adjacent to *wild-type Phem*Luc can be viewed in Figure 5.4. This image was generated by scaling the bioluminescent signal intensity cut-off such that remaining *Phem*Luc bioluminescence is represented in dark blue at the lowest end of the intensity scale in order for any higher activity colonies from the shuffle plate to be better visualized. The majority of the transformed shuffle products appear to be highly deleterious or extremely thermolabile, displaying little remaining bioluminescent activity. However, 7 colonies from the shuffle plate produced sufficient bioluminescent activity to be detected above the lower intensity cut-off. The highest bioluminescent activity colony from the shuffle (circled) was isolated in order to verify its bioluminescent activity and resistance to thermal inactivation in a secondary screen.

Secondary screening of the shuffle isolate was performed at both RT and 50 ˚C against the two parent sequences *Phem*Luc and x15 in Figure 5.5A. Sequencing of the shuffle isolate revealed a reversion of the S347G mutation in x15, producing a novel x14 mutation. Whereas the bioluminescent activity of x15 is below the intensity visualisation threshold for both conditions, the bioluminescent activity of the x14 shuffle isolate was observed to be significantly higher than the x15 parent at RT, but below that of *Phem*Luc. However, following the 1-hour incubation at 50 ˚C, the remaining bioluminescent activity of x14 exceeded the remaining bioluminescence signal produced by *Phem*Luc. Bar chart plots of the bioluminescence data from the secondary screen shown in Figure 5.5B reveal that whilst x14 produced 51% of the total bioluminescence signal of *Phem*Luc at RT, it displayed a significant improvement in resistance to thermal inactivation such that following the 50 ˚C incubation the remaining activity was 230% of the activity of *Phem*Luc under the same conditions.

Due to the high number of sequence variants and library size calculated in Equation 5.1, it is highly improbable that the x14 mutant identified is the highest thermostability candidate that could theoretically be identified through screening. However, as the generation and manual screening of 438282 colonies would be unfeasible, the bioluminescence and thermostability properties displayed by the x14 mutant were sufficient to proceed to further engineering efforts. As S347G had originally been included to explore a secondary property of improved activity with infraluciferin, the future work to engineer the x14 reversion focused only on its thermostability, and a separate engineering rational was explored for infraluciferin activity (see Chapter 4).

Equation 5.1.

Calculating variants:

$$n^r$$

n = Number of variants at each position (2)

r = Total number of positions (15)

$$= 32768 \textit{ unique combinations}$$

Calculating library:

$$L = -V \ln\left(-\frac{\ln Pc}{V}\right)$$

L = Library containing a number of clones/colonies (Unknown)

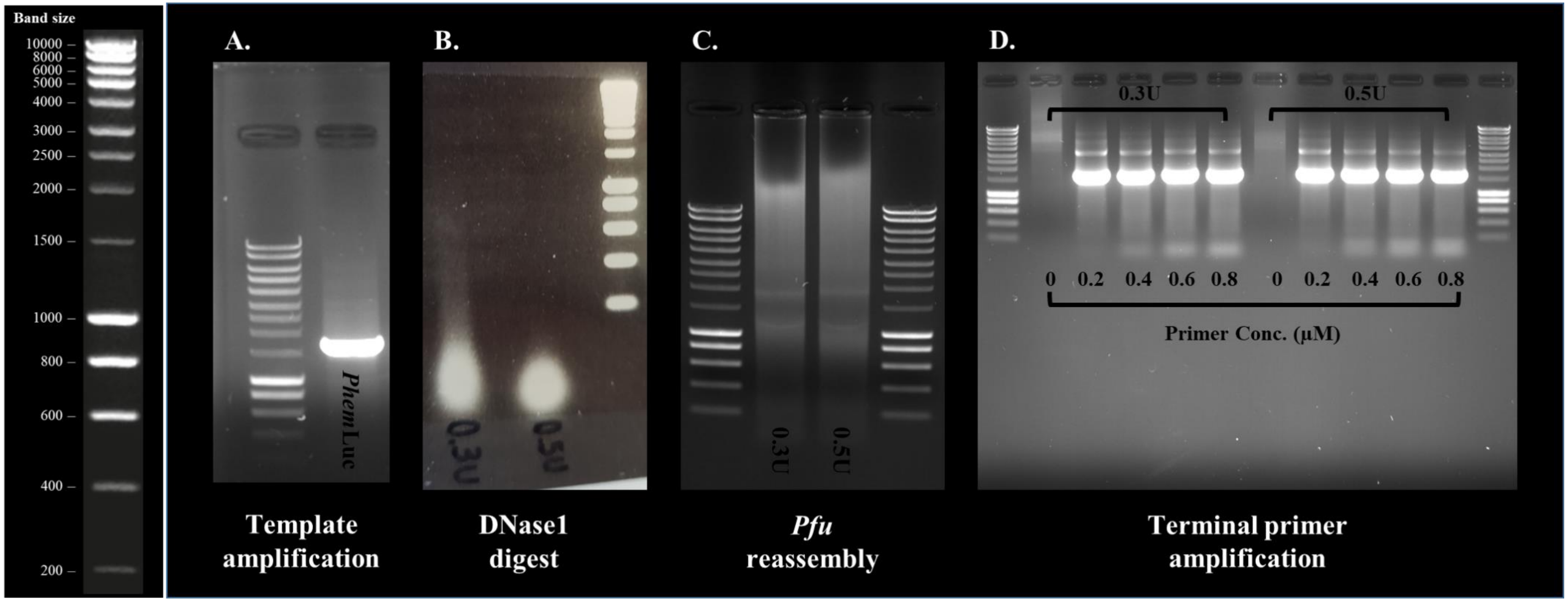
V = Total number of sequence variants (32786 possibilities from n^r)

Pc = Probability of complete library (95% confidence)

$$L = -32768 \ln\left(-\frac{\ln(0.95)}{32768}\right) \approx \mathbf{438282} \text{ (rounded)}$$

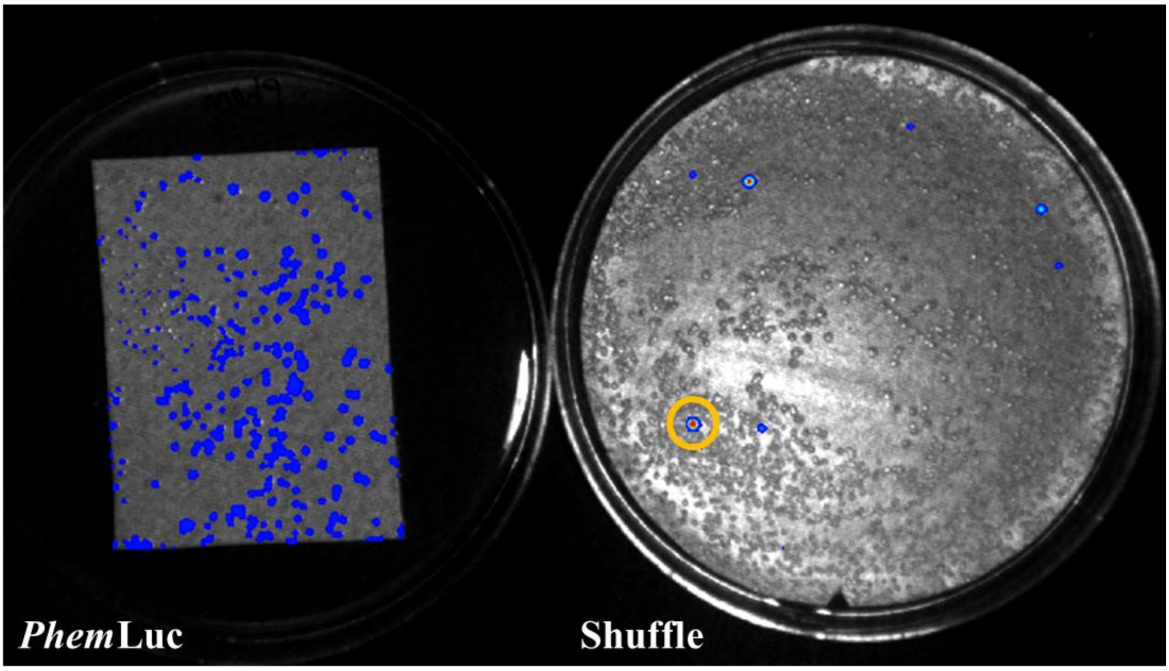
Estimating shuffle library size for complete diversity of sequence variants. n^r describes how the shuffling of *Phem*Luc with x15 would produce 32786 unique sequence variants. The second equation from Patrick *et al* (2003) calculates the size of library required to have a 95% chance of containing every possible sequence variant. The degree of over-sampling required positively correlates with the number of sequence variants, meaning to obtain a 95% chance of screening all variants, 438282 (rounded) colonies would need to be screened in the given shuffle library.

Figure 5.3.



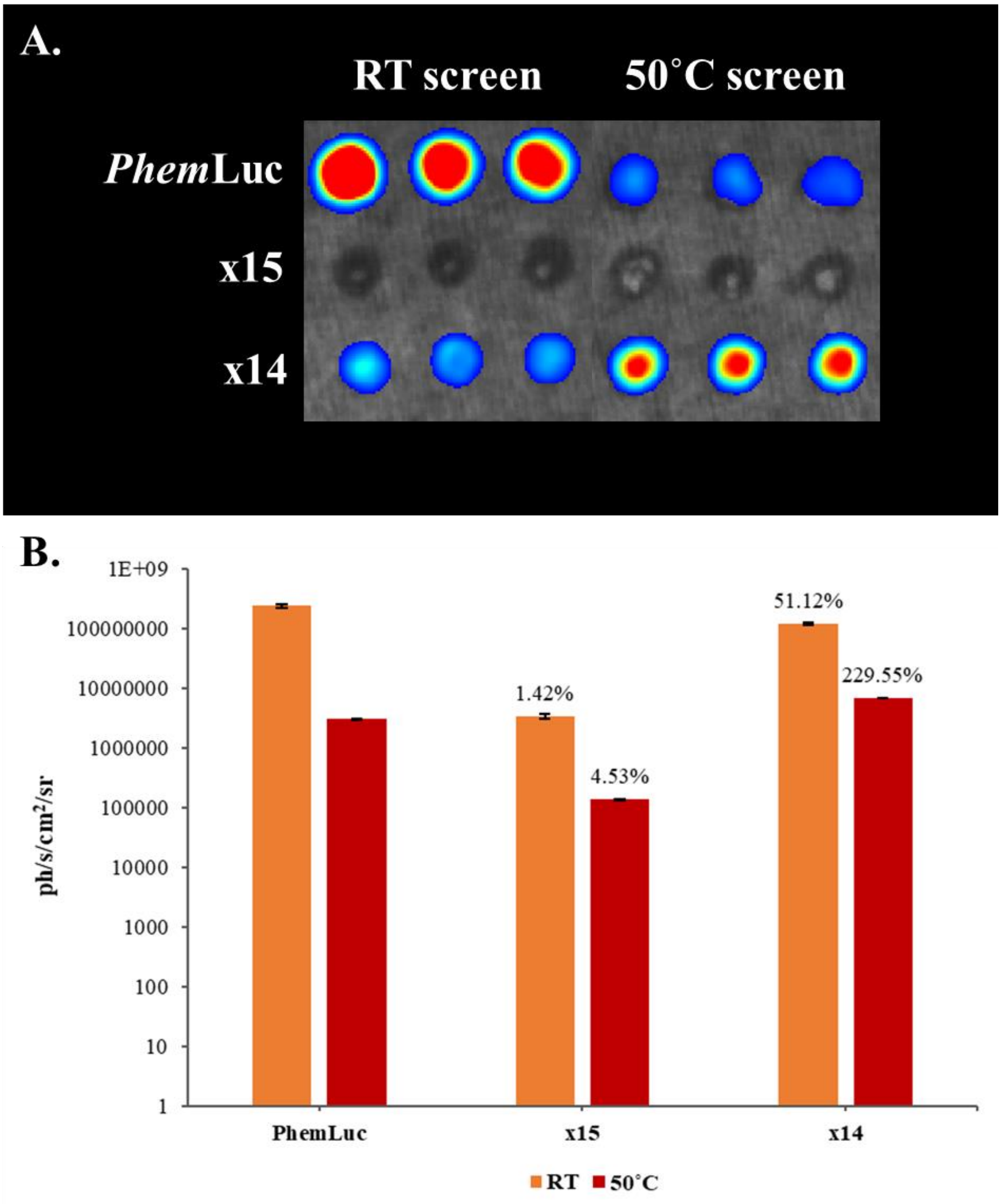
Gel electrophoresis analysis of the DNA shuffling process. The products of each DNA shuffling stage as visualised by gel electrophoresis. For clarity only *Phem*Luc is shown in A and B. (**A**) Primary amplification of shuffle templates, as demonstrated with *Phem*Luc. (**B**) DNase1 digest of primary template across two different concentrations. (**C**) Reassembly shuffle of DNase1 fragments from *Phem*Luc and x15 with *Pfu* polymerase in the absence of terminal primers. (**D**) Amplification of reassembly shuffle products with terminal primers. Ladder band sized in base pairs are displayed on the left.

Figure 5.4.



Bioluminescence of *Phem*Luc and shuffle library incubated at 50 ˚C. The colony formations from *E. coli* BL21 (DE3) transformed with pET16b plasmid containing *Phem*Luc (left) or the x15 shuffle products (right), here transferred onto nitrocellulose membranes (Chapter 2). Colony transferred membranes were induced for 3-4 hours at RT with IPTG (1 mM), prior to incubation for 60 minutes at 50 ˚C. Both plates were subsequently screened with 500 µM LH_2 and imaged adjacently in the PhotonIMAGER Optima (Biospace Labs, Paris, France), over an integrating period of 60 seconds. The highest activity shuffle product circled was isolated. Signal intensity is represented relative to the maximum signal detected, with minimum intensity indicated in blue and maximum intensity indicated in red.

Figure 5.5.



Bioluminescence of *Phem*Luc, x15, and x14, at RT and following 50 ˚C incubation. (**A**) Nitrocellulose membranes with *E. coli* BL21 (DE3) regrown from primary screen colonies of *Phem*Luc, x15 and thermostable isolate x14. Bioluminescence intensity scaling is non-comparable between RT and 50 ˚C screens. Signal intensity is represented relative to the maximum signal detected, with minimum intensity indicated in blue and maximum intensity indicated in red. Screens were performed as in Figure 5.2. (**B**) Log bar chart displaying the averaged bioluminescence activity measured across triplicate colonies of *Phem*Luc, x15 and x14, in Figure 5.5A. Percentages above the bars indicate the activity as a function of *Phem*Luc activity under the same condition. Error bars represent the standard error of the mean (SEM). Figure produced from a subset of data from Figure 5.9. Statistical analysis can be found in Figure 5.9.

5.3.3. Directed evolution of x14 for improved thermostability

Directed evolution comprises multiple methodologies which mimic the process of natural selection in order to generate proteins possessing desired traits through iterative rounds of diversification and selection. Whereas selection pressures associated with natural selection are dependent on natural environmental conditions, the selection pressure in any directed evolution process is user-defined in order to identify proteins displaying the desired traits, which may be improvements to existing characteristics or entirely novel properties (Cobb *et al.* 2013). The definitive advantage permitted by directed evolution is in the generation of improved variants where structural knowledge is limited and comparable adulterations to enzyme function would be prohibitively difficult to establish through rational design (Arnold 1998). The source of the molecular diversity can be varied, but commonly involves error-prone PCR (epPCR), oligonucleotide-directed randomization, DNA shuffling, and passing the cloned genes of interest through mutator stains (Cadwell and Joyce 1992; Stemmer 1994; Dale and Belfield 1996; Greener *et al.* 1996). Although no consensus exists for which technique is the most effective, epPCR is the most widely used for generation of mutagenic diversity *in vitro* (Labrou 2010).

5.3.3.1 epPCR as a function of polymerase fidelity

Multiple variations of the original epPCR method developed by Cadwell and Joyce (1994) are routinely used in order to generate a diverse library from a target sequence. Whilst most modern PCR applications make use of high-fidelity DNA polymerase which possesses 3' → 5' exonuclease activity (commonly referred to as proof-reading) to reduce base substitution events, epPCR makes use of the error-rate from lower fidelity polymerases such as the native DNA polymerase from *Thermus aquaticus (Taq)*, or engineered low fidelity variants. The error rate of *Taq* polymerase is amongst the highest known for *wild-type* thermostable DNA polymerases, ranging from $2x10^{-4}$ to $<1.2x10^{-5}$ mutations per nucleotide per cycle, totalling a cumulative error-rate of $\sim 10^{-3}$ per nucleotide, over the course of an average 20-25 cycle PCR (Eckert and Kunkel 1990). Whilst this native error-rate can prohibit the use of *Taq* in molecular cloning where maintaining accuracy is paramount, for the purposes of generating a mutagenized library *Taq* alone is insufficient at generating enough variation, whilst also being predominantly biased to A•T → G•C transitions and A•T → T•A transversions, reported at 63.2% and 16.1% of all mutations, respectively (Lin-Goerke *et al.* 1997).

5.3.3.2 Modulating error-rate through mutagenic additives

An established approach to increasing error-rate is the use of chemical additives, most commonly manganese chloride ($MnCl_2$) which reduces the substrate specificity of DNA polymerase and allows for the mis-coordination of the nucleotide binding domain and insertion of erroneous nucleotides in a concentration dependent manner (Beckman *et al.* 1985; Lin-Goerke *et al.* 1997). However, the yield of amplified products is known to decrease proportionally to the concentration of $MnCl_2$ present in the reaction (Vartanian *et al.* 1996). Although the addition of $MnCl_2$ induces an error-rate sufficient to generate a diversified sequence library through PCR, the substitution bias of *Taq* remains unchanged. Whilst epPCR strategies using only $MnCl_2$ have been widely utilised for mutagenesis and the subsequent isolation of advantageous mutant enzymes, an approach which confers higher error-rates than $MnCl_2$ alone without further decrease of amplification yield would allow for better exploration of the available sequence space through higher frequencies of all mutations including the more obscure transversions G → C and C → G which comprise only 1.4% of all mutations made by *Taq* (Lin-Goerke *et al.* 1997).

One such method capable of further increasing error-rate is substituting heavy water (D_2O) as the solvent in place of H_2O in a PCR. D_2O is substituted through centrifugal evaporation of H_2O from a reaction mixture prior to resuspension (see Chapter 2). D_2O contains the heavier isotope of hydrogen, deuterium, and induces random mutations without positional bias, template dependency, or decreased yield through an unknown mechanism which is independent of the polymerase selection or reaction composition (Minamoto *et al.* 2012; Minamoto 2017). The independent mutagenic mechanism of D_2O allows utilization in conjunction with $MnCl_2$ to produce an error-rate which exceeds the effect of either constituent alone.

5.3.3.3 Random mutagenesis and screening

A dual-mutagen approach to epPCR performed with $MnCl_2$ and D_2O was performed on the x14 shuffle isolate in order to identify novel mutations which would confer further resistance to thermal inactivation. The amplification products of the dual-mutagen epPCR as analysed by gel electrophoresis are shown in Figure 5.6A. All reactions were set up as detailed in Chapter 2 with a centrifugal evaporation process of H_2O to allow for D_2O substitution such that D_2O comprised ~99% of final reaction solvent relative to residual H_2O. A linear concentration of gradient of $MnCl_2$ between 0 – 0.3 mM was utilised to generate four libraries of consecutively increasing error-rate. The total yield of amplification product can be observed to diminish proportionally with the concentration of $MnCl_2$, as previously noted by Vartanian *et al* (1996). The purified mutagenically-diversified amplification products from each reaction were ligated into the pET16b vector and screened as transformed libraries of *E.coli* BL21 (DE3) for remaining bioluminescence activity after a 1-hour incubation at 50 ˚C as previously detailed in 5.3.1.3 and 5.3.2.2. The images generated from the bioluminescence imaging of the four mutagenized libraries of x14 are shown in Figure 5.6B. Similarly to the decline of amplification yield shown in Figure 5.6A, the proportion of colonies which retain bioluminescence activity diminishes in the libraries generated with higher concentration of $MnCl_2$. A likely explanation for this effect could be attributed to a higher error-rate correlating with an increased opportunity to incorporate deleterious mutations which would obfuscate the action of any advantageous mutations that may additionally be present and otherwise be able to modulate bioluminescence signal through improved resistance to thermal inactivation. Hence advantageous mutations would only be identified where they occur either independently as single mutations, or where their action alone is sufficient to outweigh any negative properties conferred by the presence of additional mutations present, i.e. they are additive.

Regardless of the high proportion of inactivated colonies, several mutations were observed to yield a high bioluminescence signal following 50 ˚C incubation. Isolation of high activity colonies for a secondary screen was attempted, and the screening and inactivation process performed identically to the primary. The bioluminescence image produced in the secondary screen is displayed in Figure 5.7; from this three triplicates of colonies can be observed to retain the previously recorded resistance to thermal inactivation through the subsequent high bioluminescence yield. Sanger sequencing was performed on the three high activity phenotypes and revealed two separate point mutations to the x14 sequence, L306H and I231V which occurred in two of the sequences investigated.

5.3.3.4 Development series comparative screening

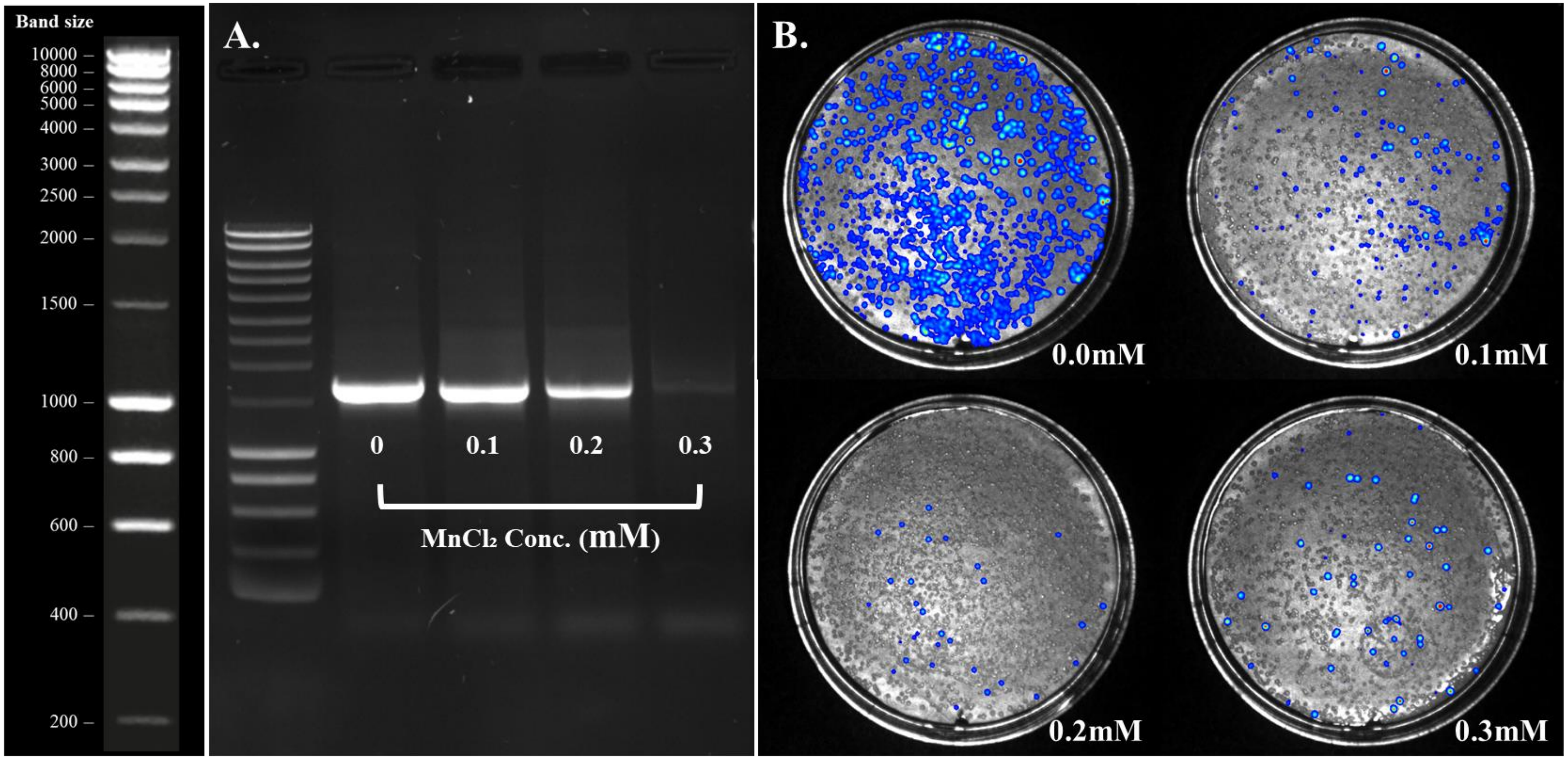
Prior to screening the ep-PCR mutants x14-I231V and x14-L306H against controls of *Phem*Luc and x14, a combinatorial mutant of the two mutations was constructed through restriction digest and subsequent ligation to generate a final x16 mutant to investigate the possibility of additive effects from the two novel mutations, alongside the original fourteen. A model indicating the positions of x16 mutations in *Phem*Luc is available in Appendices Figure 9.6. A final *E. coli* BL21 (DE3) screen of bioluminescence activity at RT and the activity remaining following a 1-hour incubation at 50 ˚C are shown in Figure 5.8. This final screen was conducted on the development series of mutants from the *wild-type Phem*Luc through x15, x14, x14-I231V, x14-L306H, and x16. A visual comparison of the bioluminescence signals at RT suggest that the activity of x16 is comparable to *wild-type Phem*Luc, and considerably improved relative to the primary functional iteration, x14. Following thermal inactivation at 50 ˚C, the bioluminescence activity of x16 significantly out performs the remaining activity of *Phem*Luc, or any of the mutants assessed in the process of development.

The bioluminescence data acquired in the final screen was used to construct the bar plots of Figure 5.9 – a data summary is shown in Table 5.2. The averaged bioluminescence data from triplicate colonies indicated that at RT x16 produced 119.47% (P=0.5549) and 233.7% (P=0.0007) of the bioluminescence signal of *Phem*Luc and x14, respectively. Following incubation at 50 ˚C, the remaining bioluminescence activity of x16 outperformed both *Phem*Luc and x14 significantly by 929.44% (P<0.0001) and 404.9% (P<0.0001) under the same conditions, respectively.

Of the two point mutations x14-I231V and x14-L306H, only x14-L306H was found to significantly improve the activity of x14 at RT (P=0.8055 and P=0.0099). However, both x14-I231V and x14-L306H conferred significant improvements to the bioluminescence activity of x14 following 50 ˚C incubation (P<0.0001 determined for the difference between all enzymes during the thermal inactivation screen). The incubated performances of x14-I231V and x14-L306H were recorded as 169.07% and 240.15% the activity of x14 under the same conditions (Table 5.2), respectively. Interestingly, the combination of the two point mutations conferred a thermostabilising potential in x16 which exceeded the sum of their individual improvements of x14, totalling 404.90% of the bioluminescence signal recorded for x14 . Relative to thermal inactivated x16, x14-I231V and x14-L306H displayed 41.76% and 59.31% of the recorded bioluminescence activity.

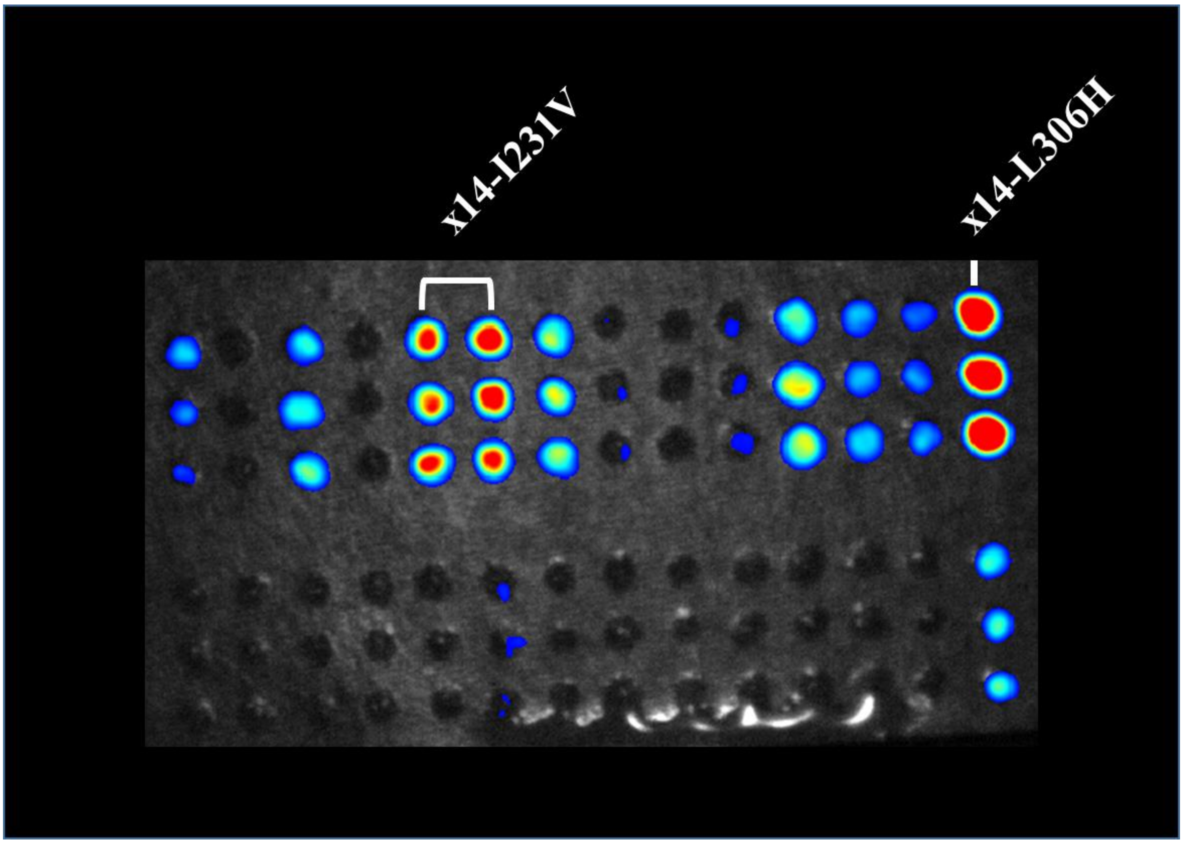
Whilst x16 yielded the greatest bioluminescence signal recorded in both conditions, a better indicator of resistance to thermal inactivation is available by calculating the activity displayed at 50 ˚C as a percentage of the total activity observed at RT. In this way, the remaining activity of x16 was calculated as 10.55%, versus 1.27% remaining activity recorded in *Phem*Luc. The remaining activity calculated in x14 (5.7%) and the increases seen in x14-I231V (7.54%) and x14-L306H (6.96%) support the conclusion that these point mutations contribute to the increased bioluminescence recorded for incubated x16 through enhancing the enzymes resistance to thermal inactivation, whilst also improving the bioluminescence activity at RT. This indicates that the mutations are additive for both bioluminescence activity improvement and thermostability.

Figure 5.6.



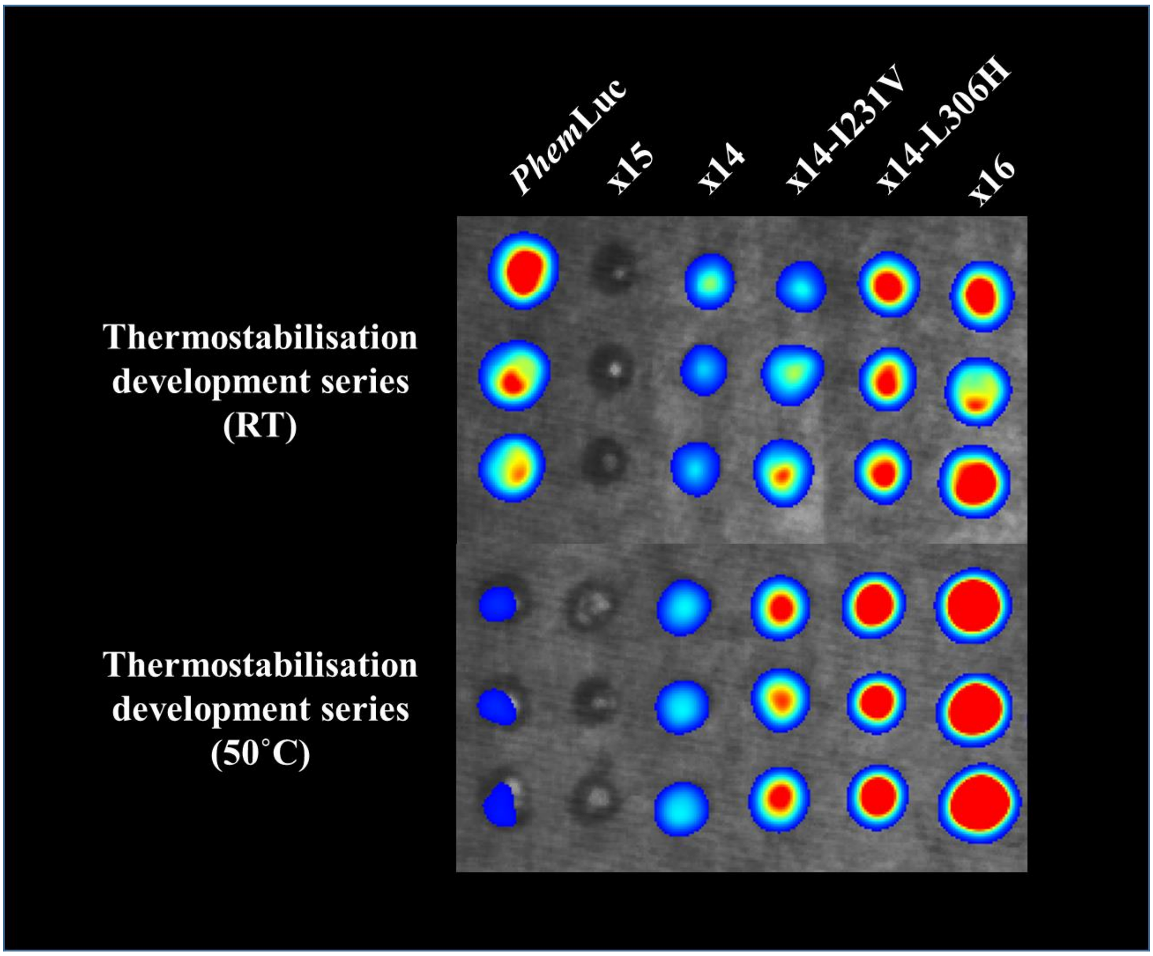
Random mutagenesis and primary screen of mutagenized x14. (**A**) Gel electrophoresis analysis of error-prone PCR products in the presence of D_2O and an increasing concentration of $MnCl_2$. Ladder band sizes in bare pairs are displayed on the left. (**B**) Random mutagenesis products of Figure 5.6A ligated into pET16b and transformed into *E. coli* BL21 (DE3). Screening was performed following a 50 ˚C incubation, as detailed in Figure 5.2. Signal intensity is represented relative to the maximum signal detected, with minimum intensity indicated in blue and maximum intensity indicated in red.

Figure 5.7.



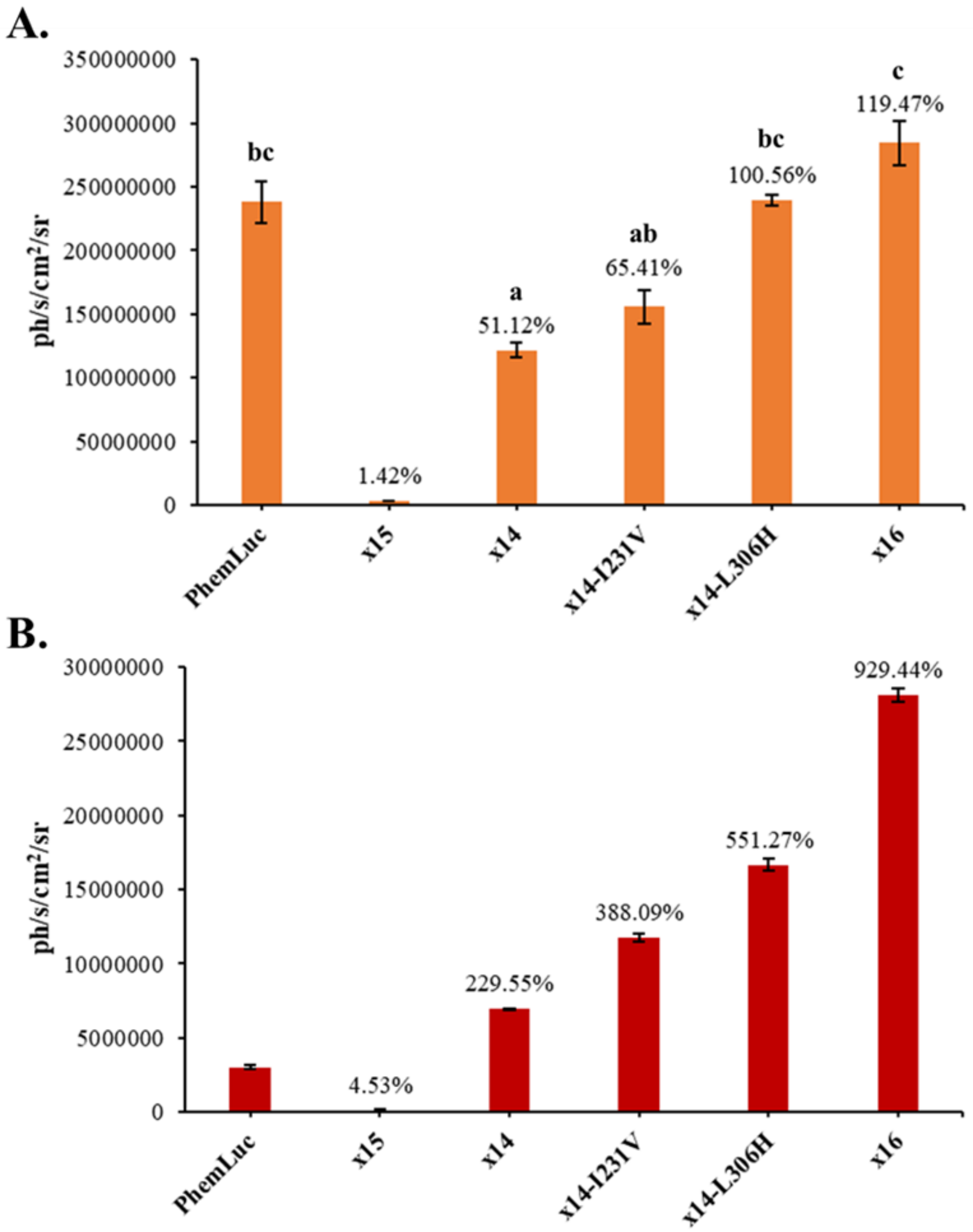
Bioluminescence of colonies selected from random mutagenesis libraries. Secondary screening of bioluminescence activity from thermostable colonies selected from primary screen in Figure 5.6B. Screening was performed following incubation at 50 ˚C, as detailed in Figure 5.2. Labels indicate the mutations incorporated into x14 as subsequently revealed by Sanger sequencing. Signal intensity is represented relative to the maximum signal detected, with minimum intensity indicated in blue and maximum intensity indicated in red.

Figure 5.8.


*Phem*Luc
x15
x14
x14-I231V
x14-L306H
x16
Thermostabilisation development series (RT)
Thermostabilisation development series (50˚C)

Comparison of bioluminescence and thermostability from *Phem*Luc and mutants leading to development of x16. Bioluminescence activity screening of *Phem*Luc and each stage of development to the final iteration of thermostable mutant x16. Screening was performed at RT and following incubation at 50 ˚C, as detailed in Figure 5.2. Bioluminescence intensity scaling is not comparable between the RT and 50 ˚C screen. Signal intensity is represented relative to the maximum signal detected, with minimum intensity indicated in blue and maximum intensity indicated in red.

Figure 5.9.



Log bioluminescence of *Phem*Luc and mutants up to development of x16 at RT and following 50 ˚C incubation. Bar chart displaying the averaged bioluminescence activity measured across triplicate colonies of *Phem*Luc and each stage of development up to the final thermostable mutant x16, from Figure 5.8. (**A**) Screen conducted at RT. (**B**) Screen conducted at 50 ˚C. Percentages above the bars indicate the activity as a function of *Phem*Luc activity under the same condition. Error bars represent the standard error of the mean (SEM). Statistical analysis performed as detailed in 2.12. The difference between all groups in **B** were determined as significantly different (P<0.05).

Table 5.2.

	ph/s/cm^2/sr		**Remaining activity**	**Percentage of *Phem*Luc activity**		**Percentage of x14 activity**		**Percentage of x16 activity**	
	RT	**50 ˚C**		**RT**	**50 ˚C**	**RT**	**50 ˚C**	**RT**	**50 ˚C**
***Phem*Luc**	238000000 ±16523196.33	3023333.33 ±104383.95	1.27%	100%	100%	195.62%	43.56%	83.70%	10.76%
x15	3386666.67 ±252144.56	137000 ±2969.06	4.05%	1.42%	4.53%	2.78%	1.97%	1.19%	0.49%
x14	121666666.67 ±6128438.54	6940000 ±42241.73	5.70%	51.12%	229.55%	100%	100%	42.79%	24.70%
x14-I231V	155666666.67 ±12775495.25	11733333.33 ±259280.09	7.54%	65.41%	388.09%	127.95%	169.07%	54.75%	41.76%
x14-L306H	239333333.33 ±3975736.77	16666666.67 ±426670.50	6.96%	100.56%	551.27%	196.71%	240.15%	84.17%	59.31%
x16	284333333.33 ±17749020.84	28100000 ±426670.50	10.55%	119.47%	929.44%	233.70%	404.90%	100%	100%

Summary of screen for resistance of bioluminescence to thermal inactivation. Data indicating average bioluminescence activity (ph/s/cm^2/sr) from induced *E. coli* screens presented in Figure 5.8-10. "RT" relates to data of screens performed at room temperature, whilst "50 ˚C" indicates the inclusion of a 1-hour incubation at 50 ˚C prior to screening. Remaining activity relates to the activity recorded following incubation at 50 ˚C as a proportion of the activity recorded at RT. ± values represent the SEM.

5.3.4. Resistance to thermal inactivation of x16 bioluminescence

The use of thermal inactivation prior to screening luciferase bioluminescence in *E. coli* is a powerful approach for comparative analysis of thermostability and subsequent isolation of advantageous mutant candidates. However, the multiple steps from transformation of *E. coli* to induced colonies can lead to significant variation whilst also being less reflective of the environment in which a thermostable luciferase may find application. Greater environmental control can be achieved with assays using purified enzymes along with additional strategies which allow for greater control of the thermal inactivation process and subsequently an improved characterisation of enzyme resistance to thermal inactivation.

As the highest activity mutants identified throughout the screens, x16 was taken forward for purification and subsequent characterisation of its thermostable conferring properties (see Chapter 2 and 6 for further purification details). Thermostability was investigated by thermal inactivation of enzyme aliquots over varied temperatures at set time points of incubation, and additionally over a narrow temperature range where bioluminescence emissions could be recorded throughout the incubation process.

5.3.4.1 Incubation assays

Primary investigation of x16 resistance to thermal inactivation was conducted by incubating 0.5 µM pre-aliquoted enzyme samples in a digital water bath preconfigured to a target temperature between 25-60 ˚C, at intervals of 5 ˚C. Enzyme aliquots were removed from incubation to ice at 15, 30, 45, and 60 minutes. Following the transfer to ice, bioluminescence activity was immediately recorded in the BMG Labtech CLARIOstar by automatic injection of room-temperature substrate mix (LH_2 and ATP) onto the chilled enzyme. The point of maximum bioluminescence intensity (I_{max}) was taken from the resulting flash kinetics to indicate the remaining bioluminescence activity. Measurements of 0-minute incubation were obtained by measuring the bioluminescence activity of non-incubated enzyme aliquots to indicate initial enzyme activity. Samples of 0-minute incubation were similarly placed on ice prior to conducting the assay, in order to equate the in-well temperature to the incubated samples which were necessarily transferred to ice to stop thermal inactivation.

The thermostability of purified x16 was assessed against its *wild-type* origin, *Phem*Luc. Additionally, the luciferase from the North American firefly *Photinus pyralis* (*Ppy* Fluc) and its engineered thermostable variant x11 were included for comparison. Values of I_{max} are presented as individual bar plots for each incubation temperature in Figures 5.10A-17A. The same values of I_{max} have been converted to percentage of initial activity in Figures 5.10B-17B, where initial activity is represented by the I_{max} measurements obtained for the 0-minute incubation sample. A summary of the I_{max} values and conversions to initial activity percentages across all time points and temperatures assessed can be found in Tables 5.3-4.

Measurements of initial activity (0-minutes) were performed in triplicate and the average used as the initial activity for each respective enzyme across each bar plot. Whereas x16 appeared to produce a higher bioluminescence yield at 119.47% of *Phem*Luc at RT during the *E. coli* screens, the activity of the purified x16 enzyme was only observed to produce a bioluminescence yield equal to 5.52% of the initial activity of *Phem*Luc. Similarly, the initial bioluminescence activity of x11 was diminished relative to its *wild-type* origin *Ppy*, at 17.26%. Comparing *Phem*Luc to the *wild-type* control *Ppy* Fluc revealed that *Phem*Luc produced 106.03% of activity under the same conditions. This shows that the luciferase from the lesser British Glow-worm is of similar activity to the North American *Ppy* Fluc.

At the lowest temperature assessed of 25 ˚C, both the thermostable engineered variants and the *wild-type* enzymes tolerate incubation well across all time points recorded. *Phem*Luc proved to be the least thermostable, producing 84.48% of initial activity by the 60-minute sample against 91.78% for *Ppy* Fluc. Interestingly, both of the thermostable variants produced a greater bioluminescent yield by the 60-minute measurement than recorded for the initial activity, recorded as 111.36% for x11, and 108.96% for x16, indicating glow-type kinetics. The *wild-type* enzymes experience a greater inactivation from incubation at 30 ˚C, with *Phem*Luc activity at 60-minutes dropping to 69.7% of initial activity ($P<0.0001$), and *Ppy* Fluc decreasing to 78.71% ($P=0.0007$). The activity of x11 still displayed a resistance to inactivation by incubation at this temperature, continuing to increase to 105.73% of initial activity by the final measurement. The activity of x16 experienced a minor instability by incubation at 30 ˚C, as the recorded activity dropped by >5% at the 15 and 45-minute samples. However, by the final measurement the bioluminescence yield had increased to 100.86%. Increasing the incubation temperature to 35 ˚C continued to reduce the bioluminescence yield recorded for the both *Phem*Luc and *Ppy* Fluc, such that final activities were 18.12% and 34.38%, respectively. This temperature proved to be a threshold region for the engineered variants, which declined only marginally to 93.32% for x11, and 90.96% for x16 by the final time point.

Increasing the incubation to 40 ˚C proved to be highly deleterious to the bioluminescence ability of both *Phem*Luc and *Ppy* Fluc, such that the activities recorded for the first time point of 15-minutes were 5.47% (P<0.0001) for *Phem*Luc, and 5.13% (P<0.0001) for *Ppy* Fluc. The final activity of both enzymes was less than 0.1%. Curiously, the lowest activity recorded for x11 occurred at the 30-minute time point, measuring 94.85% of initial activity, increasing to 104.62% by the final measurement at 60-minutes. The x16 enzyme proved less resistant to incubation at 40 ˚C, and diminished to 78.23% of initial activity by the final activity measurement (P<0.0001). Increasing the temperature to 45 ˚C drastically reduced the bioluminescence activity for the majority of enzymes. The bioluminescence signal from the *wild-type* enzymes was reduced to approximately 0.1% after only 15-minutes of incubation. Whilst x11 still displayed a high tolerance to thermal inactivation, the final bioluminescence signal was reduced to 90.68% of the initial activity (P=0.0942). The activity and thermostability of x16 continued to further diverge from the activity of x11, such that the bioluminescence signal steadily declined over each time-point and final measurements equalled 34.68% of initial activity. At this temperature all time points for x16 were determined as significantly different to each other.

From 50 ˚C incubation onwards, the bioluminescent activity of both *Phem*Luc and *Ppy* Fluc were undetectable across all time-points assessed. The degradation in x16 activity significantly increased such that the activity recorded after 15-minutes of incubation was comparable to measurements obtained for the final measurements 5 ˚C lower, at 39.27% initial activity, dropping to 6.53% by 60-minutes. Similarly to the 45 ˚C screen, all time points for x16 were determined as significantly different to each other. The bioluminescent activity of x11 was recorded as 75.46% at the 60-minute measurement (P=0.0003), indicating a high resistance to inactivation remained at 50 ˚C. However, by increasing the incubation to 55 ˚C, the bioluminescence of x11 begins to significantly diminish across all time-points, with final activity recorded at only 2.49%, indicating a key threshold exists between 50-55 ˚C. When analysis was performed at 60 ˚C, the activity of all enzymes was effectively eliminated.

5.3.4.2 Degradation analysis of x11 at 50-60 ˚C

To further investigate the thermostability of x16 against x11, a second inactivation study was conducted between 50-60 ˚C to follow continuously the kinetic profile of degradation over 30-minutes, with an interval of 2 ˚C. Bioluminescence reactions of both enzymes were initiated by manual pipetting of saturating conditions of LH_2 and ATP into plate wells of aliquoted enzyme. Initiated reactions were immediately overlaid with mineral oil and transferred to a preconfigured heat block in the LUCY imager (ERBA MDX, Ely, UK). Use of the LUCY allowed total bioluminescence emission to be integrated for 10-seconds every 20-seconds over the 30-minute acquisition window, and the resulting bioluminescence degradation curves are shown in Figure 5.18.

Whilst thermal inactivation in a water bath provides residual activity of an enzyme after incubation at different temperatures, the thermal inactivation study conducted in the LUCY allows the reduction of enzyme activity to be followed as a continuous process.

Across the range of temperature assessed, no overlap occurred between the activity curves generated for the two enzymes, such that x11 produced a greater bioluminescent signal at 60 ˚C than x16 at 50 ˚C. The bioluminescence activity peaks for both enzymes were consistent across all temperatures but the rate of degradation was greater for each consecutive 2 ˚C increase in temperature with x16.

5.3.4.3 LAMP-BART

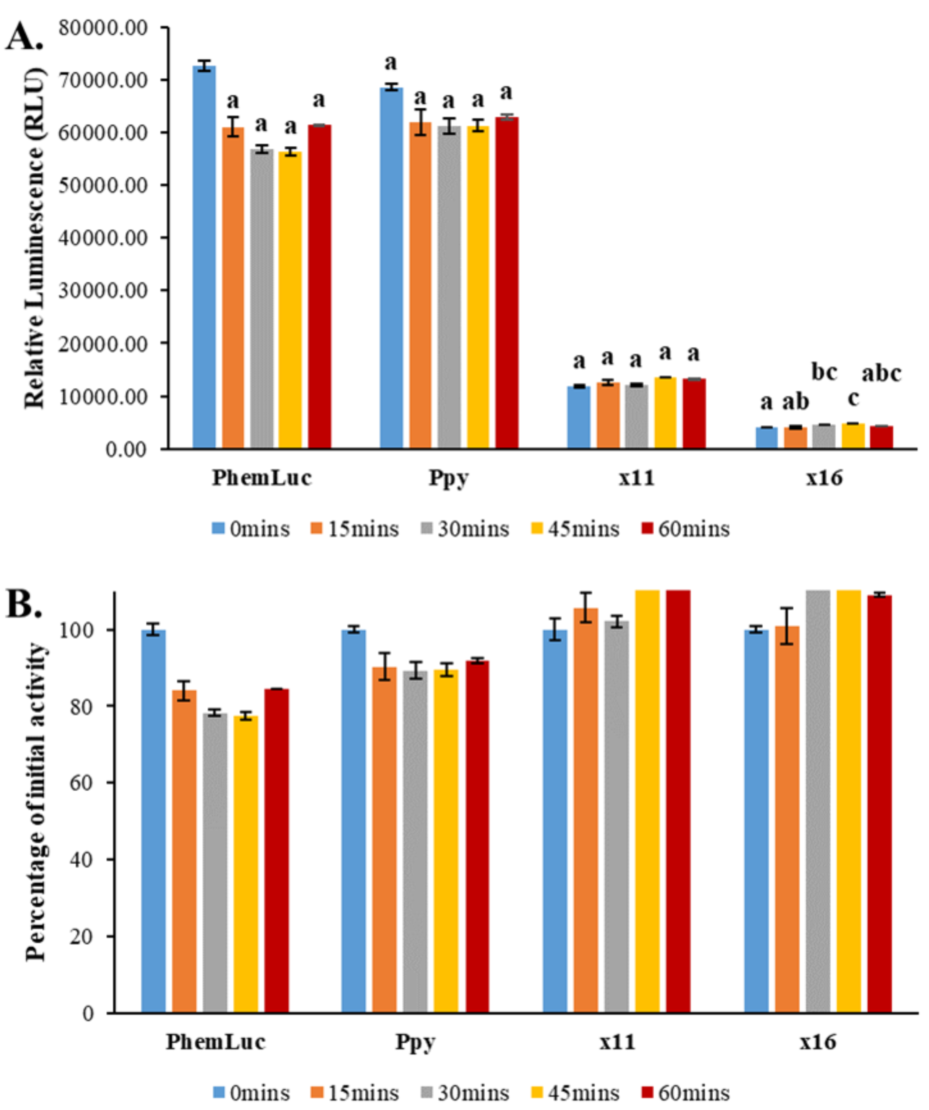
Although LAMP-BART is performed at temperatures lower than would commonly be used in PCR, the assays are most often performed above 60 ˚C to optimize for the activity of DNA polymerase and the annealing of primers. The ability for BART to function under such conditions is entirely dependent on engineered variants of luciferase which have been suitably augmented to operate at increased temperatures. Currently, the firefly luciferase with the greatest resistance to thermal inactivation is a genetically evolved variant of *Photuris pensylvanica* called Ultra-Glo^{TM} (Promega, Madison, WI, USA), which is a commercially available patented enzyme speculated to have up to 70 mutations (Hall *et al.* 1999; Hsiao *et al.* 2016). In addition to improved thermal stability, the engineering of Ultra-Glo has improved its overall robustness and resistance to ionic detergents and reductive agents. The bioluminescence emission profile has also been adulterated to replace the characteristic flash of firefly bioluminescence with a stable glow kinetic, which is more favourable than flash kinetic assays that require imaging devices equipped with injectors (Promega Corporation 2013; Promega Corporation 2015).

As indicated by the thermal inactivation assays of 5.3.4.2 and 5.3.4.3, the bioluminescence activity of x16 would not be suitable for application in a standard LAMP-BART assay. In order for x16 to function as a substitute for Ultra-Glo, the temperature would have to be reduced sufficiently to enable the bioluminescence activity whilst also maintaining compatibility with the working temperature of the DNA polymerase and the required annealing conditions of the primers. An assay temperature of 50 ˚C was selected for this purpose, which although not ideal for any single reaction component should be sufficient to enable x16 luciferase bioluminescent activity in BART, and the DNA polymerase and primer annealing required for LAMP.

LAMP-BART assays with N2 SARS-CoV-2 RNA as template were performed at 50 ˚C using Ultra-Glo, x11, and x16 in Figures 20-22 to investigate whether the engineering of *Phem*Luc into x16 provided sufficient thermostability for deployment in applications requiring a high degree of resistance to thermal inactivation. Ultra-Glo was included as a control to prove the validity of an assay conducted at lower temperature, as no precedence exists for performing LAMP-BART at 50 ˚C. x11 was included to represent an engineered variant with thermostability properties between the capabilities of x16 and Ultra-Glo.

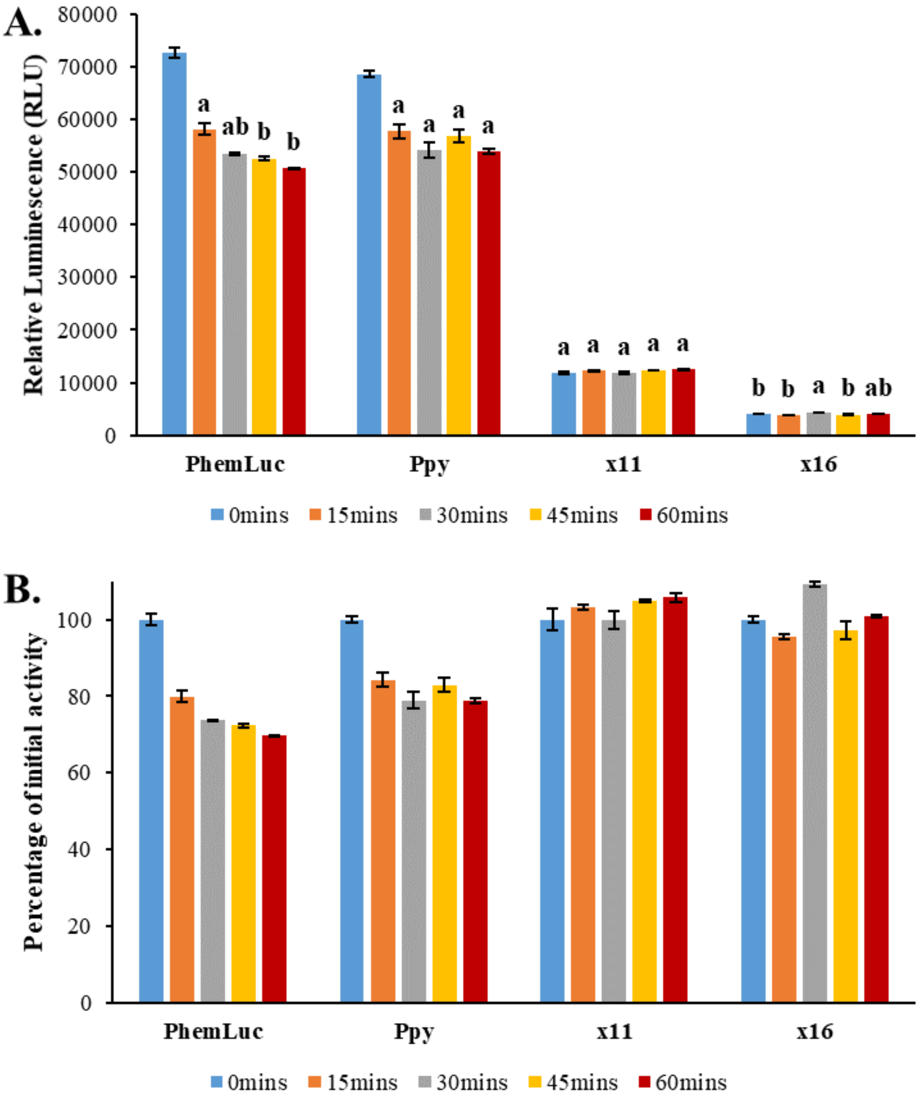
LAMP-BART performed at 50 ˚C successfully produced a bioluminescence emission peak with Ultra-Glo, taking an average of 41.11 minutes to reach T_{max} (Figure 5.19). A similar result was obtained for x11, which although emitted less bioluminescence as indicated by RLU recorded over the assay's duration, produced a very similar T_{max} of 41.37 minutes (Figure 5.20). Substitution of Ultra-Glo with x16 drastically reduced the bioluminescence emission throughout the assay and extended the time to T_{max} to an average of 46.23 minutes (Figure 5.21A). As no further adjustments to assay constituents other than the luciferase functioning in the BART reaction were made, it was assumed that the primary LAMP reaction would be performing DNA synthesis sufficiently and generating inorganic PPi to be converted to ATP at levels comparable to the assay conducted with Ultra-Glo. Therefore, insufficient luciferase activity would be responsible for the poor levels of bioluminescence emissions observed. To correct for the deficit in luciferase activity, a second LAMP-BART assay was conducted using a 20x greater concentration of x16 than was initially utilised (Figure 5.21B). The increased concentration of x16 significantly produced greater uniformity in the kinetic profiles across the replicates and increased the bioluminescence emissions throughout the duration of the assay, including at the emission peak which occurred with an earlier average T_{max} of 38.05 minutes.

Figure 5.10.



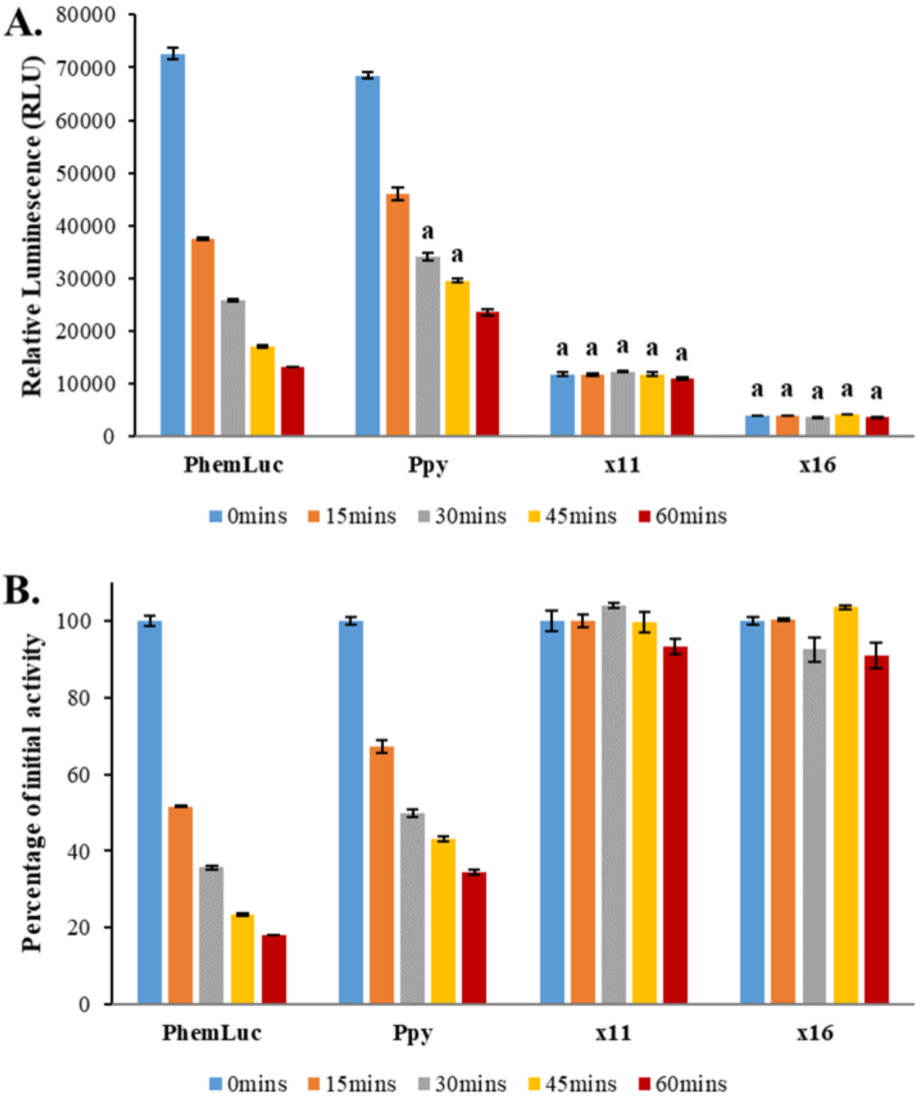
Bioluminescence following thermal inactivation at 25 ˚C. Thermal inactivation of *Phem*Luc, *Ppy*, x11 and x16 via incubation in a circulating digital water bath set at 25 ˚C. Pre-aliquoted enzymes removed from incubation to ice at set time points of 15, 30, 45, and 60 minutes. Bioluminescence activity screening was immediately conducted after each removal. Measurements obtained in a luminometer by injection of substrate mix onto each enzyme such that final assay concentrations were equal to 1 mM ATP, 500 µM LH_2, and 0.167 µM protein. Each reaction constituent previously diluted in chilled TEM buffer of pH 7.8 (±0.05) and total reaction volume equal to 150 µl. Light emission integrated over 20 ms for 1000 consecutive measurements and the point of I_{max} presented. Assays performed in triplicate for each condition, and averaged data presented. Error bars represent the standard error of the mean (SEM). (**A**) Data presented as Relative Luminescence (RLU). (**B**) Data presented as percentage of initial activity without incubation. Statistical analysis performed as detailed in 2.12. Significance groupings are discrete between individual Lucs.

Figure 5.11.



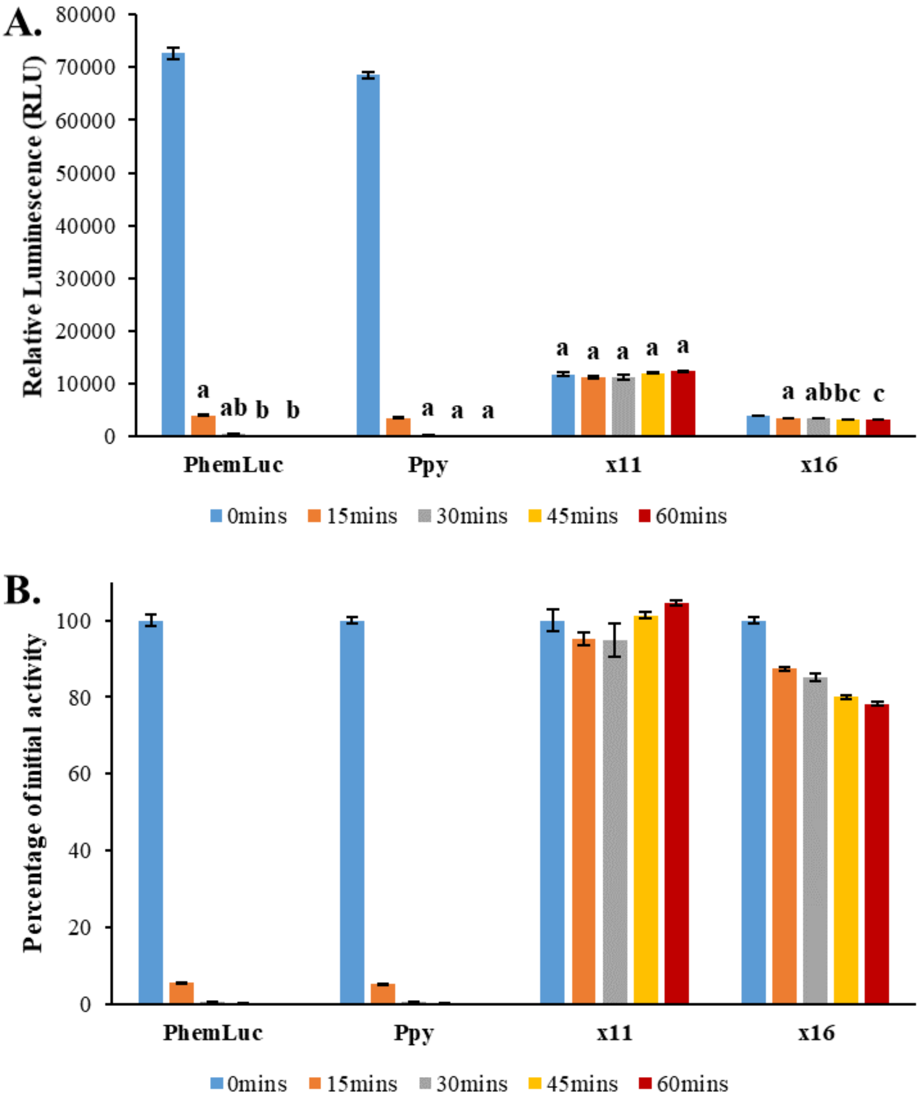
Bioluminescence following thermal inactivation at 30 ˚C. Thermal inactivation of *Phem*Luc, *Ppy*, x11 and x16 via incubation in a circulating digital water bath set at 30 ˚C. Pre-aliquoted enzymes removed from incubation to ice at set time points of 15, 30, 45, and 60 minutes. Bioluminescence activity screening was immediately conducted after each removal. Measurements performed as detailed in Figure 5.10. Error bars represent the standard error of the mean (SEM). (**A**) Data presented as Relative Luminescence (RLU). (**B**) Data presented as percentage of initial activity without incubation. Statistical analysis performed as detailed in 2.12. Significance groupings are discrete between individual Lucs.

Figure 5.12.



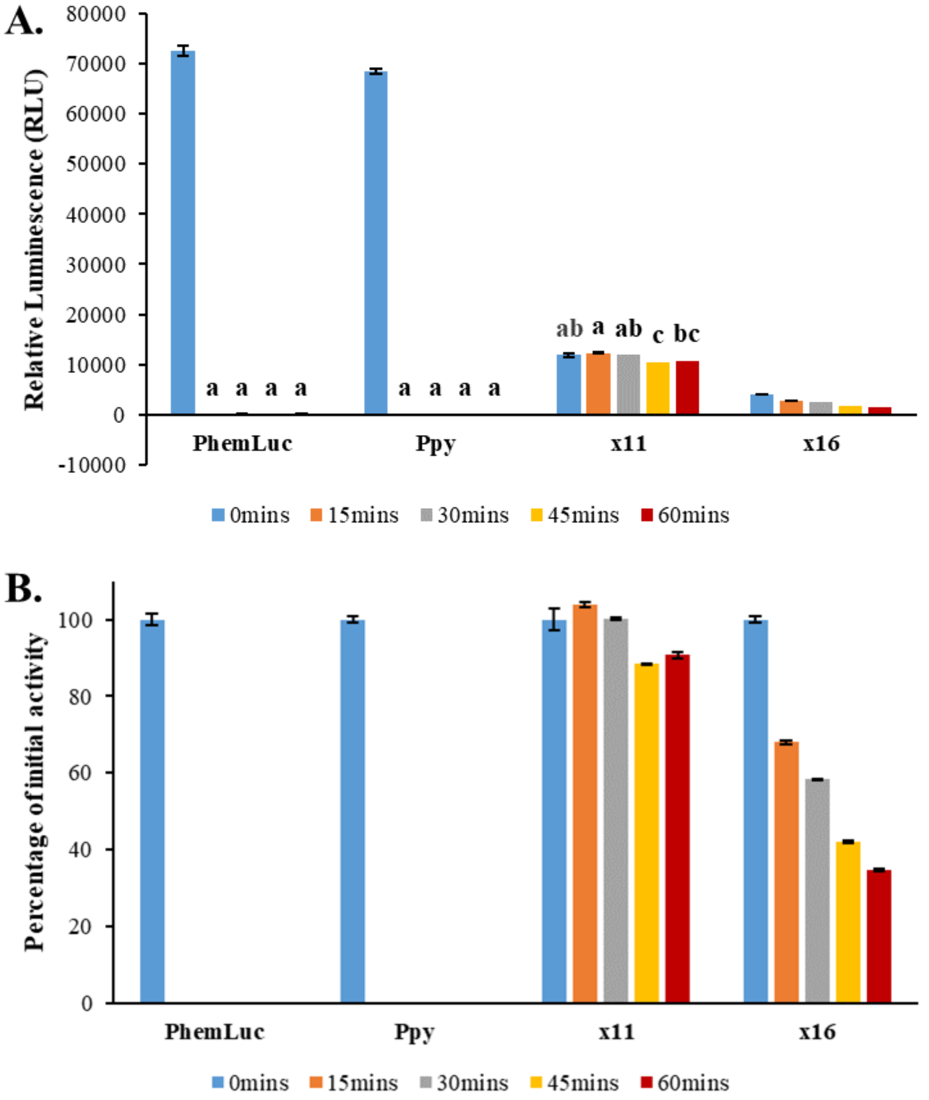
Bioluminescence following thermal inactivation at 35 ˚C. Thermal inactivation of *Phem*Luc, *Ppy*, x11 and x16 via incubation in a circulating digital water bath set at 35 ˚C. Pre-aliquoted enzymes removed from incubation to ice at set time points of 15, 30, 45, and 60 minutes. Bioluminescence activity screening was immediately conducted after each removal. Measurements performed as detailed in Figure 5.10. Error bars represent the standard error of the mean (SEM). (**A**) Data presented as Relative Luminescence (RLU). (**B**) Data presented as percentage of initial activity without incubation. Statistical analysis performed as detailed in 2.12. Significance groupings are discrete between individual Lucs.

Figure 5.13.



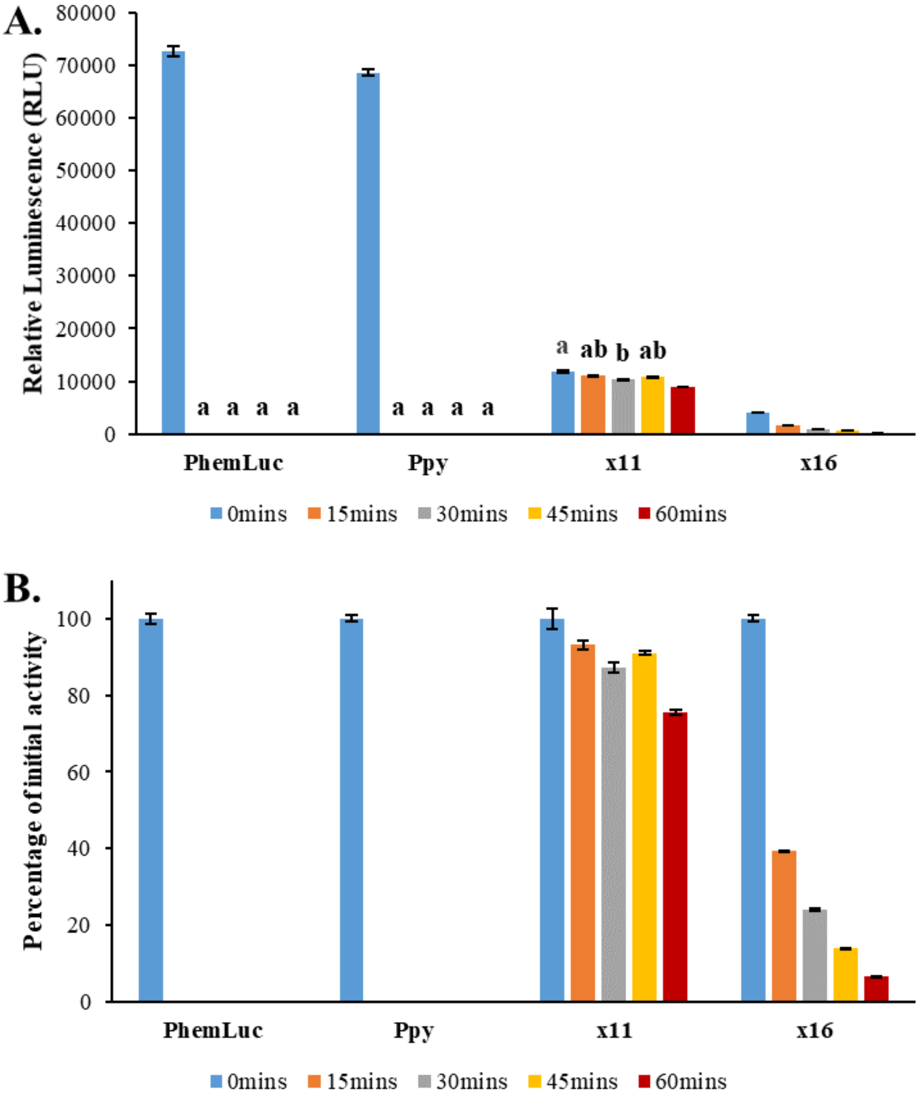
Bioluminescence following thermal inactivation at 40 ˚C. Thermal inactivation of *Phem*Luc, *Ppy*, x11 and x16 via incubation in a circulating digital water bath set at 40 ˚C. Pre-aliquoted enzymes removed from incubation to ice at set time points of 15, 30, 45, and 60 minutes. Bioluminescence activity screening was immediately conducted after each removal. Measurements performed as detailed in Figure 5.10. Error bars represent the standard error of the mean (SEM). (**A**) Data presented as Relative Luminescence (RLU). (**B**) Data presented as percentage of initial activity without incubation. Statistical analysis performed as detailed in 2.12. Significance groupings are discrete between individual Lucs.

Figure 5.14.



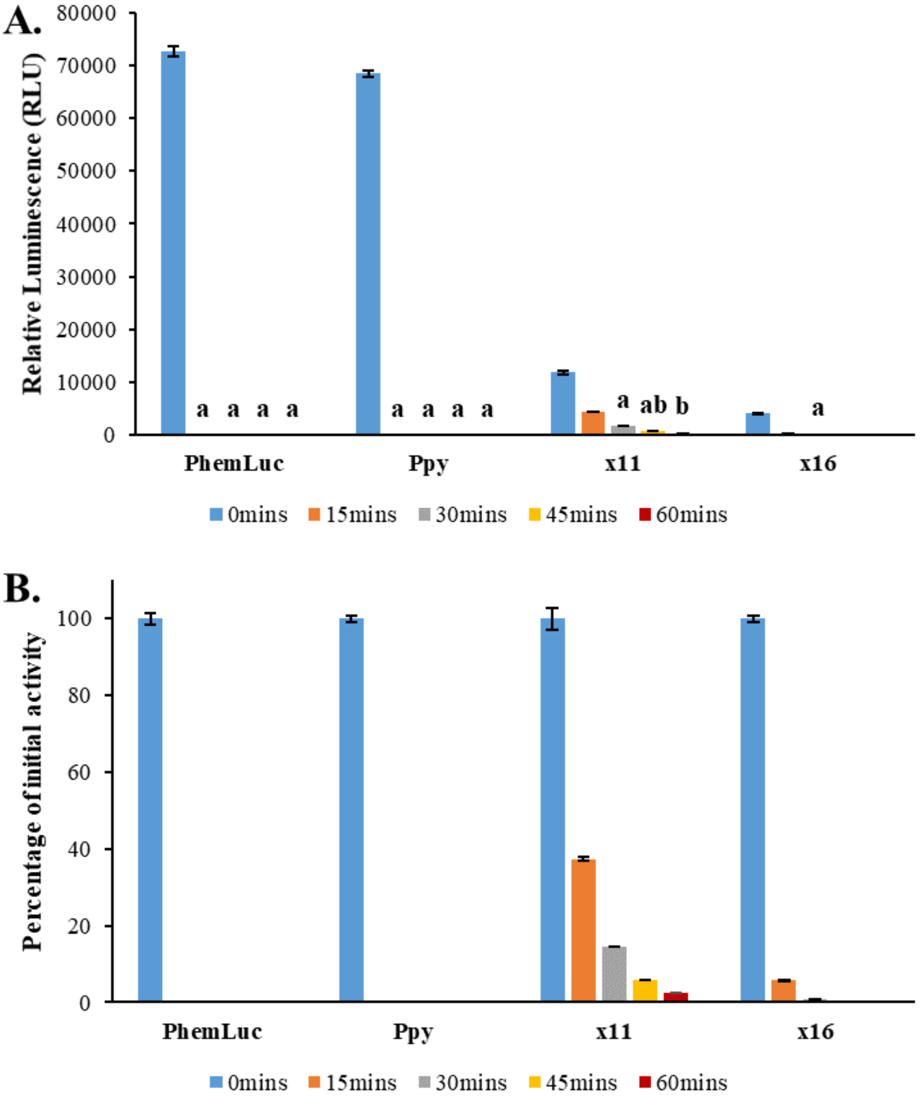
Bioluminescence following thermal inactivation at 45 ˚C. Thermal inactivation of *Phem*Luc, *Ppy*, x11 and x16 via incubation in a circulating digital water bath set at 45 ˚C. Pre-aliquoted enzymes removed from incubation to ice at set time points of 15, 30, 45, and 60 minutes. Bioluminescence activity screening was immediately conducted after each removal. Measurements performed as detailed in Figure 5.10. Error bars represent the standard error of the mean (SEM). (**A**) Data presented as Relative Luminescence (RLU). (**B**) Data presented as percentage of initial activity without incubation. Statistical analysis performed as detailed in 2.12. Significance groupings are discrete between individual Lucs.

Figure 5.15.



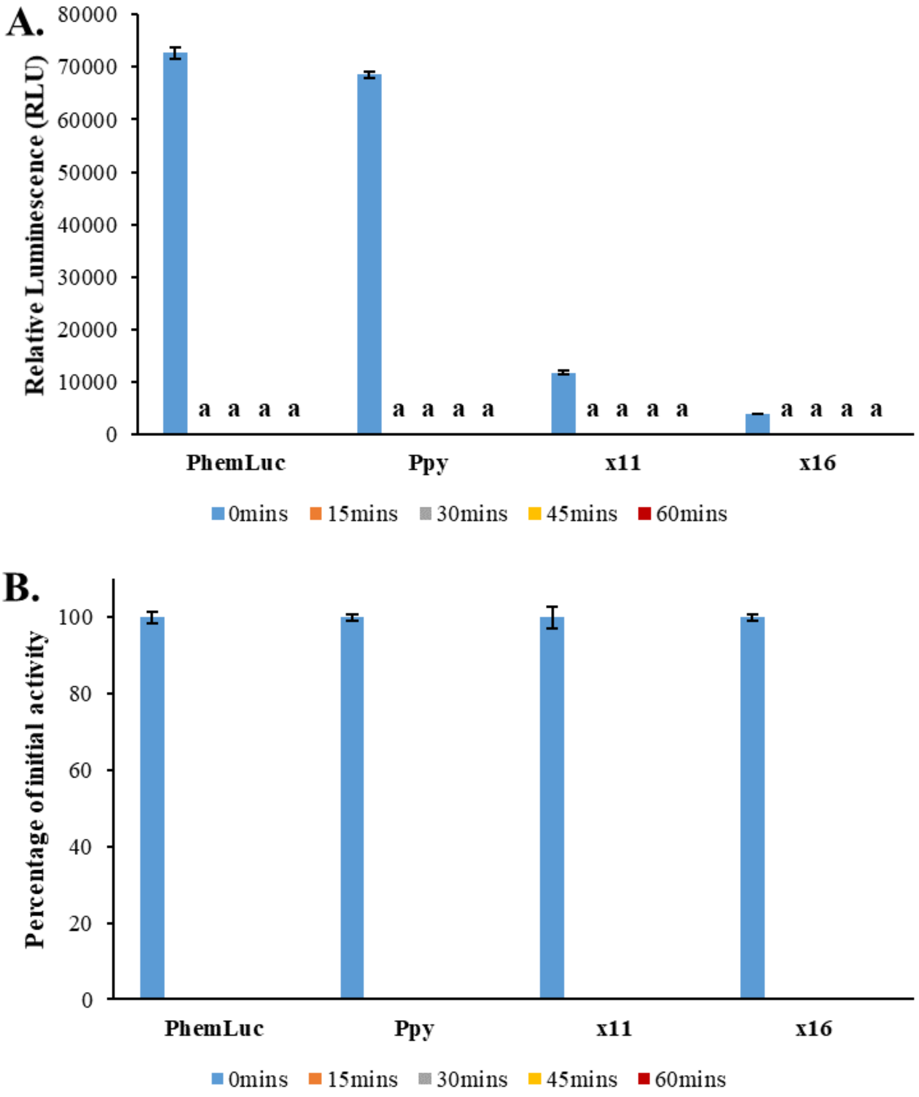
Bioluminescence following thermal inactivation at 50 ˚C. Thermal inactivation of *Phem*Luc, *Ppy*, x11 and x16 via incubation in a circulating digital water bath set at 50 ˚C. Pre-aliquoted enzymes removed from incubation to ice at set time points of 15, 30, 45, and 60 minutes. Bioluminescence activity screening was immediately conducted after each removal. Measurements performed as detailed in Figure 5.10. Error bars represent the standard error of the mean (SEM). (**A**) Data presented as Relative Luminescence (RLU). (**B**) Data presented as percentage of initial activity without incubation. Statistical analysis performed as detailed in 2.12. Significance groupings are discrete between individual Lucs.

Figure 5.16.



Bioluminescence following thermal inactivation at 55 ˚C. Thermal inactivation of *Phem*Luc, *Ppy*, x11 and x16 via incubation in a circulating digital water bath set at 55 ˚C. Pre-aliquoted enzymes removed from incubation to ice at set time points of 15, 30, 45, and 60 minutes. Bioluminescence activity screening was immediately conducted after each removal. Measurements performed as detailed in Figure 5.10. Error bars represent the standard error of the mean (SEM). (**A**) Data presented as Relative Luminescence (RLU). (**B**) Data presented as percentage of initial activity without incubation. Statistical analysis performed as detailed in 2.12. Significance groupings are discrete between individual Lucs.

Figure 5.17.



Bioluminescence following thermal inactivation at 60 ˚C. Thermal inactivation of *Phem*Luc, *Ppy*, x11 and x16 via incubation in a circulating digital water bath set at 60 ˚C. Pre-aliquoted enzymes removed from incubation to ice at set time points of 15, 30, 45, and 60 minutes. Bioluminescence activity screening was immediately conducted after each removal. Measurements performed as detailed in Figure 5.10. Error bars represent the standard error of the mean (SEM). (**A**) Data presented as Relative Luminescence (RLU). (**B**) Data presented as percentage of initial activity without incubation. Statistical analysis performed as detailed in 2.12. Significance groupings are discrete between individual Lucs.

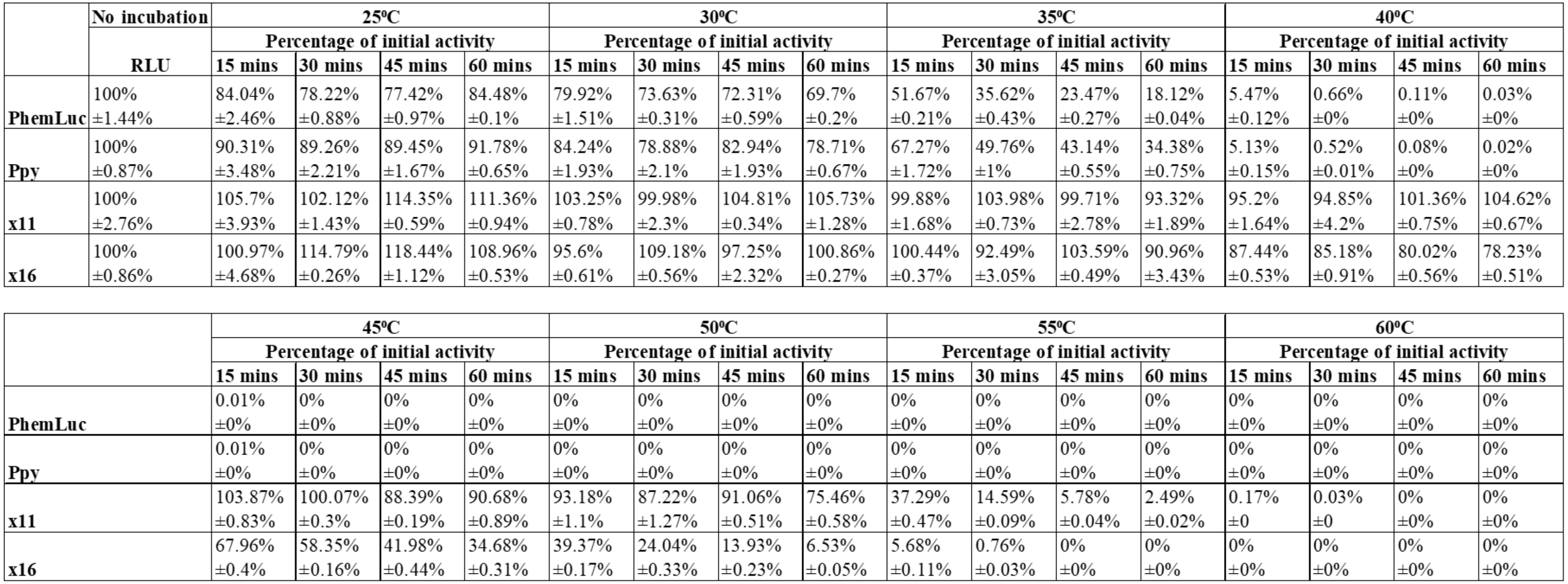
Table 5.3.

	No incubation	25ºC				30ºC				35ºC				40ºC			
		RLU				RLU				RLU				RLU			
	RLU	15 mins	30 mins	45 mins	60 mins	15 mins	30 mins	45 mins	60 mins	15 mins	30 mins	45 mins	60 mins	15 mins	30 mins	45 mins	60 mins
PhemLuc	72649 ±1046.79	61052.33 ±1787.32	56829 ±639.08	56245.33 ±705.34	61376 ±70.41	58061.67 ±1098.42	53492.67 ±225.93	52529.33 ±428.6	50638 ±147.5	37535.33 ±151.92	25880.33 ±310.33	17051.67 ±194.41	13162 ±30.81	3974 ±86.47	482 ±3.47	83.33 ±0.77	18.67 ±0.19
Ppy	68519.33 ±597.62	61880.67 ±2381.79	61158.67 ±1510.97	61289.33 ±1145.16	62887.67 ±448.52	57721 ±1322.12	54050 ±1440.97	56833 ±1322.96	53931 ±458.95	46095.33 ±1176.16	34096.33 ±687.84	29559.33 ±376.55	23555.33 ±516.62	3515.67 ±105.68	355.67 ±8.52	58 ±1.58	17 ±0.76
x11	11829 ±326	12503.33 ±465.18	12079.67 ±169.4	13526.33 ±70.04	13172.67 ±110.93	12213.67 ±92.36	11827 ±272.29	12398 ±40.73	12507 ±151.54	11814.67 ±199.01	12299.67 ±85.79	11794.33 ±328.38	11038.33 ±223.89	11261.33 ±193.72	11219.67 ±496.6	11989.67 ±89.07	12375 ±79.41
x16	4011.33 ±34.51	4050.33 ±187.6	4604.67 ±10.27	4751 ±44.85	4370.67 ±21.25	3834.67 ±24.38	4379.67 ±22.58	3901 ±93.17	4045.67 ±10.64	4029 ±14.79	3710 ±122.18	4155.33 ±19.67	3648.67 ±137.66	3507.67 ±21.42	3416.67 ±36.46	3210 ±22.39	3138 ±20.48

	45ºC				50ºC				55ºC				60ºC			
	RLU				RLU				RLU				RLU			
	15 mins	30 mins	45 mins	60 mins	15 mins	30 mins	45 mins	60 mins	15 mins	30 mins	45 mins	60 mins	15 mins	30 mins	45 mins	60 mins
PhemLuc	7 ±0.35	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0
Ppy	5 ±0.35	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0
x11	12286.33 ±98.08	11837.67 ±35.63	10455.67 ±22.06	10726.67 ±104.88	11022.33 ±130.66	10317 ±150.48	10771.33 ±59.8	8926.67 ±68.97	4411.33 ±55.94	1726 ±10.71	683.33 ±4.57	294.67 ±1.87	20 ±0	3.67 ±0.25	0 ±0	0 ±0
x16	2726 ±15.91	2340.67 ±6.3	1684 ±17.7	1391 ±12.44	1579.33 ±7.01	964.33 ±13.38	558.67 ±9.3	262 ±1.87	227.67 ±4.26	30.33 ±1	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0

Summary of I_{max} data from resistance to thermal inactivation study. Data indicating I_{max} values as Relative Luminescence (RLU) measurements presented in Figure 5.10A-18A. ± values represent the SEM.

Table 5.4.

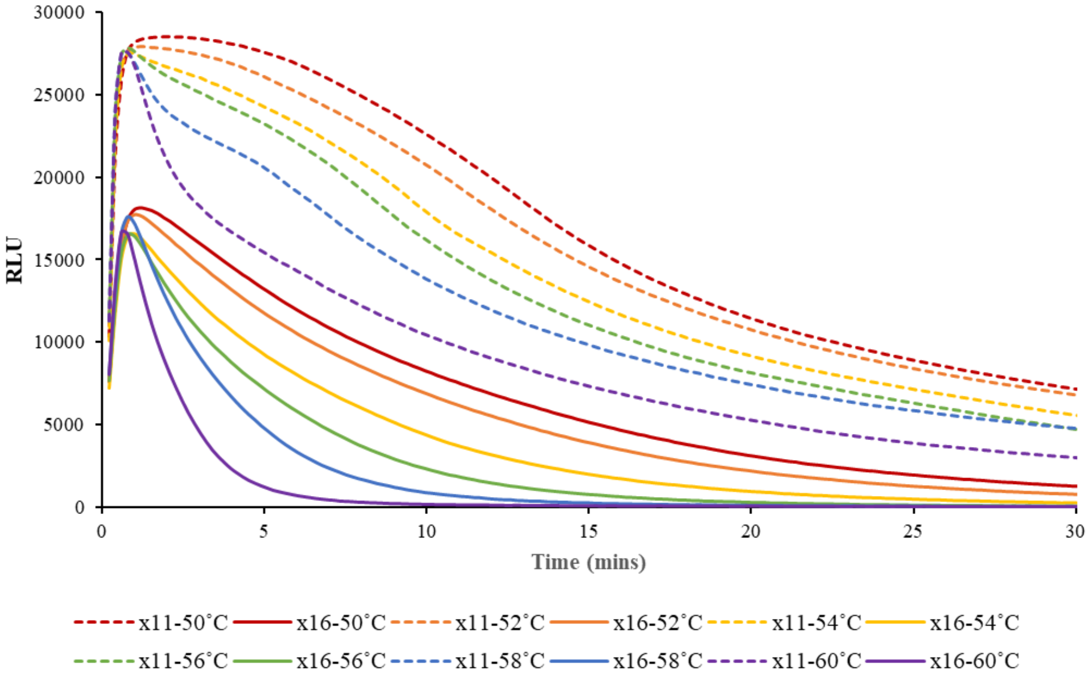


	No incubation	25ºC				30ºC				35ºC				40ºC			
		Percentage of initial activity				Percentage of initial activity				Percentage of initial activity				Percentage of initial activity			
	RLU	15 mins	30 mins	45 mins	60 mins	15 mins	30 mins	45 mins	60 mins	15 mins	30 mins	45 mins	60 mins	15 mins	30 mins	45 mins	60 mins
PhemLuc	100% ±1.44%	84.04% ±2.46%	78.22% ±0.88%	77.42% ±0.97%	84.48% ±0.1%	79.92% ±1.51%	73.63% ±0.31%	72.31% ±0.59%	69.7% ±0.2%	51.67% ±0.21%	35.62% ±0.43%	23.47% ±0.27%	18.12% ±0.04%	5.47% ±0.12%	0.66% ±0%	0.11% ±0%	0.03% ±0%
Ppy	100% ±0.87%	90.31% ±3.48%	89.26% ±2.21%	89.45% ±1.67%	91.78% ±0.65%	84.24% ±1.93%	78.88% ±2.1%	82.94% ±1.93%	78.71% ±0.67%	67.27% ±1.72%	49.76% ±1%	43.14% ±0.55%	34.38% ±0.75%	5.13% ±0.15%	0.52% ±0.01%	0.08% ±0%	0.02% ±0%
x11	100% ±2.76%	105.7% ±3.93%	102.12% ±1.43%	114.35% ±0.59%	111.36% ±0.94%	103.25% ±0.78%	99.98% ±2.3%	104.81% ±0.34%	105.73% ±1.28%	99.88% ±1.68%	103.98% ±0.73%	99.71% ±2.78%	93.32% ±1.89%	95.2% ±1.64%	94.85% ±4.2%	101.36% ±0.75%	104.62% ±0.67%
x16	100% ±0.86%	100.97% ±4.68%	114.79% ±0.26%	118.44% ±1.12%	108.96% ±0.53%	95.6% ±0.61%	109.18% ±0.56%	97.25% ±2.32%	100.86% ±0.27%	100.44% ±0.37%	92.49% ±3.05%	103.59% ±0.49%	90.96% ±3.43%	87.44% ±0.53%	85.18% ±0.91%	80.02% ±0.56%	78.23% ±0.51%

	45ºC				50ºC				55ºC				60ºC			
	Percentage of initial activity				Percentage of initial activity				Percentage of initial activity				Percentage of initial activity			
	15 mins	30 mins	45 mins	60 mins	15 mins	30 mins	45 mins	60 mins	15 mins	30 mins	45 mins	60 mins	15 mins	30 mins	45 mins	60 mins
PhemLuc	0.01% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%
Ppy	0.01% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%
x11	103.87% ±0.83%	100.07% ±0.3%	88.39% ±0.19%	90.68% ±0.89%	93.18% ±1.1%	87.22% ±1.27%	91.06% ±0.51%	75.46% ±0.58%	37.29% ±0.47%	14.59% ±0.09%	5.78% ±0.04%	2.49% ±0.02%	0.17% ±0	0.03% ±0	0% ±0%	0% ±0%
x16	67.96% ±0.4%	58.35% ±0.16%	41.98% ±0.44%	34.68% ±0.31%	39.37% ±0.17%	24.04% ±0.33%	13.93% ±0.23%	6.53% ±0.05%	5.68% ±0.11%	0.76% ±0.03%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%	0% ±0%

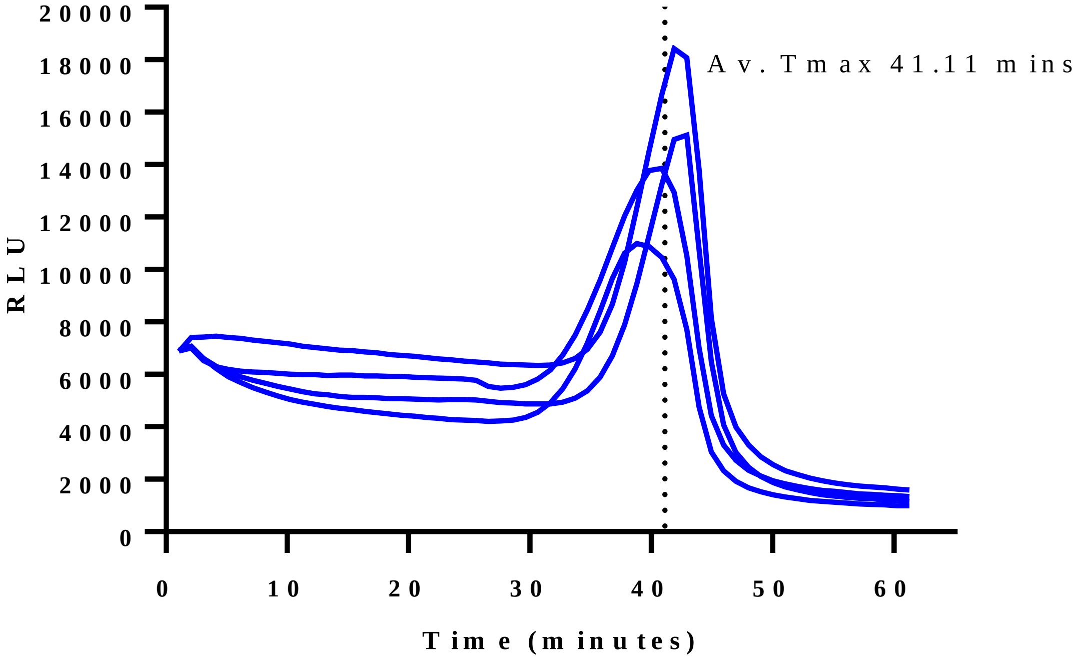
Summary of remaining activities from resistance to thermal inactivation study. Data indicating I_{max} values as percentage of remaining activity presented in Figure 5.10B-18B. ± values represent the SEM converted to a percentage.

Figure 5.18.



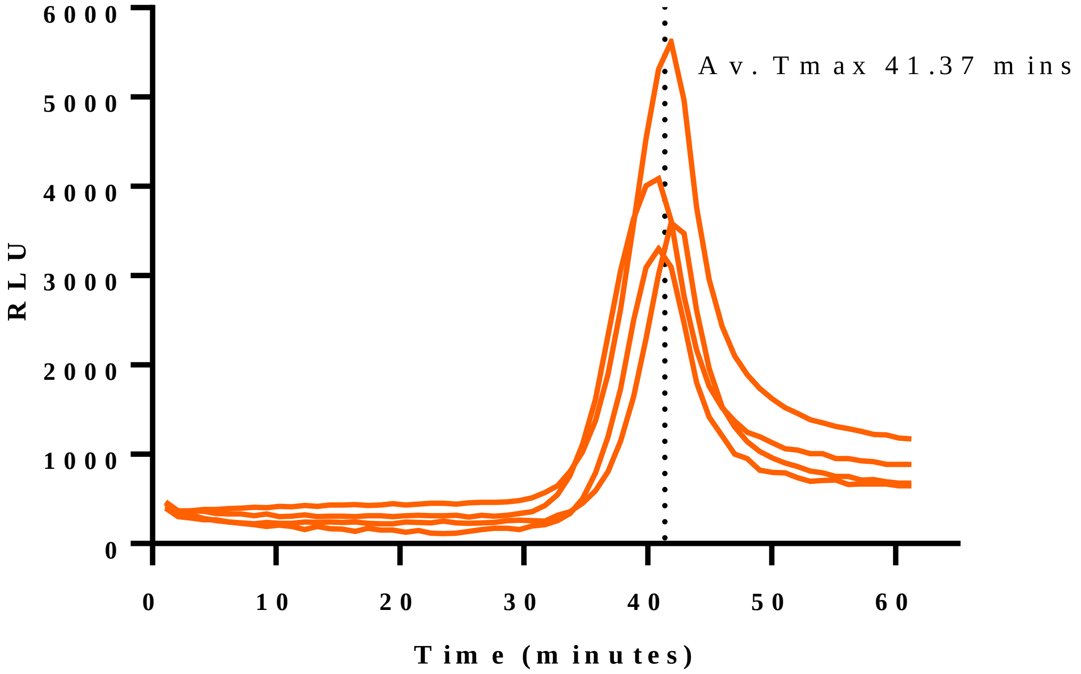
Bioluminescence emission decay between 50-60 ˚C. Bioluminescence activity degradation curves of x11 and x16 from 50-60 ˚C, with an interval of 2 ˚C. Measurements obtained in the LUCY imager (ERBA MDX, Ely, UK) by manual pipetting of substrate mix onto each enzyme such that final concentrations were equal to 1 mM ATP, 500 µM LH_2, and 0.167 µM protein. Each reaction constituent previously diluted in chilled TEM buffer at pH 7.8 (±0.05) and total reaction volume equal to 150 µl. Following substrate injection, each reaction was overlaid with mineral oil before transfer to the heat block and acquisition of the bioluminescence signal. Bioluminescence activity was recorded over a 30-minute duration by integrated 10 seconds of bioluminescence signal every 20 seconds. Assays performed in triplicate, and the curve fitted to each averaged dataset is presented.

Figure 5.19.



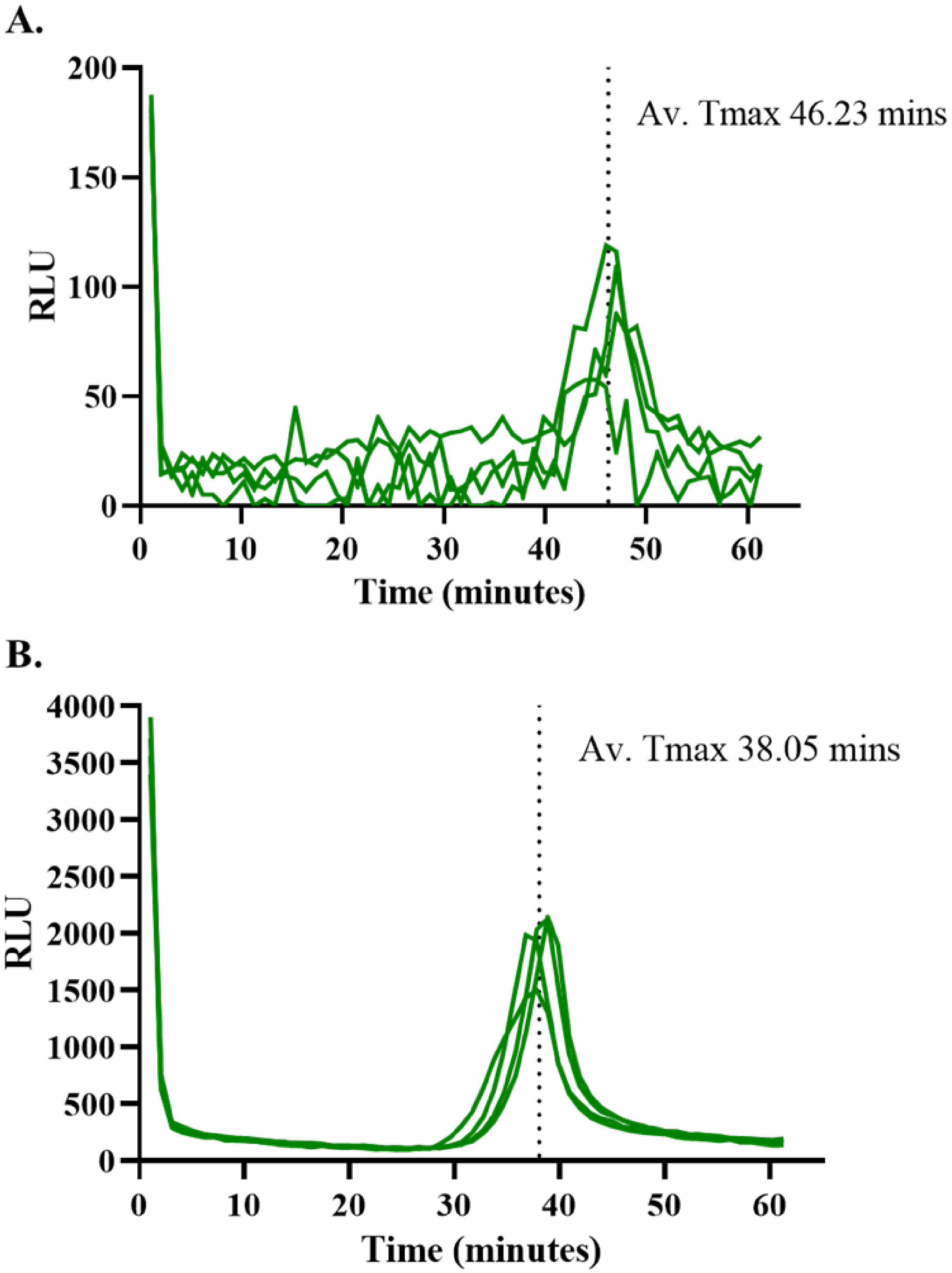
LAMP-BART at 50 ˚C with Ultra-Glo. LAMP-BART assay to demonstrate the time to peak of rLuc Ultra-Glo^{TM} (Promega, Madison, WI, USA) with SARS-CoV-2 RNA as template. Assays were performed such that Ultra-Glo concentration was 5.5 ng/µl. Reactions were set up by manual pipetting, and the LAMP-BART assay recorded in the LUCY, with the heat block set at 50 ˚C (Chapter 2 for further details). Reactions were set up in quadruplicate and Av. T$_{max}$ indicates the average time to reach the bioluminescence emission peak across the four reactions.

Figure 5.20.



LAMP-BART at 50 ˚C with x11. LAMP-BART assay to demonstrate the time to peak of x11 with SARS-CoV-2 RNA as template. Assay performed as detailed in Figure 5.19. Reactions were set up in quadruplicate and Av. T_{max} indicates the average time to reach the bioluminescence emission peak across the four reactions.

Figure 5.21.



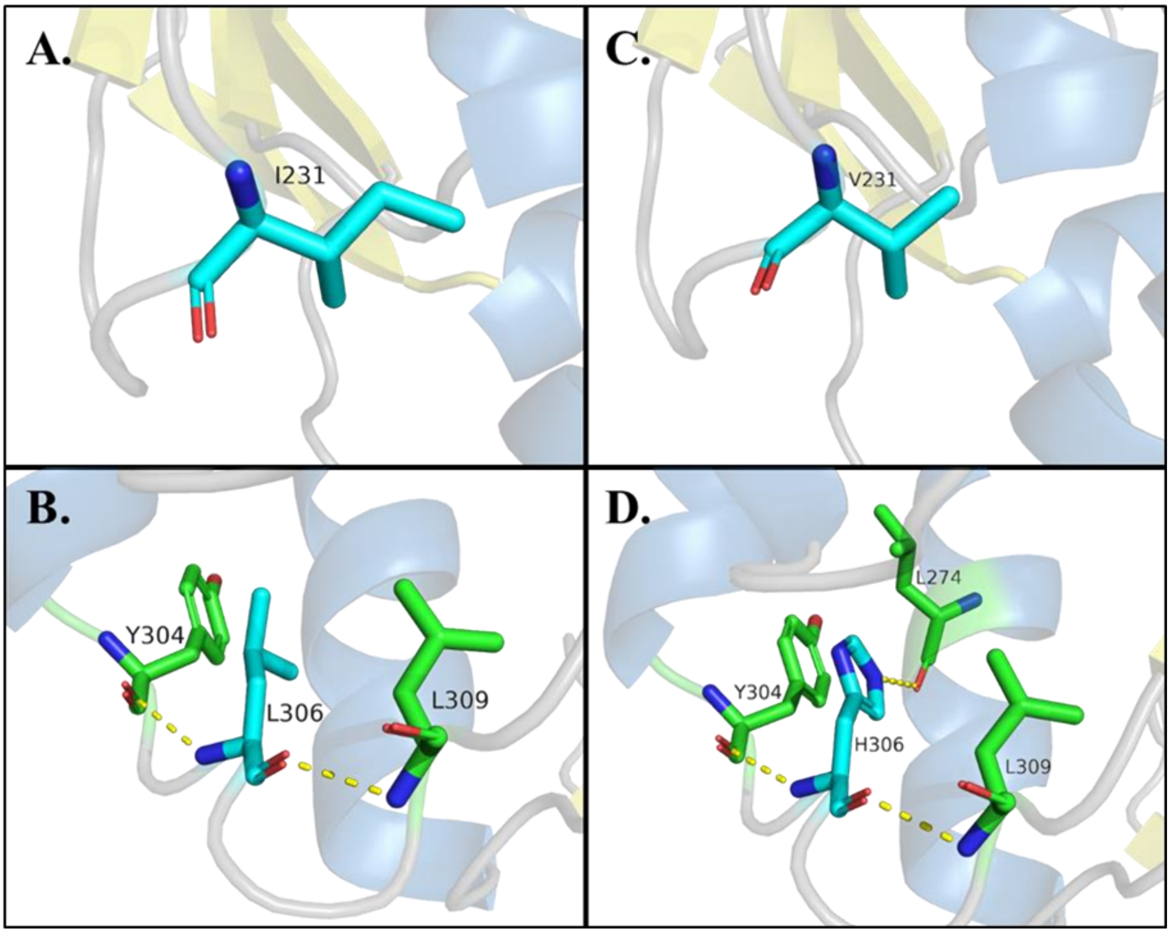
LAMP-BART at 50 ˚C with x16. LAMP-BART assay to investigate the time to peak of x16 with SARS-CoV-2 RNA as template. (**A**) Assay performed as detailed in Figure 5.19. (**B**) Assays performed with a 20x greater on x16, such that final concentration was equal to 110 ng/µl. Reactions were set up in quadruplicate and Av. T_{max} indicates the average time to reach the bioluminescence emission peak across the four reactions.

5.3.5. Modelling of I231V and L306H

A homology model of x16 was constructed following the procedures used to generate the luciferase substrate contact models of Chapter 4. To enable comparison with the previously constructed *Phem*Luc model, the x16 model was similarly constructed using the available crystal structure from PDB file 4G36.pdb of *Ppy* Fluc bound to DLSA in the adenylate-forming conformation (Sundlov et al. 2012). The resulting model was then used in The PyMOL Molecular Graphics System to analyse the novel mutations I231V and L306H relative to the *Phem*Luc *wild-type* in an attempt to propose a molecular mechanism for how they may be enhancing thermostability (Figure 5.22.).

Whilst this analysis provided no insight into how the mutation I231V might be augmenting the thermostability of *Phem*Luc, the mutation L306H was found to possess three polar contacts with proximal residues, relative to the *wild-type* L306 only possessing two. Whilst both L306 and H306 interact with residues Y304 and L309, H306 additionally interacts with L274, due to the presence of the NE2 sidechain interaction centre in histidine which is absent in leucine. Hydrogen bonds between residues in surface loops such as those produced by L306 are known to contribute to protein stability, and this contribution is further increased where a residue makes more than one hydrogen bond. In addition to this the contribution to stability is further increased when hydrogen bonds form between neighbouring turn or loops which are separated from each other in the amino acid sequence (Pokkuluri et al. 2002), such as the bond which forms between H306 and L274.

Figure 5.22.



Polar contact modelling of I231V and L306H. All polar interactions of *Phem*Luc – I231 (**A**), *Phem*Luc – L306 (**B**), x16 – V231 (**C**), and x16 – H306 (**D**). Residues I231, L306, V231, and H306 are displayed in stick form and their carbon backbones are represented as cyan. Their interacting residues are represented as green. Oxygen groups are displayed in red, and nitrogen in blue. Interaction between residues are indicated by yellow dashed lines. Model analysis and imaging performed in PyMOL.

5.4. Further Discussion

This work was undertaken to explore whether the novel luciferase enzyme from the lesser British Glow-worm *Phosphaenus hemipterus* (*Phem*Luc) could be engineered to produce variants with higher resistance to thermal inactivation based on existing mutations from homologous luciferase enzymes, which could be refined for a higher activity phenotype by DNA shuffling and further enhanced through directed evolution. This three stage process of engineering initially involved using documented mutations to generate the same mutations in *Phem*Luc. However, not all mutations advantageous to a particular luciferase are conserved for their action in other homologous luciferases (Kitayama *et al.* 2003; Koksharov and Ugarova 2011b). To account for this, DNA shuffling was implemented to select for a subset of mutations displaying improved thermal stability compared to the primary x15 mutant and *wild-type Phem*Luc, prior to the search for novel mutations by directed evolution in the final stage of engineering.

The luciferase enzymes *Phem*Luc and *Ppy* Fluc share 87.07% protein sequence identity, as indicated by the primary structural analysis conducted in Figure 4.2. Decades of research have sought to identify variants of *Ppy* Fluc through mutagenesis which possess improved thermal stability, and from these efforts a vast assortment of mutagenic target have been identified throughout the protein. To take advantage of this existing knowledge, the x15 variant was synthesized containing the fifteen mutations from Table 5.1, which originated from published studies of *Ppy* Fluc. As previously discussed, mutations which are advantageous to a particular luciferase may not produce comparable effects in other homologous enzymes, so mutations selected were limited to those with known advantageous effects in *Ppy* Fluc, on the assumption that the considerable conservation between the two enzymes increased the probability of any advantageous properties being retained. Of the fifteen mutations, 12 had previously been incorporated into a thermostable variant of *Ppy* Fluc (Jathoul *et al.* 2012), indicating that the majority of mutations selected had been demonstrated to have additive effects when deployed in an appropriate enzyme background.

Screening of bioluminescence activity of *Phem*Luc and all thermostable variants was conducted following the same principles of Chapter 4, adapted from the original luciferase screening method used by Wood and Deluca (1987) – see Chapter 2 for specific outline of screening strategies. Screens were either conducted immediately following the induction period to reflect activity at room temperature, or following incubation for 1-hour at 50 ˚C prior to screening the remaining bioluminescent activity with LH_2, in order to demonstrate the degree to which the enzymes could withstand thermal inactivation. The primary screen of the x15 enzyme was conducted at RT for an initial verification of the compatibility of the fifteen mutations with the bioluminescence function of *Phem*Luc, prior to the investigation of any thermostabilising properties. This primary screen demonstrated that the x15 enzyme had significantly diminished bioluminescence compared to *Phem*Luc at room temperature (Figure 5.2.). This result was taken to indicate that one or more of the fifteen mutations introduced into *Phem*Luc was inhibitory to the enzymes bioluminescent function and would need to be reverted in order to restore a high yield bioluminescence phenotype which could be further investigated for the production of thermostable variants.

Prior to this work, DNA shuffling had not been commonly utilized in efforts to engineer firefly luciferases. Previously, a chimeric luciferase enzyme containing the N-domain of recombinant *Ppy* Fluc and the C-domain of a recombinant luciferase from *Luciola italic* was demonstrated to produce enhanced bioluminescence that exceeded the sum of the contributions from the two luciferases (Branchini *et al.* 2014). With the optimized DNA shuffling method described here, it may be possible to develop more complex chimeric luciferases and expand the phenotypic variation that can be sourced from nature in the development of engineered variants, through advanced methods of DNA shuffling, such as family shuffling (Crameri *et al.* 1998; Kaper *et al.* 2002).

In order to conduct engineering efforts by rational design an understanding of structure-function relationships is a necessary perquisite. However, it has long been established that mutations which confer a benefit to the thermostability of a luciferase enzyme can occur on the protein surface or within the core (Tisi *et al.* 2002b), making thermostability an ideal property to engineer through directed evolution methodologies which are capable of non-targeted mutagenesis throughout the entire sequence. In this study, epPCR performed with $MnCl_2$ and D_2O were used to further enhance thermostability of the x14 enzyme, leading to the identification of two novel mutations, I231V and L306H. Whilst both of these positions are conserved in *Ppy* Fluc, no literature exists which specifically related to position I231, which occurs in a surface loop structure comprising residues 223 – 235. However, in a previous study by Viviani *et al.* (2007) residues Y227 and N229 of this loop were discovered to be buried in the protein core, fixing the loop to other structural elements participating at the bottom of the luciferin binding site. Mutagenesis of other residues within this loop was speculated to expose the active site through disrupting the interactions of these structural elements, and therefore this loop has been proposed to act as a solvent gate for the active site (Viviani *et al.* 2007). Therefore, I231V may act by a slight reduction of polarity at the surface and/ or improvement of the coordination of luciferin. Previous studies have shown that substituting surface residues of *Ppy* Fluc with less hydrophobic amino acids produces enzyme variants with greater resistance to thermal inactivation, postulated to be linked to increasing the structural stability through establishment of more favourable local interactions (Law *et al.* 2002; Law *et al.* 2006). The position L306 occurs on another surface loop structure and has previously been investigated through mutagenesis to cysteine to form disulfide bridges in an effort to develop a secreted luciferase (Nazari and Hosseinkhani 2011). As seen in Figure 5.22., L306H was found to produce an additional polar interaction with a neighbouring alpha helix residue. Whether this alone can explain the thermostabilising effect of this mutation is uncertain, but further explanations for the improvement from L306H may be provided by future investigation of whether electrostatic stabilisation of phosphates or altered active site H-bond networks could be contributing to improved coordination of ATP and AMP. The combination of I231V and L306H had an additive effect which exceeded the sum of their individual improvements to the activity of x14, when assessed in the final x16 variant. The effects of these mutations individually and as a dual-mutant in the *wild-type Phem*Luc would need to be investigated in order to verify whether their thermostabilising properties are retained, or if they are dependent on the presence of the prior mutations present in x14 *Phem*Luc. As these positions are conserved, it may also be worth investigating the effect of these mutations in *Ppy* Fluc and its own engineered variants, including the pH tolerant, thermostable variant x11.

In the process of development and screening in *E. coli,* x16 consistently displayed greater bioluminescence than *Phem*Luc, such that the activities at RT and following 50 ˚C incubation were 119.47% and 929.44% of the respective activity in *Phem*Luc (Figure 5.9.). Unexpectedly, whilst thermostability of x16 remained in the purified enzyme thermal inactivation studies, the initial activity at RT was only measured at 5.52% of the activity recorded for *Phem*Luc (Table 5.3.). However, this reduction in bioluminescent yield from a thermostable variant is in line with the result obtained for x11, which displayed 17% of the activity observed in *Ppy* Fluc. It may be that the increased signal observed during the *in vivo* screens is linked to improved protein stability or reduced turnover in the *E. coli* intracellular environment.

The lower thermostability of PhemLuc than *Ppy* Fluc, was reflected in the engineered variant of *Phem*Luc, x16, being less thermostable than the *Ppy* Fluc variant, x11. From the water bath thermal inactivation assays the activity of x16 appeared to decline at temperatures of 35 ˚C and above, with a significant reduction in activity across all time points observed at 50 ˚C. The second thermal inactivation study was set up in response to the activity drop off observed at 50 ˚C, to mimic the conditions of incubation within a test tube assay and continuously followed the activity degradation between 50-60 ˚C, as this temperature range would be relevant in assessing the utility of x16 in LAMP-BART. Although a rapid initial decline in bioluminescence activity was recorded at all temperatures for x16, the activity decay across the entire assay duration mimicked a radioactive half-life effect, where the rate of activity decay diminished with time, allowing bioluminescence activity to persist throughout the 30-minute experiment duration (Figure 5.18.).

Whilst the steady state glow kinetic of a commercial enzyme like Ultra-Glo would be an advantage in LAMP-BART, the persistence of bioluminescence activity from x16 would be sufficient for utilization in a trial assay. Although no precedence exists for conducting a similar low-temperature LAMP-BART assay, a decision was made on the advice of Dr. Patrick Hardinge to perform a trial at 50 ˚C against an x11 and Ultra-Glo control, to permit the greatest possible bioluminescence activity from x16, whilst also allowing the remaining LAMP-BART constituents which are optimized for higher temperature conditions to function correctly.

The LAMP-BART assays performed using SARS-CoV-2 RNA as template with both Ultra-Glo and x11 were successfully capable of forming an emission peak. A lower quality emission peak was also recorded with the x16 enzyme, i.e. it occurred later than registered for either Ultra-Glo or x11, and produced less bioluminescence signal throughout the duration of the assay (Figure 5.21.). This delayed and low level emission peak was attributed to insufficient remaining bioluminescence activity from the x16 enzyme, which degraded during the assay incubation. To correct for this, a second LAMP-BART assay was conducted using a 20x greater concentration of the x16 enzyme. This additional supplementation provided sufficient enzyme activity without inhibiting any other process, such that higher levels of bioluminescence were sustained throughout the duration of the assay, and a higher quality emission peak was produced, occurring earlier in the reaction.

Whilst this LAMP-BART trial was conducted to demonstrate the utility of x16 at the limit of its thermostable potential, without the steady state glow kinetics attributed to Ultra-Glo, the derivation of template copy number using the time to emission peak would be highly inaccurate and therefore the downstream applications of LAMP-BART such as the detection of GMO contamination would not be possible (Kiddle *et al.* 2012). The x16 enzyme was developed using only a single round of shuffling and directed evolution by epPCR. It is highly likely that further improved variants await to be discovered in the novel *Phem*Luc enzyme, using subsequent rounds of DNA shuffling and the directed evolution techniques utilised throughout this work. Additionally, exploration of deletion mutagenesis may enable further improvements to thermostability, as has previously been demonstrated in the x11 Fluc (Halliwell *et al.* 2018). As it stands with the degree of thermostability demonstrated by x16 in its current state, exploring its viability as a reporter for *in vivo* bioluminescence imaging may be promising.

5.5. Conclusions

The thermostability of *Phem*Luc was demonstrated to be improved by the incorporation of multiple thermostabilising mutations identified in previous luciferase engineering studies. The application of DNA shuffling on a firefly luciferase gene was demonstrated, and was proven as an effective strategy for the isolation of improved activity reversion mutants by the identification of x14. However, the theoretical screen size required to ascertain that all sequence variants have been sampled limited its usage to isolating an improved mutant relative to the original x15, rather than a definitive ‘optimum’ enzyme from the theoretical variant pool. The application of DNA shuffling in firefly luciferase engineering efforts exceeds the usage demonstrated here, and promotes the exploration of more advanced methods such as family shuffling to generate chimeric luciferase libraries. Mutagenesis of *Phem*Luc x14 with the dual-mutagen approach of $MnCl_2$ and D_2O facilitated the discovery of two novel thermostabilising mutations, I231V and L306H. Whilst the underlying mechanisms by which these mutations increased thermotolerance is unknown, an investigation of their independent activities in *Phem*Luc and *Ppy* Fluc may enable further understanding. The final x16 mutant of *Phem*Luc was capable of functioning in a modified LAMP-BART assay using SARS-CoV-2 RNA, performed at a reduced temperature. Whilst x16 is currently outperformed in thermostability by enzymes including x11 and Ultra-Glo, it remains a potential candidate for further improvement by continued application of the engineering strategies practised here.

Chapter 6

Biochemical Characterisation of *Wild-type* and Engineered Variants of Firefly Luciferases

6.1. Chapter Summary

This Chapter builds on the previous work to bioprospect for a novel luciferase gene from museum Coleoptera and to develop variants of the novel luciferase from *Phosphaenus hemipterus* which possess improved activity with a synthetic substrate analogue or resistance to thermal inactivation. These engineered variants were developed with screening strategies chosen to select specifically for a single characteristic of interest, and therefore the underlying biochemical properties had not been assessed, along with any secondary advantageous characteristics that may have emerged. Therefore, in this Chapter all firefly luciferases of the study were overexpressed, purified, and concentrations normalised prior to investigation of bioluminescence spectra, specific activity, pH-tolerance and enzyme kinetics.

6.2. Introduction

Firefly luciferases and the bioluminescence activity which they possess have enabled the development of a wide range of applications, which commonly include their use as a reporter gene to follow cells *in vivo*, the expression from a gene of interest, or as a high sensitivity ATP detection system (Kuzikov *et al.* 2003; Noguchi and Golden 2017). In order to advance these applications a constant requirement exists for novel enzyme properties and improved characteristics, whether these discoveries arise from natural sources or mutagenic exploration by means of rational design and directed evolution. The previous work of this study sought to identify novel *wild-type* Flucs, in addition to developing engineered Fluc variants which possessed improved bioluminescence activity with a synthetic substrate analogue or improved resistance to thermal inactivation. However, the core enzymatic properties of the bioprospected Costa Rican firefly luciferase (CRLuc) and the novel luciferase from the lesser British Glow-worm *Phosphaenus hemipterus* (*Phem*Luc) remain unknown. Additionally, the engineered variants of *Phem*Luc, x2 Infra and x16, were developed using engineering strategies which did not directly control for how enzyme properties not under selection pressure could be affected. Without the ability to control for all enzyme properties whilst selecting for a primary characteristic of interest, all other enzyme properties can be subject to change, which may ultimately be a disadvantage, but could also enable the discovery of additional secondary properties of interest such as pH-tolerance and improved substrate affinities.

The aims of this chapter are firstly to determine and compare the properties of the bioprospected luciferase from the unidentified Costa Rican firefly (CRLuc) and the luciferase of the lesser British Glow-worm, *Phosphaenus hemipterus* (*Phem*Luc) to the well characterised luciferase from the North American firefly, *Photinus pyralis* (*Ppy* Fluc). The second aim is to complete the biochemical characterisation for the engineered variants of *Phem*Luc, x16 and x2 Infra in comparison to their *wild-type* origin and the *Ppy* Fluc thermostable and pH-tolerant variant termed x11. To carry out characterisation of their biochemical properties, all enzymes were overexpressed, purified, and concentrations normalised prior to investigation of bioluminescence spectra, specific activity, pH-tolerance and enzyme kinetics.

6.3. Results and Discussion

6.3.1. Overexpression and purification of *wild-type* and engineered variants

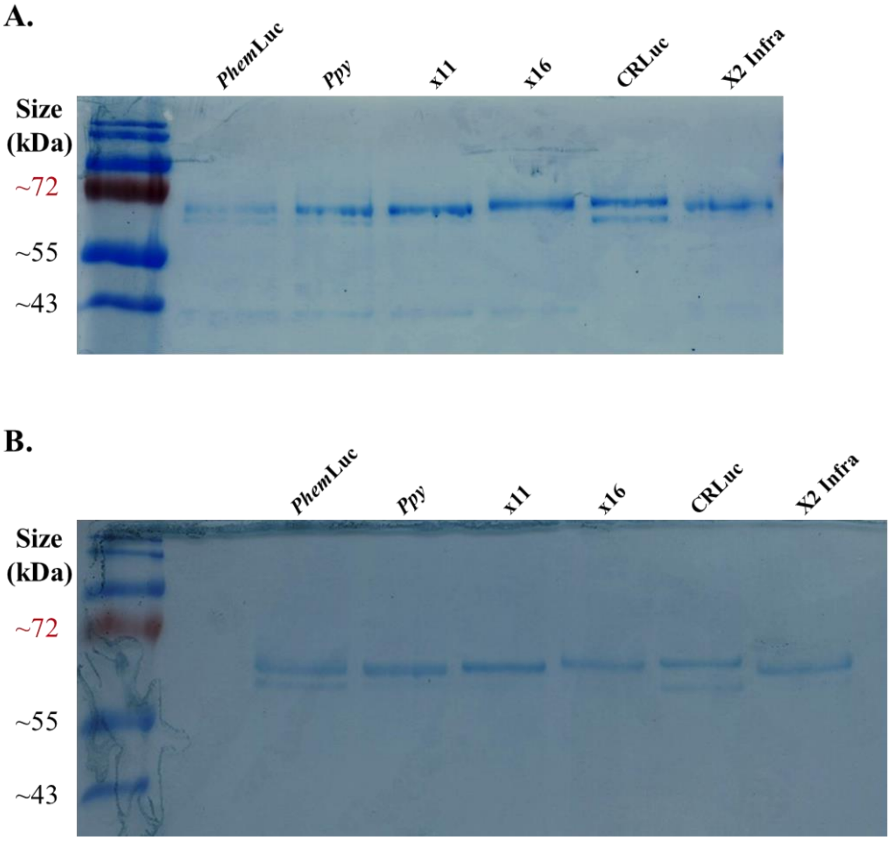
To attain pure protein, all enzymes were overexpressed in *E. coli* BL21 liquid cultures. Purification was performed by exploiting the affinity between an N-terminal 10x His-tag introduced into each protein, and nickel-nitrilotriacetic acid (Ni-NTA) resin. Detailed purification methods are discussed in Chapter 2, as previously conducted for recombinant firefly luciferases (Law *et al.* 2006). Fractions of bound protein were eluted by iteratively increased concentrations of imidazole (IMD) washes between 50 mM-500 mM. These elutions were assessed for bioluminescence activity by saturation of 50 µl aliquots with LH_2 and ATP and subsequent screening in the PhotonIMAGER Optima (Biospace Labs, Paris, France). The three highest bioluminescence activity protein fractions for each sample were desalted into storage buffer (Chapter 2) using disposable PD10 desalting columns (GE Healthcare, WI, USA). The three retained desalted fractions were combined and homogenized before immediately storing at -80 ˚C in pre-labelled aliquots. The concentration of each protein was determined via Bradford assay (Table 6.1.). These concentrations varied between 0.057 mg/ml to 1.17 mg/ml, and were subsequently analysed by diluting to the lowest recorded concentration of 0.057 mg/ml and analysed by SDS-PAGE. The primary quantification of proteins by SDS-PAGE (Figure 6.1A.) exhibited greater than anticipated variation of band intensity and secondary product bands for some of the samples, likely the cause for variation from expected band intensity. Subsequent ImageJ (Schneider *et al.* 2012) analysis of band intensity was conducted to correct the concentrations for the protein band of interest and SDS-PAGE was reperformed with the modified dilution factors (Figure 6.1B.). Analysis of the final gel by ImageJ showed the band intensities of the luciferases to be comparable, and all proteins were advanced to characterisation of the biochemical properties.

Table 6.1.

Purified and desalted Fluc	***Phem*Luc**	***Ppy* Fluc**	**x11**	**x16**	**CRLuc**	**x2 Infra**
Size (KD)	60.8	60.75	60.56	60.93	60.71	60.84
Bradford (mg/ml)	0.398	0.82	0.974	1.17	0.057	0.087
Bradford (µM)	6.55	13.51	16.09	19.19	0.93	1.44
SDS-PAGE corrected (mg/ml)	0.21	0.566	1.026	1.419	0.057	0.077
SDS-PAGE corrected (µM)	3.45	9.32	16.94	23.3	0.93	1.26

Summary of average protein concentration. The concentration of each PD10 desalted purified protein as determined via Bradford assay, and subsequent corrections from imageJ analysis of SDS-PAGE scans (Chapter 2). Fluc sizes in KD are displayed as calculated from protein sequences. Concentrations are displayed in mg/ml and the corresponding µM concentration are shown here shaded.

Figure 6.1.

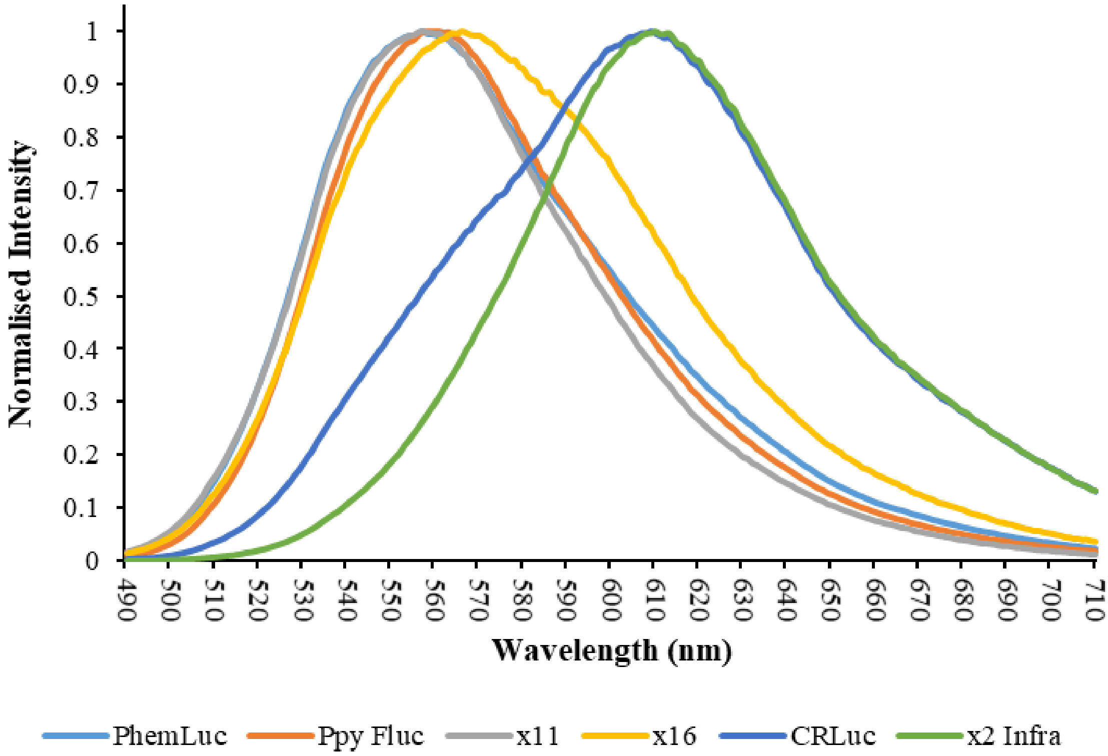


SDS-PAGE analysis for protein quantification. All Flucs prepared by diluting as described in A) or B), and mixing 3:1 in 4x protein sample buffer. **A)** All Flucs diluted to equal the lowest concentration as measured by Bradford assay (CRLuc – 0.057 mg/ml). **B)** All Flucs diluted to equal CRLuc concentration following correction by imageJ analysis of band size and intensity (Chapter 2). Both images edited by -40% brightness and +40% contrast for enhanced visualization. A non-edited equivalent figure of the original gel images is available in appendices Figure 9.4

6.3.2. Bioluminescence spectra of *wild-type* and engineered variants

High resolution bioluminescence spectra of purified enzymes were measured using the CLARIOstar Plus Microplate reader (BMG LABTECH, Ortenberg, Germany) in the presence of saturating conditions of LH_2 and ATP. All enzyme and reagent dilutions were made in pH 7.8 (±0.05) TEM. The CLARIOstar possesses a monochromator which allows for the collection of spectra with a resolution of up to 1 nm. The bioluminescence spectra λ_{max} for the control enzymes *Ppy* Fluc and x11 were ca. 558 nm and ca. 557 nm, respectively. These measurements were in broad agreement with the recorded λ_{max} for both of these enzymes from available literature (Jathoul *et al.* 2012), and thus were taken to indicated that the spectral properties recorded for the remaining enzymes were similarly accurate. The spectra of CRLuc exhibited a red-shifted emission peak (λ_{max} = ca. 609 nm) (Figure 6.2.) with a broad bandwidth (FWHM = 95 nm) (Table 6.2.) extending toward the green region. The previously uncharacterised enzyme *Phem*Luc showed similar bioluminescence spectral properties (λ_{max} = ca. 557 nm) to the *wild-type* control *Ppy* Fluc. The engineered thermostable variant x16 exhibited a subtle redshift to the emission peak (λ_{max} = ca. 566 nm), with a broader bandwidth (FWHM = 88 nm) extending toward the red region. *Phem*Luc variant x2 Infra which had been rationally engineered for improved compatibility with infraluciferin was shown to exhibit a significant bathochromic shift to the emission peak (λ_{max} = ca. 610 nm) with LH_2, whilst also retaining a similarly narrow FWHM (78 nm vs 77 nm for *Phem*Luc).

Figure 6.2.



Bioluminescence spectra of *Phem*Luc, *Ppy*, x11, x16, CRLuc, and x2 Infra with LH_2. Measurements obtained in a luminometer by injection of substrate mix onto each enzyme solution such that final assay concentrations were equal to 1 mM ATP, 500 µM LH_2, and 0.167 µM protein. Each reaction constituent was previously diluted in chilled TEM buffer of pH 7.8 (±0.05) and total reaction volume was equal to 150 µl. Following substrate injection, each reaction held at RT for 30 seconds prior to acquisition of spectra. Light emissions were integrated over 2 seconds for 221 wavelength scanpoints between 490 nm and 710 nm, with a stepwidth of 1 nm. Assays were performed in triplicate and averaged data presented. Data is normalised such that each point is presented as intensity relative to λ_{max}.

Table 6.2.

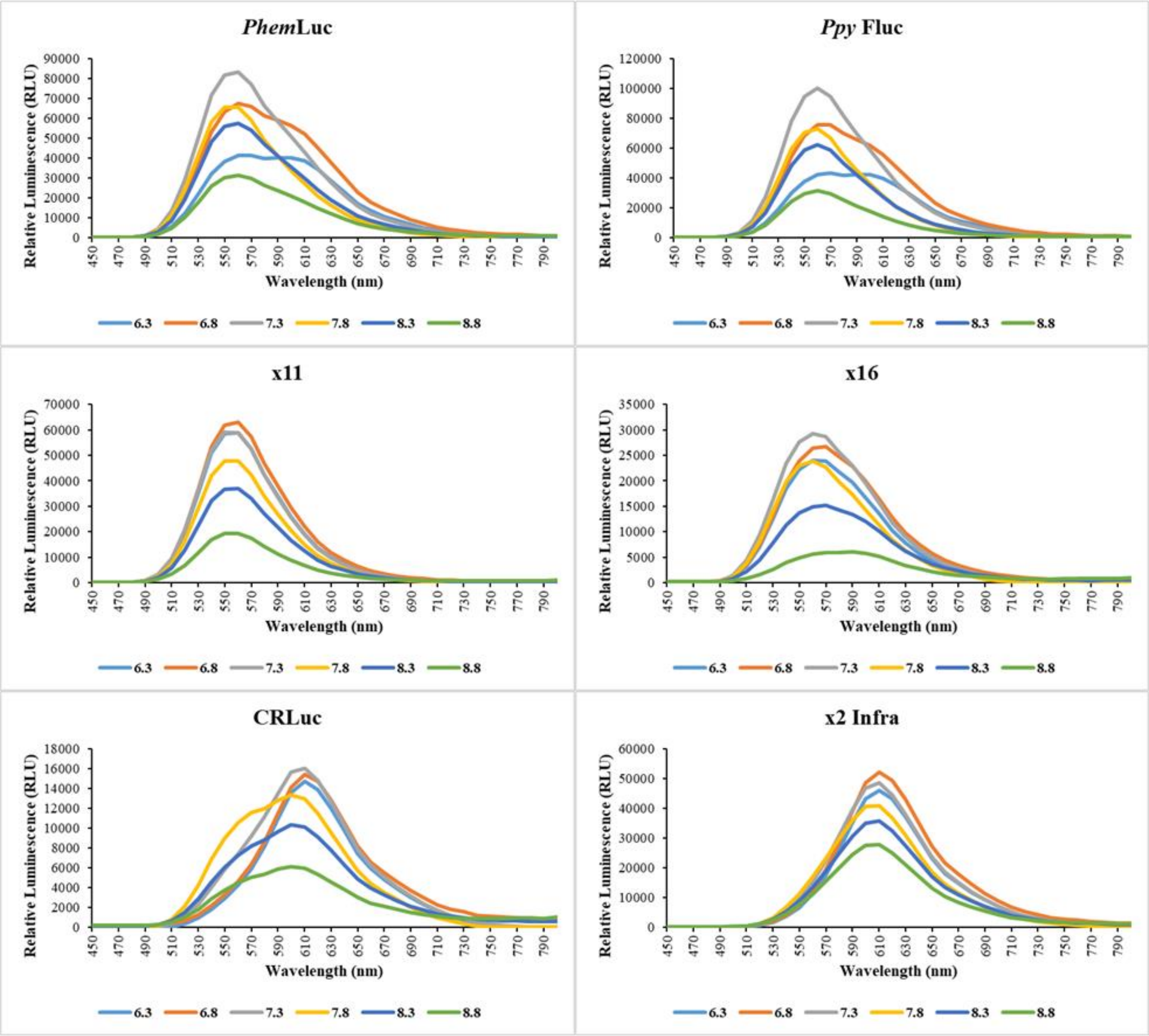
	***Phem*Luc**	***Ppy* Fluc**	**x11**	**x16**	**CRLuc**	**x2 Infra**
λ_{max} (nm)	557	558	557	566	609	610
FWHM (nm)	77	72	72	88	95	78

Summary of bioluminescence λ_{max} and FWHM with LH_2. Values derived from data presented in Figure 6.3., where data has been normalised and smoothed within Microsoft Excel. Average values across replicates shown. Full Width Half Maximum (FWHM) is the width of spectra measured between the two opposite points of the curve at half maximum intensity. Experimental conditions are as indicated in the legend of Figure 6.3. Shading in the background of λ_{max} values is indicative of respective visual colour.

6.3.3. pH dependence of bioluminescence spectra of *wild-type* and engineered variants

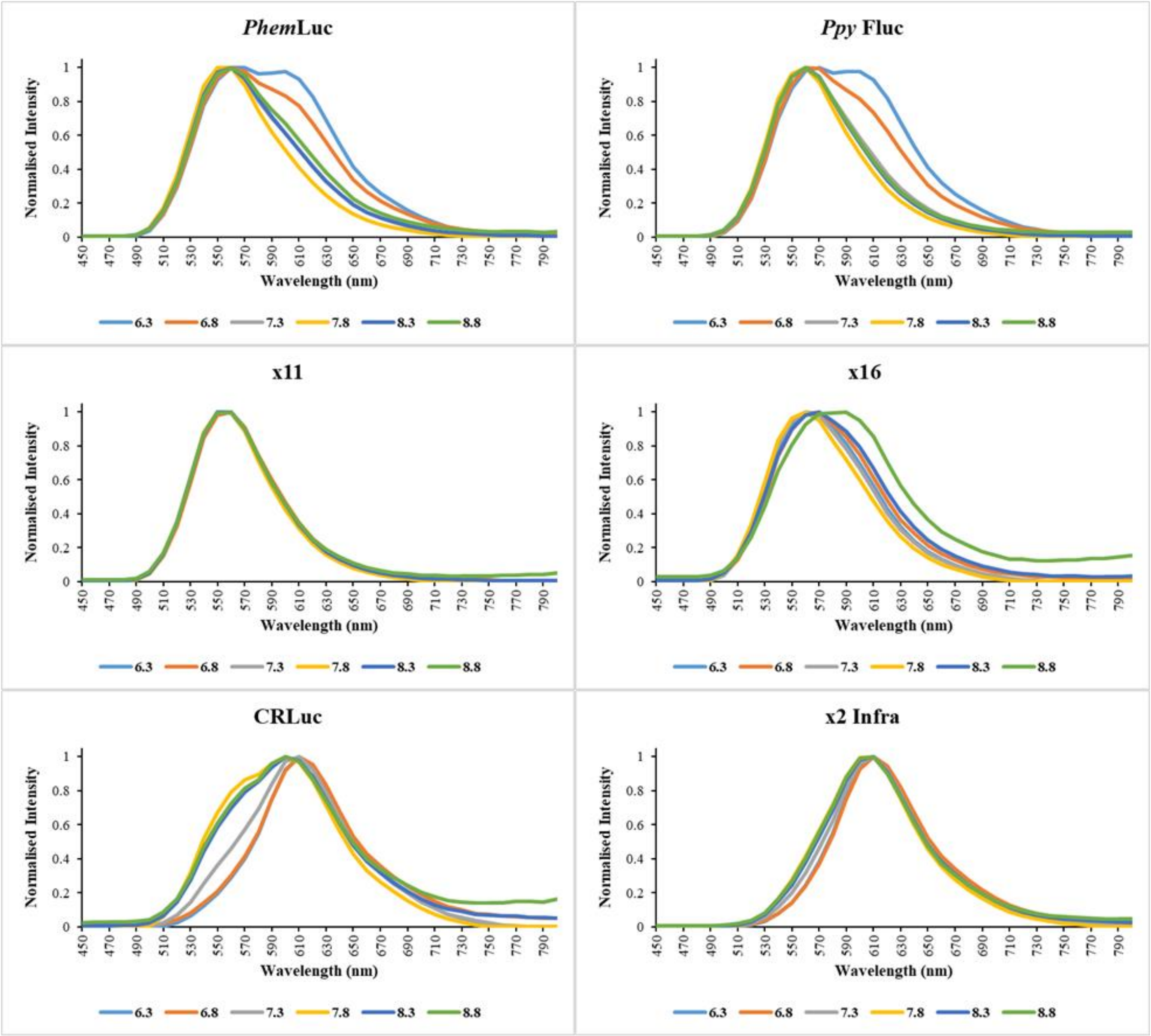
The bioluminescence spectra for all enzymes were recorded across a range of pH conditions at a resolution of 10 nm (Figure 6.3.-6.4. and Table 6.3.). Amongst the *wild-type* enzymes, *Phem*Luc and *Ppy* Fluc were found to have extremely similar responses of their spectral properties to variation in pH, at all conditions assessed. However, where *Phem*Luc and *Ppy* Fluc were observed to exhibit a bathochromic shift under acidic conditions, CRLuc in contrast appeared to display more stability at lower pH, but was hypsochromic shifted under alkali conditions. The engineered variants of *Phem*Luc were both shown to possess improved resistance to spectral shifts across the range of pH surveyed. x16 exhibited a small bathochromic shift at the most alkali pH condition assessed, but also had a lower percentage of integrated activity under the higher pH conditions than *wild-type Phem*Luc (Table 6.3.). x2 Infra was shown to have significant tolerance to bathochromic or hypsochromic shift across all pH conditions assessed, exhibiting a similar profile to x11 which is a more complex enzyme specifically engineered for pH tolerance. Both x2 Infra and x11 were shown to produce their minimum integrated activity from bioluminescence spectra at the most alkali condition tested, pH 8.8. However, x2 Infra retained 58% of its maximum integrated activity at this condition, in contrast to a 32% retention observed for x11.

Figure 6.3.



pH dependence of bioluminescence spectra from *Phem*Luc, *Ppy*, x11, x16, CRLuc, and x2 Infra with LH_2. Measurements obtained in a luminometer by injection of substrate mix onto each enzyme such that final assay concentrations were equal to 1 mM ATP, 500 µM LH_2, and 0.167 µM protein. Each reaction constituent was previously diluted in chilled TEM buffer of the appropriate pH (±0.05) and total reaction volume was 150 µl. Following substrate injection, each reaction was held at RT for 30 seconds prior to acquisition of spectra. Light emissions were integrated over 2 seconds for 36 wavelength scanpoints between 450 nm and 800 nm, with a stepwidth of 10 nm. Assays performed in triplicate for each pH condition and averaged data presented.

Figure 6.4.



pH dependence of normalised bioluminescence spectra from *Phem*Luc, *Ppy* Fluc, x11, x16, CRLuc, and x2 Infra with LH_2. Measurements obtained in a luminometer by injection of substrate mix onto each enzyme such that final assay concentrations were equal to 1 mM ATP, 500 µM LH_2, and 0.167 µM protein. Each reaction constituent was previously diluted in chilled TEM buffer of the appropriate pH (±0.05) and total reaction volume was 150 µl. Following substrate injection, each reaction was held at RT for 30 seconds prior to acquisition of spectra. Light emissions were integrated over 2 seconds for 36 wavelength scanpoints between 450 nm and 800 nm, with a stepwidth of 10 nm. Assays were performed in triplicate for each pH condition and averaged data presented. Data normalised such that each point is presented as intensity relative to λ_{max}.

Table 6.3.

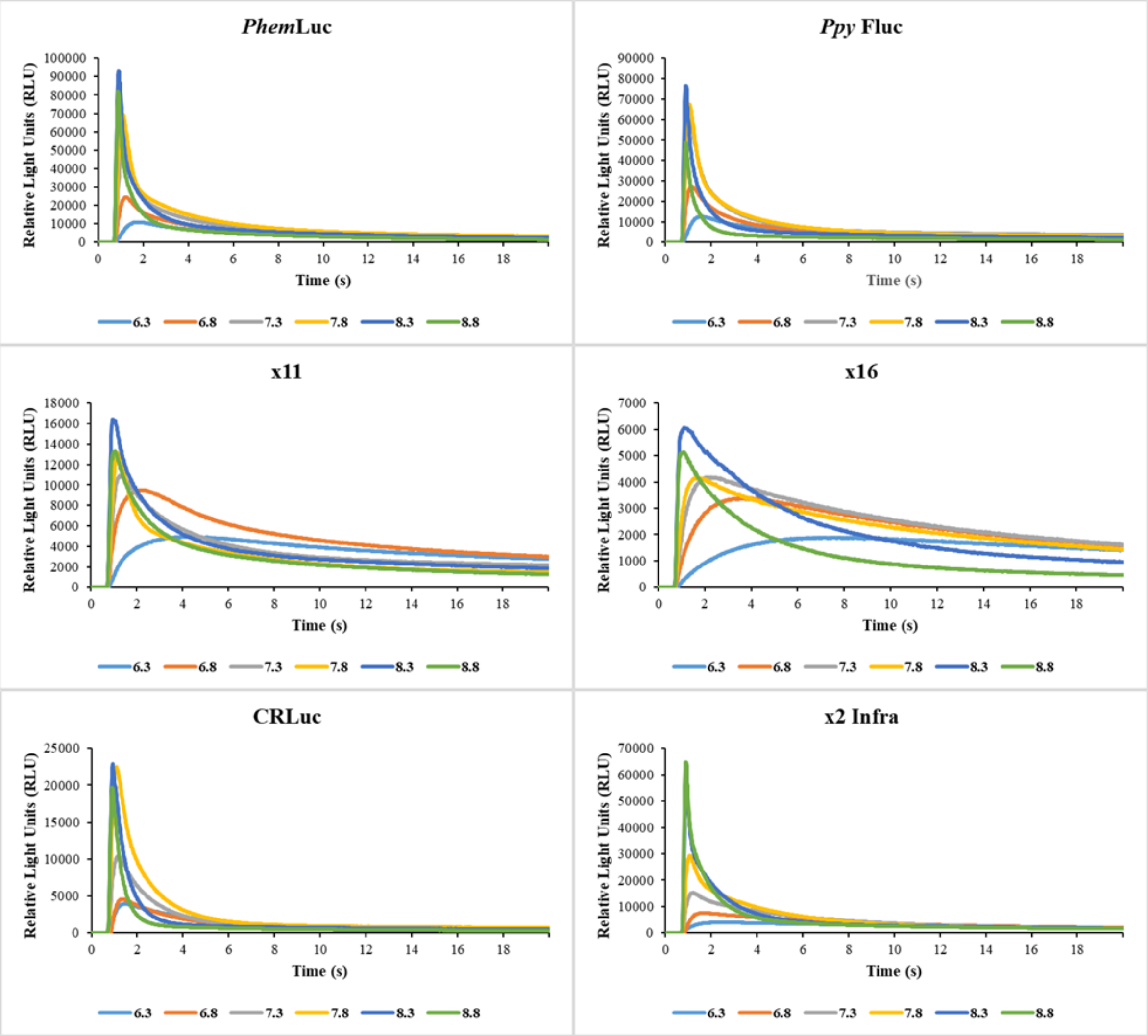
	pH 6.3		**pH 6.8**		**pH 7.3**		**pH 7.8**		**pH 8.3**		**pH 8.8**	
	Integrated Activity (RLU)	**Percentage of Max**	**Integrated Activity (RLU)**	**Percentage of Max**	**Integrated Activity (RLU)**	**Percentage of Max**	**Integrated Activity (RLU)**	**Percentage of Max**	**Integrated Activity (RLU)**	**Percentage of Max**	**Integrated Activity (RLU)**	**Percentage of Max**
***Phem*Luc**	512066.33	65.69%	770467.67	98.84%	779535.33	100%	561406	72.02%	541031.67	69.4%	316961.67	40.66%
***Ppy* Fluc**	516596.67	59.04%	816031.67	93.27%	874928.33	100%	583578	66.7%	528487	60.4%	276993.67	31.66%
x11	456472.67	89.91%	507698.33	100%	458200.33	90.25%	366433.67	72.18%	295790.33	58.26%	164350.33	32.37%
x16	221860.33	82.61%	263803.67	98.23%	268557	100%	208622	77.68%	158727.33	59.1%	76562.33	28.51%
CRLuc	130963.67	82.55%	150919.67	95.13%	158640	100%	147299.33	92.85%	120633.67	76.04%	79124	49.88%
x2 Infra	404276.33	85.04%	475404	100%	440298.67	92.62%	379246.67	79.77%	341545	71.84%	273503	57.53%

Summary of integrated activity from bioluminescence spectra acquisition at varied pH. Integrated activity reflect the total RLU recorded throughout the duration of the bioluminescence spectra acquisition. Activities are adjacently displayed as a percentage of the maximum activity condition, shown here shaded for each enzyme.

6.3.4. pH dependence of flash kinetics for *wild-type* and engineered variants

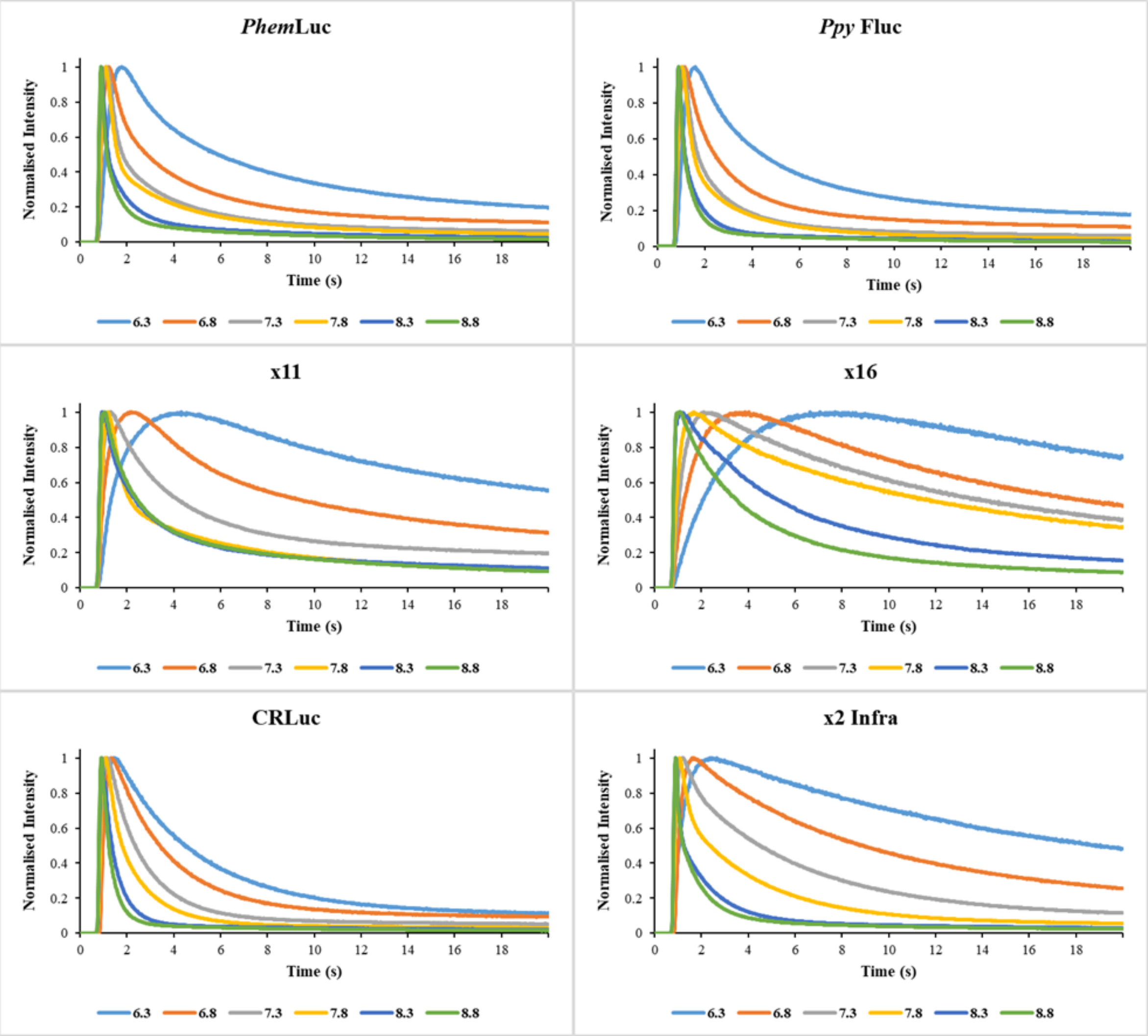
The flash kinetic of all enzymes was assessed across varied pH conditions in the presence of saturating LH_2 and ATP. Assays were performed such that the first 20 seconds of bioluminescence reaction were recorded by integrating light emissions over 20 ms for 1000 consecutive measurements (Figure 6.5.-6.6.). Of particular note were the flash heights of the reaction (Figure 6.7. and Table 6.5.), otherwise known as the maximum intensity (I_{max}), the time to peak intensity and subsequent decay to half intensity (Table 6.6.). Amongst all of the enzymes assessed, a common relationship was observed between the given pH condition and resulting I_{max} (Figure 6.7.). Lower pH conditions had a significant inhibitory effect on all kinetics flash heights. I_{max} measurements increased for each enzyme as the pH condition was raised, up to a shared optimum of pH 8.3, with the exception of x2 Infra which continued to increase recorded I_{max} at pH 8.8. A similar pH dependency was shown for the rise and decay of the flash kinetic for all enzymes (Table 6.6.). Under acidic conditions, an increased duration was required for I_{max} to be reached. A correlated increase in the decay time to half intensity was also shown under the same low pH conditions. The decay times at pH 6.3 for x11 and x16 could not be established as a point of half maximum intensity was not reached within the assays 20 second duration.

Figure 6.5.



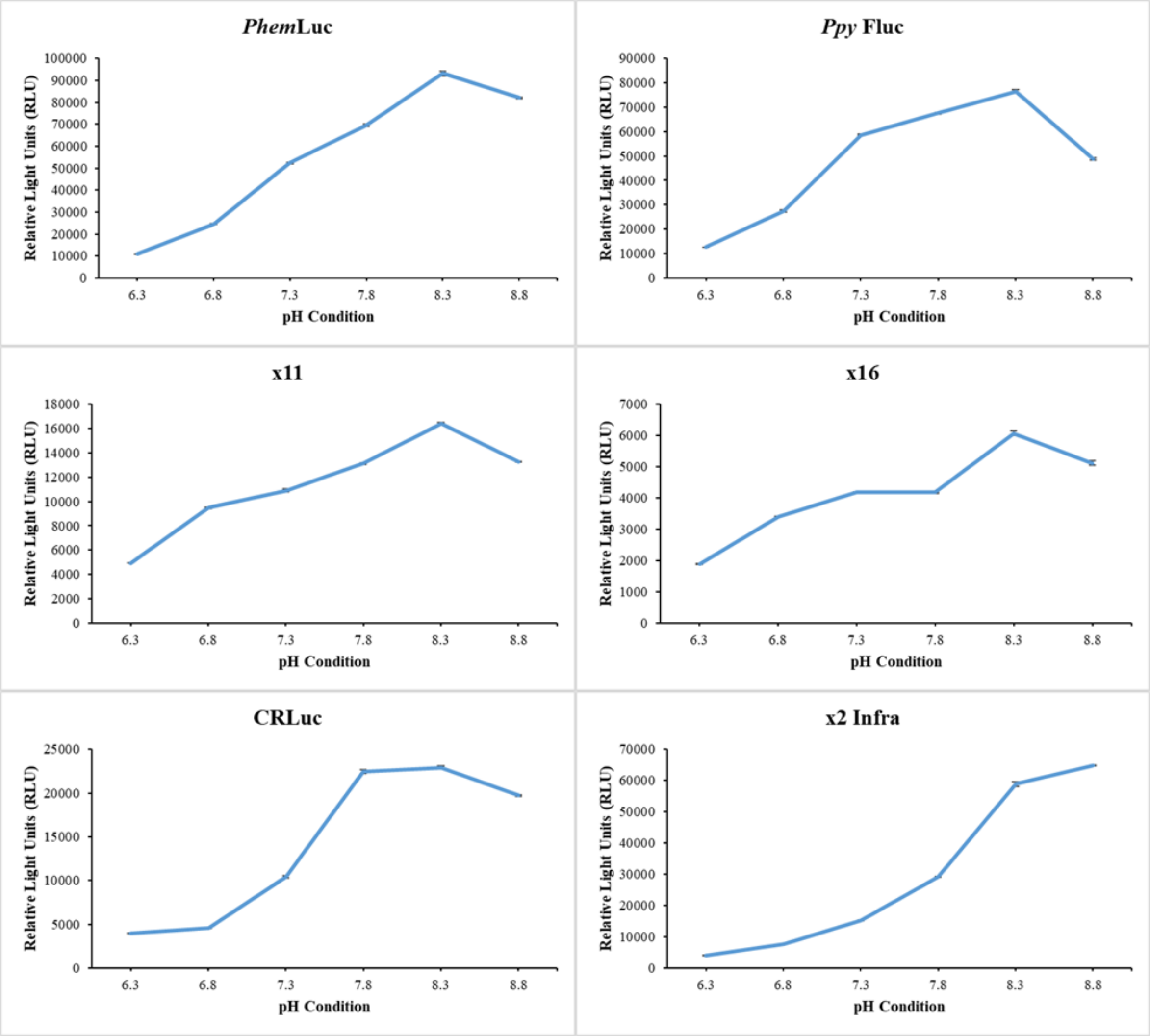
pH dependence of bioluminescence activity from *Phem*Luc, *Ppy* Fluc, x11, x16, CRLuc, and x2 Infra with LH_2 over 20 seconds. Measurements obtained in a luminometer by injection of substrate mix onto each enzyme such that final assay concentrations were equal to 1 mM ATP, 500 µM LH_2, and 0.167 µM protein. Each reaction constituent was previously diluted in chilled TEM buffer of the appropriate pH (±0.05) and total reaction volume was 150 µl. Light emission was integrated over 20 ms for 1000 consecutive measurements. Assays performed in triplicate for each pH condition and averaged data presented.

Figure 6.6.



pH dependence of normalised bioluminescence activity from *Phem*Luc, *Ppy* Fluc, x11, x16, CRLuc, and x2 Infra with LH_2 over 20 seconds. Measurements obtained in a luminometer by injection of substrate mix onto each enzyme such that final assay concentrations were equal to 1 mM ATP, 500 µM LH_2, and 0.167 µM protein. Each reaction constituent was previously diluted in chilled TEM buffer of the appropriate pH (±0.05) and total reaction volume was 150 µl. Light emission was integrated over 20 ms for 1000 consecutive measurements. Assays were performed in triplicate for each pH condition and averaged data presented. Data normalised such that each point is presented as intensity relative to emission peak.

Figure 6.7.



pH dependence for I_{max} from flash kinetic of *Phem*Luc, *Ppy* Fluc, x11, x16, CRLuc, and x2 Infra with LH_2. Measurements obtained in a luminometer by injection of substrate mix onto each enzyme mix such that final assay concentrations were equal to 1 mM ATP, 500 µM LH_2, and 0.167 µM protein. Each reaction constituent was previously diluted in chilled TEM buffer of the appropriate pH (±0.05) and total reaction volume was 150 µl. Light emission was integrated over 20 ms for 1000 consecutive measurements. Assays were performed in triplicate for each pH condition and averaged data for flash kinetic peak (I_{max}) presented. Error bars represent the standard error of the mean (SEM).

Table 6.4.

	pH 6.3		**pH 6.8**		**pH 7.3**		**pH 7.8**		**pH 8.3**		**pH 8.8**	
	Integrated Activity (RLU)	**Percentage of Max**	**Integrated Activity (RLU)**	**Percentage of Max**	**Integrated Activity (RLU)**	**Percentage of Max**	**Integrated Activity (RLU)**	**Percentage of Max**	**Integrated Activity (RLU)**	**Percentage of Max**	**Integrated Activity (RLU)**	**Percentage of Max**
***Phem*Luc**	4206983.33 ±10827.01	44.3%	5807565 ±37654.11	61.16%	8143851.33 ±76471.88	85.76%	9495542.67 ±81250.07	100%	7746073.33 ±54568.47	81.58%	5464984 ±60094.47	57.55%
***Ppy* Fluc**	4207319.33 ±37991.84	52.55%	5822525 ±98404.05	72.72%	8006674.33 ±49761.12	100%	8004730 ±41382.05	99.98%	5629845.67 ±50010.39	70.31%	3149041.33 ±33443.87	39.33%
x11	3558790.67 ±21707.31	72.83%	4886445.67 ±22434.85	100%	3646426.33 ±32347.87	74.62%	2920107.33 ±10428.87	59.76%	3627632.67 ±67373.71	74.24%	2909691 ±8556.77	59.55%
x16	1514108 ±2125.23	59.97%	2297232.33 ±6897.78	90.98%	2524883.33 ±9220.76	100%	2306454 ±17189.52	91.35%	2161595.33 ±21458.32	85.61%	1290003 ±12665.12	51.09%
CRLuc	1147128.67 ±14457.04	49.23%	1040551 ±13500.03	44.66%	1536541 ±19301.13	65.94%	2330187.67 ±23916.54	100%	1474057.67 ±9889.22	63.26%	1006889 ±5506.06	43.21%
x2 Infra	2707783 ±19720.85	50.98%	3634348.33 ±12696.79	68.42%	4611765.33 ±13970.65	86.82%	5254419 ±22537.82	98.92%	5311732.67 ±20159.28	100%	4956553.33 ±33499	93.31%

Summary of pH dependence of integrated bioluminescence activity from flash kinetics. Integrated activity reflect the total RLU recorded throughout the duration of the flash kinetic assay. Activities are adjacently displayed as a percentage of the maximum activity condition, shown here shaded for each enzyme. Errors for the integrated activity (RLU) values are the standard error of the mean (SEM).

Table 6.5.

-	pH 6.3		pH 6.8		pH 7.3		pH 7.8		pH 8.3		pH 8.8	
	I_{max} (RLU)	% I_{max} of Maximum	I_{max} (RLU)	% I_{max} of Maximum	I_{max} (RLU)	% I_{max} of Maximum	I_{max} (RLU)	% I_{max} of Maximum	I_{max} (RLU)	% I_{max} of Maximum	I_{max} (RLU)	% I_{max} of Maximum
***Phem*Luc**	10880 ±21.57	11.68%	24499 ±162.06	26.29%	52318 ±416.31	56.14%	69444.67 ±484.12	74.52%	93187.67 ±1077.5	100%	81982.67 ±326.2	87.98%
***Ppy* Fluc**	12535 ±97.56	16.39%	27299.67 ±463.39	35.69%	58633.33 ±316.13	76.65%	67573.33 ±289.39	88.34%	76490.67 ±546.06	100%	48844 ±500.7	63.86%
x11	4945.33 ±23.51	30.11%	9483.33 ±53.97	57.75%	10905.67 ±122.96	66.41%	13118 ±41.57	79.88%	16421.67 ±94.99	100%	13286.67 ±34.45	80.91%
x16	1893.33 ±10.55	31.17%	3396.33 ±14.41	55.92%	4185.67 ±15.27	68.92%	4176.33 ±25.61	68.77%	6073.33 ±69.63	100%	5123 ±77.58	84.35%
CRLuc	3928.33 ±45.84	17.15%	4575.67 ±54.42	19.98%	10385.67 ±140.44	45.35%	22452 ±184.29	98.04%	22901.33 ±189.3	100%	19706 ±82.59	86.05%
x2 Infra	4004.67 ±15.77	6.18%	7556.67 ±18.43	11.66%	15196.67 ±81.35	23.44%	29199.33 ±90.87	45.04%	58858 ±699.77	90.8%	64825 ±174.74	100%

Summary of pH dependence of I_{max} bioluminescence activity from flash kinetics. I_{max} indicates the RLU recorded at the peak intensity of the flash kinetic assay. I_{max} values are adjacently displayed as a percentage of the maximum recorded I_{max}, shown here shaded for each enzyme. Errors for the I_{max} (RLU) values are the standard error of the mean (SEM).

Table 6.6.

Luc	pH	Rise time(s)	Significance grouping	Decay time(s)	Significance grouping
***Phem*Luc**	6.3	1.76±0.04		5.91±0.17	
	6.8	1.21±0.01		2.83±0.01	
	7.3	1.12±0		1.81±0.01	
	7.8	1.07±0.01		1.59±0.01	
	8.3	0.91±0.01	a	1.2±0.02	a
	8.8	0.88±0	a	1.13±0.01	a
***Ppy* Fluc**	6.3	1.56±0.02		4.57±0.03	
	6.8	1.12±0.02		2.47±0.01	
	7.3	1.06±0.02	a	1.73±0.03	
	7.8	1.03±0.01	a	1.55±0.01	
	8.3	0.89±0.01	b	1.17±0.03	a
	8.8	0.87±0.01	b	1.15±0.01	a
x11	6.3	4.65±0.43		-	
	6.8	2.18±0.16		9.37±0.07	
	7.3	1.26±0.02	a	4.21±0.13	a
	7.8	1.19±0.01	a	2.13±0.01	ab
	8.3	0.95±0.03	a	2.3±0.08	b
	8.8	1.05±0.01	a	2.44±0.06	
x16	6.3	7.44±0.38		-	
	6.8	3.75±0.15		17.97±0.35	
	7.3	2.21±0.09		13.83±0.45	
	7.8	1.65±0.05		11.37±0.05	
	8.3	1.15±0.09	a	5.25±0.05	
	8.8	1.03±0.11	a	3.47±0.09	
CRLuc	6.3	1.45±0.07		4.47±0.11	
	6.8	1.33±0.05		3.39±0.03	
	7.3	1.15±0.01	a	2.34±0	
	7.8	1.07±0.01	a	1.79±0.01	
	8.3	0.92±0	b	1.35±0.01	
	8.8	0.91±0.01	b	1.22±0.02	
x2 Infra	6.3	2.51±0.17		18.52±0.28	
	6.8	1.62±0.04		8.78±0.1	
	7.3	1.18±0.02	a	4.47±0.03	
	7.8	1.04±0.02	ab	2.35±0.01	
	8.3	0.9±0	b	1.25±0.03	a
	8.8	0.89±0.01	b	1.21±0.01	a

Summary of the rise and decay time for Luc flash kinetics at varied pH. Rise times are the times taken from the point of injection to reach the maximum intensity (I_{max}). Decay times reflect the half-life of the emission, taken as the time required for intensity to diminish to half I_{max}. – indicates that a decay time was not reached within the duration of the assay. Flash kinetic assays were performed in triplicate for each pH condition, and the average times here presented. Errors reflect the maximum variation from the average within the respective triplicate. Statistical analysis performed as detailed in 2.12. Significance groupings across pH conditions are discrete between individual Lucs.

6.3.5. Michaelis-Menten kinetic characterisation of *wild-type* and engineered variants

As conventional methodologies to determine kinetic parameters rely upon a steady-state of reaction, they cannot be directly applied to the bioluminescence reaction and its characteristic flash kinetic. Nevertheless, it has been shown that kinetic parameters can still be derived for Flucs by interpreting the peak intensity (I_{max}) as a proxy for the pre steady-state of maximal light intensity, which can be used to extrapolate kinetic parameter values using the Michaelis-Menten (MM) equation. The point of I_{max} can be processed in this way as it is reached following a single turnover of the enzyme and is the only period of the reaction free from complicating factors including significant accumulation of inhibitory products. Consequently, when conducting an assay under saturating conditions of all but one reaction constituents (in this study, LH_2 or ATP), I_{max} is considered proportional to the initial rate (v) of LO* formation (Ugarova 1989; Brovko *et al.* 1994). With this consideration taken, the kinetic parameters of Flucs for either LH_2 or ATP can be established by plotting I_{max} against the respective substrate concentration (S).

To determine the kinetic parameters by luminometry as discussed, a fixed concentration of each enzyme was exposed by reagent injection to a variable concentration of the investigated substrate comprising ca.0.1 x K_M to ca.10 x K_M, in the presence of an invariable saturation of the additional substrate, roughly ca.10 x K_M. As such, the final concentrations of substrate in the reactions ranged from 0.1 µM-200 µM for LH_2, and 0.1 µM-1000 µM for ATP. The resulting I_{max} values for each concentration scale were plotted by implementing a linearized rearrangement of the Michaelis-Menten plot, commonly referred to as a Hanes-Woolf plot (Hanes 1932; Hofstee 1952). From the Hanes-Woolf plots of substrate concentration (S) against S/v (Substrate concentration over initial rate[I_{max}]) (Figures 6.9. and 6.11.), the kinetic constants K_M (Michaelis-Menten constant) and the V_{max} (maximal reaction velocity) were derived by regression analysis. The K_{cat} (catalytic constant) which represents the turnover number of the enzyme can be further calculated from the V_{max} using the calculated number of moles for the given enzyme. The overall catalytic efficiency of each enzyme is summarised by the ratio of K_{cat}/K_M (Table 6.7.).

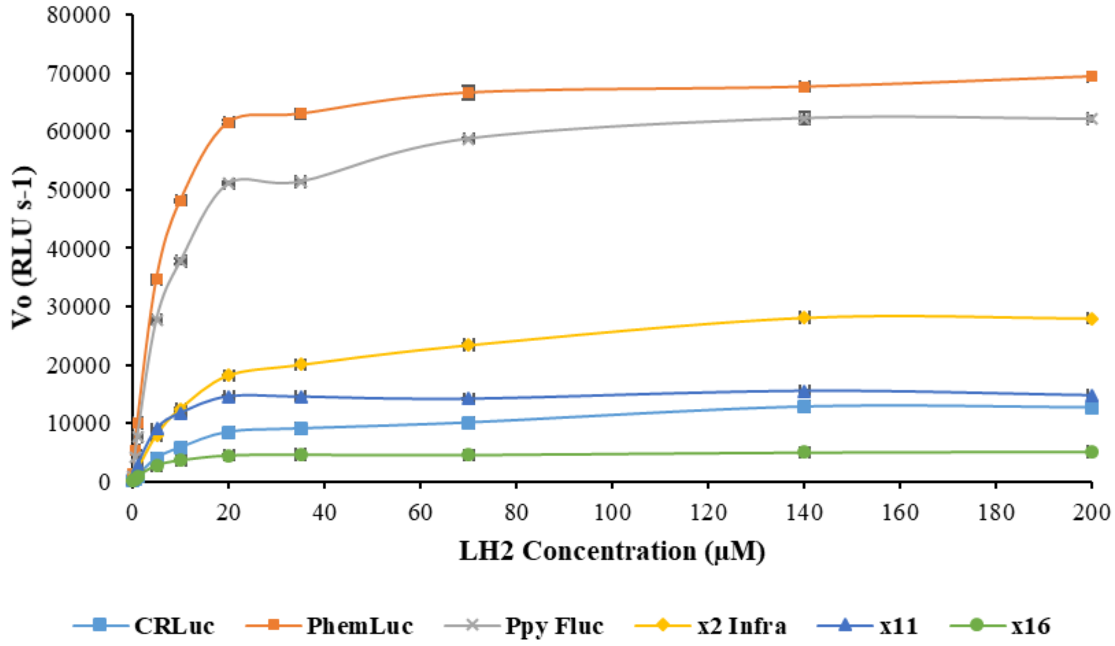
The kinetic parameters determined for the control enzymes *Ppy* Fluc and x11 were comparable but consistently lower than has been previously reported (Figures 6.8.-6.11. and Table 6.7.). In regards to *Ppy* Fluc, the K_M range for LH_2 reported in previous studies extends 10 µM to 20 µM (Maloshenok and Ugarova 2002; Tisi *et al.* 2002a; Branchini *et al.* 2003), although it has also been recorded as low as 6.6 µM for recombinant enzyme from Promega (Law *et al.* 2006). Here, *Ppy* Fluc displayed a K_M for LH_2 of 5 µM, and for ATP this value was determined as 45 µM, in contrast to the range of 56 µM to 250 µM that has previously been reported (Hirokawa *et al.* 2002; Maloshenok and Ugarova 2002; Branchini *et al.* 2003; Viviani *et al.* 2006). The kinetic parameters determined for x11 were 2.86 µM for LH_2 and 43.75 µM for ATP, whereas the reported ranges for these substrates are 3.7 µM to 7.5 µM, and 56.1 µM to 75.8 µM, respectively (Jathoul 2008; Jathoul *et al.* 2012; Halliwell 2015; Halliwell *et al.* 2018).

Across all enzyme assays performed to determine kinetic parameters for both control enzymes and all other Flucs assessed, all of the plotted data points have a good fit to their respective straight lines used for regression analysis, and minimal variation as indicated by the narrow range for all SEM values. Whilst this suggests a high reproducibility for all kinetic parameters derived, the indication from the lower than expected kinetic parameters for *Ppy* Fluc and x11 is that the adjustments made to the protein concentration by ImageJ analysis of SDS-PAGE still conceded a degree of error from the true concentration values for each protein. Disregarding purity, this issue is further compounded by the variation amongst the purified enzyme samples for the fraction of each which is composed of active protein. Whilst substrate K_M values for *Ppy* Fluc and x11 were lower than the reported ranges, it does not necessarily indicate that K_M values determined for all the project Flucs are similarly low. The K_M values recorded for *Ppy* Fluc and x11 serve only as an indication that there may be similar variation for the K_M values derived for the other Flucs assessed. Nonetheless, the determined values for K_M for *Ppy* Fluc and x11 were all recorded within 20 – 24% of their reported standard ranges, and thus the values determined for the remaining previously uncharacterised Flucs from this work should in turn serve as similarly indicative of their respective kinetic parameters. Ultimately, slight variation in the determined protein concentration from the actual protein concentration is less detrimental to the determination of K_M than issues of purity, whereas determination of K_{cat} is more dependent on accurate protein concentration. Values of K_{cat} are similarly less useful as they are calculated from V_{max}, which is itself proportional to the percentage of active enzyme in a given concentration which here is unknown, and consequently K_{cat} cannot be considered as a fundamental property of an enzyme in the same way as K_M.

The bioprospected Costa Rican firefly luciferase, CRLuc, was shown to have a higher K_M for both substrates than the values here reported for *Ppy* Fluc, with 14.29 µM calculated for LH_2 and 81.11 µM for ATP. These CRLuc K_M values do however align with the *Ppy* Fluc ranges reported elsewhere, suggesting that the substrate affinities of both enzymes are highly comparable. The previously uncharacterised *Phem*Luc displayed a K_M value for LH_2 of 7 µM, with a more dissimilar value from *Ppy* Fluc for its ATP Km of 20 µM. The *Phem*Luc engineered variant x2 Infra was shown to have <2-fold increased K_M value of 16.67 µM for LH_2 ($P<0.0001$), although its K_M for ATP remained similar to *Phem*Luc, at 26.67 µM ($P=0.98$). Considering its greater extent of modification, the thermostable engineered variant of *Phem*Luc, x16, displayed less variation from the kinetic parameters reported for *Phem*Luc than that observed from x2 Infra. The x16 K_M value for LH_2 was observed to be significantly decreased to 5 µM ($P=0.0151$), whilst the K_M for ATP was determined as an insignificant decrease to 14.5 µM ($P=0.8288$). This observed effect of only limited deviance from the *wild-type* origin enzyme in the kinetic parameter profile is consistent to x11 (of *Ppy* Fluc *wild-type* origin), with which x16 shares common mutations. The more marked deviance in kinetic parameters for x2 Infra than x16 is perhaps explained by the positon of the x2 Infra mutations, selected from modelling which indicated their involvement in the enzyme active site (Chapter 5).

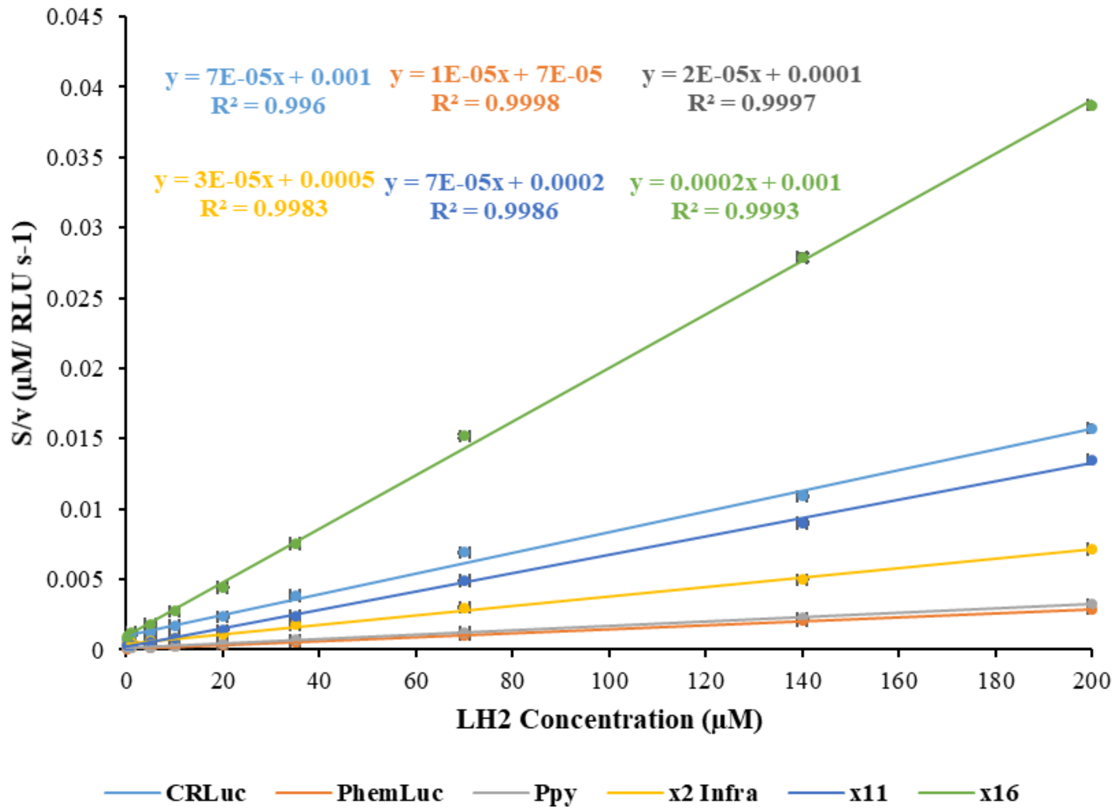
The ratio of K_{cat}/K_M is often referred to as the catalytic efficiency. A high ratio of K_{cat}/K_M suggests that a given enzyme works well in the presence of limited substrate, which can be understood as the enzyme not requiring a high concentration of substrate to achieve a high reaction rate. *Phem*Luc was determined to possess a significantly increased K_{cat}/K_M ratio relative to x2 Infra and x16 for both D-LH_2 ($P<0.0001$ determined for both) and ATP ($P=0.0005$ and $P<0.0001$, respectively). This indicates that whist x2 Infra and x16 were successfully engineered for their respective purposes, this work incidentally significantly reduced the catalytic efficiency of both enzymes variants.

Figure 6.8.



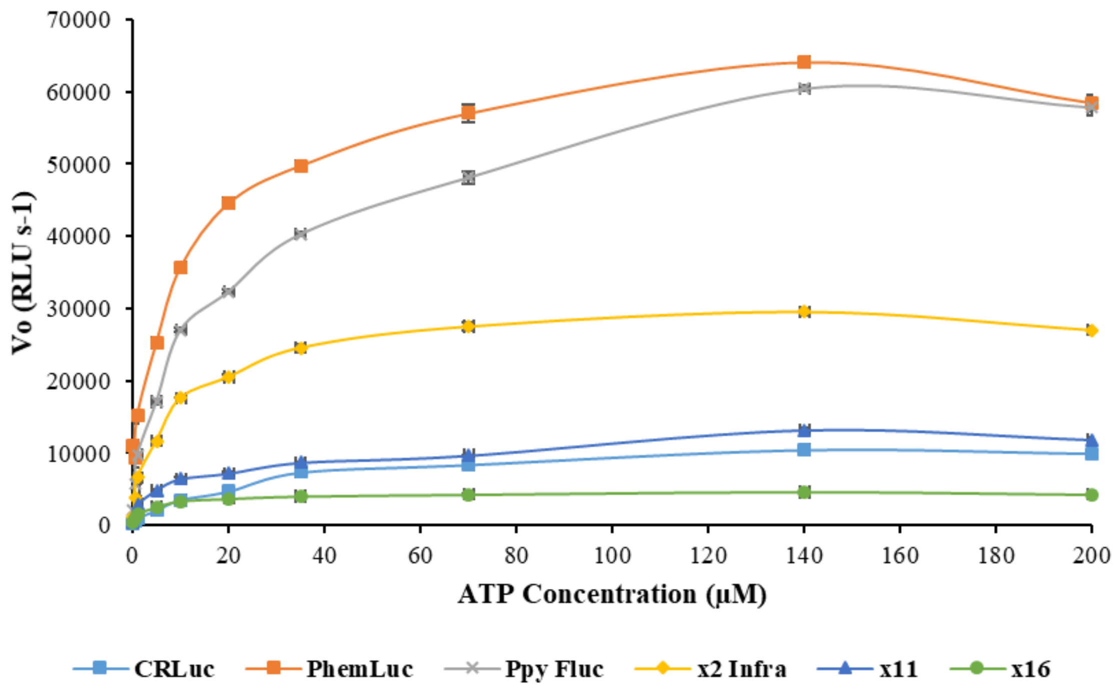
Michaelis-Menten plots of *Phem*Luc, *Ppy* Fluc, x11, x16, CRLuc, and x2 Infra with LH_2. Measurements obtained in a luminometer by injection of substrate mix into each enzyme mix such that final assay concentrations were equal to 1 mM ATP and 0.167 µM protein. Concentration of LH_2 were varied such that the substrate range included 0.1-10x the K_M concentrations. Each reaction constituent was previously diluted in chilled TEM buffer of pH 7.8 (±0.05) and total reaction volume was150 µl. Light emission was integrated over 20 ms for 1000 consecutive measurements. Assays were performed in triplicate for each substrate concentration and averaged flash-height measurements (I_{max}) are presented as an estimation of initial velocities (v_o).

Figure 6.9.



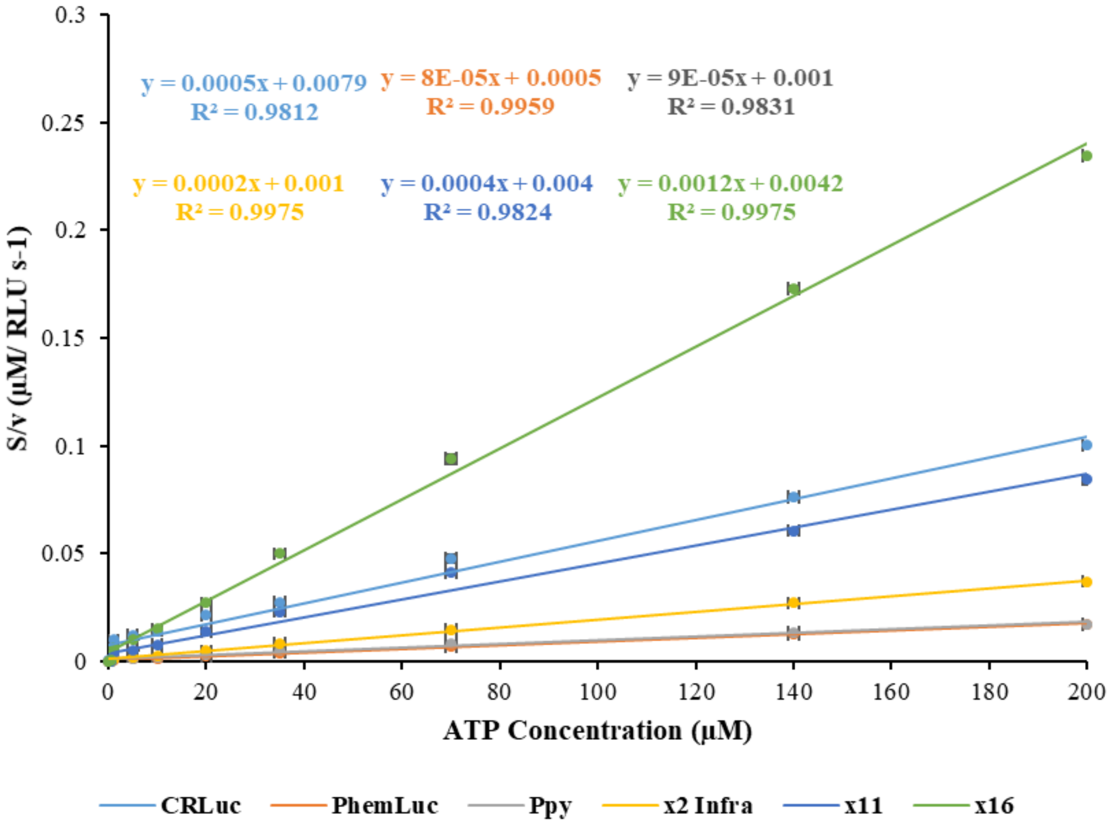
Hanes-Woolf plots of *Phem*Luc, *Ppy* Fluc, x11, x16, CRLuc, and x2 Infra with LH_2. Plots derived from data presented in Figure 6.8. Data is plotted as [S]/v against [S], where S is substrate concentration and v is estimated initial rate at that concentration, as indicated by I_{max}. Kinetic parameters are calculated by linear regression of plots (see Table 6.7.).

Figure 6.10.



Michaelis-Menten plots of *Phem*Luc, *Ppy* Fluc, x11, x16, CRLuc, and x2 Infra with ATP. Measurements obtained in a luminometer by injection of substrate mix into each enzyme mix such that final assay concentrations were equal to 500 µM LH_2 and 0.167 µM protein. Concentration of ATP was varied such that the substrate range spanned from 0.1-10x the K_M concentration. Each reaction constituent was previously diluted in chilled TEM buffer of pH 7.8 (±0.05) and total reaction volume was 150 µl. Light emission was integrated over 20 ms for 1000 consecutive measurements. Assays were performed in triplicate for each substrate concentration and averaged flash-height measurements (I_{max}) are presented as an estimation of initial velocities (v_o).

Figure 6.11.



Hanes-Woolf plots of *Phem*Luc, *Ppy* Fluc, x11, x16, CRLuc, and x2 Infra with ATP. Plots derived from data presented in Figure 6.10. Data plotted as [S]/v against [S], where S is substrate concentration and v is estimated initial rate at that concentration, as indicated by I_{max}. Kinetic parameters were calculated by linear regression of plots (see Table 6.7.).

Table 6.7.

	***D*-LH$_2$**						**ATP**					
	CRLuc	***Phem*Luc**	***Ppy* Fluc**	**x2 Infra**	**x11**	**x16**	**CRLuc**	***Phem*Luc**	***Ppy* Fluc**	**x2 Infra**	**x11**	**x16**
K_M (μM)	14.29 ±0.51	7.00 ±0.38	5.00 ±0.00	16.67 ±0.00	2.86 ±0.09	5.00 ±0.25	81.11 ±3.06	20.00 ±1.92	45.00 ±0.96	26.67 ±1.67	43.75 ±2.39	14.50 ±1.36
Significance grouping			a			a		b	a	b	a	b
K_{cat} (RLU s-1)(x10^{14})	5.71 ±0.21	40.00 ±0.00	20.00 ±0.00	13.33 ±0.00	5.71 ±0.18	2.00 ±0.00	4.44 ±0.00	20.00 ±0.00	20.00 ±0.00	13.30 ±1.01	5.00 ±0.14	2.00 ±0.00
Significance grouping	a				a		b	a	a		b	
K_{cat}/K_M (RLU s-1 M-1)(x10^{13})	4.00 ±0.11	57.14 ±3.08	40.00 ±0.00	8.00 ±0.00	20.00 ±0.00	4.00 ±0.23	0.55 ±0.02	10.00 ±0.96	4.44 ±0.09	5.00 ±0.16	1.14 ±0.04	1.38 ±0.15
Significance grouping	a			a		a	c		ab	a	c	bc

Summary of kinetic parameters for *wild-type* and engineered variants. Average kinetic parameter as derived from triplicate measurements of I_{max} across a substrate range including approximate coverage of 0.1-10x the K_M concentrations for ATP and LH$_2$. The final in-well concentrations of enzyme was estimated to be 0.167 µM in all assays. To derive the kinetic parameters for LH$_2$, assays were performed with final in-well concentration between 0.1 µM-200 µM, with 1 mM ATP to saturate. For measuring the kinetic parameters of ATP, assays were performed with final in-well concentrations between 0.1 µM-1000 µM, with 500 µM LH$_2$ to saturate (see Chapter 2). Errors are standard error of the mean (SEM). Statistical analysis performed as detailed in 2.12. Significance groupings are discrete between substrates.

6.4. Further Discussion

The aim of this chapter was to characterise the novel *wild-type* Flucs *Phem*Luc and CRLuc, whilst also evaluating how the engineered variants of *Phem*Luc, x2 Infra and x16, had been modified for core enzyme properties which had not been controlled for during the engineering selection process. For this purpose, N-terminal His-tagged *Phem*Luc, CRLuc, x2 Infra, and x16 were purified alongside controls of *Ppy* Fluc and x11. All enzymes were subsequently assessed for properties including bioluminescence spectra, pH dependence of activities and kinetic parameters.

A variation in total protein yield and purity was observed across all enzymes regardless of the strict adherence to the procedure for protein purification outlined in Chapter 2. Purity variation as indicated by the presence of lower molecular weight protein bands observed for some samples could be explained by the methodology utilised of selecting IMD elutions for desalting and retention based on activity observed in luminometry alone. As previously discussed, the highest bioluminescence activity IMD elutions were retained, without consideration for their purity at this stage. The lower purity observed in SDS-PAGE suggests that although fractions containing a large concentration of active luciferase protein may have been retained, these same fractions also contained non-specific proteins which had failed to be discarded by the increasing concentration of IMD washes.

As a result of the impurity, protein concentrations as determined via Bradford assay were not reliable for the purpose of diluting known concentrations of protein for bioluminescence assays. In order to proceed to biochemical characterisation and comparison, Fluc concentrations were corrected based on SDS-PAGE luciferase protein band size and intensity as determined by ImageJ. This analysis was complicated by the non-specific lower molecular weight bands which proved difficult to separate out from the target Fluc band.

Regardless, all protein concentrations were adjusted to better reflect their true Fluc concentrations in comparison to the results as determined by Bradford. Whilst the corrected concentrations were sufficient for investigation across the majority of assays performed, the remaining deviation from the accurate concentrations and concerns with purity introduce a degree of uncertainty to the interpretation of the derived kinetic parameters, as indicated by the lower than reported ranges of the kinetic parameters determined for the control enzymes *Ppy* Fluc and x11. For this reason, the kinetic parameters determined for CRLuc, *Phem*Luc, x2 Infra and x16 can be interpreted as indicative of their respective ranges, prior to verification with fresh protein purifications.

Nevertheless, as the other assays conducted here have a lower reliance on definitive protein concentrations, their results can be more readily accepted as an indication of the biochemical properties of each respective protein. For instance, the bioluminescence spectra for the control enzymes *Ppy* Fluc and x11 were in broad agreement with their reported λ_{max}, here measured at ca. 558 nm and ca. 557 nm, respectively (Jathoul *et al.* 2012). The implication from the control spectra being correctly observed regardless of the discussed concentration variation is that spectral properties recorded for the remaining Flucs would be similarly reliable.

The only beetle luciferase known to naturally emit true red bioluminescence (termed PxRE) is produced by the closely related species *Phrixotrix viviani* and *P. hirtus* (Viviani *et al.* 2006). These two species of Coleoptera are from the Phengodidae family, and thus distinct from fireflies and their family of Lampyridae, although their luciferases are largely similar and share the common substrate *D*-luciferin. The λ_{max} recorded for PxRE is 623 nm, with a bandwidth of only 55 nm, making it a highly specific emitter. The λ_{max} here determined for CRLuc was recorded at 609 nm with a far greater bandwidth of 95 nm. This result would make CRLuc the most red-shifted naturally produced bioluminescence from a firefly identified to date. However, red-shifted bioluminescence spectra can be a consequence of partial enzyme denaturation or instability, and therefore this result is a possible indication that the sequence retrieved for CRLuc by bioprospecting in Chapter 3 deviated from the true *wild-type* sequence. An additional unusual observation was made in the pH dependence of CRLuc bioluminescence spectra, which displayed a higher degree of spectral robustness at lower pH, but exhibited a hypsochromic shift under alkali conditions. This is in contrast to the other *wild-type* enzymes *Phem*Luc and *Ppy* Fluc which instead exhibit a bathochromic shift under acidic conditions, a common characteristic of pH-sensitive firefly luciferases (Viviani *et al.* 2008). This unusual property might be worth investigating further to determine whether CRLuc might be useful as a pH sensor.

Although a lower bioluminescence emission alone did not directly indicate that the bioprospected sequence for CRLuc was imperfect, another suggestion that the CRLuc sequence might not represent its true *wild-type* sequence is provided by the I_{max} across all pH ranges being reduced in comparison to the other *wild-type* enzymes, being recorded at only 33% of the I_{max} of *Ppy* Fluc at pH 7.8. Additionally, the K_M values recorded for CRLuc were approximately double those of *Phem*Luc and *Ppy* Fluc for both LH_2 and ATP, indicating ≈50% less affinity for both substrates. CRLuc was also determined to have a reduced K_{cat} for LH_2 of 5.71×10^{14} (RLU s-1) in contrast with 20×10^{14} and 40×10^{14} for *Ppy* Fluc and *Phem*Luc, respectively, with similar observations made for the K_{cat} with ATP. It has previously been show that mutations in the active site of *Ppy* Fluc which result in spectral shifts are correlated with a decrease in the catalytic activity of the enzyme (Ugarova and Brovko 2002). However, CRLuc shares 93.45% of amino acid sequence identity with *Ppy* Fluc, whilst also being entirely conserved across all 15 putative active site residues (Branchini *et al.* 2003). Therefore, the mechanism behind the red shifted bioluminescence emission and reduced catalytic activity cannot be attributed to variation in the active site residues. However, no single observation of CRLuc activity can determine whether the sequence derived in Chapter 3 is erroneous, and may instead only indicate that the pure protein of CRLuc is particularly unstable.

The enzyme properties recorded for *Phem*Luc were similar to those recorded for *Ppy* Fluc with which it shares 87.07% amino acid sequence identity. The bioluminescence spectra λ_{max} for *Phem*Luc was recorded at 557 nm compared to 558 nm for *Ppy* Fluc, which is in agreement with the observed λ_{max} from previous observations (Jathoul *et al.* 2012). Further to this, the pH dependence of spectra observed for both enzymes were near-identical, with both enzymes exhibiting a bathochromic shift under acidic conditions typical to pH-sensitive firefly luciferases (Viviani *et al.* 2008). The highest bioluminescence activity from each was recorded at pH 8.3, where the I_{max} of *Phem*Luc was 21.8% greater than that of *Ppy* Fluc. At pH 7.8 however, the I_{max} of *Phem*Luc was only 2.8% greater than *Ppy* Fluc. Across all pH conditions assessed the rise and decay times of the bioluminescence flash kinetic from *Phem*Luc were slightly slower than those of *Ppy* Fluc. As previously discussed, the Michaelis-Menten parameters derived here can only be taken as broadly indicative of enzyme ranges. Nonetheless, the K_M of *Phem*Luc with LH_2 was recorded as 7 µM, in broad agreement with 5 µM recorded for *Ppy* Fluc. However, the K_M with ATP was 20 µM for *Phem*Luc, relative to 45 µM for *Ppy* Fluc, suggesting that a key difference between the two enzymes is an affinity for ATP in *Phem*Luc which is approximately twice that of *Ppy* Fluc.

The x2 Infra variant of *Phem*Luc engineered for potential improved activity with the synthetic substrate analogue Infraluciferin displayed significant differences in enzyme properties. Relative to *Phem*Luc, the λ_{max} was increased by 53 nm to 610 nm, shifting from a green bioluminescence emission to the red region of the visible light spectrum. In contrast to the bathochromic shift effect of *Phem*Luc under acidic conditions, an unexpected characteristic of x2 Infra was the observation of significant pH-tolerance of the bioluminescence spectra which mimicked the profile observed in the pH-tolerant control x11 which had purposefully been engineered for this quality, and the bioluminescence emission of the pH-insensitive beetle luciferases from click beetles and railroadworms (Viviani *et al.* 2008; Jathoul *et al.* 2012). However, whilst the point of I_{max} from x11 possesses a degree of resilience to changes in pH, the I_{max} of x2 Infra, displayed a clear pH dependence and exhibited a significant positive correlation with increasing pH conditions, such that the highest measurements of I_{max} were obtained at pH 8.8. Regardless, the significant spectral pH-insensitivity conferred from only two mutations suggests the respective positions of H245W and A313G from x2 Infra could be investigated to enhance the pH-tolerance properties of the x11 Fluc. As these mutations were discovered through active-site mutagenesis to improve the bioluminescence activity of *Phem*Luc with the synthetic substrate analogue Infraluciferin, the investigation of Michaelis-Menten parameters of K_M suggested that x2 Infra affinity for LH_2 had been reduced to less than half of that from *Phem*Luc, with a less severe reduction in the affinity for ATP. Determination of the x2 Infra K_M for infraluciferin was not possible due to the infraluciferin available to the project existing as a racemic mix of dextrorotary and levorotatory stereoisomers (see Chapter 4).

Although the x16 thermostable variant λ_{max} of 566 nm was similar to the 557 nm λ_{max} from *Phem*Luc, the FWHM was 16 nm greater (88 nm vs 72 nm) than that of *Phem*Luc. Additionally, in contrast to the bathochromic shift from *Phem*Luc under acidic conditions, x16 was observed to possess a reduced pH-sensitivity at lower pH, but at the most alkaline condition assessed (pH 8.8) underwent a bathochromic shift. The overall pH-tolerance of x16 with respect to bioluminescence spectra and emission yields was improved relative to the *wild-type Phem*Luc, but not to the extent of the characteristic pH-tolerance of x11, with which x16 shares eleven common mutations (see Chapter 5). Similarly to *Phem*Luc, the greatest measurement of I_{max} for x16 was recorded at pH 8.3, but was only 6.5% of the respective I_{max} measurement obtained in *Phem*Luc. Although this is a significant reduction in overall bioluminescence yield, a similar effect can be observed in the heavily engineered x11 producing only 21.5% I_{max} of its *wild-type* origin *Ppy* Fluc at pH 8.3. The K_M values of x16 with LH_2 and ATP were both lower than the measurements obtained in *Phem*Luc, indicating that x16 possessed a higher affinity for both substrates. Combined with the thermostability exhibited by x16 in Chapter 5, it may be worth investigating whether the increased affinity for ATP enables x16 to function as a high sensitivity ATP detection system.

6.5. Conclusions

This Chapter sought to provide the primary characterisation of *Phem*Luc and CRLuc, whilst also aiming to ascertain how the engineered variants of *Phem*Luc, x2 Infra and x16, had been affected for properties beyond their screening selection pressures in development, and whether any secondary advantageous characteristics had emerged. Although concerns with the purity of each protein limited measurements of enzyme kinetics to serve only as indicative of the range for each Fluc, the remaining assays depended less on a precise determination of active protein concentration and could be more readily accepted as accurate representations of enzyme properties. Whilst *Phem*Luc was observed to be largely "similar" across all enzyme properties to *Ppy* Fluc, the bioprospected CRLuc possessed several unusual enzyme properties. It is unclear whether these represent normal properties of this enzyme or arise because the derived sequence contains errors relative to the unknown *wild-type* sequence. Ascertaining this would require further work to obtain the genomic sequence for this species. The bioluminescence spectra from the x2 Infra variant of *Phem*Luc displayed a significant pH-tolerance which mimicked the profile of the thermostable pH-tolerant x11. Further investigation is required to understand how the mutations H245W and A313G influence pH-tolerance in *Phem*Luc, and whether the effects of these mutations are translatable to enhance the capability of the x11 Fluc. The thermostable variant x16 produced a significantly reduced bioluminescence signal relative to *Phem*Luc. Although the enzyme kinetics could only be interpreted as broad indications, the observations made for x16 suggested that alongside its targeted improvement in thermostability, its affinity for both LH_2 and ATP had been significantly improved, suggesting that a possible role as a high-sensitivity ATP detection system may be worth future consideration.

Chapter 7

General Discussion

7.1. Chapter Summary

The purpose of this project was divided across three key aims which sought to discover and develop novel luciferase variants by i) bioprospecting for novel luciferase gene sequences using dry-preserved Coleoptera from museum collections, ii) engineering the luciferase from the lesser British Glow-worm *Phosphaenus hemipterus* (*Phem*Luc) for improved bioluminescence activity with the synthetic substrate analogue infraluciferin, and iii) developing a variant of *Phem*Luc possessing significantly improved thermostability. Through these efforts a novel luciferase from an unidentified Costa Rican firefly was discovered, and two variants of *Phem*Luc generated: x2 Infra which displayed improved activity with infraluciferin, and x16 which was capable of significantly increased resistance to thermal inactivation. This chapter reflects on the key experimental strategies deployed across these research areas and includes discussion on the advantages and future implications of these approaches, in addition to how with the benefit of retrospective consideration these strategies could have been refined. The current condition of each Fluc is summarised, along with the opportunities it may afford for existing bioluminescence applications and research. Finally, the future directions available to advance the understanding and development of these Flucs are discussed.

7.2. Reflection on Experimental Strategy

7.2.1. Bioprospecting

The yields recovered from the non-destructive DNA extraction method were highly variable between samples but nonetheless were sufficient to prepare Illumina sequencing libraries from all five unidentified fireflies of interest. Whilst an alternative approach of removing individual insect legs for destructive DNA extraction would have preserved the remaining tissue for future analyses, it would have likely provided insufficient DNA for the library preparation due to the lower DNA yields of aged samples, and this approach is typically only used for the purposes of genotyping by PCR. However, it would have been of value to perform a standard destructive extraction on a less valuable firefly sample which had previously been processed using the non-destructive extraction in order to verify the efficiency of the non-destructive method and whether any genomic material remained for future analyses. Furthermore, due to the near certainty that DNA extraction protocols and more importantly future DNA sequencing technologies will continue to advance, it may become possible to routinely recover high-quality genomic data from ultra-low DNA concentrations (Green *et al.* 2017), such as the residual genomic material from the fireflies investigated here.

Affinity enrichment was confirmed as a necessary process in the undertaking of this work since luciferase gene sequence could only be recovered from bioinformatic analysis of the enriched Costa Rican firefly library and not the non-enriched equivalent. However, even with enrichment, the bioinformatic process was only successful in recovering a luciferase gene in one of the five total libraries, so it is worth considering how this approach might be refined. Analysis of the CRLuc gene indicated that ≈90% sequence identity was shared with voucher specimens of *Ppy* luciferase complete coding sequences. As cross-species affinity enrichment using biotin probes has only been demonstrated to be capable of enriching sequences up to 10-13% divergence (Mason *et al.* 2011), it may be that the unsuccessful four libraries possessed sequence identities of greater divergence, and the use of only the *Ppy* Fluc gene for affinity purification was a key limiting factor that could be refined. If this were indeed the case, a better strategy would have been the use of biotin probe pools constructed from a variety of inter-species luciferase genes.

As discussed in Chapter 3, validating the enrichment using CODEHOP qPCR amplification initially appeared successful, but was later complicated due to the unaccounted 49 bp region present in all Sanger sequences of CODEHOP products from all unidentified museum fireflies. Whilst this invalidated the enrichment seen by CODEHOP qPCR, a secondary option to investigate enrichment would have been provided by comparing amplification of the ‘mini-barcode’ sequences in the enriched and non-enriched libraries such that the decrease in the relative abundance of non-target sequences could have been assessed.

The Sanger sequencing of PCR products using amplification primers was highly variable in sequence data quality and in some cases whether products could be sequenced at all. This could have perhaps been mitigated by routinely TA subcloning all PCR products into an appropriate vector prior to sequencing with a more typical approach such as T7 priming, subject to the availability of priming sites in the selected vector.

The enrichment strategy was employed to increase the opportunity of realising the primary ambition to recover novel luciferase gene sequences. However, this approach is not without its drawbacks as enrichment significantly limits the ability to derive further information from the sequencing data, relative to whole genome amplification and *de novo* assembly of non-enriched libraries. Nonetheless, luciferase gene sequences were only recoverable in the enriched library of the Costa Rican firefly and not the equivalent non-enriched library, demonstrating that for the purposes of this work that enrichment was the appropriate action. Furthermore, recent efforts to expand the availability of published insect genomes have concluded that *de novo* assembly of high-quality insect genomes by second-generation sequencing techniques such as Illumina HiSeq is complicated by a high degree of heterozygosity throughout insect genomes (Li *et al.* 2019). A common strategy to circumvent this issue is to combine the accuracy of second generation sequencing with the longer read lengths (>10kb) of third generation sequencing platforms such as NanoPore and PacBio to act as scaffolds for assembly. However, due to the degraded nature of the museum firefly DNA extracts, such an approach could not be utilised here.

Although sequencing data was successfully produced from all libraries, it remains uncertain whether the failure to recover luciferase gene sequences for four libraries was a failure independent of the enrichment strategy or the bioinformatics process, or furthermore whether both were contributing factors. If it was indeed a failure in the bioinformatics process, the reference genome mapping strategy utilised here would have been susceptible to the same issues speculated to have negated the enrichment design such that the unsuccessful firefly libraries in this case would be excessively divergent from the three reference genomes available, whereas the Costa Rican firefly possessed sufficient complementarity with the *Ppy* reference genome to enable recovery of the luciferase gene. An attempt was made to modify the Bowtie2 alignment to map to a collection of reference luciferase genes, but these efforts were entirely unsuccessful for all five libraries. Whether the existing bioinformatics pipeline could be successfully optimized in this way without further firefly reference genomes is unlikely. With the continuous development of bioinformatics tools there is a high probability that multiple strategies could have been designed capable of recovering the CRLuc gene independently to the methods explored here. However, no such methods were found during the limited bioinformatics work of this study and it therefore remains a possibility that bioinformatics tools exist which are capable of recovering the luciferase gene sequences from the four unsuccessful libraries.

7.2.2. Engineering infraluciferin compatibility

In the absence of any structural data available on *Phem*Luc, homology modelling against *Ppy* Fluc was explored in order to investigate the question of whether amino acid residues in close proximity to the bound luciferin substrates could be mutagenized to substitute residues which improve the bioluminescence activity with the synthetic substrate analogue infraluciferin. Although protein residue interactions with bound substrates are not limited to within 4 Å, this distance was selected to identify only the positions within the most immediate contact with the bound substrate, as expansion out to a 5 Å region would significantly increase the residues identified for mutagenesis beyond the scope of this current study. Existing crystal structures of *Ppy* Fluc presented a unique opportunity due to the high shared protein sequence identity of 87.07% with *Phem*Luc, which is a known reliable predictor of the achievable model quality (Rodrigues *et al.* 2013), and most critically that structures were available bound to structural analogues of luciferyl-AMP for both luciferin and infraluciferin. The key limitation of this approach was that only models in the adenylate conformation were explored. As a consequence, residues which are brought into substrate proximity only following the 140˚ rotation of the C-domain to participate in oxidation catalysis, were excluded from targeted mutagenesis. Although models of *Ppy* Fluc are available in the second catalytic conformation, these were not explored due to their availability being limited to in complex with only DLSA, and not iDLSA. Nonetheless, as the majority of targeted residues identified across all four homology models were conserved between DLSA and iDLSA models, it may have been of value to explore the oxidation model of *Ppy* Fluc to identify additional target residues, even if some residues unique to infraluciferin catalysis could not have been identified in this way.

The infraluciferin assays performed with pure protein in Chapter 4 were heavily restricted due to the limited availability of infraluciferin remaining at this stage of the project. The decision was made to utilize a selection of assays which could indicate the activity of x2 Infra relative to *Phem*Luc and x11, but which could also be performed using the same bioluminescence reactions for the acquisition of all data. For this reason, infraluciferin activity data with pure protein was limited to a spectral acquisition, with unfiltered bioluminescence yields recorded immediately before and after. If availability of infraluciferin had not been an issue, further assay options would have been explored.

7.2.3. Thermostability engineering

The fifteen existing mutations originally incorporated into *Phem*Luc were a combination of fourteen mutations known to produce thermostabilising effects in *Ppy* Fluc and an additional mutation S347G which was originally included in the pursuit of a dual function thermostable mutant possessing improved activity with infraluciferin, based on observations made in *Ppy* Fluc by Dr Amit Jathoul. The expectation from the simultaneous introduction of fifteen cross-species mutations into *Phem*Luc was that one or more mutations could be detrimental to bioluminescence activity and a method of reversion would be required since advantageous firefly luciferase mutations are not always known to conserve their effect once introduced into the luciferase of a distinct firefly species (Kitayama *et al.* 2003; Koksharov and Ugarova 2011b). Ultimately, the S347G mutation was found to be the culprit of the diminished bioluminescence emission and the decision was made to pursue engineering of infraluciferin activity independently elsewhere (see Chapter 4).

Whilst it would have been possible to identify S347G as the source of the diminished bioluminescence activity by systematic reversion of each mutated position, at the time it was unclear whether the deleterious effect might be arising from the inclusion or interactions of multiple mutations. For this reason, DNA shuffling was explored as a strategy to backcross the x15 *Phem*Luc with the *wild-type* gene and generate a sub-collection of mutations in which bioluminescence activity was restored and further thermostability engineering could be pursued. Although these efforts were successful, the full potential of a firefly luciferase gene DNA shuffling method was not explored. Primarily, not all the 32,786 unique potential combinations possible from the fifteen mutations were screened, but going beyond this more advanced directed evolution methods were in reach such as the shuffling of thermostable variants from diverse firefly species. Engineered thermostable luciferases exist for multiple firefly species (Kajiyama and Nakano 1994; Koksharov and Ugarova 2011a; Mortazavi and Hosseinkhani 2011; Jathoul *et al.* 2012). As discussed previously, the effects of some mutations can be conserved between species, but others can fail to reproduce the desired effect in all but the original Fluc it was discovered. Even when considering only the conserved mutations, the effects do not always perform to the same degree between species. With this method of DNA shuffling it would be possible to explore whether shuffled chimeras of multiple thermostable Fluc variants from diverse species could possess thermostability properties which exceed those of the parent templates. Furthermore, this strategy would not be limited to only thermostability engineering, and could be applied to any Fluc characteristic of interest for which multiple species variants are available.

Although the $MnCl_2$ and D_2O epPCR approach was successfully used to identify two novel thermostabilising mutations, this method was also not fully explored due to time constraints. The original intended strategy had been to incorporate several round of epPCR in order to discover further mutations capable of incrementally improving the thermostability of *Phem*Luc with each round. This would have additionally been paired with further rounds of DNA shuffling to refine the acquired mutations into an improved activity subset where possible.

7.3. Overview of Project Flucs

The recovery of the luciferase gene sequence from the unidentified Costa Rican firefly in Chapter 3 demonstrates that Nagoya compliant dry-preserved insect specimens can serve as valuable repositories of relevant genomic data which go beyond phylogenetic analyses. It was speculated in Chapter 3 that the bioprospected CRLuc gene sequence may contain erroneous positions within its sequence. This was later supported by the bioluminescence emission observations made in Chapter 6, including a significantly red-shifted spectral λ_{max} and a pH-dependence profile which differed from the typical sensitivity of firefly luciferases (Viviani *et al.* 2008). Further work would be required to confirm the sequence and potentially recover a CRLuc gene sequence which is true to the *wild-type* sequence from the unidentified Costa Rican Firefly.

The bioluminescence enzyme assays performed in this study (Chapter 6) serve as the primary characterisation of the luciferase from the lesser British Glow-worm *Phosphaenus hemipterus*, *Phem*Luc. Whilst this Fluc appears to share many highly similar characteristics with *Ppy* Fluc including bioluminescence emission I_{max} and spectral λ_{max}, it presents an opportunity to investigate a new sequence landscape and to engineer variants possessing novel enzyme properties to expand the toolbox of bioluminescent enzymes available for modern applications, such as what was explored here.

The x2 Infra variant was proven to possess improved bioluminescence activity with infraluciferin compared to *Phem*Luc using screening in *E. coli* and as pure protein across the limited assays performed. However, the true advantage of infraluciferin bioluminescence systems is only apparent in bioluminescence imaging of mammalian tissues (Anderson *et al.* 2019; Stowe *et al.* 2019). As such investigations were not conducted here, it is difficult to accurately assess the advantages x2 Infra might offer without an understanding of its performance in the role for which it was originally designed. Furthermore, the secondary property of reduced spectral pH-sensitivity discovered in Chapter 6 might additionally directly benefit multispectral bioluminescence imaging of mammalian tissues where any shift is undesirable (Chaudhari *et al.* 2005; Mezzanotte *et al.* 2010).

The thermostable variant x16 exhibited significant improvements in resistance to thermal inactivation (Chapter 5) and pH-tolerance (Chapter 6) relative to *Phem*Luc. However, even with these improvements x16 was less capable across both of these properties than the x11 control Fluc. Nonetheless, the identification of the two novel thermostabilising mutations I321V and L306H in a single application of the directed evolution method suggests that further thermostabilising mutations await to be discovered. Ultimately, the identification and incorporation of further thermostabilising mutations through continued cycling of the directed evolution method is required to construct a thermostable variant of *Phem*Luc which is able to compete with the thermostable Flucs which are already available.

7.4 Future Directions

With the suggestion that the CRLuc gene sequence may be erroneous based on bioluminescence emission observations made in Chapter 6, efforts would have been made to use the existing sequence information to discover the true *wild-type* sequence if the project time remaining had permitted. Although the fragmented nature of the genomic DNA extract from the unidentified Costa Rican firefly likely precludes amplification with terminal primers against the current known sequence as an option, a more achievable approach would be to attempt the amplification of short overlapping sequence fragments of ≈200 bp in length and to assemble the overlapping sequence data derived into what would perhaps be an accurate *wild-type* sequence for CRLuc. Beyond CRLuc, the luciferase gene sequences from the four additional enriched libraries Illumina data remains undiscovered. As there is currently no suggestion that luciferase gene reads are entirely absent in all four libraries, it may be worth reinvestigating with a new bioinformatic strategy.

Purification of all project Flucs would need to be repeated to obtain higher quality protein samples in order to confirm the bioluminescence observations made during the assays of Chapter 6, and more importantly to derive enzyme kinetic parameters of greater accuracy. Improved purifications would additionally enable future ventures of crystallography. With the increased accuracy of crystal protein models over homology modelling, the effects of the existing and future mutations in *Phem*Luc could perhaps be better understood.

Although the engineering goals of this project to generate variants of *Phem*Luc with improved infraluciferin activity and enhanced thermostability were fulfilled, neither of these ambitions had predefined activity threshold targets, and there is no existing suggestion that either of these variants cannot be improved further. The engineering of x2 Infra was intentionally restricted to rational design, whereas the engineering of the thermostable mutant x16 was in part pursued with the intention of designing a directed evolution pipeline for *Phem*Luc, where thermostability could be substituted for any desired property of the luciferase enzyme that can be selected in a bioluminescence screen. Hence, if a renewed supply of infraluciferin were secured, the directed evolution process could be applied to x2 Infra to access the unconstrained mutagenic potential of the entire protein for the purpose of improving infraluciferin bioluminescence activity. Furthermore, this renewed supply of infraluciferin would enable the completion of refined infraluciferin assays which go beyond what was possible with the limited supply in Chapter 4. Ultimately, x2 Infra or any improved derivatives would need to be investigated in their intended role of bioluminescence imaging, and therefore codon optimized and transfected into mammalian cell lines prior to subsequent implantation in immunocompromised mice for *in vivo* imaging.

The x16 thermostable variant serves as a platform from which further novel thermostabilising mutations of *Phem*Luc should be discovered though use of the paired epPCR and DNA shuffling directed evolution process developed in Chapter 5. If time had permitted, continued cycling of mutagenesis and screening would have been explored until the acquisition of advantageous mutations stagnated.

7.5 Concluding Remarks

Advancing the applications of firefly luciferase bioluminescence depends upon the discovery of novel phenotypes and their respective sequences, regardless of whether these are derived from nature or mutagenic approaches. For this purpose, this study explored the acquisition of novel luciferases from the three distinct areas of bioprospecting natural resources, engineering by rational design, and directed evolution. The prospects enabled from the successful bioprospecting of the CRLuc gene sequence perhaps go beyond the acquisition of a novel firefly luciferase and instead alludes to the utility and relevance of biological materials sourced from the collections of museums in the pursuit of future genetic discoveries. Although each endeavour was successful, a natural point of conclusion for all three investigations could not be reached due to the open nature of each undertaking. Regardless, each of the investigations conducted here offer a clear route of continuation for the development of the novel luciferases of this study, in addition to suggestions of broader lines of enquiry which could be pursued independently of this studies ambitions.

References

Adams, S.T. and Miller, S.C. 2014. Beyond D-luciferin: expanding the scope of bioluminescence imaging in vivo. *Current Opinion in Chemical Biology* 21, pp. 112–120. doi: 10.1016/J.CBPA.2014.07.003.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tool. *Journal of molecular biology* 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Amadeo, F. et al. 2021. Firefly luciferase offers superior performance to AkaLuc for tracking the fate of administered cell therapies. *European Journal of Nuclear Medicine and Molecular Imaging* , pp. 1–13. doi: 10.1007/s00259-021-05439-4.

Anderson, J.C., Chang, C.-H., Jathoul, A.P. and Syed, A.J. 2019. Synthesis and bioluminescence of electronically modified and rotationally restricted colour-shifting infraluciferin analogues. *Tetrahedron* 75(3), pp. 347–356. doi: 10.1016/J.TET.2018.11.061.

Anderson, J.C., Grounds, H., Jathoul, A.P., Murray, J.A.H., Pacman, S.J. and Tisi, L. 2017. Convergent synthesis and optical properties of near-infrared emitting bioluminescent infra-luciferins. *RSC Advances* 7(7), pp. 3975–3982. doi: 10.1039/C6RA19541E.

Arnold, F.H. 1998. When blind is better: Protein design by evolution. *Nature Biotechnology* 16(7), pp. 617–618. doi: 10.1038/NBT0798-617.

Asghar, U., Malik, M.F., Anwar, F., Javed, A. and Raza, A. 2014. DNA Extraction from Insects by Using Different Techniques: A Review. 3(3), pp. 132–138. doi: 10.4236/ae.2015.34016.

Baggett, B., Roy, R., Momen, S., Morgan, S., Tisi, L., Morse, D. and Gillies, R.J. 2004. Thermostability of Firefly Luciferases Affects Efficiency of Detection by In Vivo Bioluminescence. *Molecular Imaging* 3(4), pp. 324–332. doi: 10.1162/1535350042973553.

Bahar, I. and Jernigan, R.L. 1996. Coordination geometry of nonbonded residues in globular proteins. *Folding & design* 1(5), pp. 357–370. doi: 10.1016/S1359-0278(96)00051-X.

Bankevich, A. et al. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* 19(5), p. 455. doi: 10.1089/CMB.2012.0021.

Beattie, A.J., Hay, M., Magnusson, B., de Nys, R., Smeathers, J. and Vincent, J.F.V. 2011. Ecology and bioprospecting. *Austral Ecology* 36(3), pp. 341–356. doi: 10.1111/J.1442-9993.2010.02170.X.

Bechara, E.J.H. and Stevani, C. V. 2018. Brazilian bioluminescent beetles: Reflections on catching glimpses of light in the Atlantic Forest and Cerrado. *Anais da Academia Brasileira de Ciencias* 90(1), pp. 663–679. doi: 10.1590/0001-3765201820170504.

Beckman, R.A., Mildvan, A.S. and Loeb, L.A. 1985. On the Fidelity of DNA Replication: Manganese Mutagenesis in Vitro†. *Biochemistry* 24(21), pp. 5810–5817. doi: 10.1021/BI00342A019.

Berraud-Pache, R. and Navizet, I. 2016. QM/MM calculations on a newly synthesised oxyluciferin substrate: new insights into the conformational effect. *Physical chemistry chemical physics : PCCP* 18(39), pp. 27460–27467. doi: 10.1039/C6CP02585D.

Bessetti, J. 2007. An Introduction to PCR Inhibitors. *Promega Corporation*

Bolger, A.M., Lohse, M. and Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* 30(15), pp. 2114–2120. doi: 10.1093/BIOINFORMATICS/BTU170.

Borel, J.F., Kis, Z.L. and Beveridge, T. 1995. The History of the Discovery and Development of Cyclosporine (Sandimmune®). *The Search for Anti-Inflammatory Drugs* , pp. 27–63. doi: 10.1007/978-1-4615-9846-6_2.

Boyce, R., Chilana, P. and Rose, T.M. 2009. iCODEHOP: a new interactive program for designing COnsensus-DEgenerate Hybrid Oligonucleotide Primers from multiply aligned protein sequences. *Nucleic acids research* 37(Web Server issue), pp. W222-8. doi: 10.1093/nar/gkp379.

Boyle, R. 1668. New Experiments concerning the relation between light and air (in shining wood and fish;) made by the honourable Robert Boyle, and by him addressed from Oxford to the publisher, and so communicated to the Royal Society. *Philosophical Transactions of the Royal Society of London* 2(31), pp. 581–600. doi: 10.1098/rstl.1666.0060.

Bradford, M. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical biochemistry*

72(1–2), pp. 248–254. doi: 10.1006/ABIO.1976.9999.

Branchini, B.R., Ablamsky, D.M., Murtiashaw, M.H., Uzasci, L., Fraga, H. and Southworth, T.L. 2007. Thermostable red and green light-producing firefly luciferase mutants for bioluminescent reporter applications. *Analytical Biochemistry* 361(2), pp. 253–262. doi: 10.1016/J.AB.2006.10.043.

Branchini, B.R., Magyar, R.A., Murtiashaw, M.H., Anderson, S.M. and Zimmer, M. 1998. Site-directed mutagenesis of histidine 245 in firefly luciferase: A proposed model of the active site. *Biochemistry* 37(44), pp. 15311–15319. doi: 10.1021/BI981150D.

Branchini, B.R., Murtiashaw, M.H., Carmody, J.N., Mygatt, E.E. and Southworth, T.L. 2005a. Synthesis of an N-acyl sulfamate analog of luciferyl-AMP: A stable and potent inhibitor of firefly luciferase. *Bioorganic & Medicinal Chemistry Letters* 15(17), pp. 3860–3864. doi: 10.1016/J.BMCL.2005.05.115.

Branchini, B.R., Southworth, T.L., Fontaine, D.M., Davis, A.L., Behney, C.E. and Murtiashaw, M.H. 2014. A photinus pyralis and Luciola italica chimeric firefly luciferase produces enhanced bioluminescence. *Biochemistry* 53(40), pp. 6287–6289. doi: 10.1021/bi501202u.

Branchini, B.R., Southworth, T.L., Murtiashaw, M.H., Boije, H. and Fleet, S.E. 2003. A mutagenesis study of the putative luciferin binding site residues of firefly luciferase. *Biochemistry* 42(35), pp. 10429–10436. doi: 10.1021/bi030099x.

Branchini, B.R., Southworth, T.L., Murtiashaw, M.H., Wilkinson, S.R., Khattak, N.F., Rosenberg, J.C. and Zimmer, M. 2005b. Mutagenesis evidence that the partial reactions of firefly bioluminescence are catalyzed by different conformations of the luciferase C-terminal domain. *Biochemistry* 44(5), pp. 1385–1393. doi: 10.1021/BI047903F.

Brovko, L., Gandel'man, O.A., Polenova, T.E. and Ugarova, N.N. 1994. Kinetics of bioluminescence in the firefly luciferin-luciferase system. *Biochemistry* 59(2), pp. 195–201.

Buchan, D.W.A. and Jones, D.T. 2019. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Research* 47(W1), pp. W402–W407. doi: 10.1093/NAR/GKZ297.

Buck, J.B. 1948. The anatomy and physiology of the light organ in fireflies. *Annals of the New York Academy of Sciences* 49(3), pp. 397–485. doi: 10.1111/J.1749-

6632.1948.TB30944.X.

Cadwell, R.C. and Joyce, G.F. 1992. Randomization of genes by PCR mutagenesis. *Genome Research* 2(1), pp. 28–33. doi: 10.1101/gr.2.1.28.

Chaudhari, A.J. et al. 2005. Hyperspectral and multispectral bioluminescence optical tomography for small animal imaging. *Physics in Medicine and Biology* 50(23), pp. 5421–5441. doi: 10.1088/0031-9155/50/23/001.

Cheng, Y.Y. and Liu, Y.J. 2019. Luciferin Regeneration in Firefly Bioluminescence via Proton-Transfer-Facilitated Hydrolysis, Condensation and Chiral Inversion. *Chemphyschem : a European journal of chemical physics and physical chemistry* 20(13), pp. 1719–1727. doi: 10.1002/CPHC.201900306.

Choy, G., O'Connor, S., Diehn, F.E., Costouros, N., Alexander, H.R., Choyke, P. and Libutti, S.K. 2003. Comparison of noninvasive fluorescent and bioluminescent small animal optical imaging. *BioTechniques* 35(5), pp. 1022–1030. doi: 10.2144/03355rr02.

Cobb, R.E., Chao, R. and Zhao, H. 2013. Directed evolution: Past, present, and future. *AIChE Journal* 59(5), pp. 1432–1440. doi: 10.1002/AIC.13995.

Conti, E., Franks, N.P. and Brick, P. 1996. Crystal structure of firefly luciferase throws light on a superfamily of adenylate-forming enzymes. *Structure* 4(3), pp. 287–298. doi: 10.1016/S0969-2126(96)00033-0.

Cormier, M.J. and Karkhanis, Y.D. 1971. Isolation and properties of Renilla reniformis luciferase, a low molecular weight energy conversion enzyme. *Biochemistry* 10(2), pp. 317–326. doi: 10.1021/bi00778a019.

Crameri, A., Raillard, S.-A., Bermudez, E. and Stemmer, W.P.C. 1998. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* 391(6664), pp. 288–291. doi: 10.1038/34663.

Dale, S.J. and Belfield, M. 1996. Oligonucleotide-Directed Random Mutagenesis Using the Phosphorothioate Method. *Methods in molecular biology (Clifton, N.J.)* 57, pp. 369–374. doi: 10.1385/0-89603-332-5:369.

Day, J.C., Chaichi, M.J., Najafil, I. and Whiteley, A.S. 2006. Genomic structure of the luciferase gene from the bioluminescent beetle, Nyctophila cf. caucasica. *Journal of insect* *science (Online)* 6, pp. 1–8. doi: 10.1673/031.006.3701.

Day, J.C., Tisi, L.C. and Bailey, M.J. 2004. Evolution of beetle bioluminescence: the origin of beetle luciferin. *Luminescence* 19(1), pp. 8–20. doi: 10.1002/bio.749.

DeLuca, M. and McElroy, W.D. 1974. Kinetics of the firefly luciferase catalyzed reactions. *Biochemistry* 13(5), pp. 921–925. doi: 10.1021/bi00702a015.

Dillon, N., Austin, A.D. and Bartowsky, E. 1996. Comparison of preservation techniques for DNA extraction from hymenopterous insects. *Insect molecular biology* 5(1), pp. 21–4. doi: 10.1111/j.1365-2583.1996.tb00036.x.

Drews, A. et al. 2016. The two worlds of Nagoya ABS legislation in the EU and provider countries: discrepancies and how to deal with them.

Dubuisson, M., Marchand, C. and Rees, J.-F. 2004. Firefly luciferin as antioxidant and light emitter: the evolution of insect bioluminescence. *Luminescence* 19(6), pp. 339–344. doi: 10.1002/bio.789.

Eckert, K.A. and Kunkel, T.A. 1990. High fidelity DNA synthesis by the Thermus aquaticus DNA polymerase. *Nucleic Acids Research* 18(13), pp. 3739–3744. doi: 10.1093/NAR/18.13.3739.

Fallon, T.R. et al. 2018. Firefly genomes illuminate parallel origins of bioluminescence in beetles. *eLife* 7. doi: 10.7554/ELIFE.36495.

Fraga, H. 2008. Firefly luminescence: A historical perspective and recent developments. *Photochemical and Photobiological Sciences* 7(2), pp. 146–158. doi: 10.1039/b719181b.

Fraga, H., Fernandes, D., Novotny, J., Fontes, R. and Esteves Da Silva, J.C.G. 2006. Firefly luciferase produces hydrogen peroxide as a coproduct in dehydroluciferyl adenylate formation. *ChemBioChem* 7(6). doi: 10.1002/cbic.200500443.

Gandelman, O.A. et al. 2010. Novel Bioluminescent Quantitative Detection of Nucleic Acid Amplification in Real-Time. Ho, P. L. ed. *PLoS ONE* 5(11), p. e14155. doi: 10.1371/journal.pone.0014155.

Gandelman, O.A., Church, V.L., Moore, C.A., Carne, C., Jalal, H., Murray, J.A.H. and Tisi, L.C. 2007. BART – Bioluminescent alternative to Real-Time PCR. In: *Bioluminescence and Chemiluminescence*. WORLD SCIENTIFIC, pp. 95–98. doi:

10.1142/9789812770196_0023.

Georgescu, R., Bandara, G. and Sun, L. 2003. Saturation Mutagenesis. In: *Directed Evolution Library Creation*. Humana Press, pp. 75–83. doi: 10.1385/1-59259-395-X:75.

Gilbert, M.T.P., Moore, W., Melchior, L. and Worebey, M. 2007. DNA extraction from dry museum beetles without conferring external morphological damage. *PLoS ONE* . doi: 10.1371/journal.pone.0000272.

Green, E.D., Rubin, E.M. and Olson, M. V. 2017. The future of DNA sequencing. *Nature* 550(7675), pp. 179–181. doi: 10.1038/550179A.

Greener, A., Callahan, M. and Jerpseth, B. 1996. An Efficient Random Mutagenesis Technique Using an *E. coli* Mutator Strain. *Methods in molecular biology (Clifton, N.J.)* 57, pp. 375–385. doi: 10.1385/0-89603-332-5:375.

Gulick, A.M. 2009. Conformational dynamics in the acyl-CoA synthetases, adenylation domains of non-ribosomal peptide synthetases, and firefly luciferase. *ACS Chemical Biology* 4(10), pp. 811–827. doi: 10.1021/CB900156H.

Haddock, S.H.D., Moline, M.A. and Case, J.F. 2010. Bioluminescence in the Sea. *Annual Review of Marine Science* 2(1), pp. 443–493. doi: 10.1146/annurev-marine-120308-081028.

Hajibabaei, M., Janzen, D.H., Burns, J.M., Hallwachs, W. and Hebert, P.D.N. 2006a. DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences* 103(4), pp. 968–971. doi: 10.1073/PNAS.0510466103.

Hajibabaei, M., Smith, M.A., Janzen, D.H., Rodriguez, J.J., Whitfield, J.B. and Hebert, P.D.N. 2006b. A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes* 6(4), pp. 959–964. doi: 10.1111/J.1471-8286.2006.01470.X.

Hall, M.P., Gruber, M.G., Hannah, R.R., Jennens-Clough, M.L. and Wood, K. V. 1999. Stabilisation of firefly luciferase using directed evolution. In: Roda, A., Pazzagli, M., Kricka, L. J., and Stanley, P. E. eds. *Bioluminescence and Chemiluminescence: Perspectives for the 21st Century*. Wiley, pp. 392–395.

Halliwell, L.M. 2015. *Protein Engineering Utilising Single Amino Acid Deletions Within Photinus Pyralis Firefly Luciferase*. Cardiff University.

Halliwell, L.M., Jathoul, A.P., Bate, J.P., Worthy, H.L., Anderson, J.C., Jones, D.D. and

Murray, J.A.H. 2018. ΔFlucs: Brighter Photinus pyralis firefly luciferases identified by surveying consecutive single amino acid deletion mutations in a thermostable variant. *Biotechnology and Bioengineering* 115(1), pp. 50–59. doi: 10.1002/bit.26451.

Hanes, C.S. 1932. The effect of starch concentration upon the velocity of hydrolysis by the amylase of germinated barley. *Biochemical Journal* 26(5), pp. 1406–1421. doi: 10.1042/BJ0261406.

Hardinge, P. 2014. *Low copy number quantification of DNA utilising Loop-mediated Amplification (LAMP) with Bioluminescent Assay in Real-Time (BART) reporter*. Cardiff University.

Harvey, E.N. 1957. *A history of luminescence from the earliest times until 1900.* doi: 10.5962/bhl.title.14249.

Hastings, J.W. and Wilson, T. 1998. Bioluminescence. *Annu. Rev. Cell Dev. Biol* , pp. 197–230.

Hirokawa, K., Kajiyama, N. and Murakami, S. 2002. Improved practical usefulness of firefly luciferase by gene chimerization and random mutagenesis. *Biochimica et Biophysica Acta - Protein Structure and Molecular Enzymology* 1597(2), pp. 271–279. doi: 10.1016/S0167-4838(02)00302-3.

Hofstee, B.H.J. 1952. On the evaluation of the constants Vm and KM in enzyme reactions. *Science* 116(3013), pp. 329–331. doi: 10.1126/SCIENCE.116.3013.329.

Hsiao, K., Zegzouti, H. and Goueli, S.A. 2016. Methyltransferase-Glo: a universal, bioluminescent and homogenous assay for monitoring all classes of methyltransferases. *Epigenomics* 8(3), pp. 321–339. doi: 10.2217/EPI.15.113.

Hulet, W.H. and Musil, G. 1968. Intracellular Bacteria in the Light Organ of the Deep Sea Angler Fish, Melanocetus murrayi. *Copeia* 1968(3), p. 506. doi: 10.2307/1442019.

Iwano, S. et al. 2013. Development of simple firefly luciferin analogs emitting blue, green, red, and near-infrared biological window light. *Tetrahedron* 69(19), pp. 3847–3856. doi: 10.1016/j.tet.2013.03.050.

Jathoul, A., Law, E., Gandelman, O., Pule, M., Tisi, L. and Murray, J. 2012. Development of a pH-Tolerant Thermostable Photinus pyralis Luciferase for Brighter In Vivo Imaging.

In: *Bioluminescence - Recent Advances in Oceanic Measurements and Laboratory Applications*. InTech. doi: 10.5772/37170.

Jathoul, A.P. 2008. *Activity of Firefly Luciferase with 6'-Amino-D-Luciferin*. University of Cambridge.

Jathoul, A.P., Grounds, H., Anderson, J.C. and Pule, M.A. 2014. A dual-color far-red to near-infrared firefly luciferin analogue designed for multiparametric bioluminescence imaging. *Angewandte Chemie (International ed. in English)* 53(48), pp. 13059–63. doi: 10.1002/anie.201405955.

Jawhara, S. and Mordon, S. 2004. In Vivo Imaging of Bioluminescent Escherichia coli in a Cutaneous Wound Infection Model for Evaluation of an Antibiotic Therapy. *Antimicrobial Agents And Chemotherapy* 48(9), pp. 3436–3441. doi: 10.1128/AAC.48.9.3436-3441.2004.

Jazayeri, F.S., Amininasab, M. and Hosseinkhani, S. 2017. Structural and dynamical insight into thermally induced functional inactivation of firefly luciferase. *PLOS ONE* 12(7), p. e0180667. doi: 10.1371/JOURNAL.PONE.0180667.

Joern, J.M. 2003. DNA Shuffling. In: *Directed Evolution Library Creation*. New Jersey: Humana Press, pp. 85–90. doi: 10.1385/1-59259-395-X:85.

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* 292(2), pp. 195–202. doi: 10.1006/JMBI.1999.3091.

Jones, K.A., Porterfield, W.B., Rathbun, C.M., McCutcheon, D.C., Paley, M.A. and Prescher, J.A. 2017. Orthogonal Luciferase–Luciferin Pairs for Bioluminescence Imaging. *Journal of the American Chemical Society* 139(6), pp. 2351–2358. doi: 10.1021/jacs.6b11737.

Kahlke, T. and Umbers, K.D. 2016. *Current Biology*. doi: 10.1016/j.cub.2016.01.007.

Kajiyama, N. and Nakano, E. 1993. Thermostabilization of Firefly Luciferase by a Single Amino Acid Substitution at Position 217. *Biochemistry* 32(50), pp. 13795–13799. doi: 10.1021/BI00213A007.

Kajiyama, N. and Nakano, E. 1994. Enhancement of Thermostability of Firefly Luciferase from Luciola lateralis by a Single Amino Acid Substitution. *Bioscience, Biotechnology, and Biochemistry* 58(6), pp. 1170–1171. doi: 10.1271/BBB.58.1170.

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H. and Xu, J. 2012. Template-based protein structure modeling using the RaptorX web server. *Nature Protocols* 7(8), pp. 1511–1522. doi: 10.1038/NPROT.2012.085.

Kaper, T., Brouns, S.J.J., Geerling, A.C.M., De Vos, W.M. and Van der Oost, J. 2002. DNA family shuffling of hyperthermostable β-glycosidases. *Biochemical Journal* 368(2), pp. 461–470. doi: 10.1042/BJ20020726.

Kaskova, Z.M., Tsarkova, A.S. and Yampolsky, I. V. 2016. 1001 lights: luciferins, luciferases, their mechanisms of action and applications in chemical analysis, biology and medicine. *Chemical Society Reviews* 45(21), pp. 6048–6077. doi: 10.1039/C6CS00296J.

Kazantsev, S. V. 2015. Protoluciola albertalleni gen.n., sp.n., a new luciolinae firefly (Insecta: Coleoptera: Lampyridae) from burmite amber. *Russian Entomological Journal* 24(4), pp. 281–283. doi: 10.15298/rusentj.24.4.02.

Khan, R.H., Siddiqi, M.K. and Salahuddin, P. 2017. Protein Structure and Function. In: *Basic Biochemistry*. Austin Publishing Group, pp. 1–39.

Kiddle, G. et al. 2012. GMO detection using a bioluminescent real time reporter (BART) of loop mediated isothermal amplification (LAMP) suitable for field use. *BMC Biotechnology* 12(1). doi: 10.1186/1472-6750-12-15.

Kim-Choi, E., Danilo, C., Kelly, J., Carroll, R., Shonnard, D. and Rybina, I. 2006. Creating a mutant luciferase resistant to HPV chemical inhibition by random mutagenesis and colony-level screening. *Luminescence* 21(3), pp. 135–142. doi: 10.1002/BIO.897.

Kitayama, A., Yoshizaki, H., Ohmiya, Y., Ueda, H. and Nagamune, T. 2003. Creation of a thermostable firefly luciferase with pH-insensitive luminescent color. *Photochemistry and photobiology* 77(3), pp. 333–8. doi: 10.1562/0031-8655(2003)077<0333:coatfl>2.0.co;2.

Koksharov, M.I. and Ugarova, N.N. 2011a. Thermostabilization of firefly luciferase by in vivo directed evolution. *Protein Engineering, Design and Selection* 24(11), pp. 835–844. doi: 10.1093/protein/gzr044.

Koksharov, M.I. and Ugarova, N.N. 2011b. Triple substitution G216N/A217L/S398M leads to the active and thermostable Luciola mingrelica firefly luciferase. *Photochemical & photobiological sciences : Official journal of the European Photochemistry Association and* *the European Society for Photobiology* 10(6), pp. 931–8. doi: 10.1039/c0pp00318b.

Koksharov, M.I. and Ugarova, N.N. 2012. Approaches to engineer stability of beetle luciferases. *Computational and structural biotechnology journal* 2, p. e201209004. doi: 10.5936/csbj.201209004.

Koswatta, T., Samaraweera, P. and Sumanasinghe, V. 2012. A Simple Comparison between Specific Protein Secondary Structure Prediction Tools. *Tropical Agricultural Research* 23(1), p. 91. doi: 10.4038/TAR.V23I1.4636.

Kuzikov, A.N., Bondarenko, V.M. and Latkin, A.T. 2003. Use of the bioluminescent method for the determination of bacterial adenosinetriphosphate (ATP-metry) in microbiology. *Zhurnal mikrobiologii, epidemiologii, i immunobiologii* (1), pp. 80–89.

Labrou, N. 2010. Random Mutagenesis Methods for In Vitro Directed Enzyme Evolution. *Current Protein & Peptide Science* 11(1), pp. 91–100. doi: 10.2174/138920310790274617.

Langmead, B. and Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* 9(4), p. 357. doi: 10.1038/NMETH.1923.

Law, G.H., Gendelman, O.A., Tisi, L.C., Lowe, C.R. and Murray, J. 2002. Altering the surface hydrophobicity of firefly luciferase. In: *Bioluminescence and Chemiluminescence*. WORLD SCIENTIFIC, pp. 37–40. doi: 10.1142/9789812776624_0007.

Law, G.H.E., Gandelman, O.A., Tisi, L.C., Lowe, C.R. and Murray, J.A.H. 2006. Mutagenesis of solvent-exposed amino acids in Photinus pyralis luciferase improves thermostability and pH-tolerance. *Biochemical Journal* 397(2), pp. 305–312. doi: 10.1042/BJ20051847.

Lee, J. 2008. Bioluminescence: the First 3000 Years (Review). *Journal of Siberian Federal University. Biology* 3(1), pp. 194–205. doi: 10.17516/1997-1389-0264.

Lemasters, J.J. and Hackenbrock, C.R. 1977. Kinetics of product inhibition during firefly luciferase luminescence. *Biochemistry* 16(3), pp. 445–447. doi: 10.1021/bi00622a016.

Li, F. et al. 2019. Insect genomes: progress and challenges. *Insect Molecular Biology* 28(6), pp. 739–758. doi: 10.1111/IMB.12599.

Li, H. et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16), pp. 2078–2079. doi: 10.1093/BIOINFORMATICS/BTP352.

Lin-Goerke, J.L., Robbins, D.J. and Burczak, J.D. 1997. PCR-based random mutagenesis using manganese and reduced DNTP concentration. *BioTechniques* 23(3), pp. 409–412. doi: 10.2144/97233BM12.

Lindahl, T. 1993. Instability and decay of the primary structure of DNA. *Nature* 362(6422), pp. 709–715. doi: 10.1038/362709a0.

Lloyd, J.E. 1983. Bioluminescence and Communication in Insects. *Annual Review of Entomology* 28, pp. 131–160. doi: 10.1146/annurev.en.28.010183.001023.

Lower, S.S., Johnston, J.S., Stanger-Hall, K.F., Hjelmen, C.E., Hanrahan, S.J., Korunes, K. and Hall, D. 2017. Genome Size in North American Fireflies: Substantial Variation Likely Driven by Neutral Processes. *Genome biology and evolution* . doi: 10.1093/gbe/evx097.

Lutz, S. 2010. Beyond directed evolution-semi-rational protein engineering and design. *Current Opinion in Biotechnology* 21(6), pp. 734–743. doi: 10.1016/J.COPBIO.2010.08.011.

Maloshenok, L.G. and Ugarova, N.N. 2002. *Catalytic properties and bioluminescence spectra of recombinant firefly luciferase luciola mingrelica with point mutations out of the enzyme active site*. Stanley, P. E. and Kricka, L. J. eds. World Scientific, Singapore.

Markova, S. V., Golz, S., Frank, L.A., Kalthof, B. and Vysotski, E.S. 2004. Cloning and expression of cDNA for a luciferase from the marine copepod Metridia longa: A novel secreted bioluminescent reporter enzyme. *Journal of Biological Chemistry* 279(5), pp. 3212–3217. doi: 10.1074/jbc.M309639200.

Markova, S. V., Larionova, M.D., Burakova, L.P. and Vysotski, E.S. 2015. The smallest natural high-active luciferase: Cloning and characterization of novel 16.5-kDa luciferase from copepod Metridia longa. *Biochemical and Biophysical Research Communications* 457(1), pp. 77–82. doi: 10.1016/J.BBRC.2014.12.082.

Markova, S. V., Larionova, M.D. and Vysotski, E.S. 2019. Shining Light on the Secreted Luciferases of Marine Copepods: Current Knowledge and Applications. *Photochemistry and Photobiology* 95(3), pp. 705–721. doi: 10.1111/php.13077.

Mason, V.C., Li, G., Helgen, K.M. and Murphy, W.J. 2011. Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. *Genome Research* 21(10), p. 1695. doi: 10.1101/GR.120196.111.

Mcelroy, W.D., Seliger, H.H. and White, E.H. 1969. Mechanism of bioluminescence, chemi-luminescence and enzyme function in the oxidation of firefly luciferin. *Photochemistry and Photobiology* 10(3), pp. 153–170. doi: 10.1111/j.1751-1097.1969.tb05676.x.

Meusnier, I., Singer, G.A.C., Landry, J.F., Hickey, D.A., Hebert, P.D.N. and Hajibabaei, M. 2008. A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics* 9(1), pp. 1–4. doi: 10.1186/1471-2164-9-214.

Meyer, A.J., Ellefson, J.W., Ellington, A.D. and Meyer, A.J,Ellefson, J.W, Ellington, A.D. 2015. Library generation by gene shuffling. *Current Protocols in Molecular Biology* 105(Figure 1), pp. 1–10. doi: 10.1002/0471142727.mb1512s105.

Mezzanotte, L., Fazzina, R., Michelini, E., Tonelli, R., Pession, A., Branchini, B. and Roda, A. 2010. In vivo bioluminescence imaging of murine xenograft cancer models with a red-shifted thermostable luciferase. *Molecular imaging and biology : MIB : the official publication of the Academy of Molecular Imaging* 12(4), pp. 406–414. doi: 10.1007/S11307-009-0291-3.

Minamoto, T. 2017. Random Mutagenesis by Error-Prone Polymerase Chain Reaction Using a Heavy Water Solvent. Humana Press, New York, NY, pp. 491–495. doi: 10.1007/978-1-4939-6472-7_33.

Minamoto, T., Wada, E. and Shimizu, I. 2012. A new method for random mutagenesis by error-prone polymerase chain reaction using heavy water. *Journal of Biotechnology* 157(1), pp. 71–74. doi: 10.1016/j.jbiotec.2011.09.012.

Moradi, A., Hosseinkhani, S., Naderi-Manesh, H., Sadeghizadeh, M. and Alipour, B.S. 2009. Effect of Charge Distribution in a Flexible Loop on the Bioluminescence Color of Firefly Luciferases. *Biochemistry* 48(3), pp. 575–582. doi: 10.1021/bi802057w.

Mortazavi, M. and Hosseinkhani, S. 2011. Design of thermostable luciferases through arginine saturation in solvent-exposed loops. *Protein Engineering, Design and Selection* 24(12), pp. 893–903. doi: 10.1093/protein/gzr051.

Nakajima, Y. et al. 2005. Multicolor luciferase assay system: one-step monitoring of multiple gene expressions with a single substrate. *BioTechniques* 38(6), pp. 891–894. doi: 10.2144/05386ST03.

Nakatsu, T., Ichiyama, S., Hiratake, J., Saldanha, A., Kobashi, N., Sakata, K. and Kato, H. 2006. Structural basis for the spectral difference in luciferase bioluminescence. *Nature* 440(7082), pp. 372–376. doi: 10.1038/nature04542.

Naumov, P. and Kochunnoonny, M. 2010. Spectral-structural effects of the keto-enol-enolate and phenol-phenolate equilibria of oxyluciferin. *Journal of the American Chemical Society* 132(33), pp. 11566–11579. doi: 10.1021/JA102885G.

Nazari, M. and Hosseinkhani, S. 2011. Design of disulfide bridge as an alternative mechanism for color shift in firefly luciferase and development of secreted luciferase. *Photochemical & Photobiological Sciences* 10(7), p. 1203. doi: 10.1039/c1pp05012e.

Noguchi, T. and Golden, S. 2017. Bioluminescent and fluorescent reporters in circadian rhythm studies., pp. 1–24.

Notomi, T., Okayama, H., Masubuchi, H., Yonekawa, T., Watanabe, K., Amino, N. and Hase, T. 2000. Loop-mediated isothermal amplification of DNA. *Nucleic Acids Research* 28(12), pp. 63e – 63. doi: 10.1093/nar/28.12.e63.

Oba, Y., Konishi, K., Yano, D., Shibata, H., Kato, D. and Shirai, T. 2020. Resurrecting the ancient glow of the fireflies. *Science Advances* 6(49), p. 5705. doi: 10.1126/SCIADV.ABC5705.

Oba, Y., Ojika, M. and Inouye, S. 2003. Firefly luciferase is a bifunctional enzyme: ATP-dependent monooxygenase and a long chain fatty acyl-CoA synthetase. *FEBS letters* 540(1–3), pp. 251–4. doi: 10.1016/S0014-5793(03)00272-2.

Oba, Y., Ojika, M. and Inouye, S. 2004. Characterization of CG6178 gene product with high sequence similarity to firefly luciferase in Drosophila melanogaster. *Gene* 329(1–2), pp. 137–145. doi: 10.1016/j.gene.2003.12.026.

Oba, Y., Sato, M. and Inouye, S. 2006. Cloning and characterization of the homologous genes of firefly luciferase in the mealworm beetle, Tenebrio molitor. *Insect Molecular Biology* 15(3), pp. 293–299. doi: 10.1111/j.1365-2583.2006.00646.x.

Oba, Y., Sato, M., Ojika, M. and Inouye, S. 2005. Enzymatic and genetic characterization of firefly luciferase and Drosophila CG6178 as a fatty acyl-Coa synthetase. *Bioscience, Biotechnology and Biochemistry* 69(4), pp. 819–828. doi: 10.1271/bbb.69.819.

Patrick, W.M. and Firth, A.E. 2005. Strategies and computational tools for improving randomized protein libraries. *Biomolecular Engineering* 22(4), pp. 105–112. doi: 10.1016/J.BIOENG.2005.06.001.

Patrick, W.M., Firth, A.E. and Blackburn, J.M. 2003. User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. *Protein Engineering Design and Selection* 16(6), pp. 451–457. doi: 10.1093/protein/gzg057.

Pokkuluri, P.R., Raffen, R., Dieckman, L., Boogaard, C., Stevens, F.J. and Schiffer, M. 2002. Increasing protein stability by polar surface residues: domain-wide consequences of interactions within a loop. *Biophysical journal* 82(1 Pt 1), pp. 391–398. doi: 10.1016/S0006-3495(02)75403-9.

Pongsupasa, V., Anuwan, P., Maenpuen, S. and Wongnate, T. 2022. Rational-Design Engineering to Improve Enzyme Thermostability. In: *Methods in Molecular Biology*. Humana, New York, NY, pp. 159–178. doi: 10.1007/978-1-0716-1826-4_9.

Prebble, S., Price, R., Lingard, B., Tisi, L. and White, P. 2001. Protein Engineering and Molecular Modelling of Firefly Luciferase., pp. 181–184. doi: 10.1142/9789812811158_0045.

Promega Corporation 2013. Glo^{TM} Luciferase Assay Systems Stable, Sensitive Bioluminescence Detection.

Promega Corporation 2015. Kinase-Glo® Luminescent Kinase Assay Platform Technical Bulletin.

Ratnasingham, S. and Hebert, P.D.N. 2007. bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes* 7(3), pp. 355–364. doi: 10.1111/J.1471-8286.2007.01678.X.

Rees, J.-F., De Wergifosse, B., Noiset, O., Dubuisson, M., Janssens, B. and Thompson, E.M. 1998. The origins of marine bioluminescence: turning oxygen defence mechanisms into deep-sea communication tools. *The Journal of Experimental Biology* 201, pp. 1211–1221. doi: 10.1242/jeb.201.8.1211.

Ribeiro, C. and Esteves da Silva, J.C.G. 2008. Kinetics of inhibition of firefly luciferase by oxyluciferin and dehydroluciferyl-adenylate. *Photochemical & Photobiological Sciences* 7(9), p. 1085. doi: 10.1039/b809935a.

Rice, B.W., Cable, M.D. and Nelson, M.B. 2002. In vivo imaging of light-emitting probes. *Journal of Biomedical Optics* 6(4), p. 432. doi: 10.1117/1.1413210.

Rice, B.W. and Contag, C.H. 2009. The importance of being red. *Nature Biotechnology* 27(7), pp. 624–625. doi: 10.1038/nbt0709-624.

Rodrigues, J.P.G.L.M. et al. 2013. Defining the limits of homology modeling in information-driven protein docking. *Proteins* 81(12), pp. 2119–2128. doi: 10.1002/PROT.24382.

Rose, T.M., Henikoff, J.G. and Henikoff, S. 2003. CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. *Nucleic acids research* 31(13), pp. 3763–6. doi: 10.1093/nar/gkg524.

Rose, T.M., Schultz, E.R., Henikoff, J.G., Pietrokovski, S., McCallum, C.M. and Henikoff, S. 1998. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Research* 26(7), pp. 1628–1635. doi: 10.1093/nar/26.7.1628.

Sandalova, T.P. and Ugarova, N.N. 1999. Model of the active site of firefly luciferase. *Biochemistry. Biokhimiia* 64(8), pp. 962–7.

Schneider, C.A., Rasband, W.S. and Eliceiri, K.W. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* 9(7), pp. 671–675. doi: 10.1038/NMETH.2089.

Selan, L., Berlutti, F., Passariello, C., Thaller, M.C. and Renzini, G. 1992. Reliability of a bioluminescence ATP assay for detection of bacteria. *Journal of clinical microbiology* 30(7), pp. 1739–42.

Shimomura, O. 1995. A short story of aequorin. *The Biological bulletin* 189(1), pp. 1–5. doi: 10.2307/1542194.

Shimomura, O. 2006. *Bioluminescence: Chemical principles and methods*. World Scientific Publishing Co. doi: 10.1142/6102.

Si, M., Xu, Q., Jiang, L. and Huang, H. 2016. SpyTag/SpyCatcher Cyclization Enhances the Thermostability of Firefly Luciferase. van Raaij, M. J. ed. *PLOS ONE* 11(9), p. e0162318.

doi: 10.1371/journal.pone.0162318.

Sivinski, J. [no date]. The Nature and Possible Functions of Luminescence in Coleoptera Larvae. *The Coleopterists Bulletin* 35, pp. 167–179. doi: 10.2307/4007935.

Staheli, J.P., Boyce, R., Kovarik, D. and Rose, T.M. 2011. CODEHOP PCR and CODEHOP PCR primer design. *Methods in molecular biology (Clifton, N.J.)* 687, pp. 57–73. doi: 10.1007/978-1-60761-944-4_5.

Stanger-Hall, K.F., Lloyd, J.E. and Hillis, D.M. 2007. Phylogeny of North American fireflies (Coleoptera: Lampyridae): Implications for the evolution of light signals. *Molecular Phylogenetics and Evolution* 45(1), pp. 33–49. doi: 10.1016/J.YMPEV.2007.05.013.

Stanley, P.E. 1989. A review of bioluminescent ATP techniques in rapid microbiology. *Journal of Bioluminescence and Chemiluminescence* 4(1), pp. 375–380. doi: 10.1002/bio.1170040151.

Stemmer, W.P. 1994. DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America* 91(22), pp. 10747–51. doi: 10.1073/pnas.91.22.10747.

Stowe, C.L. et al. 2019. Near-infrared dual bioluminescence imaging in mouse models of cancer using infraluciferin. *eLife* 8. doi: 10.7554/ELIFE.45801.

Sun, Y.Q., Liu, J., Wang, P., Zhang, J. and Guo, W. 2012. D-luciferin analogues: A multicolor toolbox for bioluminescence imaging. *Angewandte Chemie - International Edition* 51(34), pp. 8428–8430. doi: 10.1002/ANIE.201203565.

Sundlov, J.A., Fontaine, D.M., Southworth, T.L., Branchini, B.R. and Gulick, A.M. 2012. Crystal Structure of Firefly Luciferase in a Second Catalytic Conformation Supports a Domain Alternation Mechanism. *Biochemistry* 51(33), pp. 6493–6495. doi: 10.1021/bi300934s.

Takenaka, Y., Ikeo, K. and Shigeri, Y. 2016. Molecular cloning of secreted luciferases from marine planktonic copepods. In: *Methods in Molecular Biology*. Humana Press Inc., pp. 33–41. doi: 10.1007/978-1-4939-3813-1_3.

Takenaka, Y., Masuda, H., Yamaguchi, A., Nishikawa, S., Shigeri, Y., Yoshida, Y. and Mizuno, H. 2008. Two forms of secreted and thermostable luciferases from the marine copepod crustacean, Metridia pacifica. *Gene* 425(1–2), pp. 28–35. doi: 10.1016/j.gene.2008.07.041.

Thompson, E.M., Nafpaktitis, B.G. and Tsuji, F.I. 1988. Dietary uptake and blood transport of Vargula (crustacean) luciferin in the bioluminescent fish, Porichthys notatus. *Comparative Biochemistry and Physiology -- Part A: Physiology* 89(2), pp. 203–209. doi: 10.1016/0300-9629(88)91079-1.

Thompson, E.M., Nagata, S. and Tsuji, F.I. 1989. Cloning and expression of cDNA for the luciferase from the marine ostracod Vargula hilgendorfii. *Proceedings of the National Academy of Sciences of the United States of America* 86(17), pp. 6567–71. doi: 10.1073/pnas.86.17.6567.

Thomsen, P.F. et al. 2009. Non-destructive sampling of ancient insect DNA. *PLoS ONE* . doi: 10.1371/journal.pone.0005048.

Timmins, G.S., Jackson, S.K. and Swartz, H.M. 2001. The evolution of bioluminescent oxygen consumption as an ancient oxygen detoxification mechanism. *Journal of Molecular Evolution* 52(4), pp. 321–332. doi: 10.1007/s002390010162.

Tisi, L.., White, P.., Squirrell, D.., Murphy, M.., Lowe, C.. and Murray, J.A.. 2002a. Development of a thermostable firefly luciferase. *Analytica Chimica Acta* 457(1), pp. 115–123. doi: 10.1016/S0003-2670(01)01496-9.

Tisi, L.C., Law, G.H., Gandelman, O., Lowe, C.R. and Murray, J. 2002b. The Basis of the Bathochromic Shift in the Luciferase From Photinus Pyralis., pp. 57–60. doi: 10.1142/9789812776624_0012.

Trowell, S.C., Dacres, H., Dumancic, M.M., Leitch, V. and Rickards, R.W. 2016. Molecular basis for the blue bioluminescence of the Australian glow-worm Arachnocampa richardsae (Diptera: Keroplatidae). *Biochemical and Biophysical Research Communications* 478(2), pp. 533–539. doi: 10.1016/J.BBRC.2016.07.081.

Tu, S.-C. and Mager, H.I.X. 1995. Biochemistry of Bacterial Bioluminescence. *Photochemistry and Photobiology* 62(4), pp. 615–624. doi: 10.1111/j.1751-1097.1995.tb08708.x.

Ugarova, N.N. 1989. Luciferase of Luciola mingrelica fireflies. Kinetics and regulation mechanism. *Journal of bioluminescence and chemiluminescence* 4(1), pp. 406–418. doi: 10.1002/BIO.1170040155.

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

Underwood, T.J., Tallamy, D.W. and Pesek, J.D. 1997. Bioluminescence in firefly larvae: A test of the aposematic display hypothesis (Coleoptera: Lampyridae). *Journal of Insect Behavior* 10(3), pp. 365–370. doi: 10.1007/BF02765604.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. and Rozen, S.G. 2012. Primer3—new capabilities and interfaces. *Nucleic Acids Research* 40(15), p. e115. doi: 10.1093/NAR/GKS596.

Vartanian, J.P., Henry, M. and Wain-Hobson, S. 1996. Hypermutagenic PCR Involving All Four Transitions and a Sizeable Proportion of Transversions. *Nucleic Acids Research* 24(14), pp. 2627–2631. doi: 10.1093/NAR/24.14.2627.

Vencl, F. V., Luan, X., Fu, X. and Maroja, L.S. 2017. A day-flashing Photinus firefly (Coleoptera: Lampyridae) from central Panamá: An emergent shift to predator-free space? *Insect Systematics and Evolution* 48(5), pp. 512–531. doi: 10.1163/1876312X-48022162.

Verhaegen, M. and Christopoulos, T.K. 2002. Recombinant Gaussia Luciferase. Overexpression, Purification, and Analytical Application of a Bioluminescent Reporter for DNA Hybridization. doi: 10.1021/AC025742K.

Viviani, V.R., Arnoldi, F.G.C., Neto, A.J.S., Oehlmeyer, T.L., Bechara, E.J.H. and Ohmiya, Y. 2008. The structural origin and biological function of pH-sensitivity in firefly luciferases. *Photochemical and Photobiological Sciences* 7(2), pp. 159–169. doi: 10.1039/B714392C.

Viviani, V.R., Arnoldi, F.G.C., Venkatesh, B., Neto, A.J.S., Ogawa, F.G.T., Oehlmeyer, A.T.L. and Ohmiya, Y. 2006. Active-Site Properties of Phrixotrix Railroad Worm Green and Red Bioluminescence-Eliciting Luciferases. *J. Biochem* 140, pp. 467–474. doi: 10.1093/jb/mvj190.

Viviani, V.R., Bechara, E.J.H. and Ohmiya, Y. 1999. Cloning, sequence analysis, and expression of active Phrixothrix railroad-worms luciferases: Relationship between bioluminescence spectra and primary structures. *Biochemistry* 38(26), pp. 8271–8279. doi:

10.1021/bi9900830.

Viviani, V.R., Silva Neto, A.J., Arnoldi, F.G.C., Barbosa, J.A.R.G. and Ohmiya, Y. 2007. The Influence of the Loop between Residues 223-235 in Beetle Luciferase Bioluminescence Spectra: A Solvent Gate for the Active Site of pH-Sensitive Luciferases. *Photochemistry and Photobiology* 84(1), pp. 138–144. doi: 10.1111/j.1751-1097.2007.00209.x.

Vysotski, E.S. and Lee, J. 2004. Ca 2+ -Regulated Photoproteins: Structural Insight into the Bioluminescence Mechanism. *Cheminform* 35(34), pp. 405–415. doi: 10.1002/chin.200434297.

Wang, Z., Zhao, F., Peng, J. and Xu, J. 2011. Protein 8-class secondary structure prediction using conditional neural fields. *Proteomics* 11(19), pp. 3786–3792. doi: 10.1002/PMIC.201100196.

Waterhouse, A. et al. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* 46(W1), pp. W296–W303. doi: 10.1093/NAR/GKY427.

Watts, P.C., Thompson, D.J., Allen, K.A. and Kemp, S.J. 2007. How useful is DNA extracted from the legs of archived insects for microsatellite-based population genetic analyses? *Journal of Insect Conservation* . doi: 10.1007/s10841-006-9024-y.

De Wergifosse, B., Dubuisson, M., Marchand-Brynaert, J., Trouet, A. and Rees, J.F. 2004. Coelenterazine: A two-stage antioxidant in lipid micelles. *Free Radical Biology and Medicine* 36(3), pp. 278–287. doi: 10.1016/j.freeradbiomed.2003.11.008.

De Wet, J.R., Wood, K. V, Deluca, M., Helinski, D.R. and Subramani1, S. 1987. Firefly Luciferase Gene: Structure and Expression in Mammalian Cells. *Molecular And Cellular Biology* 7(2), pp. 725–737. doi: 10.1128/mcb.7.2.725-737.1987.

White, E.H., Rapaport, E., Seliger, H.H. and Hopkins, T.A. 1971. The chemi- and bioluminescence of firefly luciferin: An efficient chemical production of electronically excited states. *Bioorganic Chemistry* 1(1–2), pp. 92–122. doi: 10.1016/0045-2068(71)90009-5.

White, E.H., Steinmetz, M.G., Miano, J.D., Wildes, P.D. and Morland, R. 1980. Chemi- and bioluminescence of firefly luciferin. *Journal of the American Chemical Society* 102(9), pp.

3199–3208. doi: 10.1021/ja00529a051.

White, E.H., Wörther, H., Seliger, H.H. and McElroy, W.D. 1966. Amino Analogs of Firefly Luciferin and Biological Activity Thereof. *Journal of the American Chemical Society* 88(9), pp. 2015–2019. doi: 10.1021/ja00961a030.

White, P.J., Leslie, R.L., Lingard, B., Williams, J.R. and Squirrell, D.J. 2002. Novel in Vivo Reporters Based on Firefly Luciferase., pp. 509–512. doi: 10.1142/9789812776624_0115.

White, P.J., Squirrell, D.J., Arnaud, P., Lowe, C.R. and Murray, J.A. 1996. Improved thermostability of the North American firefly luciferase: saturation mutagenesis at position 354. *The Biochemical journal* 319(Pt 2), pp. 343–350. doi: 10.1042/bj3190343.

Willey, T.L., Squirrell, D.J. and White, P.J. 2001. Design and selection of Firefly Luciferases with novel in vivo and in vitro properties. In: *Bioluminescence and Chemiluminescence*. World Scientific, pp. 201–204. doi: 10.1142/9789812811158_0050.

Williams, C.J. et al. 2018. MolProbity: More and better reference data for improved all-atom structure validation. *Protein science : a publication of the Protein Society* 27(1), pp. 293–315. doi: 10.1002/PRO.3330.

Wilson, I.G. 1997. Inhibition and facilitation of nucleic acid amplification. *Applied and Environmental Microbiology* 63(10), pp. 3741–3751.

Wong, T., Zhurina, D. and Schwaneberg, U. 2006. The Diversity Challenge in Directed Protein Evolution. *Combinatorial Chemistry & High Throughput Screening* 9(4), pp. 271–288. doi: 10.2174/138620706776843192.

Wood, K. V. and DeLuca, M. 1987. Photographic detection of luminescence in Escherichia coli containing the gene for firefly luciferase. *Analytical Biochemistry* 161(2), pp. 501–507. doi: 10.1016/0003-2697(87)90480-5.

Wood, K. V., Lam, Y.A. and McElroy, W.D. 1989. Introduction to beetle luciferases and their applications. *Journal of Bioluminescence and Chemiluminescence* 4(1), pp. 289–301. doi: 10.1002/bio.1170040141.

Zambito, G., Chawda, C. and Mezzanotte, L. 2021. Emerging tools for bioluminescence imaging. *Current Opinion in Chemical Biology* 63, pp. 86–94. doi: 10.1016/J.CBPA.2021.02.005.

Zeale, M.R.K., Butlin, R.K., Barker, G.L.A., Lees, D.C. and Jones, G. 2011. Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces. *Molecular Ecology Resources* 11(2), pp. 236–244. doi: 10.1111/j.1755-0998.2010.02920.x.

Zhang, R. et al. 2020a. Genomic and experimental data provide new insights into luciferin biosynthesis and bioluminescence evolution in fireflies. *Scientific Reports* 10(1). doi: 10.1038/S41598-020-72900-Z.

Zhang, Y., Ren, G., Buss, J., Barry, A.J., Patton, G.C. and Tanner, N.A. 2020b. Enhancing colorimetric loop-mediated isothermal amplification speed and sensitivity with guanidine chloride. *BioTechniques* 69(3), pp. 179–185. doi: 10.2144/BTN-2020-0078.

Zhao, H. and Arnold, F.H. 1997. Optimization of DNA shuffling for high fidelity recombination. *Nucleic Acids Research* 25(6). doi: 10.1093/nar/25.6.1307.

9.0. Appendices

9.1. Supplementary Figures to Chapter 3

Table 9.1.

DNA Sample	Successful Amplification	Sequencing Successful?	Sequencing Results Clipped Length	Top BLAST Description	Accession	Percent Identity
Lampyris noctiluca gDNA	Yes	Yes	162 bp	Lampyris noctiluca mitochondrion, partial genome	MN122858.1	98.11%
Photinus pyralis gDNA	No	-	-	-	-	-
Costa Rica firefly Illumina library	Yes	Yes	164 bp	Photinus australis cytochrome c oxidase subunit I (COI) gene, partial sequence, mitochondrial gene	EU009298.1	92.31%
Indonesia firefly Illumina library	No	-	-	-	-	-
USA – unk. firefly Illumina library	Yes	Yes	91 bp	Lucidota sp. 18-RU-1B1-2609 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	MK091857.1	98.86%
USA – Maryland firefly Illumina library	Yes	Yes	92 bp	Photinus sp. 1 LSM-2017 isolate L2 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	KX909936.1	95.24%
USA – Pennsylvania firefly Illumina library	Yes	Yes	167bp	Lucidota atra voucher BIOUG20382-H08 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	MF636105.1	98.73%

Sequencing of COI mini-barcodes. Details of sequencing results for amplifications with Zeale *et al* (2011) primers ZBJ-ArtF1c and ZBJ-ArtR2c. The description, accession, and percent identity of the top BLAST match are provided.

Table 9.2.

Species	**Accession number**
Photinus pyralis	AAA29795.1
Luciola lateralis	AAN73267.1
Luciola mingrelica	AAB26932.1
Luciola cruciata	BAE80731.1
Luciola terminalis	ABZ88151.1
Luciola parvula	BAU71688.1
Luciola tsushimana	AAN40979.1
Lampyris turkestanicus	AAU85360.1
Lampyris noctiluca	AAW72003.1
Cratomorphus distinctus	AAV32457.1
Photuris pennsylvanica	BAA05005.1

Luciferase genes used in CODEHOP design. Eleven Coleopteran luciferase protein sequences used in the design of CODEHOP primers DKYD-F and GYG-R.

Table 9.3.

DNA Sample	**Successful Amplification**	**Sequencing Successful?**	**Sequencing Results Clipped Length**	**Top BLAST Description**	**Accession**	**Percent Identity**
Lampyris noctiluca gDNA	Yes	Yes	175 bp	Lampyris noctiluca clone LanLUC luciferase gene, partial cds	EU684100.1	99.28%
Photinus pyralis gDNA	Yes	Yes	179 bp	Photinus pyralis voucher KSH 11022 luciferase 1 (LUC1) gene, complete cds	MH759196.1	94.67%
Eluc in pET16b vector	Yes	Yes	132 bp	Cloning vector pLR6-Eluc, complete sequence	KU756582.1	95.87%
CBR in pET16b vector	Yes	No	-	-	-	-
Costa Rica firefly sequencing library	Yes	Yes	175 bp	Photinus pyralis voucher KSH 11022 luciferase 1 (LUC1) gene, complete cds	MH759196.1	86.83%
Indonesia firefly sequencing library	Yes	Yes	127 bp	PREDICTED: Photinus pyralis luciferin 4-monooxygenase (LOC116160065), mRNA	XM_031473197.1	93.16%
USA – unk. firefly sequencing library	Yes	Yes	102 bp	PREDICTED: Photinus pyralis luciferin 4-monooxygenase (LOC116160065), mRNA	XM_031473197.1	91.30%
USA – Maryland firefly sequencing library	Yes	Yes	124 bp	PREDICTED: Photinus pyralis luciferin 4-monooxygenase (LOC116160065), mRNA	XM_031473197.1	93.16%
USA – Pennsylvania firefly sequencing library	Yes	Yes	122 bp	PREDICTED: Photinus pyralis luciferin 4-monooxygenase (LOC116160065), mRNA	XM_031473197.1	92.98%

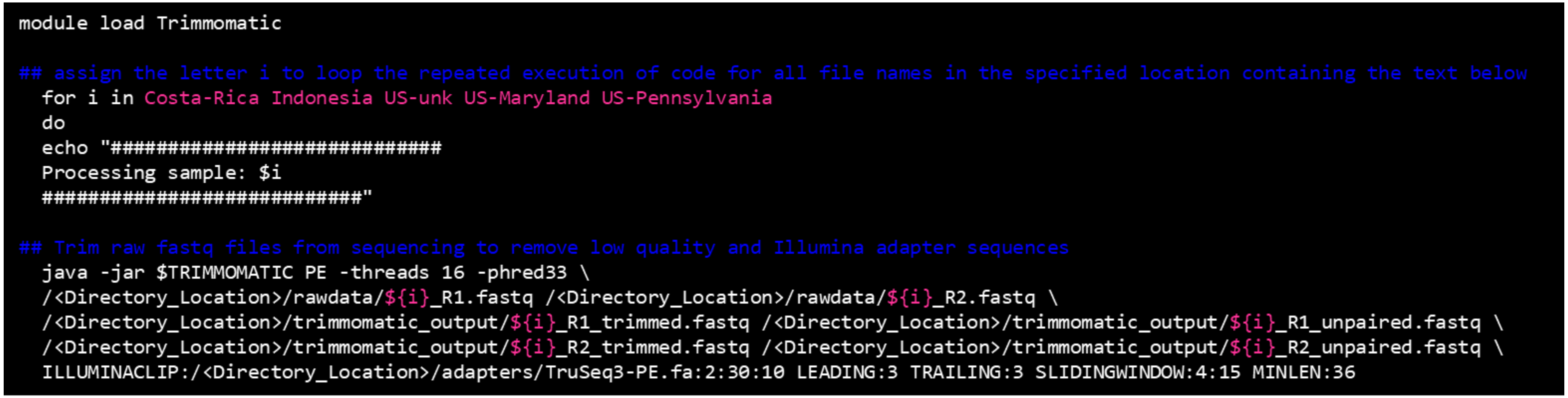
Sequencing from CODEHOP DKYD-F > GYG-R amplification. Details of sequencing results for amplifications with CODEHOP primers DKYD-F and GYG-R. The description, accession, and percent identity of the top BLAST match are provided.

Table 9.4.

Sample	**gDNA Extract Concentration**	**Pre Library-prep Average Fragment Size**	**Post Library-prep Average Fragment Size**	**Library DNA Concentration**
Photinus pyralis, 1996	99.2 ng/µl	217 bp	369 bp	4.32 ng/µl
Lampyris noctiluca, 2006	164 ng/µl	17273 bp	500 bp	9.42 ng/µl
Costa Rica, 2012	10.4 ng/µl	881 bp	323 bp	9.8 ng/µl
Indonesia, 1985	2.74 ng/µl	146 bp	259 bp	9.44 ng/µl
USA – unk., 2013	16.6 ng/µl	172 bp	274 bp	17.8 ng/µl
USA – Maryland, 2015	14.5 ng/µl	146 bp	267 bp	18.2 ng/µl
USA – Pennsylvania, 2015	5.52 ng/µl	178 bp	255 bp	17.4 ng/µl

DNA quality pre and post library preparation. Details of DNA concentration and average fragment size before and after NEXTFLEX Illumina library preparation. DNA concentration as measured by Fluorometric Quantification on the Qubit 4 Fluorometer (ThermoFisher, MA, USA) and average fragment size of DNA extracts measured by analysis on the 4200 TapeStation System (Agilent, CA, USA).

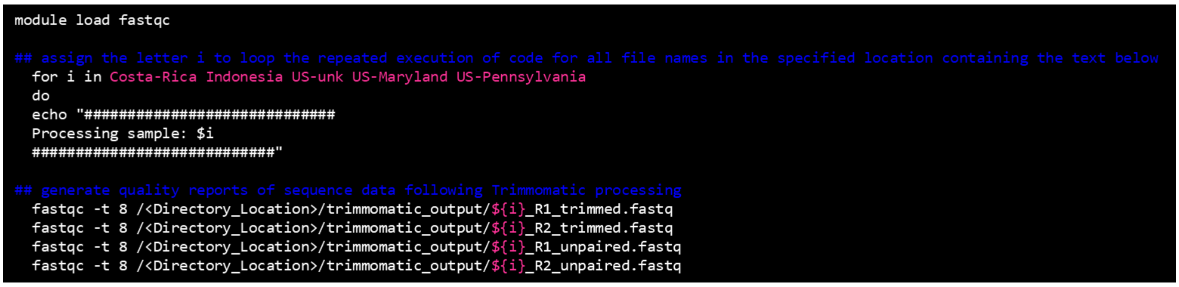
Figure 9.1.



```
module load Trimmomatic

## assign the letter i to loop the repeated execution of code for all file names in the specified location containing the text below
  for i in Costa-Rica Indonesia US-unk US-Maryland US-Pennsylvania
  do
  echo "#############################
  Processing sample: $i
  ############################"

## Trim raw fastq files from sequencing to remove low quality and Illumina adapter sequences
  java -jar $TRIMMOMATIC PE -threads 16 -phred33 \
  /<Directory_Location>/rawdata/${i}_R1.fastq /<Directory_Location>/rawdata/${i}_R2.fastq \
  /<Directory_Location>/trimmomatic_output/${i}_R1_trimmed.fastq /<Directory_Location>/trimmomatic_output/${i}_R1_unpaired.fastq \
  /<Directory_Location>/trimmomatic_output/${i}_R2_trimmed.fastq /<Directory_Location>/trimmomatic_output/${i}_R2_unpaired.fastq \
  ILLUMINACLIP:/<Directory_Location>/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

Annotated Trimmomatic script. Script used to process sequenced libraries with Trimmomatic (Bolger *et al.* 2014). Annotations describing the function of the following section of code are shown in blue. Sample names and their designation of '${i}' are shown in purple to indicate where in the script each sample name will be substituted in as the script is looped to process each sample. R1 and R2 files are forward and reverse paired reads, respectively. Singletons are reads which have no identified pairing. Trimmomatic is available at https://github.com/usadellab/Trimmomatic.

Figure 9.2.



```
module load fastqc

## assign the letter i to loop the repeated execution of code for all file names in the specified location containing the text below
  for i in Costa-Rica Indonesia US-unk US-Maryland US-Pennsylvania
  do
  echo "#############################
  Processing sample: $i
  ############################"

## generate quality reports of sequence data following Trimmomatic processing
  fastqc -t 8 /<Directory_Location>/trimmomatic_output/${i}_R1_trimmed.fastq
  fastqc -t 8 /<Directory_Location>/trimmomatic_output/${i}_R2_trimmed.fastq
  fastqc -t 8 /<Directory_Location>/trimmomatic_output/${i}_R1_unpaired.fastq
  fastqc -t 8 /<Directory_Location>/trimmomatic_output/${i}_R2_unpaired.fastq
```

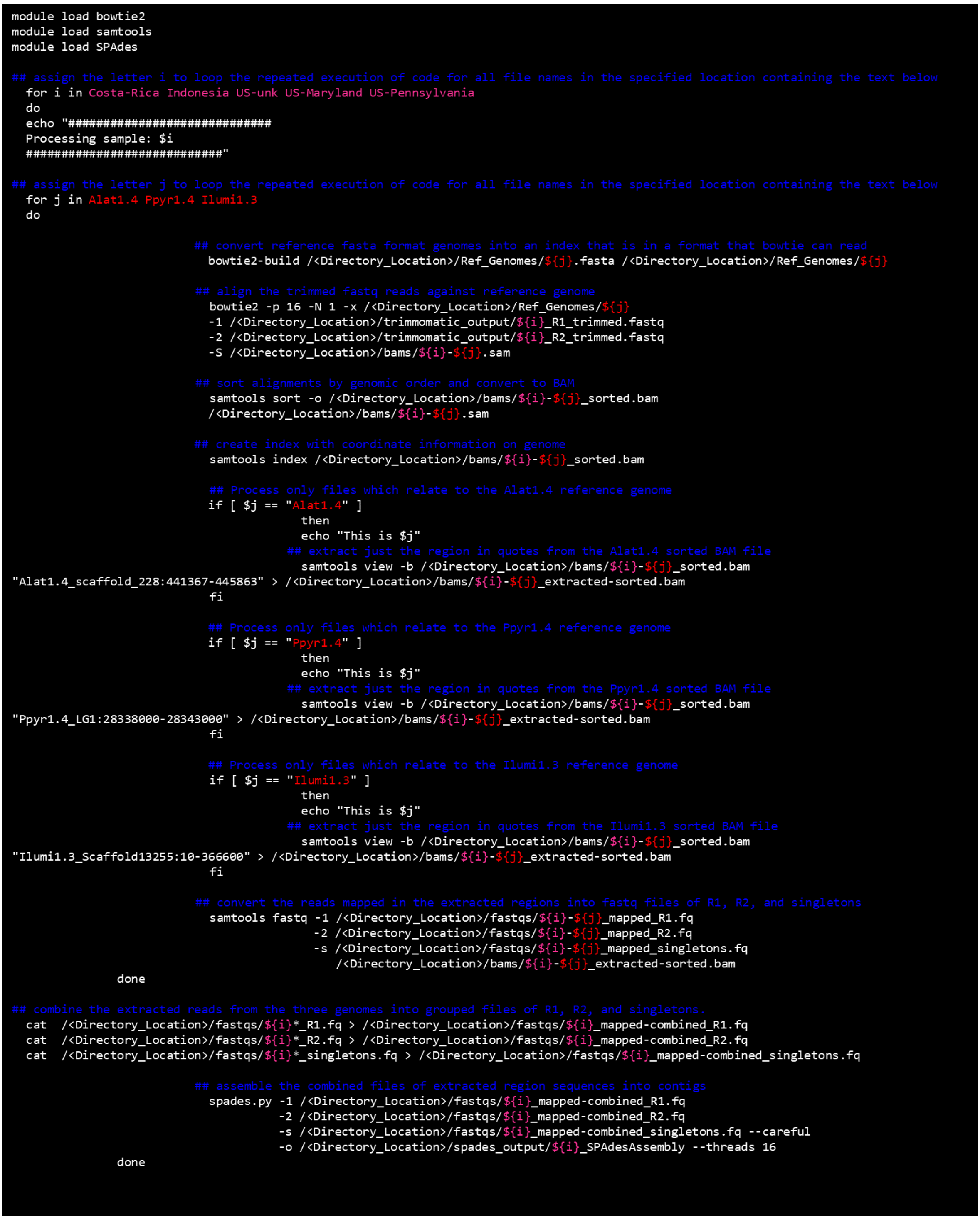
Annotated FastQC script. Script used to generate quality reports on trimmed libraries with FastQC. Annotations describing the function of the following section of code are shown in blue. Sample names and their designation of '${i}' are shown in purple to indicate where in the script each sample name will be substituted in as the script is looped to process each sample. R1 and R2 files are forward and reverse paired reads, respectively. Singletons are reads which have no identified pairing. FastQC is available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Table 9.5.

Sample Library	**Total Paired Sequences**	**GC Content**
Costa Rica, 2012	114752	42%
Indonesia, 1985	48543	35%
USA - unk., 2013	61143	39%
USA - Maryland, 2015	57829	37%
USA - Pennsylvania, 2015	49603	37%
Costa Rica, 2012 (Non-enriched)	403846	39%

Overview of FastQC reports for paired data. Total paired sequences and GC content are detailed.

Figure 9.3.



```
module load bowtie2
module load samtools
module load SPAdes

## assign the letter i to loop the repeated execution of code for all file names in the specified location containing the text below
  for i in Costa-Rica Indonesia US-unk US-Maryland US-Pennsylvania
  do
  echo "#############################
  Processing sample: $i
  #############################"

## assign the letter j to loop the repeated execution of code for all file names in the specified location containing the text below
  for j in Alat1.4 Ppyr1.4 Ilumi1.3
  do

                          ## convert reference fasta format genomes into an index that is in a format that bowtie can read
                            bowtie2-build /<Directory_Location>/Ref_Genomes/${j}.fasta /<Directory_Location>/Ref_Genomes/${j}

                          ## align the trimmed fastq reads against reference genome
                            bowtie2 -p 16 -N 1 -x /<Directory_Location>/Ref_Genomes/${j}
                            -1 /<Directory_Location>/trimmomatic_output/${i}_R1_trimmed.fastq
                            -2 /<Directory_Location>/trimmomatic_output/${i}_R2_trimmed.fastq
                            -S /<Directory_Location>/bams/${i}-${j}.sam

                          ## sort alignments by genomic order and convert to BAM
                            samtools sort -o /<Directory_Location>/bams/${i}-${j}_sorted.bam
                            /<Directory_Location>/bams/${i}-${j}.sam

                          ## create index with coordinate information on genome
                            samtools index /<Directory_Location>/bams/${i}-${j}_sorted.bam

                            ## Process only files which relate to the Alat1.4 reference genome
                            if [ $j == "Alat1.4" ]
                                         then
                                         echo "This is $j"
                                       ## extract just the region in quotes from the Alat1.4 sorted BAM file
                                         samtools view -b /<Directory_Location>/bams/${i}-${j}_sorted.bam
"Alat1.4_scaffold_228:441367-445863" > /<Directory_Location>/bams/${i}-${j}_extracted-sorted.bam
                            fi

                            ## Process only files which relate to the Ppyr1.4 reference genome
                            if [ $j == "Ppyr1.4" ]
                                         then
                                         echo "This is $j"
                                       ## extract just the region in quotes from the Ppyr1.4 sorted BAM file
                                         samtools view -b /<Directory_Location>/bams/${i}-${j}_sorted.bam
"Ppyr1.4_LG1:28338000-28343000" > /<Directory_Location>/bams/${i}-${j}_extracted-sorted.bam
                            fi

                            ## Process only files which relate to the Ilumi1.3 reference genome
                            if [ $j == "Ilumi1.3" ]
                                         then
                                         echo "This is $j"
                                       ## extract just the region in quotes from the Ilumi1.3 sorted BAM file
                                         samtools view -b /<Directory_Location>/bams/${i}-${j}_sorted.bam
"Ilumi1.3_Scaffold13255:10-366600" > /<Directory_Location>/bams/${i}-${j}_extracted-sorted.bam
                            fi

                          ## convert the reads mapped in the extracted regions into fastq files of R1, R2, and singletons
                            samtools fastq -1 /<Directory_Location>/fastqs/${i}-${j}_mapped_R1.fq
                                           -2 /<Directory_Location>/fastqs/${i}-${j}_mapped_R2.fq
                                           -s /<Directory_Location>/fastqs/${i}-${j}_mapped_singletons.fq
                                              /<Directory_Location>/bams/${i}-${j}_extracted-sorted.bam
               done

## combine the extracted reads from the three genomes into grouped files of R1, R2, and singletons.
  cat  /<Directory_Location>/fastqs/${i}*_R1.fq > /<Directory_Location>/fastqs/${i}_mapped-combined_R1.fq
  cat  /<Directory_Location>/fastqs/${i}*_R2.fq > /<Directory_Location>/fastqs/${i}_mapped-combined_R2.fq
  cat  /<Directory_Location>/fastqs/${i}*_singletons.fq > /<Directory_Location>/fastqs/${i}_mapped-combined_singletons.fq

                          ## assemble the combined files of extracted region sequences into contigs
                            spades.py -1 /<Directory_Location>/fastqs/${i}_mapped-combined_R1.fq
                                      -2 /<Directory_Location>/fastqs/${i}_mapped-combined_R2.fq
                                      -s /<Directory_Location>/fastqs/${i}_mapped-combined_singletons.fq --careful
                                      -o /<Directory_Location>/spades_output/${i}_SPAdesAssembly --threads 16
               done
```

Annotated luciferase gene extraction script. Script used to extract reads corresponding to luciferase gene sequences from trimmed libraries through the execution of Bowtie2 (Langmead and Salzberg 2012), SAMtools (Li *et al.* 2009), and SPAdes (Bankevich *et al.* 2012). Annotations describing the function of the following section of code are shown in blue. Sample names and their designation of '${i}' are shown in purple to indicate where in the script each sample name will be substituted in as the script is looped to process each sample. Similarly, reference genomes and their designation of '${j}' are shown in red to indicate where in the script each reference genome name will be substituted in as the script is looped to process each sample. R1 and R2 files are forward and reverse paired reads, respectively. Singletons are reads which have no identified pairing. Bowtie2 is available at https://sourceforge.net/projects/bowtie-bio/. SAMtools is available at http://samtools.sourceforge.net. SPAdes is available at https://github.com/ablab/spades. Firefly reference genomes Alat1.4, Ppyr1.4, and Ilumi1.3 are available at http://fireflybase.org/.

Table 9.6.

Sample Library	**Aligned Genome**	**Read Type**	**Total Sequences**	**Top BLAST Description**	**Accession**	**Percent Identity**
Costa Rica, 2012	Ppyr1.4	Paired	182	-	-	-
		Singletons	889	-	-	-
Indonesia, 1985	Ppyr1.4	Paired	2	Photinus pyralis voucher KSH 11044 luciferase 1 (LUC1) gene, complete cds	MH759210.1	88.08%
		Singletons	0	N/A	N/A	N/A
USA - unk., 2013	Ppyr1.4	Paired	7	Photinus pyralis voucher KSH 11044 luciferase 1 (LUC1) gene, complete cds	MH759210.1	96.60%
		Singletons	4	PREDICTED: Photinus pyralis luciferin 4-monooxygenase (LOC116160065), mRNA	XM_031473197.1	97.35%
USA - Maryland, 2015	Ppyr1.4	Paired	2	Photinus pyralis voucher KSH 11044 luciferase 1 (LUC1) gene, complete cds	MH759210.1	92.70%
		Singletons	0	N/A	N/A	N/A
USA - Pennsylvania, 2015	Ppyr1.4	Paired	1	PREDICTED: Photinus pyralis luciferin 4-monooxygenase (LOC116160065), mRNA	XM_031473197.1	88.20%
		Singletons	0	N/A	N/A	N/A
Costa Rica, 2012 (Non-enriched)	Ppyr1.4	Paired	2	PREDICTED: Photinus pyralis uncharacterized LOC116173352 (LOC116173352), transcript variant X3, mRNA	XM_031490758.1	82.12%
		Singletons	0	N/A	N/A	N/A
	Ilumni1.3	Paired	1	PREDICTED: Photinus pyralis thyroid transcription factor 1 (LOC116171563), mRNA	XM_031488517.1	100%
		Singletons	0	N/A	N/A	N/A

Total reads mapped to reference genome region of interest. Total paired and singletons sequences extracted by the script in Figure 9.3., prior to SPAdes assembly. Top BLASTn match description, accessions and percent identity are provided. Costa Rica is omitted as SPAdes was able to assemble contigs for further analysis in Chapter 3.

Table 9.7.

Library Name	**Run**	**BioSample**	**Experiment**
USA - Pennsylvania, 2015	SRR17886597	SAMN25554923	SRX14045690
USA - Maryland, 2015	SRR17886598	SAMN25554922	SRX14045689
USA - unk., 2013	SRR17886599	SAMN25554921	SRX14045688
Indonesia, 1985	SRR17886601	SAMN25554920	SRX14045687
Costa Rica, 2012	SRR17886602	SAMN25554919	SRX14045686

NGS data accession. Individual accessions of the five dataset under the BioProject accession PRJNA802557. Accessions are made available for access using the NCBI SRA Run Selector (available at https://www.ncbi.nlm.nih.gov/Traces/study/), NCBI BioSample (available at https://www.ncbi.nlm.nih.gov/biosample), and an overview of experiment details at NCBI SRA (available at https://www.ncbi.nlm.nih.gov/sra).

9.2. Clipped Sanger Sequencing Results from Chapter 3

9.2.1. Trial DNA extractions mini-barcodes

>*Lnoc* mini-barcode 24..185 of sequence

GTACATCATTTAGATTGCTAATTCGAGCAGAATTAGGAAGGGCTGGAACCTTA
ATTGGAAATGACCATATTTTTAATGTTATTGTAACAAGTCATGCATTTATTATA
ATTTTTTTTATAGTTATACCTATTATAATTGGAGGATTTGGTAATTGATTAGTAA

>*Ppy* mini-barcode

N/A

9.2.2. Trial DNA extractions CODEHOP amplifications

>*Lnoc* CODEHOP 17..191 of sequence

CCACTTACATGAATTGCGTCTGGTGGAGCTCCCCTCGCGAAAGAAGTTGGAGA
AGCTGTAGCAAAACGGTAAGTCACGATACCAAGTACTCAGTGCCTATTAAGGC
TTTGTAGTTTTAAGCTGCCGGGAATACGACAAGGCTACGGCCTGACCGAGACC
ACCTCCGCTATCAAAA

>*Ppy* CODEHOP 25..203 of sequence

ACACGAATTGCTTCTGGGGGCGCACCTCTTTCGAAAGAAGTCGGGGAAGCGGT
TGCAAAACGGTGAGTTAAGCGCATTGCTAGTATTTCAAGGCTCTAAAACGGCG
CGTAGCTTCCATCTTCCAGGGATACGACAAGGCTACGGCCTGACCGAGACCAC
CTCCGCTATCAATGTGGTTG

9.2.3. Enriched libraries mini-barcodes

>Costa Rica mini-barcode 18..181 of sequence

CTAGGACATCTTTTAGATTACTAATTCGTGCAGAATTAGGGAGACCTGGATCTTTAATTGGAAATGACCACATTTTTAATGTAATTGTAACTAGTCATGCTTTTATCATAATTTTTTTCATAGTAATACCTATTATAATTGGAGGATTTGGTAATTGATTAATA

>Indonesia mini-barcode

N/A

>USA – unk. mini-barcode 54..144 of sequence

GGAACCCTGGATCATTAATTGGAAATGATCATATTTTTAATGAATTGTTACAAGTCATGCATTCATCATAATTTTCTTTATAGTAATACCA

>USA – Maryland mini-barcode 33..124 of sequence

GAATTATCTAATCTCGAGACAGAATTAGGTAATCCCATGGTATCATTAATTGGTAAATGATCATATTTATTAATGTAATTGTATACAACCCA

>USA – Pennsylvania mini-barcode 27..193 of sequence

GTTTCATCTTTTAGTCTACTAATTCGAACAGAATTAGGGATCCCTGGATCATTAATTGGAAATGATCATATTTTTAATGTAATTGTTACAAGTCATGCATTCATCATAATTTTCTTTATAGTAATACCAATTATAATTGGAGGATTTGGTAATTGATTAGTAAAAAT

9.2.4. Enriched libraries CODEHOP amplification

>Costa Rica Enriched CODEHOP 27..201 of sequence

CAAATTGCCTCTGGGCGGCAGCACCTCTTTCAAAAGAAAGTTGGAGAAGCGGT
TGCAAAACGGTGAGTTAAGGGCATTGCTTGTTCTCCAAGGCTCTAAAGCGGCG
TGTAGCTTCCATCTTCCAGGGATACGACAAGGCTACGGCCTGACCGAGACCAC
TCCGCTATCATATTTC

>Indonesia Enriched CODEHOP 11..137 of sequence

CCGAAATTGCTTCTGGGGGCGCACCTCTTTCGAAAGAAGTCGGGGAAGCGGTT
GCAAAACGCTTCCATCTTCCAGGGATACGACAAGGCTACGGCCTGACCGAGAC
CACCTCCGCTATCAAGTGCCT

>USA – unk. Enriched CODEHOP 42..143 of sequence

CTTTCGAAAGAAGTCGGGGAAGCGGTTGCAAAACGCTTCCATCTTCCAGGGAT
ACGACAAGGCTACGGCCTGACCGAGACCACCTCCGCTATCAAGTTCGGA

>USA – Maryland Enriched CODEHOP 11..134 of sequence

CCGAAATTGCTTCTGGGGGCGCACCTCTTTCGAAAGAAGTCGGGGAAGCGGTT
GCAAAACGCTTCCATCTTCCAGGGATACGACAAGGCTACGGCCTGACCGAGAC
CACCTCCGCTATCATGCC

>USA – Pennsylvania Enriched CODEHOP 14..135 of sequence

CGAATTGCTTCTGGGGGCGCACCTCTTTCGAAAGAAGTCGGGGAAGCGGTTGC
AAAACGCTTCCATCTTCCAGGGATACGACAAGGCTACGGCCTGACCGAGACCA
CCTCCGCTATCACACT

>Conserved 49bp section

GCTTCCATCTTCCAGGGATACGACAAGGCTACGGCCTGACCGAGACCAC

9.2.5. Non-enriched libraries CODEHOP amplification

>Costa Rica Non-enriched CODEHOP 56..197 of sequence

GCTAGCAATTGGAGAAGCGGTTGCAAAACGGTGAGTTAAGGGCATTGCTTGTT
CTCCAAGGATCTAAAGCGGCGTGTAGCTTCCATCTTCCAGGGATACGACAAGG
CTACGGCCTGACCGAGACCACCCTCCGCTATCAAAA

>Indonesia Non-enriched CODEHOP 132..197 of sequence

AGGTTCAAAATTAAATACGTTCGACAAGGCTACGGCCTGACCGAGACCACCCT
CCGCTATCAAAAT

> USA – unk. Non-enriched CODEHOP 55..100 of sequence

TGCCATACGGCCTGACCGAGACCACCTCCGCTATCAAAGAGACTCT

>USA – Maryland Non-enriched CODEHOP

N/A

> USA – Pennsylvania Non-enriched CODEHOP 55..94 of sequence

CCTACGGCCTGACCGAGACCACCTCCGCTATCAAGATTTT

9.3. Supplementary Figures to Chapter 4

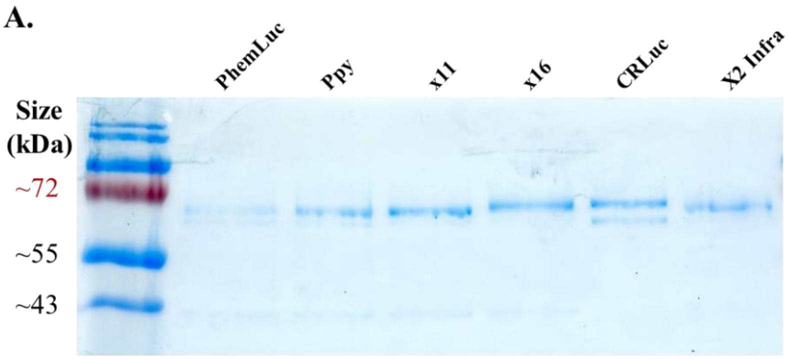
Table 9.8

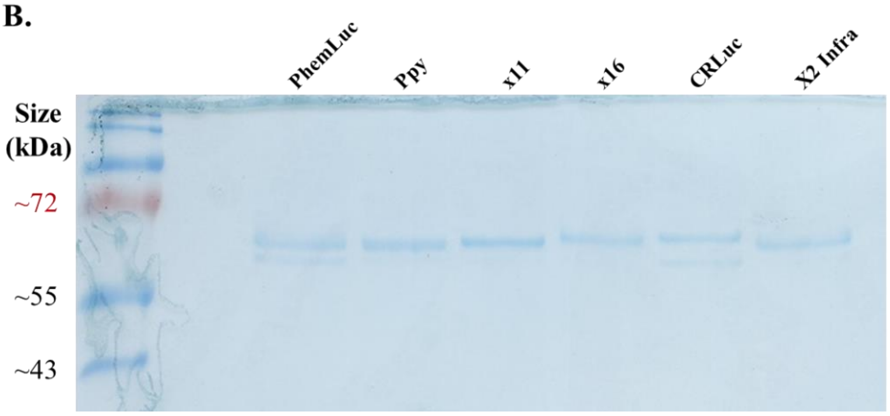
	Clashscore	**Molprobity score**
4G36 (*Ppy* DLSA)	7.95	2.48
***Phem*Luc DLSA**	5.76	1.70
6HPS (*Ppy* iDLSA)	3.34	2.26
***Phem*Luc iDLSA**	3.41	2.18
x2 Infra	7.05	1.78

Molprobity assessment of model quality. The clashscores and Molprobity score for each model. All scores were reported as "Good" by molprobity, which indicates a result ≥ 66th percentile. Analysis performed at http://molprobity.biochem.duke.edu/.

9.4. Supplementary Figures to Chapter 6

Figure 9.4.

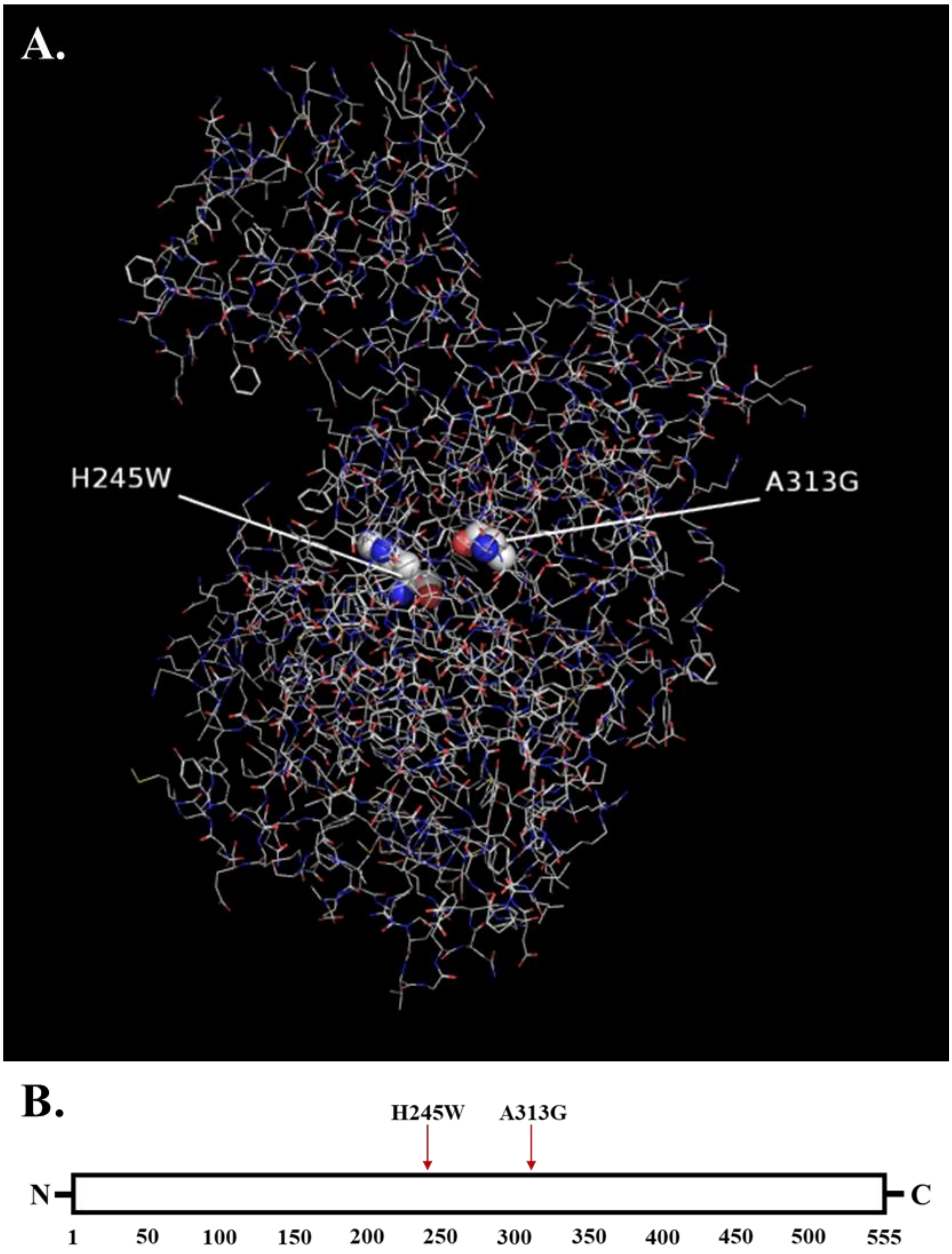




SDS-PAGE analysis for protein quantification. All Flucs prepared by diluting as described in A) or B), and mixing 3:1 in 4x protein sample buffer. **A)** All Flucs diluted to equal the lowest concentration as measured by Bradford assay (CRLuc – 0.057 mg/ml). **B)** All Flucs diluted to equal CRLuc concentration following correction by imageJ analysis of band size and intensity (Chapter 2).

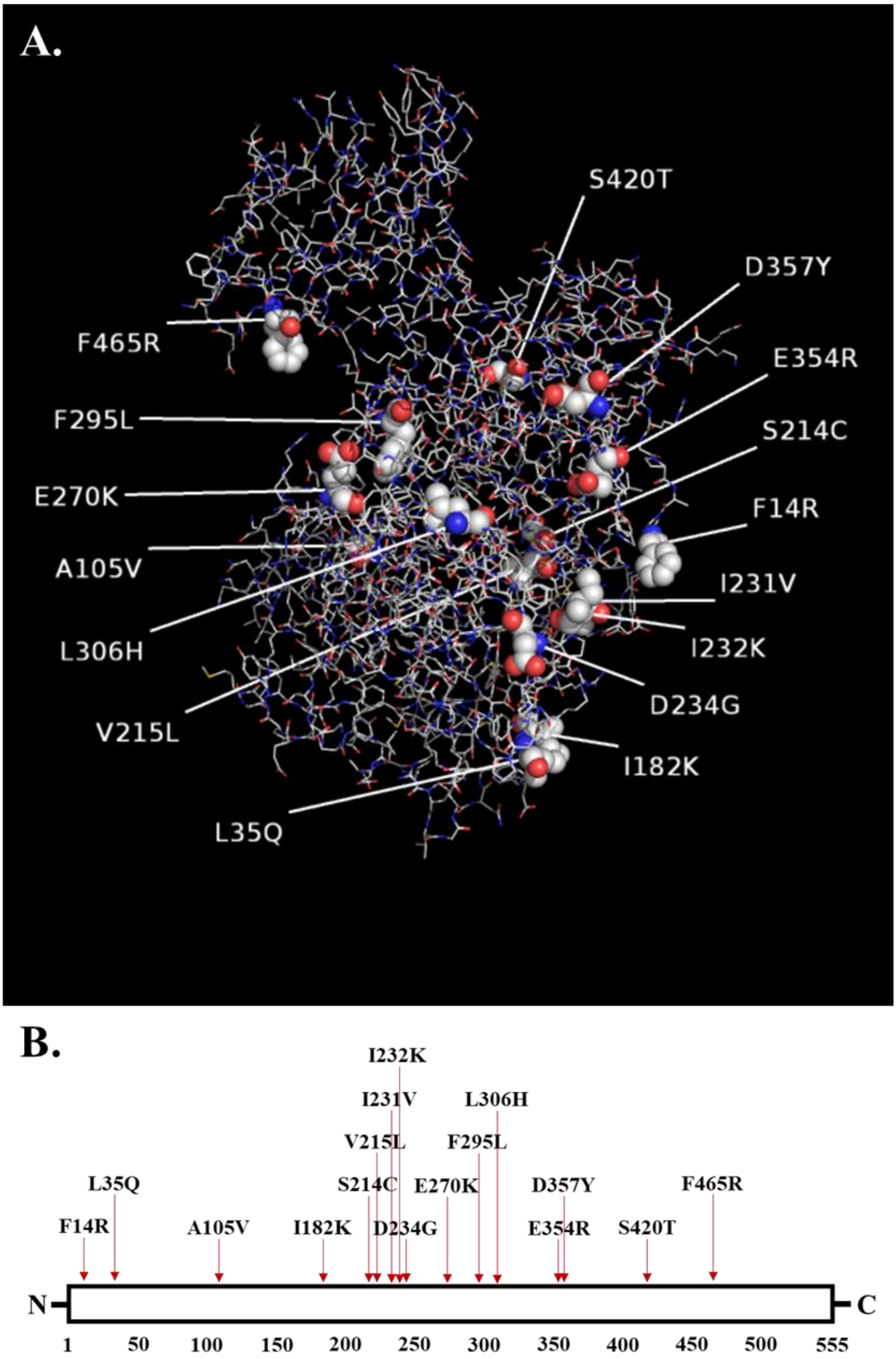
9.5. Supplementary Models of Fluc Mutants

Figure 9.5.



Positions of x2 Infra mutations in *Phem*Luc. (**A**) 3D stick model of *Phem*Luc with the positions mutated in x2 Infra represented as spheres and labelled. Model produced using PyMOL. (**B**) Linear block diagram indicating the relative locations of x2 Infra mutations in *Phem*Luc.

Figure 9.6.



Positions of x16 mutations in *Phem*Luc. (**A**) 3D stick model of *Phem*Luc with the positions mutated in x16 represented as spheres and labelled. Model produced using PyMOL. (**B**) Linear block diagram indicating the relative locations of x16 mutations in *Phem*Luc.