The evolution and specialized metabolism of beetle bioluminescence.

by

Timothy Robert Fallon

A.B. Biochemistry and Molecular Biology
Rollins College, 2012

Submitted to the Department of Biology in Partial Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2019

Signature redacted

Signature of Author ________ ____________________

Department of Biology
May 22nd, 2019

Signature redacted

Certified By ____________ ____________________

Jing-Ke Weng
Assistant Professor of Biology
Thesis Supervisor

Signature redacted

Accepted By __________ ____________________

Stephen Bell
Uncas and Helen Whitaker Professor of Biology
Investigator, Howard Hughes Medical Institute
Co-Director, Biology Graduate Committee

The evolution and specialized metabolism of beetle bioluminescence.

by

Timothy Robert Fallon

Submitted to the Department of Biology on May 22nd, 2019, in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Biology

ABSTRACT

Fireflies (Lampyridae) and certain other families of beetles including the American railroad worms (Phengodidae), Asian starworms (Rhagophthalmidae), and American click-beetles (Elateridae), produce light in a process known as bioluminescence. The bioluminescent systems of beetles, natively used for the purposes of mating communication and/or an aposematic warning signal, are now well understood and have been widely applied in biotechnology and biomedical research. There have been considerable advancements in the engineering of the luciferin substrate, and the luciferase enzyme, for beneficial characteristics such as altered emission wavelength, improved thermostability, and improved catalytic parameters, but despite this substantial effort focused on the biotechnological applications of beetle bioluminescence, major questions remain regarding its natural biochemistry and evolutionary origins. Four major questions that were unanswered at the beginning of this PhD study were: (1) Do fireflies possess a storage form of their luciferin? (2) What is the evolutionary relationship of bioluminescence amongst the bioluminescent beetles families, and has this trait independently evolved multiple times? (3) How is firefly luciferin biosynthesized? And (4) Are there accessory genes from the bioluminescent beetles which act in bioluminescent metabolism, and might these genes be useful for biotechnological applications? Here I describe the discovery and characterization of the presumed storage form of luciferin in fireflies, sulfoluciferin, and the enzyme which produces it, luciferin-sulfotransferase. Furthermore, I describe the sequencing, assembly, and characterization of the genome of the North American "Big Dipper" firefly *Photinus pyralis*, along with the Japanese "heike" firefly *Aquatica lateralis* genome, and the genome of the Puerto Rican bioluminescent click beetle or "cucubano" *Ignelater luminosus*. Genomic comparisons amongst these three species support the hypothesis that firefly and click beetle luciferase evolved independently, suggesting an independent evolutionary origin of the bioluminescent systems between these fireflies and click beetles. I also describe stable isotope tracing experiments in live fireflies, establishing that adult and larval fireflies likely do not *de novo* biosynthesize firefly luciferin, and may instead rely on a "recycling" pathway to re-synthesize luciferin from the luminescence product oxyluciferin. Lastly, I discuss the future directions resulting from this thesis, and the yet unanswered questions.

Thesis supervisor: Jing-Ke Weng
Title: Assistant Professor of Biology

ACKNOWLEDGEMENTS

As the saying goes, it takes a village to raise a child. I now also of the firm belief that it takes a village to train a PhD. Here I would like to thank some of those villagers that helped me along the way.

First and foremost, I'd like to thank my thesis advisor, Dr. Jing-Ke Weng. In my PhD thesis, I've had the gift of flexibility to dive into a research area that fascinated me, but I did not have a research background in. So above all I sincerely I thank Jing-Ke for that gift. Jing-Ke has always been a well-spring of support. Despite what you might think, my love of fireflies and other bioluminescent critters, while arguably quite substantial, does not in of-itself power the completion of experiments and ultimately this thesis. Jing-Ke's support and enthusiasm has kept me going in those times where the difficulty of working on a non-model organisms with limited availability and few tools, had dampened my spirit to continue on this project. Furthermore, Jing-Ke's mentorship in the fields of natural products biochemistry, evolutionary biology, and of the general process of science, has also been invaluable, and has shaped the scientific questions I would like to address in my ongoing career. From Jing-Ke, I've also learned the sometimes hard-fought lessons of patience, and of leadership, which are truly invaluable.

I am grateful to my thesis committee, Dr. Matthew Vander Heiden, Dr. David Sabatini, and Dr. Stephen Miller. I am grateful to Matt Vander Heiden for his continual support of the project. His comment in committee meeting #1 of the interest and value of my, at that time quite tentative, goal of working on firefly bioluminescence was more impactful than he probably realizes. His insight into enzymology and metabolism, including activity guided fractionation, has been quite valuable. I am grateful for David Sabatini's comments in thesis committee meetings, including his emphasis on doing impactful science, and his insights into metabolism. I'd also like to thank the professors who served in some capacity on my thesis committee over the years, including Dr. Barbara Imperiali, and Dr. Terry Orr-Weaver, who both gave me important advice and lessons during the process of my PhD.

I'd like to thank Mr. Geoff Liou. First off, for being a good sport when my often untidy desk would creep across the maginot line of our adjoining desks, and for being an understanding roommate and an unwavering friend over the course of our PhDs. It is some incredible stroke of luck that Geoff and I met and ended up as roommates, given what I presume must be the extraordinarily rare combination of our shared interest in Japanese culture & large-municipal-infrastructure. I've greatly enjoyed our time in our PhD, and hope I can come visit you in Japan. I'd also like to thank Dr. Olesya Levsh, who, along with Geoff and I, made up the first three graduate students in the Weng lab, and is another wonderful friend. Olesya in particular, was always inspiring with her impressive organization and work-ethic, and her ability to call a spade a spade. True to form, Olesya graduated several months before Geoff or I, and her templates and advice were quite helpful in my thesis writing process.

I'd like to thank members of the Weng laboratory. I feel that I've learned something from essentially everyone in the lab, but I'd like to highlight a few people that stand out in my mind. First of, I'd like to thank Dr. Fu-Shuang Li, who has been my constant support in all things chemistry, and beyond. From performing synthesis on my behalf while I watched with interest, to being a sounding board for my own synthesis and forays into chemistry, the opportunity to work Fu-Shuang has given insight into all those aspects of chemistry that you can only learn from an experienced chemist. Fu-Shuang has also been the most accomplished supporter of my field work. From firefly collection trips to New Jersey, to the Chatannoche National Forest, I am eternally grateful for Fu-Shuang's willingness to take these multi-day, and often quite strenuous, collection trips with me. I'd also like to thank Dr. Michael Torrens-Spence, for being the resident protein expression guru, who was always willing to help as I fumbled my way through the early days of expressing protein. Mike is also thanked for being a general goofball who does a great job of keeping up spirits in the lab, and for being my sounding board for our sometimes similar sense of humor of jokes with extraordinary convoluted situational premises. Finally, I'd like to thank Dr. Tomás Pluskal, for being the guru in computational metabolomics, software development and general bioinformatic wizardry. I've learned a ton from Tomas, and I now agree Tomas, symbolic links are worth using. I still prefer Nextflow over Snakemake though. Finally, I'd like to thank Ms. Valentina Carballo, for her true kindness to everyone in the lab. It's no exaggeration to say that Valentina has been a surrogate mother to many people in the lab during hard times, including myself, and her continual kindness has been an inspiration to me for the ideal model of how to treat people in life.

I'd like to thank my collaborators. Namely, Dr. Sara Lewis, who, from the very first email, has been an unwavering support in my quest to understand the firefly. Dr. Sarah Lower, also from the very first email, has been a constant wellspring of support, and sound advice on science. I'm grateful for the opportunity to meet, work, or get advice with other collaborators including, Dr. Manabu Bessho-Uehara, Dr. Yuichi Oba, Mr. Gavin Martin, Dr. Seth Bybee, Dr. Humberto Debat, Ms. Megan Behringer, Dr. Kathrin Stanger-Hall, Dr. Shuji Shigenobu, Dr. Amanda Larracuente, Ms. Avalon Owens, Mr. Oliver Keller, Ms. Lynn Faust, Ms. Lu Yang, Dr. Marc Branham, and Dr. Raphaël DeCock, Dr. Adam South, Dr. D. Winston Bellott, Dr. David Jenkins, and Mr. Haruyoshi Ikeya. From viruses, to repeat biology, to biochemistry, to systematics and evolutionary biology, my scientific training and the quality of my scientific thinking has been deepened and broadened by my experiences with you. I look forward to hopefully continuing these collaborations in the years to come.

I'd like to thank the staff of Whitehead Institute. Namely, Ms. Parimala Arokiaraj, and Mr. James Manganello. Pari and Jim had the unenviable job of helping me sort out how and if to keep the tightly regulated Asian *Aquatica* spp. fireflies here at Whitehead, along with trying to determine how to consider the other odd bioluminescent organisms I had in the lab from time to time. Furthermore, I am grateful to Mr. Haruyoshi Ikeya, Dr. Yuichi Oba, and Mr. Richard Cowan, for enabling and coordinating the quite complicated process of importing *Aquatica lateralis* fireflies to Whitehead.

I'd like to thank my parents, Robert and Joanne Fallon, my sisters Alexandra and Lindsay Fallon, and my brothers in-law, Michael Nelson and Richard Peterson-Cremer, for their support throughout my PhD. From applications, to acceptance, to ongoing projects, I am grateful for your continuous support, and your willingness to ask questions about my research, and/or, listen with interest when I just couldn't help but share some of my interesting findings from my research. I'd also like to thank my family on Tina's side, Kiran Udhwani, Alisha Udhwani, and Punit Udhwani, for the kindness they've continually showed me, and their support and interest in my research.

Finally, I'd like to thank Tina, my partner, and my love. Tina, you are kind, intelligent, beautiful, no-nonsense and incredibly driven, and an incredibly special person to put up all those long-hours of my being a PhD student. We've grown up together, even despite the fact that we lived apart for all of your medical training, and all of my scientific training, we've made it through, stronger than ever. I am incredibly lucky to have met you. I can't want to see what we can accomplish as we continue our life together as MD-PhD power couple. You also deserve extra special credit for your willingness to collect bioluminescent critters with me, despite your justified fear of the often very dark places we would collect from. From tromping at night on St. Croix, to tromping around in Virginia, to our moonless night in the Chattahoochee National Forest, I am extraordinarily grateful that you conquered your fear to support me, both in science, and in life and love.

TABLE OF CONTENTS

CHAPTER 1.

Introduction

Authors
Timothy R. Fallon, and Jing-Ke Weng,[1,2]

Author Affiliations
[1]Whitehead Institute for Biomedical Research, 455 Main Street, Cambridge, Massachusetts 02142, United States
[2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

Author contributions:
I wrote Chapter 1 with input and minor edits from Jing-Ke Weng.

Bioluminescence

As darkness falls in the forests, grasslands, and oceans of the world, a striking display of evolution's ingenuity can be observed. Ranging from kaleidoscope flashes of light as quick as the blink of an eye, to near-imperceptible glows, *bioluminescence*, the production of light by living organisms, is an unique adaptation of a diverse set of species. While fireflies (Order Coleoptera; Family Lampyridae), an entirely bioluminescent family comprised of over 2000 species, and found on every continent except Antarctica (Stanger-Hall et al., 2007), are likely the most commonly experienced example of bioluminescence, these charismatic beetles represent only a fraction of the bioluminescent diversity present in nature. The majority of bioluminescent diversity is found in the ocean (Herring, 1987), but beetles are especially well represented amongst the more rare examples of terrestrial bioluminescence. There are four families of bioluminescent beetles including the fireflies, bioluminescent click beetles (Elateridae), American railroad worms (Phengodidae) and Asian starworms (Rhagophthalmidae) (Martin et al., 2017). Across the diverse bioluminescence lineages the function of the emitted light varies. Intraspecific functions of bioluminescence include its use as a mating signal, such as it is with adult fireflies (Lloyd, 1966). Interspecific functions include defensive functions such as an aposematic warning sign of toxic chemical defenses, as firefly larvae are thought to advertise (De Cock and Matthysen, 1999), as well as offensive functions, such as a lure to entice prey, as seen in the luminescent barb of an anglerfish (Haddock et al., 2010). Diverse in its taxonomic distribution, emitted color and fundamental biochemical mechanisms, the ability to produce light has independently evolved several times across the tree of life.

Through over a century of research, most notably starting with the pioneering experiments of French physiologist Raphael Dubois (1849-1929) establishing the enzymatic basis of bioluminescence in bioluminescent click beetles (Dubois, 1886, 1885a, 1885b), an international group of biologists, chemists, and physicists has elucidated much of the biochemistry of bioluminescence. Their research revealed that the stereotypical bioluminescent reaction follows a relatively simple scheme: oxidation of a small molecule, dubbed *luciferin*, by a specialized enzyme, dubbed *luciferase*, using molecular oxygen, producing an oxidized product *oxyluciferin*, along with a photon of light (Shimomura, 2012). This general scheme has some minor variations, such as systems that utilize H_2O_2 instead of O_2, or systems which utilize addition co-substrates in addition to luciferin such as ATP, but the general statement that bioluminescence consists of an enzymatic oxidation of a small molecule substrate holds for all known systems. Many of the diverse luciferin molecules have been structurally characterized, often in sustained efforts requiring extensive field collection of bioluminescent organisms and specialized purification techniques to combat the often notable air instability of luciferins. The chemical diversity revealed by these experiments was substantial. There are currently 10 different structurally characterized luciferin molecules (Figure 1) (Hirata et al., 1959; Inoue et al., 1976a; Nakamura et al., 1989; Ohtsuka et al., 1976; Purtov et al., 2015; Strehler and Cormier, 1953; Watkins et al., 2018; White et al., 1963), and there remain some systems with undescribed systems that are known not to use any existing luciferin, such as the bioluminescent systems of springtails (Collembola) (Oba et al., 2011) or the Appalachian foxfire fly (*Orfelia fultoni*) (Viviani et al., 2002).

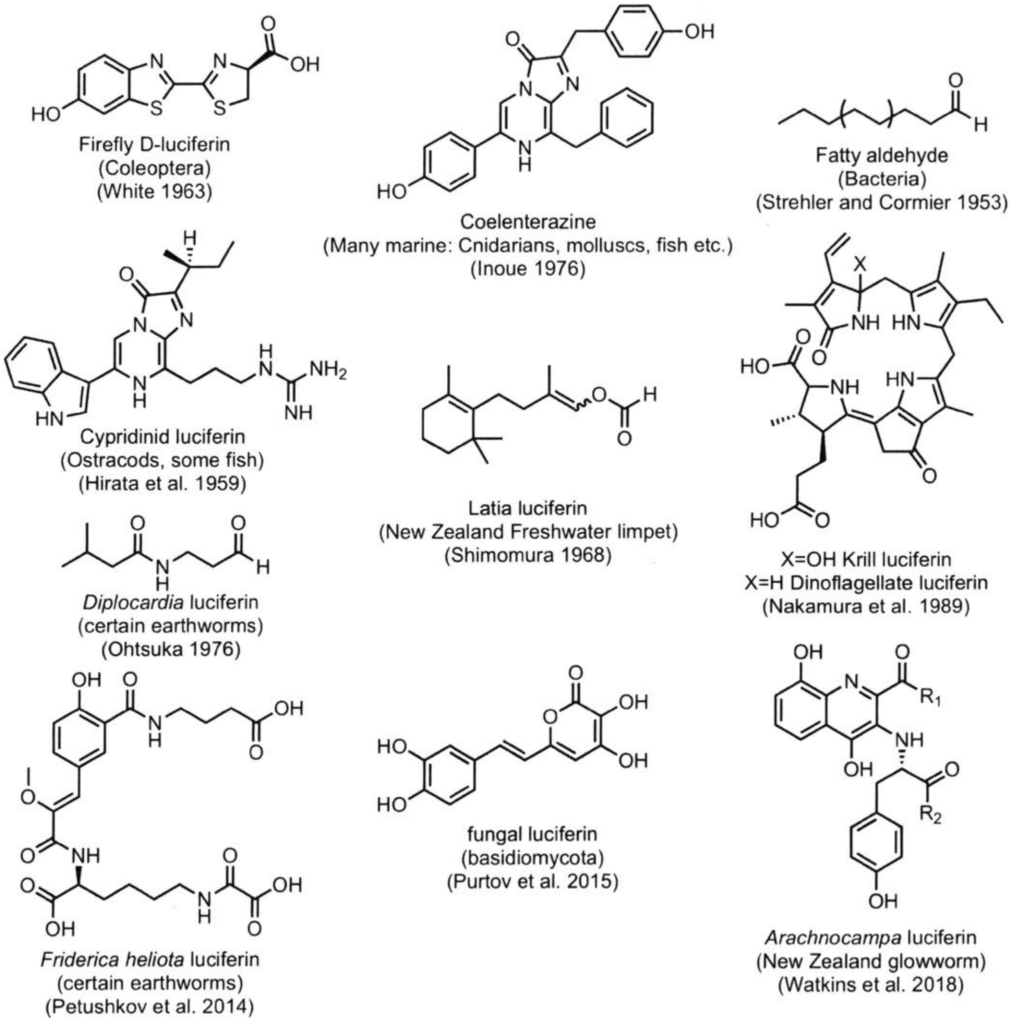


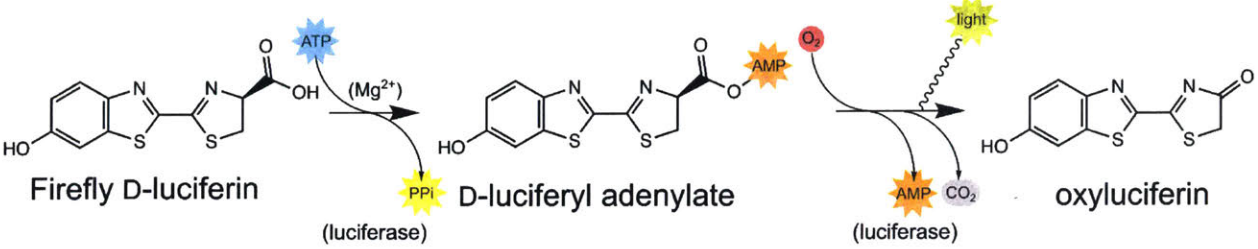
Figure 1: The ten currently known luciferin structures.
Common names are given underneath the chemical structure, while the taxonomic range these luciferins are used in given in parentheses. Citation of the description of the chemical structure is given below.

Some luciferins are used by evolutionarily distant lineages, such as the luciferin coelenterazine in over six phyla of marine organisms including cnidarians (Hori et al., 1977), ctenophores (comb jellies), copepods (Herring, 1988), decapods (shrimp) (Inoue et al., 1976b), molluscs (Inoue et al., 1976a), and vertebrates (fish) (Inoue et al., 1977). It is hypothesized that presence of luciferins in food-webs has enabled evolution of luciferases without prerequisite of luciferin biosynthesis (Haddock et al., 2010), and several of these unrelated and independently evolved luciferase enzymes are now cloned and structurally characterized. These challenging research projects produced tools that are now ubiquitous in biological laboratories, including the luciferase reporters from fireflies (de Wet et al., 1985) and other organisms such as the soft-coral *Renilla* (Lorenz et al., 1991). The study of bioluminescence was also responsible for the discovery of two major biotechnological tools, the calcium sensitive light producing protein *aequorin* from the North Pacific medusan jellyfish *Aequorea victoria* (Shimomura et al., 1962), and the "bioluminescence resonance energy transfer" (BRET) acceptor for aequorin in *A. victoria*, the green fluorescent protein GFP (Chalfie et al., 1994; Prasher et al., 1992; Shimomura et al., 1962). Directed evolution and engineering of these natural tools have been extensive, with many dozen of fluorescent proteins variants (Lambert, 2019), luciferase variants (Pozzo et al., 2018)(Halliwell et al., 2018), and luciferin analogs (Reddy et al., 2010)(Kuchimaru et al., 2016) reported in the literature. Amongst the cloned bioluminescent systems, the systems of the bioluminescent beetles have received the most study.

Phylogeny and biochemistry of beetle bioluminescence

Each of the bioluminescent beetle families contains hundreds to thousands of bioluminescent species, with families like the Lampyridae (fireflies), Phengodidae (American railroad worms) and Rhagophthalmidae (Asian starworms) thought to be comprised of only bioluminescent species (Lloyd, 1983), while in contrast the bioluminescent Elateridae only make up a small proportion of the species-rich and generally non-luminous Elateridae family (Martin et al., 2017). Within the Lampyridae, the major subdivision are the subfamilies Lampyrinae, found in the Americas and Europe, and the Luciolinae, generally found in Africa, Asia, and Australasia. The Luciolinae and Lampyrinae are ancient lineages, and are estimated to have diverged 100 million years ago (Fallon et al., 2018). In the Elateridae, the vast majority of the bioluminescent species are found in the tribe Pyrophorini, of the subfamily Agrypinae (Rosa, 2007). All bioluminescent Elateridae are found in the Americas and Caribbean, excepting a few species found on Pacific islands such as Fiji (Mitani et al., 2013). For the few species outside Pyrophorini that are bioluminescent such as *Campyloxenus pyrothorax* (of the monotypic subfamily Campyloxeninae), and *Alampoides alychnus* (Agrypninae; Euplinthini) (Rosa et al., 2015; Rosa and Costa, 2013), although the current taxonomic structure suggests otherwise, they are probably relatively closely related to the Pyrophorini (Costa, 1975a), suggesting a single origin of bioluminescence in the Elateridae (Fallon et al., 2018).

The biochemical basis for firefly bioluminescence was established through decades of study of the American firefly *Photinus pyralis* by the laboratory of William D. McElroy at Johns Hopkins University. Just six years after the first hypotheses that ATP was a biological energy carrier (Lipmann 1941), ATP was determined to be an essential reactant in the firefly bioluminescence reaction (McElroy 1947). Nine years later, luciferase was successfully crystallized (Green and McElroy 1956), although its structural determination by X-ray crystallography was not completed for some decades (Conti et al., 1996). Firefly luciferin was crystallized and its chemical formula determined a year after the first report of luciferase crystallization, from an extraction of over 15,000 fireflies (Bitler and McElroy 1957). It then took several years for a full synthesis and structural elucidation of luciferin (White et al. 1961; White et al. 1963). Luciferin was found to be an unusual bicyclic nitrogen-sulfur heterocycle, consisting of benzothiazole and thiazoline moieties. Due to the difficulty of chemically synthesizing the reaction product oxyluciferin, there was some some confusion regarding its structure, however a revised synthesis protocol did ultimately reveal the true structure of oxyluciferin (Suzuki and Goto, 1972). At that point the scheme of firefly bioluminescence was fully described as the enzymatic monooxygenation of a specialized substrate, firefly D-luciferin, using Mg^{2+}, ATP and O_2 as co-substrates, producing oxyluciferin, CO_2, and a photon of light as products (Scheme 1).



Scheme 1: Biochemical scheme of beetle bioluminescence
Firefly D-luciferin is adenylated then monooxygenated producing the oxidized product oxyluciferin, CO_2, and a photon of light.

Cross-reactivity tests demonstrated that synthesized firefly luciferin could produce light with the luciferase extracts of click beetles (Mcelroy et al., 1965), railroad worms (Viviani and Becham, 1993), and starworms (Ohmiya et al., 2000). With the cloning of luciferases from the these four families, it is now known that these beetle luciferases are specialized, monomeric, soluble, and (based on the presence of the C-terminal -SKL PTS1 targeting signal) peroxisome-targeted enzymes (Hanna et al., 1976). The two-step reaction catalyzed by firefly luciferase: adenylation, followed by monooxygenation is roughly analogous to the two-step adenylation, followed by coenzyme A ligation catalyzed by the ATP-dependent CoA synthetase super family enzymes (McElroy et al., 1967), and indeed, it is now known that beetle luciferase are homologous members of this enzyme superfamily, and retain CoA synthetic activity on certain substrates (Oba et al., 2003).

Life history, physiology, and function of beetle bioluminescence

The bioluminescent beetles are luminous in multiple life stages. In the case of fireflies, the full life-cycle including eggs, larvae, pupae, and adults, have all been reported to be bioluminescent (Harvey, 1952; Oba et al., 2013a). Fireflies typically produce a yellow to green luminescence, although some species (e.g. *Photinus scintillans*) produce a orange light (Branchini et al., 2017). There are two types of light production in fireflies. The first is neuronally controlled light production from specialized organs known as the lantern, light organ, or photophore. These organs are present in larvae, pupae, and adults. The second type of light production found in fireflies is a diffuse, continuous glow found in all life stages.

Firefly eggs and pupae (Oba et al., 2013a) glow at an intensity that is either just perceptible to the naked eye (e.g. in the subfamily Luciolinae) (Harvey, 1952), or imperceptible and only detectable with specialized cameras (e.g. *P. pyralis* eggs) (Fallon et al., 2018). The function of this diffuse egg and pupal glowing is not known, but given that many fireflies are chemically defended, it may be an aposematic (warning) signal advertising their chemical defenses.

The adult and larval light organs are thought to be developmentally derived from the insect fat body (Hess, 1922), a metabolically active tissue that might be considered analogous to both mammalian fat and liver. Firefly larvae typically possess two, but sometimes more, relatively small disc-shaped light organs on the lateral edges of the abdomen (Buck, 1948). Control of luminescence in the larval light organs is slow and relatively crude, with the typical kinetics of light intensity consisting of sporadic slow-rising (~1 second), then relatively constant (~10 seconds) and slow-falling glow. The luminescence of the larval light organs appear to be associated with a defensive, aposematic role, as larvae are often stimulated to luminescence in response to physical stimulation (De Cock and Matthysen, 1999). The larval light organ is made up of two tissues, the ventral layer of light-producing, or photocyte, cells and the dorsal layer of opaque cells thought to serve as a reflective layer (Oertel et al., 1975). Respiratory structures (trachea and tracheoles) and nerves branch profusely in the larval light organ, but do not appear to have a well-ordered structure. The neurons within the larval light organ terminate directly on the photocytes (Oertel et al., 1975), but do not appear to form tight synapses (Peterson, 1970). The effector neurotransmitter of the larval light organ is the monoamine octopamine (Carlson and Jalenak, 1986). The physiological mechanism regulating light production in the larval light organ is not fully understood, but it involves cyclic AMP production in response to octopamine (Nathanson and Hunnicutt, 1979).

The adult light organs of fireflies are relatively thin organs found on the ventral surface of the last few abdominal segments, typically on abdominal segments six and seven (Buck, 1948; Harvey, 1952). The function of the adult light organs is clear: adult fireflies use their light production to find mates (Buck and Case, 2002). In Luciolinae fireflies, the mating communication is simple, with relatively uncoordinated flashing of both males and females leading to a mutual attraction (Ohba, 1983). In Lampyrinae fireflies, notably the North American genera *Photinus*, *Photuris*, and *Pyractomena*, males and females have stereotyped flash patterns, where females respond only if they observe a male flash with precisely timed parameters, such as number of flashes, and delay between flashes (Buck and Case, 2002; Lloyd, 1966).

The adult light organ, especially in the males of many species, often spreads across the entire ventral surface of the abdominal segment (Buck, 1948). The capability of adult fireflies to control their light-emission kinetics is significantly greater than that of the larvae. While some adult fireflies produce simple glows from their light organs, the so called "lightning-bug" fireflies, common in North America can have as short as 70 millisecond rise and fall times of their light emission intensity, allowing for complex flash patterns (Ghiradella and Schmidt, 2004). The photocyte cells of the adult light organ comprise a tissue-layer of ~4-20 cells deep lying underneath a non-cellular cuticle and single layer of the hypodermal cells (Buck, 1948). Akin to the larval light organ, a layer of opaque presumed-reflective cells lies behind the photocyte layer. In flashing fireflies, such as the North American fireflies in genus *Photinus*, the respiratory structure of the adult light organ is extensive and highly organized, with the photocytes forming a regular cylindrical structure around the oxygen-carrying tracheoles and trachea which permeate the photocyte layer. Mitochondria in the photocytes are almost exclusively clustered along the edges of the photocytes that face the respiratory structures. The rest of the photocyte cell is filled with luciferase-containing peroxisomes. Neuronal innervation in the adult light organ does not terminate directly on the photocytes, but instead terminates on the tracheal system (Ghiradella and Schmidt, 2004; Smith, 1963). Like the larval light organ, the effector neurotransmitter of these neurons is the neurotransmitter octopamine (Copeland and Robertson, 1982). The control of luminescence in the adult light organ is not fully understood, but clearly oxygen is the limiting factor which controls light production. The current major model of physiological O_2 control in the adult firefly lantern is the gatekeeper hypothesis. The gatekeeper hypothesis stipulates that the respiratory activity of the abundant mitochondria on the periphery of the photocytes actively prevents oxygen diffusion to the luciferase-containing peroxisomes in the interior of the photocytes (Ghiradella and Schmidt, 2004), and that transient octopamine-stimulated nitric oxide (NO) production directly inhibits mitochondrial respiration and allows for O_2 diffusion into the photocyte interior (Trimmer et al., 2001).

Bioluminescent click beetles are luminescent in their egg, larval, pupal, and adult stage (Colepicolo-Neto et al., 1986; Dubois, 1886; Harvey, 1952; Seaman, 1891). Similar to the firefly, bioluminescent click beetle eggs and pupae have a diffuse glow, whereas larvae and adults have organs under neuronal control. In adult bioluminescent click beetles, there is some evidence for a role in mating communication, but this has not been well studied (Kretsch, 2000). In the larvae, luminescence appears to play a defensive role, as larvae will not typically luminescence unless extensively disturbed, at which point a transition to an aggressive defensive behavior takes place with active biting and bright luminescence (Colepicolo-Neto et al., 1986). Larval light organs are found in the head of the larvae, and along the lateral sides of the abdominal segments. The cellular anatomy of the larval or adult light organs is little described, and Dubois's 1886 thesis remains the most comprehensive source (Dubois, 1886). The adult light organs are found as two paired spots on the dorsal surface of the prothorax, with an additional single ventral light organ on the anterior edge of the first abdominal segment. The typical color of elaterid luminescence is green, but some in some species the color of the dorsal and ventral light organ varies (Colepicolo-Neto et al., 1986; Stolz et al., 2003). There is no mechanistic understanding of how luminescence is controlled in click beetles, and to my knowledge there is no evidence on the identity of the effector neurotransmitters of their light organs.

Evolution of beetle bioluminescence

In the chapter "Difficulties of the Theory" in the Origin of Species (Charles Darwin, 1872), Darwin highlights examples of biological complexity which his theory of natural selection seemed ill-equipped to explain. His discussion of the evolution of the vertebrate camera eye is often cited by those who believe or want to mislead that the theory of natural selection was unable to account for the evolution of complex traits, but in the following sections Darwin does in fact describe a stepwise evolutionary path for the eye (Charles Darwin, 1872; Fishman, 2008). But in the next section, however, the "Special Difficulties of the Theory of Natural Selection", Darwin highlights some difficult cases of evolution where complex analogous traits arose in parallel lineages, but in contrast to situations like the eye, these were situations for which he had no explanation of their evolutionary path. As Darwin notes: "The luminous organs which occur in a few insects, belonging to widely different families, and which are situated in different parts of the body, under our present state of ignorance, a [serious case of difficulty] ... there is no reason to suppose that they have been inherited from a common progenitor". In Darwin's time, the four families of bioluminescent beetles had been described, while the other families of bioluminescent insects, such as the flies *Arachnocampa* and *Orfelia*, and the bioluminescent Springtails (Collembola), had either not yet described or were barely known (Harvey, 1952). Furthermore, during the voyage of the Beagle, while at Bahia Brazil, Darwin directly observed and described abundant bioluminescent click beetles which he identified as "*Pyrophorus luminosus*" (now *Ignelater luminosus*) (Darwin, 1898). Although the geographic distribution of *Ignelater luminosus* is not known to extend to Brazil (Fallon et al., 2018), Darwin's description of "*Pyrophorus luminosus*" unambiguously identifies it as a bioluminescent click beetle. It can then be clearly inferred that the "luminous organs" that Darwin considered in the Origin of Species, were the luminous organs of the bioluminescent beetles.

Darwin clearly indicated that he believed the light organs on bioluminescent beetles had evolved independently, however as he said in his "present state of ignorance", he was unable to describe the mechanistic basis for the independent evolution of beetle bioluminescence. How have modern studies amended Darwin's hypothesis? In the modern literature, the discovery that bioluminescent beetles used identical luciferins, and had homologous luciferases, combined with the long confused phylogenetic relationships of these families, led to the belief that the beetle bioluminescent systems had a single origin, rather than multiple origins as Darwin had hypothesized (Day et al., 2004; Wood, 1995). Although several more recent papers supported Darwin's claim of independent evolution (Bocakova et al., 2007; Branham and Wenzel, 2003; Charles Darwin, 1872; Costa, 1975b; Day, 2013; Oba, 2009; Sagegami-Oba et al., 2007), each of them were based on ancestral state inference from species phylogenetic analyses, which given the constant flux of the early species phylogenies, did not lead to a strong conclusion in the field as to the nature of the origins of bioluminescence. Although the luciferases of all bioluminescent beetles are homologous in the sense that they arose from the same peroxisomal acyl-CoA synthetase superfamily, luciferase activity is now thought to have independently evolved at least twice. In particular, genomic evidence has indicated that the luciferases of the click beetles and fireflies are evolutionarily independent (Fallon et al., 2018). The most recent species phylogenies have confirmed the relatively distant relationship of the bioluminescent Elateridae with the fireflies, supporting the conclusion of independent origins (Kusy et al., 2018). On the other hand, these recent phylogenies have also supported the sister relationship of the Lampyridae with the Phengodidae+Rhagophthalmidae, suggesting there may be a single origin of Lampyridae+Phengodidae+Rhagophthalmidae luminescence. Arguing against a single origin, Phengodidae light organs have a theorized ectodermal secretory "oenocyte" origin (Bassot, 1974; Makki et al., 2014), as compared to the theorized mesodermal origin of firefly light organs from the fat body (Hess, 1922; Li et al., 2019). But genomic evidence supporting or rejecting the independent origins of luciferase amongst the Phengodidae, Rhagophthalmidae, and Lampyridae, is not yet available.

But even unambiguous evidence supporting the independent evolution of luciferase amongst these families does not necessarily require the entire bioluminescent systems to be independent. A complex trait like bioluminescence is made up of multiple subtraits, such as a neofunctionalized luciferase, the presence or biosynthesis of luciferin, specialized cells to produce light, and the organ and/or neuronal architecture to mediate control of light production. Each of these subtraits has its own evolutionary history proceeding in a stepwise fashion and intersecting with the evolution of the other subtraits. While independent evolution of luciferase implies independent evolution of subtraits which were dependent on an already specialized luciferase, traits which luciferase neofunctionalization depended on, such as the presence of luciferin, presumably preceed luciferase neofunctionalization. Firefly luciferin itself does not have a clear evolutionary path, as none of the biosynthetic enzymes have been definitively identified. Luciferin is known to be naturally found only in the luminous beetles families (Oba et al., 2008), although there is a preliminary report of firefly luciferin in the unrelated Australasian bioluminescent fly *Arachnocampa* (Trowell et al., 2016). The recent report that luciferin can be produced by the non-enzymatic reaction of cysteine and benzoquinone (Kanie et al., 2016), makes it possible that firefly luciferin exists at a low level in many lineages of insects. So while we now understand enough to mechanistically describe how luciferase arose between some of the bioluminescent beetle lineages, supporting the independent evolution Darwin first hypothesized, the full story, including the origin of luciferin, and the confirmation or rejection of shared luciferase evolution amongst all the bioluminescent beetle families, has not yet been written.

Applications of beetle bioluminescence

The chemical synthesis of firefly luciferin is straightforward (Santaniello et al., 2009; White et al., 1963), and with the cloning of over 50 beetle luciferase genes to date (Oba and Hoffmann, 2014), light production via recombinant beetle luciferases and chemically synthesized luciferin is now commonly used in biomedical research and biotechnology (Paley, 2014; Stanley, 1989). The luciferase of the common North American firefly *Photinus pyralis* was the first cloned luciferase (de Wet et al., 1985), and *P. pyralis* luciferase plus its engineered or evolved variants (Branchini et al., 2014; Groskreutz et al.,

1995) are the most widely used beetle luciferases. The requirement for ATP in the beetle bioluminescence reaction has led to several specialized applications, including a highly sensitive and rapidly applied *in vitro* test for ATP, commonly used as a measure of sterility in industrial or medical applications (Hastings and Johnson, 2003). This luciferase-mediated sterility test became the recommended sterility determination method during the assembly of the Mars Rover "Curiosity" (Benardini and Venkateswaran, 2016). Firefly luciferases are also used *in vivo* as reporter genes in heterologous hosts. One especially useful application is mammalian live-imaging, where disease progression in cancer xenograft or microorganism infection models can be monitored non-invasively by injection of luciferase-expressing cells and imaging of these cells through the skin of the infected animal with sensitive cameras (Mezzanotte et al., 2017; Welsh and Noguchi, 2012). More specialized applications have also been described, such as the measurement of protein-protein interactions through luciferase complementation experiments (Kato and Jones, 2010), or through the short-distance transfer of the oxyluciferin excited state energy to a fluorescent acceptor molecule through a mechanism analogous to Förster resonance energy transfer (FRET) (Arai et al., 2002).

Perhaps due to this widespread use, firefly luciferases have been very well studied from an enzymological, and biotechnological perspective. There are multiple literature descriptions of luciferase kinetics and detailed reaction mechanisms (Branchini et al., 2015; da Silva and da Silva, 2011; DeLuca and McElroy, 1974; Niwa et al., 2010; Ribeiro and Esteves da Silva, 2008), the mechanism and modulation of the color of light emission via site directed mutagenesis (Branchini et al., 2004; Nakatsu et al., 2006), and mutagenesis of luciferases with advantageous properties such as enhanced thermostability or kinetics (Halliwell et al., 2018; Pozzo et al., 2018). Furthermore, synthetic analogs of luciferin, with enhanced solubility, transmembrane transport characteristics, or modified light emission colors have been developed (Kaskova et al., 2016; Reddy et al., 2010). These engineering efforts are perhaps exemplified by the recently described AkaLumine-HCl luciferin analog (Kuchimaru et al., 2016), that, when combined with a directed evolution variant of *Photinus pyralis* luciferase variant known as Akaluc, produces a 677 nm near-infrared emission well suited to deep tissue imaging, which even allows for the detection of luminescence from single cells in freely moving animals (Iwano et al., 2018; Nasu and Campbell, 2018).

The metabolism of firefly bioluminescence

Although there have been substantial advances in the application of firefly luciferase itself, there has been almost no description of other enzymes from the firefly bioluminescent system which might act in luciferin metabolism. Uncovering these enzymes could potentially enhance biotechnological uses of firefly bioluminescence, and would help untangle the evolution of beetle bioluminescence. One such enzyme, the so-called firefly luciferin-regenerating-enzyme (LRE) (Gomi and Kajiyama, 2001), was reported to catalyze the degradation of oxyluciferin into the nitrile compound 2-cyano-6-hydroxybenzothiazole (CHBT), which, in the presence of D-cysteine would regenerate D-luciferin (Figure 3). This luciferin recycling activity would be highly valuable to biotechnology, but to date, there has not been a convincing confirmation of LRE's activity (Hosseinkhani et al., 2017), and LRE does not appear to be highly or specifically expressed in the firefly light organ (Fallon et al., 2018; Gomi et al., 2002). Oxyluciferin has been described as a weak competitive inhibitor of luciferin (Ribeiro and Esteves da Silva, 2008), and therefore in the absence of an LRE, oxyluciferin must presumably be either recycled into luciferin, catabolized, or transported outside the light organ to allow light production to continue efficiently. Early radioactive tracing experiments demonstrated radiolabeled oxyluciferin could be converted back into luciferin (Okada et al., 1974), but this may be due to oxyluciferin's tendency to non-enzymatically slowly degrade to CHBT, which can the non-enzymatically couple with cysteine to produce luciferin.

The *de novo* biosynthesis of luciferin is hypothesized to arise from cysteine and benzoquinone (McCapra and Razavi, 1975; Oba et al., 2013b; Okada et al., 1976), but to date no enzymes have been described (Figure 3).

The recently described firefly luciferin-sulfotransferase (LST) (Figure 3) (Fallon et al., 2016), which catalyzes the interconversion of luciferin to the putative storage form sulfoluciferin, is to date the only non-luciferase enzyme with an unambiguous activity in luciferin metabolism. LST is highly and differentially expressed in the adult male light organ of fireflies from both of the major subfamilies (Lampyrinae and Luciolinae), supporting a role relevant to bioluminescence (Fallon et al., 2018). Some LST orthologs (e.g., that of *P. pyralis*) appear to have a peroxisomal targeting signal, but there is not yet direct confirmation that LST is localized to the peroxisome along with luciferase and luciferin (Fallon et al., 2018; Smalley et al., 1980).

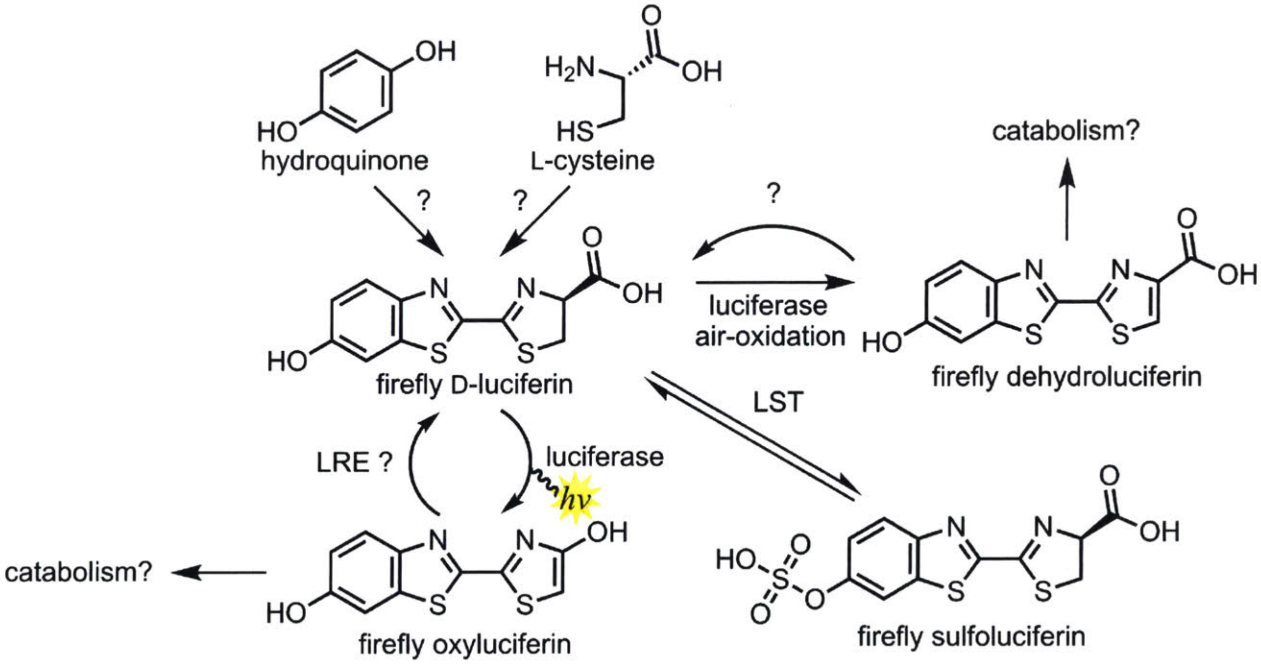


Figure 3: Known and hypothesized transformations in firefly bioluminescence
Firefly luciferin is monooxygenated to oxyluciferin, producing light in the process. A recycling pathway from oxyluciferin to D-luciferin is hypothesized (Okada et al., 1974). Luciferin sulfotransferase (LST) catalyzes the transformation of D-luciferin and its sulfonated form sulfoluciferin (Fallon et al., 2016). Dehydroluciferin is produced in a low-level side reaction of luciferase (Fontes et al., 1997). As both oxyluciferin and dehydroluciferin are inhibitors of luciferase, it seems likely that oxyluciferin and/or dehydroluciferin are either recycled into luciferin, or are catabolized or removed from the light organ via transport.

Another described activity in firefly bioluminescence, but one still without known enzymes, involves the chiral biosynthesis of D-luciferin. The stereocenter of luciferin -- where its "D" name is derived from -- lies on the thiazoline ring of luciferin. This thiazoline ring is structurally analogous to a cysteine that cyclized after a nucleophilic substitution onto a carboxylic acid. This structural similarity led to the early hypotheses that cysteine was the biosynthetic precursor of the luciferin thiazoline (McCapra and Razavi, 1975). However, the natural stereochemistry of all amino acids is "L", corresponding to a (S) configuration of the stereocenter, so, if luciferin is biosynthesized from L-cysteine, at some point the L-stereocenter must be epimerized to "D". Niwa and colleagues demonstrated in the Japanese firefly *Aquatica lateralis* that cysteine extracted from firefly lanterns was almost entirely the "L" epimer, and furthermore, demonstrated a robust ATP, CoA, and Mg^{2+} dependent activity which could rapidly convert L-luciferin directly to D-luciferin (Niwa et al., 2006). They hypothesized that luciferase, which can catalyze the formation of L-luciferyl-CoA, but not of D-luciferyl-CoA, combined with the propensity of CoA thioesters to non-enzymatically epimerize, followed by hydrolysis of the resulting D-luciferyl-CoA with a thioesterase, was performing this activity in the firefly lantern. More recent results have been able to reconstitute a D-luciferin epimerization pathway from luciferase, the *E. coli* thioesterase TESB, and a bacterial fatty-acyl CoA α-methyl-acyl-CoA-racemase (Maeda et al., 2017), but the enzymes which perform the epimerization activity *in vivo* in the firefly remain unknown.

Beyond these described activities, one other activity which seems biochemically plausible and likely necessary for efficient light production is the reduction or catabolism of dehydrolucierin. Dehydroluciferin (DHL) is the result of oxidation on the luciferin thiazoline ring, and can be easily produced by air oxidation of luciferin, or as a low-level side product of luciferase-mediated luciferin oxygenation (Fraga et al., 2006) (Figure 3). Dehydrolucifierin, specifically dehydroluciferyl-adenylate which luciferase can synthesize, is a strong competitive inhibitor of firefly luminescence (da Silva and da Silva, 2011). Given the ease that DHL is likely produced in the native *in vivo* context, it seems reasonable to assume that it would be either catabolised or reduced back to luciferin, to prevent its competitive inhibition of the luminescence reaction. No such activity has been described to date. Alternatively, dehydroluciferin could be exported from the photocytes, and excreted as waste or sequestered elsewhere in the body.

Conclusions

Fireflies and other bioluminescent beetles have been well studied, however there are still major questions. The work in this dissertation contributes to a fundamental advance in four questions: (1) Do fireflies possess a storage form of their luciferin? (2) What is the evolutionary relationship of bioluminescence amongst the bioluminescent beetles, and has this trait independently evolved multiple times? (3) How is firefly luciferin biosynthesized? And (4) Are there accessory genes from the bioluminescent beetles which act in bioluminescent metabolism, and might these genes be useful for biotechnological applications? First, in Chapter 2 I attempt to answer the question of whether fireflies possess a storage form for luciferin, by characterizing the newly discovery firefly metabolite sulfoluciferin, and the enzyme which produces it, luciferin sulfotransferase. In Chapter 3, I explore the question of the origins of bioluminescence, by sequencing and comparing the genomes of two fireflies representing the major subfamilies Lampyrinae and Luciolinae, with the genome of an also bioluminescent but otherwise unrelated click-beetle. In Chapter 4, I characterize if fireflies actively *de novo* biosynthesize luciferin in their light organs using stable isotope tracing of the presumed biosynthetic precursors of firefly luciferin. Finally in Chapter 5 I discuss the conclusions and interpretations of my work, and whether it could be useful for biotechnology.

REFERENCES

Arai R, Nakagawa H, Kitayama A, Ueda H, Nagamune T. 2002. Detection of protein-protein interaction by bioluminescence resonance energy transfer from firefly luciferase to red fluorescent protein. *J Biosci Bioeng* **94**:362–364.

Bassot JM. 1974. Les cellules lumineuses du coleoptere Phengodes In: Avery L, editor. Extrait de

Recherces Biologiques Contemporareus. Nancy, France: Imp. Vagner.

Benardini JN, Venkateswaran K. 2016. Application of the ATP assay to rapidly assess cleanliness of spacecraft surfaces: a path to set a standard for future missions. *AMB Express* **6**:113.

Bocakova M, Bocak L, Hunt T, Teraväinen M, Vogler AP. 2007. Molecular phylogenetics of Elateriformia (Coleoptera): evolution of bioluminescence and neoteny. *Cladistics* **23**:477–496.

Branchini BR, Behney CE, Southworth TL, Fontaine DM, Gulick AM, Vinyard DJ, Brudvig GW. 2015. Experimental Support for a Single Electron-Transfer Oxidation Mechanism in Firefly Bioluminescence. *J Am Chem Soc* **137**:7592–7595.

Branchini BR, Southworth TL, Fontaine DM, Davis AL, Behney CE, Murtiashaw MH. 2014. A Photinus pyralis and Luciola italica chimeric firefly luciferase produces enhanced bioluminescence. *Biochemistry* **53**:6287–6289.

Branchini BR, Southworth TL, Fontaine DM, Murtiashaw MH, McGurk A, Talukder MH, Qureshi R, Yetil D, Sundlov JA, Gulick AM. 2017. Cloning of the Orange Light-Producing Luciferase from Photinus scintillans—A New Proposal on how Bioluminescence Color is Determined. *Photochem Photobiol* **93**:479–485.

Branchini BR, Southworth TL, Murtiashaw MH, Magyar RA, Gonzalez SA, Ruggiero MC, Stroh JG. 2004. An alternative mechanism of bioluminescence color determination in firefly luciferase. *Biochemistry* **43**:7255–7262.

Branham MA, Wenzel JW. 2003. The origin of photic behavior and the evolution of sexual communication in fireflies (Coleoptera: Lampyridae). *Cladistics* **19**:1–22.

Buck JB. 1948. The anatomy and physiology of the light organ in fireflies. *Annals of the New York Academy of Sciences.*

Buck J, Case J. 2002. Physiological Links in Firefly Flash Code Evolution. *J Insect Behav* **15**:51–68.

Carlson AD, Jalenak M. 1986. Release of octopamine from the photomotor neurones of the larval firefly lanterns. *J Exp Biol* **122**:453–457.

Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC. 1994. Green fluorescent protein as a marker for gene expression. *Science* **263**:802–805.

Charles Darwin. 1872. The Origin of Species, 6th ed. PF Collier & Son, New York.

Colepicolo-Neto P, Costa C, Bechara EJH. 1986. Brazilian species of luminescent Elateridae: Luciferin identification and bioluminescence spectra. *Insect Biochem* **16**:803–810.

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

Copeland J, Robertson HA. 1982. Octopamine as the transmitter at the firefly lantern: presence of an octopamine-sensitive and a dopamine-sensitive adenylate cyclase. *Comp Biochem Physiol C* **72**:125–127.

Costa C. 1975a. Systematics and evolution of the tribes Pyrophorini and Heligmini, with description of Campyloxeninae, new subfamily (Coleoptera, Elateridae). *Arq Zool* **26**:49.

Costa C. 1975b. Systematics and evolution of the tribes Pyrophorini and Heligmini, with description of Campyloxeninae, new subfamily (Coleoptera, Elateridae). *Arq Zool* **26**:49.

Darwin C. 1898. Journal of researches into the natural history and geology of the countries visited during the voyage of H.M.S. Beagle round the world, under the command of Capt. Fitz Roy, R.N. doi:10.5962/bhl.title.20767

da Silva LP, da Silva JCGE. 2011. Kinetics of inhibition of firefly luciferase by dehydroluciferyl-coenzyme A, dehydroluciferin and L-luciferin. *Photochem Photobiol Sci* **10**:1039–1045.

Day JC. 2013. The role of gene duplication in the evolution of beetle bioluminescence. *Trends Entomol* **9**:55–63.

Day JC, Tisi LC, Bailey MJ. 2004. Evolution of beetle bioluminescence: the origin of beetle luciferin.

Luminescence **19**:8–20.

De Cock R, Matthysen E. 1999. Aposematism and Bioluminescence: Experimental evidence from Glow-worm Larvae(Coleoptera: Lampyridae). *Evol Ecol* **13**:619–639.

DeLuca M, McElroy WD. 1974. Kinetics of the firefly luciferase catalyzed reactions. *Biochemistry* **13**:921–925.

de Wet JR, Wood KV, Helinski DR, DeLuca M. 1985. Cloning of firefly luciferase cDNA and the expression of active luciferase in Escherichia coli. *Proc Natl Acad Sci U S A* **82**:7870–7873.

Dubois R. 1886. Les Élatérides lumineux: contribution a l'étude de la production de la lumière par les êtres vivants. la Société zoologique de France.

Dubois R. 1885a. Note sur la physiologie des pyrophores. *CR Soc Biol* **2**:559.

Dubois R. 1885b. Fonction photogénique des Pyrophores. *CR Seances Soc Biol Fil* **37**:559–562.

Fallon TR, Li F-S, Vicent MA, Weng J-K. 2016. Sulfoluciferin is Biosynthesized by a Specialized Luciferin Sulfotransferase in Fireflies. *Biochemistry* **55**:3341–3344.

Fallon TR, Lower SE, Chang C-H, Bessho-Uehara M, Martin GJ, Bewick AJ, Behringer M, Debat HJ, Wong I, Day JC, Suvorov A, Silva CJ, Stanger-Hall KF, Hall DW, Schmitz RJ, Nelson DR, Lewis SM, Shigenobu S, Bybee SM, Larracuente AM, Oba Y, Weng J-K. 2018. Firefly genomes illuminate parallel origins of bioluminescence in beetles. *Elife* **7**. doi:10.7554/eLife.36495

Fishman RS. 2008. Evolution and the eye: the Darwin bicentennial and the sesquicentennial of the origin of species. *Arch Ophthalmol* **126**:1586–1592.

Fontes R, Dukhovich A, Sillero A, Sillero MA. 1997. Synthesis of dehydroluciferin by firefly luciferase: effect of dehydroluciferin, coenzyme A and nucleoside triphosphates on the luminescent reaction. *Biochem Biophys Res Commun* **237**:445–450.

Fraga H, Fernandes D, Novotny J, Fontes R, Esteves da Silva JCG. 2006. Firefly luciferase produces hydrogen peroxide as a coproduct in dehydroluciferyl adenylate formation. *Chembiochem* 7:929–935.

Ghiradella H, Schmidt JT. 2004. Fireflies at one hundred plus: a new look at flash control. *Integr Comp Biol* **44**:203–212.

Gomi K, Hirokawa K, Kajiyama N. 2002. Molecular cloning and expression of the cDNAs encoding luciferin-regenerating enzyme from Luciola cruciata and Luciola lateralis. *Gene* **294**:157–166.

Gomi K, Kajiyama N. 2001. Oxyluciferin, a luminescence product of firefly luciferase, is enzymatically regenerated into luciferin. *J Biol Chem* **276**:36508–36513.

Groskreutz DJ, Sherf BA, Wood KV, Schenborn ET. 1995. Increased expression and convenience with the new pGL3 luciferase reporter vectors. *Promega Notes* **50**.

Haddock SHD, Moline MA, Case JF. 2010. Bioluminescence in the sea. *Ann Rev Mar Sci* **2**:443–493.

Halliwell LM, Jathoul AP, Bate JP, Worthy HL, Anderson JC, Jones DD, Murray JAH. 2018. ΔFlucs: Brighter Photinus pyralis firefly luciferases identified by surveying consecutive single amino acid deletion mutations in a thermostable variant. *Biotechnol Bioeng* **115**:50–59.

Hanna CH, Hopkins TA, Buck J. 1976. Peroxisomes of the firefly lantern. *J Ultrastruct Res* **57**:150–162.

Harvey EN. 1952. Bioluminescence. Academic Press.

Hastings JW, Johnson CH. 2003. [3] Bioluminescence and chemiluminescenceMethods in Enzymology. Academic Press. pp. 75–104.

Herring PJ. 1988. Copepod luminescence. *Hydrobiologia* **167**:183–195.

Herring PJ. 1987. Systematic distribution of bioluminescence in living organisms. *J Biolumin Chemilumin* **1**:147–163.

Hess WN. 1922. Origin and development of the light-organs of photurus pennsylvanica de geer. *Journal of Morphology*. doi:10.1002/jmor.1050360206

Hirata Y, Shimomura O, Eguchi S. 1959. The Structure of Cypridina Luciferin. *Tetrahedron Letters*. doi:10.1016/s0040-4039(01)82760-x

Hori K, Charbonneau H, Hart RC, Cormier MJ. 1977. Structure of native Renilla reinformis luciferin. *Proc Natl Acad Sci U S A* **74**:4285–4287.

Hosseinkhani S, Emamgholi Zadeh E, Sahebazzamani F, Ataei F, Hemmati R. 2017. Luciferin-Regenerating Enzyme Crystal Structure Is Solved but its Function Is Still Unclear. *Photochem Photobiol* **93**:429–435.

Inoue S, Kakoi H, Goto T. 1976a. Squid bioluminescence III. Isolation and structure of Watasenia luciferin. *Tetrahedron Lett* **17**:2971–2974.

Inoue S, Kakoi H, Goto T. 1976b. Oplophorus luciferin, bioluminescent substance of the decapod shrimps, Oplophorus spinosus and Heterocarpus laevigatus. *J Chem Soc Chem Commun* 1056–1057.

Inoue S, Okada K, Kakoi H, Goto T. 1977. FISH BIOLUMINESCENCE I. ISOLATION OF A LUMINESCENT SUBSTANCE FROM A MYCTOPHINA FISH, NEOSCOPELUS MICROCHIR, AND IDENTIFICATION OF IT AS OPLOPHORUS LUCIFERIN. *Chem Lett* **6**:257–258.

Iwano S, Sugiyama M, Hama H, Watakabe A, Hasegawa N, Kuchimaru T, Tanaka KZ, Takahashi M, Ishida Y, Hata J, Shimozono S, Namiki K, Fukano T, Kiyama M, Okano H, Kizaka-Kondoh S, McHugh TJ, Yamamori T, Hioki H, Maki S, Miyawaki A. 2018. Single-cell bioluminescence imaging of deep tissue in freely moving animals. *Science* **359**:935–939.

Kanie S, Nishikawa T, Ojika M, Oba Y. 2016. One-pot non-enzymatic formation of firefly luciferin in a neutral buffer from p-benzoquinone and cysteine. *Sci Rep* **6**:24794.

Kaskova ZM, Tsarkova AS, Yampolsky IV. 2016. 1001 lights: luciferins, luciferases, their mechanisms of action and applications in chemical analysis, biology and medicine. *Chem Soc Rev* **45**:6048–6077.

Kato N, Jones J. 2010. The split luciferase complementation assay. *Methods Mol Biol* **655**:359–376.

Kretsch E. 2000. Courtship Behavior of Ignelater luminosus.

Kuchimaru T, Iwano S, Kiyama M, Mitsumata S, Kadonosono T, Niwa H, Maki S, Kizaka-Kondoh S. 2016. A luciferin analogue generating near-infrared bioluminescence achieves highly sensitive deep-tissue imaging. *Nat Commun* **7**:11856.

Kusy D, Motyka M, Bocek M, Vogler AP, Bocak L. 2018. Genome sequences identify three families of Coleoptera as morphologically derived click beetles (Elateridae). *Sci Rep* **8**:17084.

Lambert TJ. 2019. FPbase: a community-editable fluorescent protein database. *Nat Methods* **16**:277–278.

Li S, Yu X, Feng Q. 2019. Fat Body Biology in the Last Decade. *Annu Rev Entomol* **64**:315–333.

Lloyd JE. 1983. Bioluminescence and communication in insects. *Annu Rev Entomol* **28**:131–160.

Lloyd JE. 1966. Studies on the flash communication system in Photinus fireflies.

Lorenz WW, McCann RO, Longiaru M, Cormier MJ. 1991. Isolation and expression of a cDNA encoding Renilla reniformis luciferase. *Proc Natl Acad Sci U S A* **88**:4438–4442.

Maeda J, Kato D-I, Okuda M, Takeo M, Negoro S, Arima K, Ito Y, Niwa K. 2017. Biosynthesis-inspired deracemizative production of d-luciferin by combining luciferase and thioesterase. *Biochim Biophys Acta Gen Subj* **1861**:2112–2118.

Makki R, Cinnamon E, Gould AP. 2014. The development and functions of oenocytes. *Annu Rev Entomol* **59**:405–425.

Martin GJ, Branham MA, Whiting MF, Bybee SM. 2017. Total evidence phylogeny and the evolution of adult bioluminescence in fireflies (Coleoptera: Lampyridae). *Mol Phylogenet Evol* **107**:564–575.

McCapra F, Razavi Z. 1975. A model for firefly luciferin biosynthesis. *J Chem Soc Chem Commun* 42b–43.

McElroy WD, DeLuca M, Travis J. 1967. Molecular Uniformity in Biological Catalyses: The enzymes concerned with firefly luciferin, amino acid, and fatty acid utilization are compared. *Science*. doi:10.1126/science.157.3785.150

Mcelroy WD, Seliger HH, Deluca M. 1965. Enzyme Catalysis and Color of Light in Bioluminescent Reactions In: Bryson V, Vogel HJ, editors. Evolving Genes and Proteins. Academic Press. pp. 319–340.

Mezzanotte L, van 't Root M, Karatas H, Goun EA, Clemens W G. 2017. In Vivo Molecular Bioluminescence Imaging: New Tools and Applications. *Trends in Biotechnology*. doi:10.1016/j.tibtech.2017.03.012

Mitani Y, Futahashi R, Niwa K, Ohba N, Ohmiya Y. 2013. Cloning and characterization of luciferase from a Fijian luminous click beetle. *Photochem Photobiol* **89**:1163–1169.

Nakamura H, Kishi Y, Shimomura O, Morse D, Hastings JW. 1989. Structure of dinoflagellate luciferin and its enzymic and nonenzymic air-oxidation products. *J Am Chem Soc* **111**:7607–7611.

Nakatsu T, Ichiyama S, Hiratake J, Saldanha A, Kobashi N, Sakata K, Kato H. 2006. Structural basis for the spectral difference in luciferase bioluminescence. *Nature* **440**:372–376.

Nasu Y, Campbell RE. 2018. Unnaturally aglow with a bright inner light. *Science* **359**:868–869.

Nathanson JA, Hunnicutt EJ. 1979. Neural control of light emission in Photuris larvae: Identification of Octopamine-Sensitive adenylate cyclase. *J Exp Zool* **208**:255–262.

Niwa K, Ichino Y, Kumata S, Nakajima Y, Hiraishi Y, Kato D-I, Viviani VR, Ohmiya Y. 2010. Quantum yields and kinetics of the firefly bioluminescence reaction of beetle luciferases. *Photochem Photobiol* **86**:1046–1049.

Niwa K, Nakamura M, Ohmiya Y. 2006. Stereoisomeric bio-inversion key to biosynthesis of firefly d-luciferin. *FEBS Lett* **580**:5283–5287.

Oba Y. 2009. On the origin of beetle luminescence. *Bioluminescence in focus Res Signpost, India* 277–290.

Oba Y, Branham MA, Fukatsu T. 2011. The terrestrial bioluminescent animals of Japan. *Zoolog Sci* **28**:771–789.

Oba Y, Furuhashi M, Bessho M, Sagawa S, Ikeya H, Inouye S. 2013a. Bioluminescence of a firefly pupa: involvement of a luciferase isotype in the dim glow of pupae and eggs in the Japanese firefly, Luciola lateralis. *Photochem Photobiol Sci* **12**:854–863.

Oba Y, Hoffmann KH. 2014. Insect Bioluminescence in the Post-Molecular Biology Era. *Insect Molecular Biology and Ecology* 94–120.

Oba Y, Ojika M, Inouye S. 2003. Firefly luciferase is a bifunctional enzyme: ATP-dependent monooxygenase and a long chain fatty acyl-CoA synthetase. *FEBS Lett* **540**:251–254.

Oba Y, Shintani T, Nakamura T, Ojika M, Inouye S. 2008. Determination of the luciferin contents in luminous and non-luminous beetles. *Biosci Biotechnol Biochem* **72**:1384–1387.

Oba Y, Yoshida N, Kanie S, Ojika M, Inouye S. 2013b. Biosynthesis of firefly luciferin in adult lantern: decarboxylation of L-cysteine is a key step for benzothiazole ring formation in firefly luciferin synthesis. *PLoS One* **8**:e84023.

Oertel D, Linberg KA, Case JF. 1975. Ultrastructure of the larval firefly light organ as related to control of light emission. *Cell Tissue Res* **164**:27–44.

Ohba N. 1983. Studies on the communication system of Japanese fireflies. *Sci Rept Yokosuka City Mus* **30**:1–62.

Ohmiya Y, Sumiya M, Viviani VR, Ohba N. 2000. Comparative aspects of a luciferase molecule from the Japanese luminous beetle, Rhagophthalmus ohbai. *Sci Rep Yokosuka City Mus* **47**:31–38.

Ohtsuka H, Rudie NG, Wampler JE. 1976. Structural identification and synthesis of luciferin from the bioluminescent earthworm, Diplocardia longa. *Biochemistry* **15**:1001–1004.

Okada K, Iio H, Goto T. 1976. Biosynthesis of firefly luciferin. Probable formation of benzothiazole from p-benzoquinone and cysteine. *J Chem Soc Chem Commun* 32–32.

Okada K, Iio H, Kubota I, Goto T. 1974. Firefly bioluminescence III. Conversion of oxyluciferin to luciferin in firefly. *Tetrahedron Lett* **15**:2771–2774.

Paley MA. 2014. Expanding the bioluminescent toolkit for in vivo imaging. UC Irvine.

Peterson MK. 1970. The fine structure of the larval firefly light organ. *J Morphol* **131**:103–115.

Pozzo T, Akter F, Nomura Y, Louie AY, Yokobayashi Y. 2018. Firefly Luciferase Mutant with Enhanced

Activity and Thermostability. *ACS Omega* **3**:2628–2633.

Prasher DC, Eckenrode VK, Ward WW, Prendergast FG, Cormier MJ. 1992. Primary structure of the Aequorea victoria green-fluorescent protein. *Gene* **111**:229–233.

Purtov KV, Petushkov VN, Baranov MS, Mineev KS, Rodionova NS, Kaskova ZM, Tsarkova AS, Petunin AI, Bondar VS, Rodicheva EK, Medvedeva SE, Oba Y, Oba Y, Arseniev AS, Lukyanov S, Gitelson JI, Yampolsky IV. 2015. The Chemical Basis of Fungal Bioluminescence. *Angew Chem Int Ed Engl* **54**:8124–8128.

Reddy GR, Thompson WC, Miller SC. 2010. Robust light emission from cyclic alkylaminoluciferin substrates for firefly luciferase. *J Am Chem Soc* **132**:13586–13587.

Ribeiro C, Esteves da Silva JCG. 2008. Kinetics of inhibition of firefly luciferase by oxyluciferin and dehydroluciferyl-adenylate. *Photochem Photobiol Sci* **7**:1085–1090.

Rosa SP. 2007. Análise filogenética e revisão taxonômica da tribo Pyrophorini Candeze, 1863 (Coleoptera, Elateridae, Agrypninae). Universidade de São Paulo.

Rosa SP, Albertoni FF, de Cassia Benà D. 2015. Description of the immature stages of Platycrepidius dewynteri Chassain (Coleoptera, Elateridae, Agrypninae, Platycrepidiini) from Brazil with a synopsis of the larval characters of Agrypninae tribes. *Zootaxa*. doi:10.11646/zootaxa.3914.3.5

Rosa SP, Costa C. 2013. Description of the larva of Alampoides alychnus (Kirsch, 1873), the first known species with bioluminescent immatures in Euplinthini (Elateridae, Agrypninae). *Pap Avulsos Zool* **53**.

Sagegami-Oba R, Oba Y, Ohira H. 2007. Phylogenetic relationships of click beetles (Coleoptera: Elateridae) inferred from 28S ribosomal DNA: insights into the evolution of bioluminescence in Elateridae. *Mol Phylogenet Evol* **42**:410–421.

Santaniello E, Meroni G, Ciana P, Maggi A. 2009. A New Synthesis of 2-Cyano-6-hydroxybenzothiazole, the Key Intermediate of d-Luciferin, Starting from 1,4-Benzoquinone. *Synlett*. doi:10.1055/s-0029-1217971

Seaman WH. 1891. On the Luminous Organs of Insects. *Proceedings of the American Society of Microscopists* **13**:133–162.

Shimomura O. 2012. Bioluminescence: Chemical Principles and Methods. World Scientific.

Shimomura O, Johnson FH, Saiga Y. 1962. Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, Aequorea. *J Cell Comp Physiol* **59**:223–239.

Smalley KN, Tarwater DE, Davidson TL. 1980. Localization of fluorescent compounds in the firefly light organ. *J Histochem Cytochem* **28**:323–329.

Smith DS. 1963. THE ORGANIZATION AND INNERVATION OF THE LUMINESCENT ORGAN IN A FIREFLY, PHOTURIS PENNSYLVANICA (COLEOPTERA). *J Cell Biol* **16**:323–359.

Stanger-Hall KF, Lloyd JE, Hillis DM. 2007. Phylogeny of North American fireflies (Coleoptera: Lampyridae): implications for the evolution of light signals. *Mol Phylogenet Evol* **45**:33–49.

Stanley PE. 1989. A review of bioluminescent ATP techniques in rapid microbiology. *J Biolumin Chemilumin* **4**:375–380.

Stolz U, Velez S, Wood KV, Wood M, Feder JL. 2003. Darwinian natural selection for orange bioluminescent color in a Jamaican click beetle. *Proc Natl Acad Sci U S A* **100**:14955–14959.

Strehler BL, Cormier MJ. 1953. Factors affecting the luminescence of cell-free extracts of the luminous bacterium, Achromobacter fischeri. *Arch Biochem Biophys* **47**:16–33.

Suzuki N, Goto T. 1972. Studies on firefly bioluminescence—II: Identification of oxyluciferin as a product in the bioluminescence of firefly lanterns and in the chemiluminescence of firefly luciferin. *Tetrahedron* **28**:4075–4082.

Trimmer BA, Aprille JR, Dudzinski DM, Lagace CJ, Lewis SM, Michel T, Qazi S, Zayas RM. 2001. Nitric oxide and the control of firefly flashing. *Science* **292**:2486–2488.

Trowell SC, Dacres H, Dumancic MM, Leitch V, Rickards RW. 2016. Molecular basis for the blue bioluminescence of the Australian glow-worm Arachnocampa richardsae (Diptera: Keroplatidae). *Biochem Biophys Res Commun* **478**:533–539.

Viviani VR, Becham EJH. 1993. BIOPHYSICAL AND BIOCHEMICAL ASPECTS OF PHENGODID (RAILROAD-WORM) BIOLUMINESCENCE. *Photochem Photobiol* **58**:615–622.

Viviani VR, Hastings JW, Wilson T. 2002. Two Bioluminescent Diptera: The North American Orfelia fultoni and the Australian Arachnocampa flava. Similar Niche, Different Bioluminescence Systems. *Photochem Photobiol* **75**:22–27.

Watkins OC, Sharpe ML, Perry NB, Krause KL. 2018. New Zealand glowworm (Arachnocampa luminosa) bioluminescence is produced by a firefly-like luciferase but an entirely new luciferin. *Sci Rep* **8**:3278.

Welsh DK, Noguchi T. 2012. Cellular bioluminescence imaging. *Cold Spring Harb Protoc* **2012**. doi:10.1101/pdb.top070607

White EH, McCapra F, Field GF. 1963. The Structure and Synthesis of Firefly Luciferin. *J Am Chem Soc* **85**:337–343.

Wood KV. 1995. THE CHEMICAL MECHANISM and EVOLUTIONARY DEVELOPMENT OF BEETLE BIOLUMINESCENCE. *Photochem Photobiol* **62**:662–673.

CHAPTER 2.

Sulfoluciferin is Biosynthesized by a Specialized Luciferin Sulfotransferase in Fireflies

Authors

Timothy R. Fallon,[1,2] Fu-Shuang Li,[1] Maria A. Vicent,[1,3] and Jing-Ke Weng,[1,2]

Author Affiliations

[1]Whitehead Institute for Biomedical Research, 455 Main Street, Cambridge, Massachusetts 02142, United States
[2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States
[3]Department of Biology, Williams College, Williamstown, Massachusetts 01267, United States

Published as

Fallon, T. R., Li, F.-S., Vicent, M. A., & Weng, J.-K. (2016). Sulfoluciferin is Biosynthesized by a Specialized Luciferin Sulfotransferase in Fireflies. *Biochemistry*, *55*(24), 3341–3344.

Author contributions:

I performed experiments and analyses of all figures, and wrote the paper in combination with Jing-Ke Weng. Maria Vicent assisted with cloning of these genes from cDNA, and preliminary protein expression experiments. Fu-Shuang Li assisted with chemical syntheses, and NMR analyses.

ABSTRACT

Firefly luciferin is a specialized metabolite restricted to fireflies (family Lampyridae) and other select families of beetles (order Coleoptera). Firefly luciferin undergoes luciferase-catalyzed oxidation to produce light, thereby enabling the luminous mating signals essential for reproductive success in most bioluminescent beetles. Although firefly luciferin and luciferase have become widely used biotechnological tools, questions remain regarding the physiology and biochemistry of firefly bioluminescence. Here we report sulfoluciferin to be an *in vivo* derivative of firefly luciferin in fireflies and report the cloning of luciferin sulfotransferase (LST) from the North American firefly *Photinus pyralis*. LST catalyzes the production of sulfoluciferin from firefly luciferin and the sulfo-donor PAPS. Sulfoluciferin is abundant in several surveyed firefly genera as well as in the bioluminescent elaterid beetle *Ignelater luminosus* at a low level. We propose that sulfoluciferin could serve as a luciferin storage molecule in fireflies and that LST may find use as a new tool to modulate existing biotechnological applications of the firefly bioluminescent system.

INTRODUCTION

Bioluminescence is the production of light by a chemical reaction in a biological context. In well-described cases, such as fireflies (White et al., 1963), luminous ostracods (Kishi et al., 1966), and dinoflagellates (Nakamura et al., 1989), the reaction consists of the oxidation of a small molecule, known as luciferin, by an enzyme, known as luciferase, with molecular oxygen. Despite the shared nomenclature of luciferin and luciferase, known bioluminescence consists of at least seven independently evolved systems with structurally unique luciferins and nonhomologous luciferases (Shimomura, 2012). Firefly luciferin (hereinafter luciferin) was the first luciferin to be structurally characterized (White et al., 1963), and firefly luciferase (hereinafter luciferase) was the first luciferase gene to be cloned (de Wet et al., 1985). As luciferase has no prosthetic groups and requires only D-luciferin, ATP, Mg^{2+}, and O_2 to produce light, the enzyme has been readily adapted to in vivo and in vitro applications, such as usage as a reporter gene (Ow et al., 1986), and quantification of ATP by luminometry (Lundin, 2000). Although extensive research exists on the biotechnological usage of firefly bioluminescence, key questions remain on the metabolic biochemistry of the firefly bioluminescent system. For example, it is unknown how luciferin is biosynthesized from primary metabolic precursors (Oba et al., 2013), and it is unclear how accessory enzymes function to store excess luciferin or to recycle the luminescent reaction product oxyluciferin (Gomi and Kajiyama, 2001; Niwa et al., 2006).

RESULTS

In an effort to elucidate firefly luciferin metabolism, we analyzed methanolic extracts of the posterior abdominal tissue containing the bioluminescent lantern (hereinafter referred to as “lantern” tissue) from the firefly Photinus pyralis by liquid-chromatography-high-resolution accurate-mass mass-spectrometry (LC-HRAM-MS). Under positive ion mode, this analysis detected luciferin as one of the most dominant mass features in the total ion chromatogram (TIC) (Figure S1). We also noted an identical ion to the luciferin $[M+H]^+$ ion with a well resolved retention time 2 min earlier than that of luciferin (Figure S2). In-depth analysis of the MS^1 and MS^2 scans revealed that the luciferin-matching ion was likely an in-source fragment ion from a $[M+H]^+$ precursor ion with a mass-to-charge ratio (*m/z*) of 360.9614. The 360.9614 precursor ion also had a highly similar fragmentation pattern to luciferin (Figure S3). Given the constraint of the luciferin chemical formula ($C_{11}H_8N_2O_3S_2$) and the high mass accuracy (<5 ppm) of the Q-Exactive mass spectrometer used in our study, the predicted chemical formula of the precursor ion was limited to that of luciferin with addition of a sulfo group ($C_{11}H_9N_2O_6S_3$ – expected $[M+H]^+$ *m/z* 360.9622). The same conclusion was also drawn from LC-HRAM-MS analysis under negative ion mode (Figure S4). Analysis of the MS^2 of the $[M-H]^-$ ion of the sulfo-modified luciferin peak revealed the loss of a carboxyl group (−43.99 Da) without the loss of a sulfo group (−79.96 Da), indicating that the sulfo group was bound to the hydroxyl group of luciferin (Figure 1). We dub this compound, 2-(6-sulfooxy-1,3-benzothiazol-2-yl)-4,5-dihydro-1,3-thiazole-4-carboxylic acid, firefly sulfoluciferin. The identity of putative sulfoluciferin found in the P. pyralis lantern extract was confirmed by comparing the retention time, exact mass, MS^1 isotopic pattern, and MS^2 spectra to an chemically synthesized authentic sulfoluciferin standard (Figures S3 and S5). The authentic sulfoluciferin standard was synthesized from commercial D-luciferin and sulfur trioxide pyridine complex essentially according to published protocols (Miska and Geiger, 1987; Nakamura et al., 2014), and structurally verified by ^1H NMR (Figure S6).

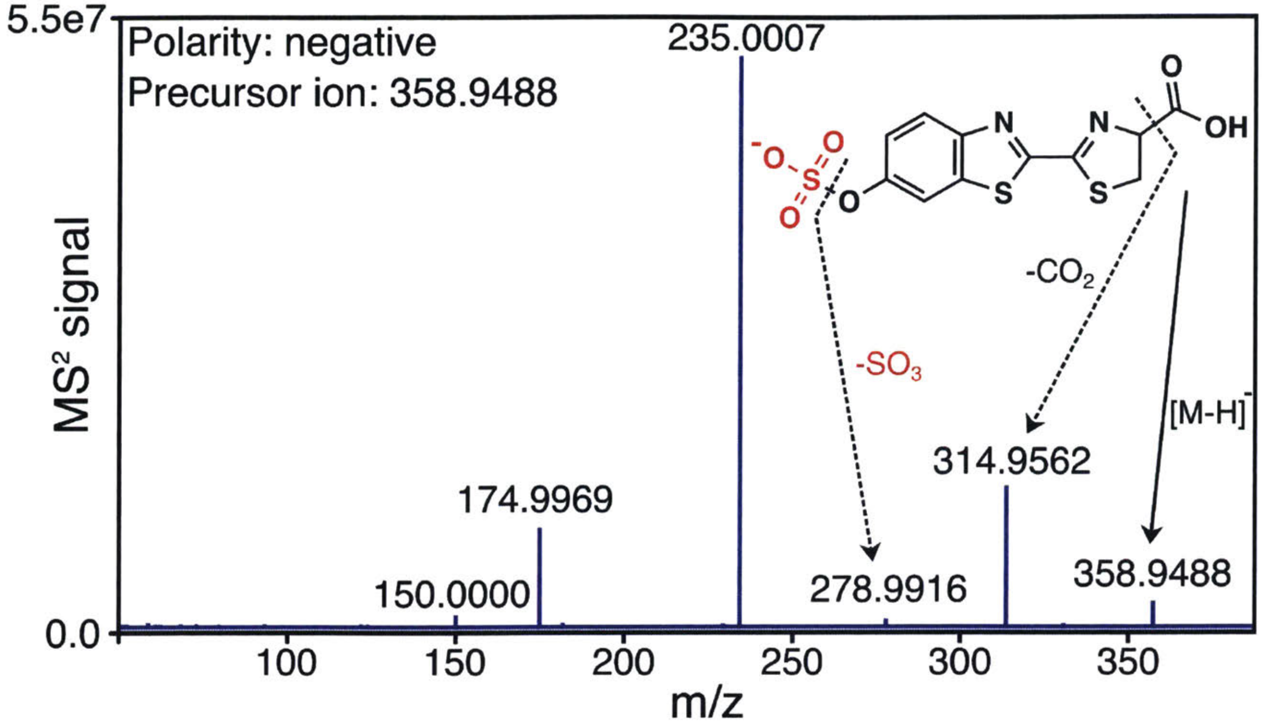


Figure 1: MS^2 fragmentation spectra of firefly sulfoluciferin under negative ion mode indicates that the sulfo-group is bound to the luciferin hydroxyl.

To assess whether the occurrence of sulfoluciferin is widespread among bioluminescent beetles, we analyzed methanolic extracts prepared from adult firefly specimens under the genera *Photinus*, *Pyractomena*, *Photuris*, *Ellychnia*, and the bioluminescent click beetle *Ignelater luminosus* by LC-HRAM-MS (Figure 2). We found comparable quantities of luciferin, and sulfoluciferin respectively in adult nocturnal fireflies of the genera *Photinus*, *Photuris*, and *Pyractomena* (Figures 2 and S7). Interestingly, *Ellychnia corrusca*, a diurnal firefly species that does not develop an adult lantern (Figure 2A), contained decreased levels of luciferin but a comparable ratio of luciferin to sulfoluciferin (Figure S7). *Ignelater luminosus*, an elaterid beetle that develops two dorsal lanterns on the adult prothorax as well as a ventral lantern on the anterior abdomen (Figure 2A), contains luciferin at a level comparable to those of the surveyed nocturnal fireflies, but relatively little sulfoluciferin (Figures 2b and S7). Measurement of the sulfoluciferin to luciferin molar ratio in the *Photinus pyralis* lantern indicated that sulfoluciferin is typically at least four times more abundant than luciferin (Table S1, Supporting Information 1).

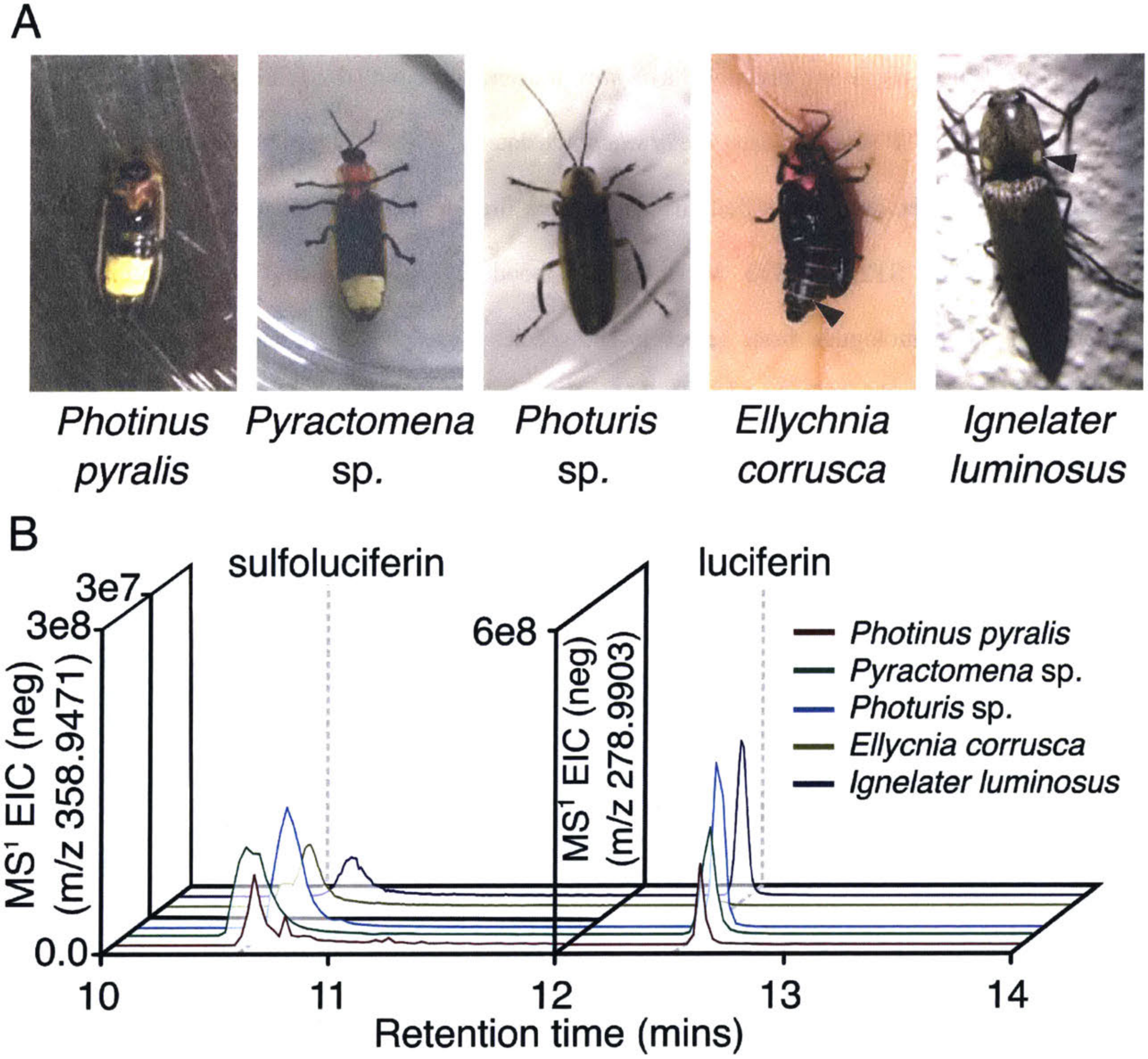
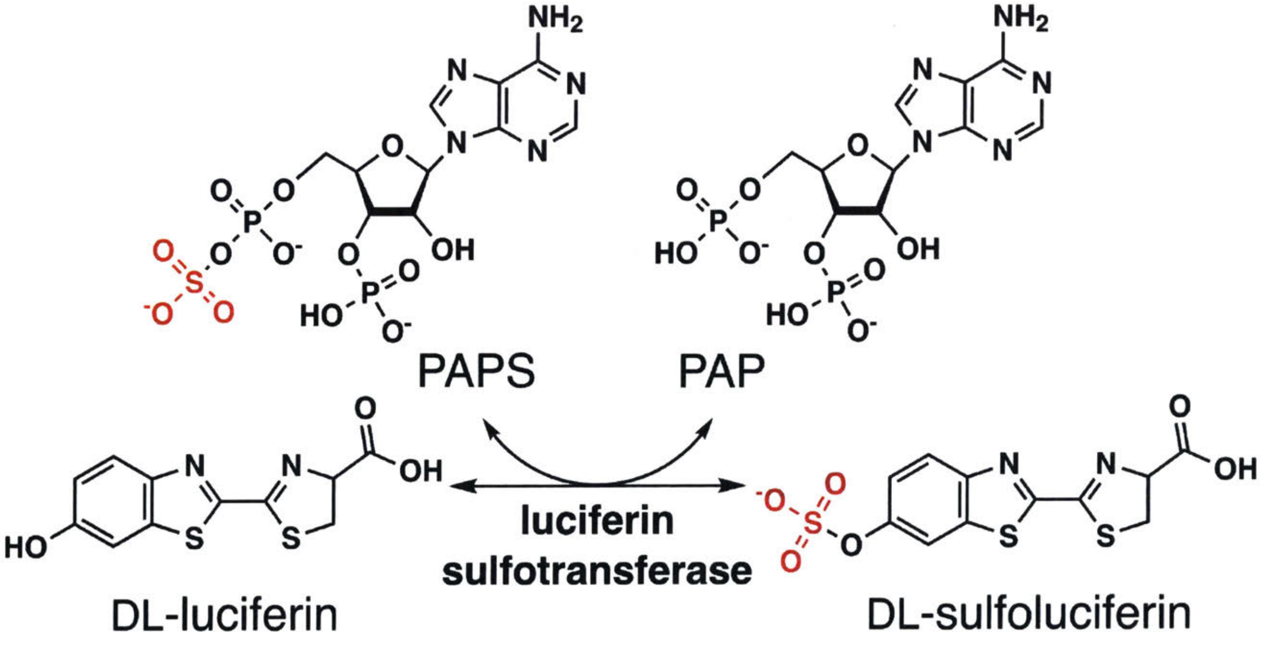


Figure 2: (A) Bioluminescent beetle species of genera *Photinus*, *Pyractomena*, *Photuris*, *Ellychnia*, and *Ignelater*, used in this study. Arrows denote the lack of the abdominal lantern in the diurnal firefly *Ellychnia corrusca*, and the presence of dorsal prothorax lanterns on the prothorax of *Ignelater luminosus*. (B) Relative quantities of sulfoluciferin, and luciferin in select bioluminescent beetle species as measured by LCHRAM- MS.

Given that sulfotransferases (STs) are a well-known class of enzymes that catalyze the transfer a sulfo group from the universal sulfo-donor 3′-phosphoadenosine-5′-phosphosulfate (PAPS) to an acceptor alcohol or amine (James, 2014), we hypothesized that the formation of sulfoluciferin could be catalyzed by a specialized luciferin sulfotransferase (LST) present in fireflies (Scheme 1). To identify candidate genes encoding LST from *P. pyralis*, we performed an RNA-Seq experiment using total RNA extracted from *P. pyralis* lantern tissue and assembled a *de novo* transcriptome with the Trinity assembler (Grabherr et al., 2011). A BLASTP search against the *in silico* translated *P. pyralis* lantern transcriptome using Mus musculus Sult1a1 as the query identified three full-length firefly ST candidates. We dub these candidate sulfotransferases ST1, ST2, and ST3. Maximum-likelihood phylogenetic analysis of these three genes together with ST homologues from select insect species highlighted a clade structure suggesting specialization of ST1 and ST2 within fireflies (Figures 3 and S12). Expression analysis of the ST transcripts from the de novo transcriptome indicated that ST1 was markedly more highly expressed (Figure S8).



Scheme 1. DL-Luciferin Is Enzymatically Interconverted to Sulfoluciferin by Luciferin Sulfotransferase (LST)Scheme 1. DL-Luciferin Is Enzymatically Interconverted to Sulfoluciferin by Luciferin Sulfotransferase (LST)

To test the biochemical function of *P. pyralis* ST1, ST2, and ST3, we cloned their corresponding open reading frame (ORF) from *P. pyralis* cDNA and produced purified recombinant STs from *Escherichia coli*. In vitro enzyme assays using the three recombinant STs revealed that only one, *P. pyralis* ST1, could catalyze the formation of sulfoluciferin from luciferin and PAPS (Figure 4). This enzyme indeed corresponds to the most highly expressed ST candidate gene in the *P. pyralis* lantern transcriptome, and one of the two candidates highlighted as possibly neofunctionalized by the phylogenetic analysis (Figures 3 and S12). We therefore dub this enzyme firefly luciferin sulfotransferase (LST). Further experiments demonstrated that ST2 and ST3 were able to convert the model ST substrate p-nitrophenol-sulfate to p-nitrophenol in the presence of 3′-phosphoadenosine-5′-phosphate (PAP), whereas LST did not have this activity (Figure S9). LST was able to catalyze the formation of luciferin from sulfoluciferin in the presence of PAP (Figure S10). The k_{cat} of sulfoluciferin formation by LST was characterized to be at least 3 s^{-1} (Supporting Information 1). Attempts to precisely determine the K_M of LST for luciferin were confounded by PAP product inhibition at higher substrate conditions. Product inhibition by PAP was previously reported for several other STs with K_d values in the sub-micromolar range (Sekura and Jakoby, 1979). As our k_{cat} value is comparable to reported (Hattori et al., 2008) k_{cat} values of other STs, and we did not observe a change in the catalytic rate at substrate concentrations down to 1 μM, we propose that the K_M of LST is in the low micromolar range or below. Analysis of the stereochemical preference of LST for D- or L-luciferin indicated the enzyme had no preference for either luciferin stereoisomer (Figure S11). Analysis of published transcriptome data for other firefly species (Sander and Hall, 2015) indicated a LST ortholog likely exists in all Lampyrids (Figure S12).

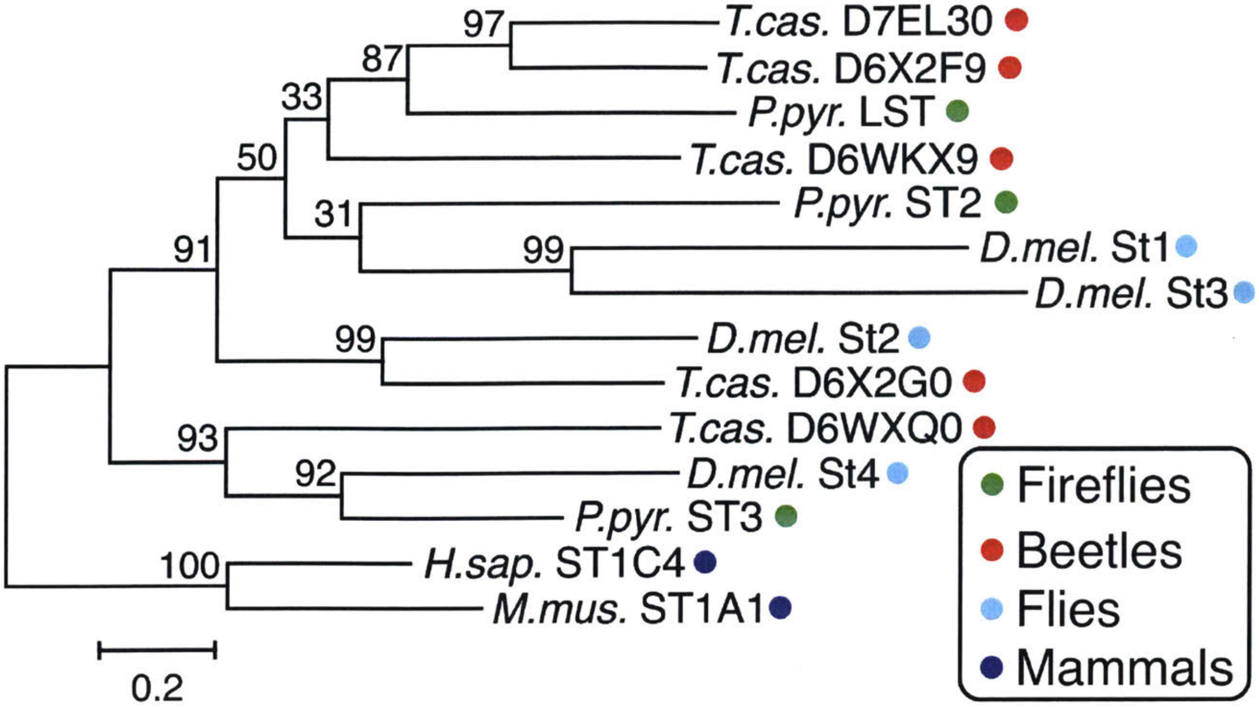


Figure 3: Maximum likelihood inference of sulfotransferase phylogeny from select insect species rooted on two mammalian STs as an outgroup. Node labels indicate bootstrap support (5000 replicates). Branch length measures substitutions per site.

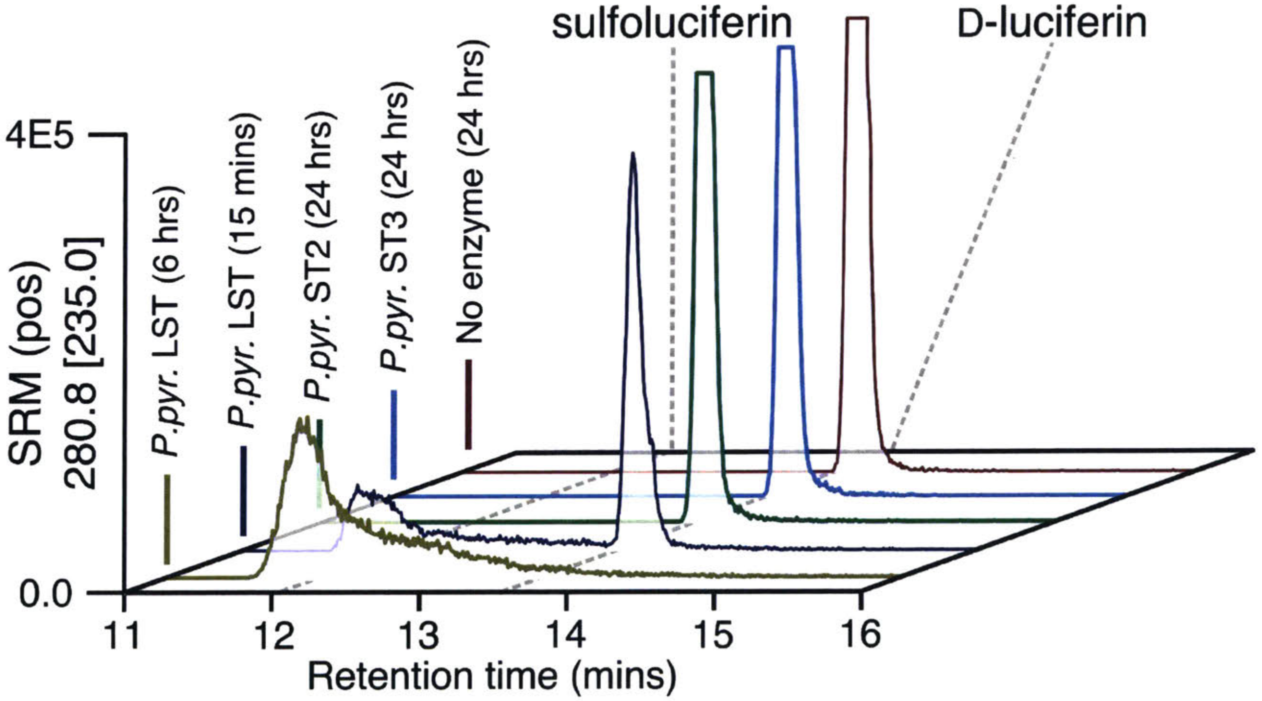


Figure 4: Enzyme activity assay of candidate firefly STs. Only LST showed detectable conversion of luciferin to sulfoluciferin by liquid chromatography-selected reaction monitoring-mass spectrometry (LC-SRM-MS).

DISCUSSION

Taken together, our results demonstrate that sulfoluciferin is biosynthesized by a specialized sulfotransferase in *P. pyralis* and likely in other Lampyrids. Sulfonation of luciferins have been reported in several bioluminescent systems of marine origin, e.g., vargulin enol-sulfate found in the bioluminescent ostracod *Vargula hilgendorfii* (Nakamura et al., 2014), coelenterazine disulfate found in the firefly squid *Watasenia scintillans* (Inoue et al., 1976), and coelenterazine enolsulfate found in the soft coral *Renilla reniformis* (Cormier et al., 1970). It has been proposed that sulfonated luciferins may serve as a luciferase inaccessible storage form in certain bioluminescent organisms (Shimomura, 2012). Previous enzymatic characterization of synthetic firefly sulfoluciferin with *Photinus pyralis* luciferase (Miska and Geiger, 1987) and our own characterization (Figure S13) demonstrate that sulfoluciferin is not utilized by *Photinus pyralis* luciferase, and hence would be suitable as a storage form. Luciferin may be released from sulfoluciferin in vivo by LST in the presence of excess PAP, or by yet uncharacterized sulfatases. The occurrence of sulfoluciferin in fireflies likely evolved from the promiscuous action of some progenitor STs. Indeed, the low levels of sulfoluciferin relative to luciferin observed in the bioluminescent click beetle *Ignelater luminous* may represent such an ancestral state, where the production of sulfoluciferin does not require a dedicated ST. In fireflies, however, our data suggest that a specialized LST has emerged, arguing that sulfoluciferin has sufficient functional relevance to drive enzyme specialization. From the perspective of biotechnology, LST represents a unique tool to sequester luciferin into a luciferase inactive but readily available form in vivo and in vitro.

Acknowledgements

The authors thank Sara Lewis for assistance with collection of *Ellychnia corrusca* and assistance with identification of firefly genera. We thank Adam South for collection assistance with *Photinus pyralis*. We

thank David Jenkins for collection assistance with *Ignelater luminosus*, and providing the image of *I. luminosus* used in Figure 2A.

Supporting Information 1

Materials and Methods

Specimen collection

Live *Photinus pyralis* were collected from private land in Allentown, PA by Dr. Adam South (Harvard School of Public Health) in July of 2015 on the basis of flash patterns. Dried adult firefly specimens were purchased commercially (P/N: FFW-5G, Sigma-Aldrich). In both cases specimens were individually verified to be *P. pyralis* before experimentation on the basis of size, pronotal pigmentation pattern, and the margin of unpigmented tissue on the anterior segment to the lantern carrying segments.

Pyractomena sp. and *Photuris* sp. specimens were collected as larvae and reared to adults from a collection in October 2014, from the Rock Meadow Conservation Area in Belmont, MA (42° 24' 6.65" N, 71° 11' 50.40" W). Firefly collections from Rock Meadow were approved by the Belmont Conservation Commission. Firefly larvae were collected from ~6-inch tall grass in the Rock Meadow at night by hand on the basis of sporadic glowing behavior. Identifications of firefly genera, both as larvae and adult, were assisted by Dr. Sara Lewis (Tufts University), and through comparisons to firefly photographs on BugGuide.net. Firefly larvae were kept in continual darkness in plastic containers with airholes & moistened kimwipes. Larvae were fed on a weekly diet of moistened cat food (Friskies), as well as occasional live Bladder snails (*Physella* sp.). Food was provided to the larvae overnight, and was removed the next day. Under these conditions firefly larvae survived for multiple months, although a minority of larvae did die during rearing. Larvae were resistant to starvation for at least 1 month. No specific manipulation was made to induce pupation of the larvae. Pupation occurred stochastically after months in captivity and not at all in some specimens. Live adult firefly specimens were maintained in the laboratory for less than 2 weeks in petri-dishes with regularly moistened kimwipes (Kimtech) and slices of apple (replaced when browned).

Ignelater luminosus specimens were collected from private land in Mayagüez, Puerto Rico (18° 13' 12.1974" N, 67° 6' 31.6866" W) with permission of the landowner by Dr. David Jenkins (USDA-ARS). *I. luminosus* specimens were captured at night on April 20th and April 28th 2015 during flight on the basis of flashing. The *I. luminosus* specimens were frozen in a -80˚C freezer, lyophilized, shipped to our laboratory on dry ice, and stored at -80˚C.

Ellychnia corrusca specimens were collected on April 15th 2015 by Dr. Sara Lewis from the Massachusetts Aubdobon Habitat in Belmont MA (42° 24' 8.7912" N, 71° 11' 1.7082" W) with permission of the Massachusetts Audubon Society. *E. corrusca* specimens were collected by hand from tree trunks in the morning. Live *E. corrusca* specimens were provided to our laboratory by Dr. Lewis.

All live fireflies were anesthetized by transient exposure to CO_2, sacrificed by flash freezing in liquid nitrogen, and either stored lyophilized at -80˚C, or used directly for experimentation.

Liquid chromatography high-resolution accurate-mass mass spectrometry (LC-HRAM-MS)

All specimens used for Figure 2 and Table S1 were collected as live specimens by the authors or their collaborators. The lantern-carrying abdominal segments of single fireflies (posterior 2 lantern segments) were removed with a razor blade at 4 ˚C. The lantern-carrying tissue was then transferred to 50% methanol (150 µL). In the case of *Ignelater luminosus*, the prothorax tissue containing the prothorax lanterns was separated from the thorax, and transferred to 50% methanol (0.5 mL). In both cases, the tissue was macerated in the solvent, and intermittently sonicated in a water bath sonicator for 30 minutes, not letting the temperature rise above 40 ˚C. These extraction conditions are intentionally mild, given that luciferin and its derivatives are prone to air oxidation, and measurements of stereochemistry are less reliable when compounds have been exposed to high temperatures. Post sonication, the extract was centrifuged in a benchtop centrifuge at 14,000 g @ 4˚C for 10 min to pellet tissue debris and other particulates. The clarified extract was filtered through a 0.2 µm PFTE filter (Filter Vial, P/No. 15530-100, Thomson Instrument Company). 20 μL of the filtered extract was separated on an UltiMate 3000 liquid chromatography system (Dionex) equipped with a 150 mm C18 Column (Kinetex 2.6 μm silica core shell C18 100Å pore, P/No. 00F-4462-Y0, Phenomenex) coupled to a Q-Exactive mass spectrometer (Thermo Scientific). Compounds were separated by reversed-phase chromatography on the C18 column by a gradient of Solvent A (0.1% formic acid in H2O) and Solvent B (0.1% formic acid in acetonitrile); 5% B for 2 min, 5-80% B over 40 min, 95% B for 4 min, and 5% B for 5 min; flow rate 0.8 mL/min. Under these conditions, the retention time of luciferin was ~12.5 min while the retention time of sulfoluciferin was ~10.6 min. As samples were run over multiple months, a retention time drift (<0.3 min) was noted as the C18 column aged.

The mass spectrometer was configured to perform 1 MS^1 scan from *m/z* 120-1250 followed by 1-3 data-dependent MS^2 scans using HCD fragmentation with a stepped collision energy of 10, 15, 25 normalized collision energy (NCE). Data was collected as profile data. The instrument was always used within 7 days of the last mass accuracy calibration. The ion source parameters were as follows: spray voltage (+) at 3000 V, spray voltage (-) at 2000 V, capillary temperature at 275 ˚C, sheath gas at 40 arb units, aux gas at 15 arb units, spare gas at 1 arb unit, max spray current at 100 (μA), probe heater temp at 350 ˚C, ion source: HESI-II. The raw data in Thermo format was converted to mzML format using ProteoWizard MSConvert (Chambers et al., 2012). Data analysis was performed with Xcalibur (Thermo Scientific) and MZmine2 (Pluskal et al., 2010).

Chemical synthesis, purification, and characterization of firefly sulfoluciferin

2-(6-sulfooxy-1,3-benzothiazol-2-yl)-4,5-dihydro-1,3-thiazole-4-carboxylic acid (InChI=1S/C11H8N2O6S3/c14-11(15)7-4-20-9(13-7)10-12-6-2-1-5(3-8(6)21-10)19-22(16,17)18/h1-3,7H,4H2,(H,14,15)(H,16,17,18), InChi Key LPKFAQWRYNJJRB-UHFFFAOYSA-N), which we dub firefly sulfoluciferin, was synthesized according to protocols described previously, with some modifications (Miska and Geiger, 1987; Nakamura et al., 2014). Dry pyridine (6 mL, P/N: 270970, Sigma-Aldrich) was transferred to a 25 mL round bottom flask using anhydrous technique. Anhydrous conditions and reagents are absolutely essential for high-yield synthesis of sulfoluciferin. A gentle stream of N_2 was used throughout the synthesis to displace air from the headspace of the flask. Free acid firefly luciferin (150 mg, 0.53 mmol, P/N: L-123, Gold Biotechnology) was added to the flask and dissolved with stirring. Sulfur trioxide pyridine complex (160 mg, 1 mmol, P/N: S7556, Sigma-Aldrich) was then added and dissolved with stirring. The flask was covered to protect from light and incubated at room temperature for 2 hours. The yield of sulfoluciferin from this reaction was ~60% as gauged by UV-HPLC at 210 nm. The volume of the crude reaction mixture was reduced by rotary evaporation to a viscous residue. The residue was diluted with 20 mL of H_2O supplemented with ammonia (120 µL, 0.5 M) before storage at -80 °C. Sulfoluciferin was purified from the crude reaction mixture by reversed-phase UV-HPLC on a preparative PFP column (Kinetex, 5 µm silica core shell PFP with TMS endcapping, 100Å pore, P/No. 00F-4602-P0-AX, Phenomenex). We observed that sulfoluciferin degrades under acidic conditions; therefore no acid additive was added to the solvents used for the chromatographic purification process. Compounds were eluted from the PFP column by a gradient consisting of Solvent A (H2O) and Solvent B (acetonitrile); 5% B for 5 min, 5-95% B over 20 min, 95%B for 5 min, 95-5% B over 1 min, 5% B, 5% B for 4 min; flow rate 15 mL/min. Under these conditions the retention time of luciferin was ~6.5 min while the retention time of sulfoluciferin was ~4 min. The fractions containing sulfoluciferin were collected and lyophilized, which yielded 3.5 mg of firefly sulfoluciferin as a white solid. Sulfoluciferin was the only peak in this sample by UV-HPLC at 210 nm and LC-MS scanning from *m/z* 100-900.

MS^2 fragmentation spectra (Figure S3), ^1H NMR spectra (Figure S6), and UV-Vis absorbance spectra (Figure S14) were obtained. All spectra matched expectations from theory. We found that sulfoluciferin was prone to oxidation when stored in DMSO, with nearly complete oxidation, largely to sulfodehydrodroluciferin (Figure S14), at 4 °C within 4 weeks. Sulfoluciferin also showed substantial degradation when stored at -20°C in PBS after 6 months.

.^{1}H NMR spectroscopy

10 mg of free acid luciferin (P/N: L-123, Gold Biotechnology) and 3.5 mg purified sulfoluciferin were dissolved in 0.75 mL DMSO-*d6*. ^{1}H NMR spectra were acquired for both compounds on a Bruker Avance III 400 MHz NMR spectrometer (MIT Department of Chemistry Instrumentation Facility), using the default pulse sequence while locked, tuned, and spinning. Spectra were analyzed with and plotted with MestReNova 10.0.2 (Mestrelab Research), and are presented in Figure S6. Measured peaks were as follows:

Luciferin:

^{1}H NMR (400 MHz, DMSO-*d6*) δ: 13.15 (1H, s, OH), 10.22 (1H, s, COOH), 7.95 (1H, d, *J* = 8.8, 1-H), 7.45 (1H, d, *J* = 2.4, 4-H), 7.06 (1H, dd, *J* = 8.8, 2.4, 2-H), 5.40 (1H, dd, *J* = 9.6, 8.0, 9-H), 3.71 (2H, m, 10-H).

Sulfoluciferin:

^{1}H NMR (400 MHz, DMSO-*d6*) δ: 8.02 (1H, d, *J* = 8.8, 1-H), 7.95 (1H, d, *J* = 2.0, 4-H), 7.36 (1H, dd, *J* = 8.8, 2.0, 2-H), 5.12 (1H, t, *J* = 9.6, 8.4, 9-H), 3.67 (2H, m, 10-H).

UV-Vis spectroscopy

UV-Vis spectra for luciferin, dehydroluciferin, sulfoluciferin, and dehydrosulfoluciferin (Figure S14) were obtained on an UltiMate 3000 liquid chromagtography system coupled to an UltiMate3000 in-line diode-array-detector (Dionex).

Preparation of *P. pyralis* total lantern RNA

Total lantern RNA was extracted from dried single adult male *P. pyralis* specimens (Sigma-Aldrich) by the acidic phenol-chloroform method. *P. pyralis* abdominal tissue containing the lantern (posterior 2 abdominal segments) was separated from the remainder of the body using a razor blade at 4°C, and placed directly into QIAzol reagent (Qiagen). Total RNA was extracted from the tissue by phenol-chloroform extraction using the RNeasy Lipid Tissue Mini Kit (Qiagen), following the manufacturer's instructions. RNA preps from two separate male *P. pyralis* individuals were used for Illumina sequencing & cDNA synthesis respectively.

Preparation of *P. pyralis* cDNA

Single-strand cDNA was prepared from *P. pyralis* total RNA extracted from the lantern by poly-T primed reverse transcriptase using the SuperScript III First-Strand Synthesis System for RT-PCR (Invitrogen), following the manufacturer's instructions.

High-throughput Illumina RNA-Seq and transcriptome assembly.

P. pyralis total RNA was submitted to Novus Genomics for strand-specific Illumina sequencing library preparation and Illumina sequencing. A single Illumina sequencing library was prepared from the total RNA by Novus Genomics using the TruSeq Stranded Total RNA with Ribo-Zero Gold kit (Illumina). The resulting Illumina sequencing library was multiplexed with unspecified libraries and sequenced on a single lane with 125x125 paired-end sequencing on a HiSeq2500 sequencer (Illumina) to a depth of 20,140,685 forward reads and 12,922,768 reverse reads passing the quality filter.

Resulting reads in FASTQ format were checked with the FastQC software package (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and Illumina TruSeq2 adaptor contamination and low quality reads were removed by the Trimmomatic software package (http://www.usadellab.org/cms/?page=trimmomatic (Bolger et al., 2014), with the following parameters "ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 SLIDINGWINDOW:4:5 LEADING:5 TRAILING:5 MINLEN:25". 10,915,359 paired reads remained post quality filtering. A de novo transcriptome was assembled from the filtered paired reads with Trinity 2.0.6 (Grabherr et al., 2011) using default parameters with "--SS_lib_type RF" for strand specific assembly, on a single high-memory server

(Whitehead Institute). Candidate ORFs were translated in silico from the de novo transcriptome using Transdecoder 2.0.1 (Haas et al., 2013), with the minimum protein length set to 20 amino acids.

Unfiltered RNA-Seq reads have been uploaded to NCBI SRA with accession number (SRR3521424). The de novo assembled transcriptome produced in this study has been uploaded to NCBI TSA with accession number (GEOW00000000). We highlight that this is an unreplicated low-coverage RNA-Seq dataset without filtering of low-confidence transcripts. The dataset should be used with appropriate caution and appreciation of the caveats of RNA-Seq and *de novo* transcriptome assembly.

De novo transcriptome expression analysis

Expression analysis was performed using Trinity by the included "align_and_estimate_abundance.pl" script, which utilizes Bowtie (Langmead et al., 2009) and RSEM (Li and Dewey, 2011) to map reads to assembled transcripts and perform transcript quantification with expectation maximization respectively. Default parameters were used with the exception of "--SS_lib_type RF" for strand specific expression analysis.

Selection of *P.pyralis* sulfotransferase candidates from the *de novo* transcriptome

Protein sequences in FASTA format were provided to SequenceServer, an open-source standalone BLAST server (Camacho et al., 2009; Priyam et al., 2015), for interactive BLAST analysis. Candidate sulfotransferases were selected from the Transdecoder-produced *P. pyralis* protein database by a BLASTP similarity search using the protein sequence of *Mus musculus* cytosolic sulfotransferase Sult1a1 as the query sequence. Unless otherwise stated, an e-value cutoff of 1e-20 was used for all BLAST queries. Only complete protein sequences containing a putative start codon and stop codon (TransDecoder description:"type:complete") were kept in the results. Unusually short proteins for the sulfotransferase family (<250 amino acids) and duplicate sequences were manually removed. Three full-length sulfotransferase candidates, LST, ST2, and ST3 remained after this filtering.

Phylogenetic analysis

LST, ST2, and ST3 were used as a BLASTP query against the model beetle *Tribolium castenum* Uniprot.org reference proteome (downloaded 2015-01-16), which yielded 5 proteins after filtering unusually large proteins (>450 amino acids) and filtering for duplicate hits. The 5 *T. castenum* sulfotransferases were then used as a query against the fruit fly *Drosophila melanogaster* Uniprot reference proteome (downloaded 2015-12-08), which yielded 5 sulfotransferase sequences after duplicate removal.

The amino acid sequences of the 3 *P. pyralis* sulfotransferase candidates, the 5 putative sulfotransferases from the *T. castenum proteome*, the 5 putative sulfotransferases from *D. melanogaster, Mus musculus* Sult1a1, and *Homo sapiens* ST1C4 were then concatenated and utilized for phylogenetic analysis. Multiple sequence alignment was performed using the online MAFFT server (http://mafft.cbrc.jp/alignment/software/) with parameters "MAFFT - L-INS-i - mafft --reorder --maxiterate 1000 --retree 1 --localpair input". Phylogenetic analyses in Figure 3 and Figure S12 were conducted in MEGA6 (Tamura et al., 2013). The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model (Jones et al., 1992). The percentage of trees (5000 bootstrap replicates) in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. For Figure 3, the analysis involved 14 amino acid sequences. There were a total of 432 positions in the final dataset, and the tree with the highest log likelihood (-9967.3035) is shown in figure 6. For Figure S12, the analysis involved 17 amino acid sequences, there were a total of 432 positions in the final dataset, and the tree with the highest log likelihood (-10700.8028) is shown.

Cloning of candidate genes

CDS sequences encoding candidate sulfotransferases were PCR amplified from the *P. pyralis* cDNA with their respective primers (Table S2) using Phusion polymerase (New England BioLabs). Amplified bands were purified by agarose gel extraction, and cloned into NcoI linearized pHis8-4 plasmid by Gibson assembly (Gibson Assembly Master Mix, NEB). pHis8-4 is an *E. coli* T7 expression plasmid descended from pHIS8-3 (Weng and Noel, 2012), that harbors an N-terminal 8xHis tag followed by a TEV protease cleavage site for His-tag removal. The Gibson assembly mix was directly transformed into DH5-α *E. coli* for propagation. Plasmid clones were sequence confirmed by dual ended fluorescent Sanger sequencing (Genewiz), and stored at -80 ˚C in a 25% glycerol stock. Some differences, such as synonymous SNPs and an in-frame triplet deletion in the case of ST2, were noted between the de novo transcriptome sequence and the cDNA-cloned sequence. These polymorphisms are to be expected as *P. pyralis* represent a heterozygous population, and separate *P. pyralis* individuals were sampled for cDNA synthesis and Illumina sequencing. The confirmed clones were designated pJKW 0643(LST), pJKW 0633(ST2), and pJKW 0690(ST3), and are available from Addgene.org with the following accession numbers LST (74121) ST2 (74122) and ST3 (74123).

Recombinant protein expression in *E. coli*

BL21(DE3) *E. coli* carrying the respective expression plasmids were seeded from glycerol stocks at -80 ˚C, grown at 37 ˚C in TB media to an optical density at 600 nm of 1.0, induced with 1 mM isopropyl-β-D-thiogalactoside (IPTG), and allowed to grow for an additional 16 h at 18 °C. Bacterial cells were harvested by centrifugation, resuspended in lysis buffer (50 mM Tris-HCl, pH 8.0, 0.5 M NaCl, 20 mM imidazole, 1% [v/v] Tween 20, 10% [v/v] glycerol, and 20 mM 2-mercaptoethanol), and incubated with ~0.05 mg/mL lysozyme and ~0.05 mg/mL DNase I with stirring for 30 min at 4˚C. The cell homogenate was then lysed by shearing (Microfluidics, Microfludizer Corporation) to produce a crude protein lysate. After clarification of the crude protein lysate by centrifugation at 30,000xg at 4 ˚C for 1 hour, the expressed protein was isolated from the lysate by affinity chromatography with nickel nitrilotriacetic acid (Ni-NTA) coupled agarose (Bio-Rad Laboratories). Ni-NTA bound protein was eluted from the column with lysis buffer containing 0.25 M imidazole. The partially purified protein was dialyzed overnight in dialysis buffer (50 mM Tris-HCl pH 8, 500 mM NaCl, 5 mM DTT), and treated with 1 mg recombinant TEV (lab-made) for His-tag cleavage. A second Ni-NTA chromatography step was used to remove the remaining His-tagged proteins, including TEV, uncleaved target proteins, and background Ni-NTA binders. The flowthrough containing the protein of interest was concentrated using an Ultra centrifugal filter (P/N UFC901024, 10,000 Da MWCO, Amicon) to ~1 mg/mL. Protein concentration was gauged by absorbance at 280 nm on a NanoDrop 2000 UV-Vis spectrophotometer (Thermo-Scientific). The molar absorptivity coefficient at 280 nm for a given enzyme was predicted from the primary protein sequence using the ProtParam online tool (http://web.expasy.org/protparam/). The protein was further purified by size exclusion chromatography (Superdex-200, GE Healthcare), with storage buffer (10mM Tris-HCl pH 8, 300mM NaCl, 5mM DTT). Protein post gel-filtration was again concentrated by ultrafiltration to the highest possible concentration before substantial precipitation, in this case ~1 mg/mL, flash-frozen in liquid nitrogen, and stored frozen at -80 °C. Proteins post-storage were assayed for purity and identity by SDS-PAGE and LC/MS intact proteomics on a Xevo Q-TOF MS electrospray-ionization time-of-flight mass spectrometer (Waters Corporation). LST, ST2, and ST3 all expressed robustly in *E. coli* and were obtained at >95% purity as gauged by SDS-PAGE and intact proteomics.

Liquid-chromatography-triple quadrupole-mass spectrometry (LC-QqQ-MS)

5-10 μL of a given sample was injected onto an Ultimate 3000 liquid chromatography system (Dionex), equipped with a 150 mm C18 Column (Kinetex 2.6 μm silica core shell C18 100Å pore, P/No. 00F-4462-Y0, Phenomenex), coupled to UltiMate 3000 diode-array-detector (DAD) in-line UV-Vis spectrophotometer (Dionex) and a TSQ Quantum Access MAX triple-quadrupole mass spectrometer

(Thermo-Scientific). Compounds were separated by reversed-phase chromatography on the C18 column by a gradient of Solvent A (0.1% formic acid in H2O) and Solvent B (0.1% formic acid in acetonitrile); 5% B for 2 min, 5-29.6% B over 15 min, 29.6-95% B over 1 min, and 95% B for 5 min, 5%B for 2 minutes; flow rate 0.8 mL/min.

The diode-array detector was configured to scan at 5 Hz at 210 nm, 254 nm, 280 nm, 312 nm, and a wavelength scan from 200 nm to 800 nm.

The mass spectrometer was configured to either full scan (LC-MS) from m/z 100-500 (Figure S10), or perform two selected-reaction-monitoring (LC-SRM-MS) scans (Figure 4, S9, S11, Table S1). Each SRM scan was 0.25 seconds, for luciferin and sulfoluciferin individually. The luciferin SRM was as follows: precursor ion selection at 280.880 m/z on positive ion mode, fragment at 20V, and product ion selection at 234.920 m/z. The sulfoluciferin SRM was as follows: precursor ion selection on negative ion mode at 358.912 m/z, fragment at 21V, and product ion selection at 234.920 m/z. As both SRMs select the same product ion at 234.920 (decarboxylation of luciferin), there was minor cross-talk between SRM scans for luciferin and sulfoluciferin. The m/z resolution of Q1 was set to 0.7 FWHM, the argon collision gas pressure of Q2 was set to 1.5 mTorr, and the Q3 scan width was set to 0.100 m/z for both SRM scans.

Enzymatic assays

Unless otherwise stated, all enzymatic assays utilized single-use enzyme aliquots which had been stored at -80 ˚C post flash-freezing. LST, ST2, and ST3 were frozen at a stock concentration of 0.2 mg/mL, 0.6 mg/mL, and 1.8 mg/mL respectively. Enzyme aliquots were diluted in fresh enzymology buffer (PBS buffer pH 7.4, 1 mM DTT) as a working stock for experiments. A 2 mM working stock of 3'-phosphoadenosine-5'-phosphosulfate (PAPS) in PBS was prepared from the PAPS lithium salt (P/N: sc-210759, Santa Cruz Biotechnology), and stored at -80˚C. A 2 mM working stock of 3'-Phosphoadenosine-5'-phosphate (PAP) in PBS was prepared from the PAPS disodium salt (P/N A5763, Sigma-Aldrich), and stored at -20˚C. A 100 mM stock of firefly D-luciferin was prepared from its sodium salt (P/N LUCK, Gold Biotechnology) in water, and stored at -80°C. Working stocks of firefly luciferin were prepared from the 100 mM stock by serial dilution and stored at -80°C. A 2 mM stock of *p*-nitrophenol sulfate in PBS was prepared from the *p*-nitrophenol sulfate potassium salt (P/N N3877, Sigma-Aldrich), and stored at -20 °C. An estimated 1 mM stock of sulfoluciferin was prepared from ~90% pure lyophilized solid (purity estimate by UV-HPLC, major contaminant luciferin) in PBS, and stored at -20 °C.

For luciferin sulfonation enzyme assays (Figure 4), a reaction mix was prepared from PAPS (50 μL, 2 mM), D-luciferin (10 μL, 250 μM), and fresh enzymology buffer (30 μL PBS, 1 mM DTT). The reaction mix and 1:100 dilution of enzyme (LST, ST2, and ST3) were equilibrated to 25 °C, and 10 μL of a single enzyme was added to the reaction mix to start the reaction. The final reaction volume was 100 μL, with an assay concentration of 1 mM PAPS and 25 μM luciferin. The final enzyme concentration was 0.2 μg/mL, 0.6 μg/mL and 1.8 μg/mL, for LST, ST2, and ST3 respectively. Enzymes were added with a 30 second interval between samples to ensure accurate timing, and were incubated at 25 °C in the dark. 20 μL aliquots were removed at 15 minutes, 6 hours, and 24 hours, and quenched 1:1 with 100% methanol. 10 μL of the quenched sample was analyzed by LC-SRM-MS.

For the *p*-nitrophenol sulfate desulfonation enzyme assays (Figure S9), a reaction mix was prepared from PAP (25 μL, 2 mM), *p*-nitrophenol sulfate (2.5 μL, 2 mM), and fresh enzymology buffer (17.5 μL PBS, 1 mM DTT). The reaction mix and a 1:100 dilution of enzyme (LST, ST2, and ST3) were equilibrated to 25 °C, and 5 μL of a single enzyme was added to the reaction mix to start the reaction. The final reaction volume was 50 μL, with an assay concentration of 1 mM PAP, and 100 μM *p*-nitrophenol sulfate. The final enzyme concentration was 0.2 μg/mL, 0.6 μg/mL and 1.8 μg/mL, for LST, ST2, and ST3 respectively. Enzymes were added with a 30 second interval between samples to ensure accurate timing, and were incubated at 25 °C in the dark. 10 μL aliquots were removed at 15 minutes, 6 hours, and

24 hours, measured from the start time of that particular sample, and quenched 1:1 with 100% MeOH. 10 μL of the quenched sample was analyzed by LC-MS.

For the sulfoluciferin desulfonation enzyme assays (Figure S10), a reaction mix was prepared from PAP (25 μL, 2 mM), sulfoluciferin (5 μL, ~1 mM), and fresh enzymology buffer (60 μL, PBS, 1 mM DTT). The reaction mix and 1:100 dilution of enzyme (LST, ST2, and ST3) were equilibrated to 25 °C, and 10 μL of a single enzyme was added to the reaction mix to start the reaction. The final reaction volume was 100 μL, with an assay concentration of 1 mM PAP, ~100 μM sulfoluciferin. The final enzyme concentration was 0.2 μg/mL, 0.6 μg/mL and 1.8 μg/mL, for LST, ST2, and ST3 respectively. Enzymes were added with a 30 second interval between samples to ensure accurate timing, and were incubated at 25 °C in the dark. 20 μL aliquots were removed at 15 minutes, and 16 hours, measured from the start time of that particular sample, and quenched 1:1 with 100% MeOH. 10 μL of the quenched sample was analyzed by LC-MS.

Estimate of molar ratio of sulfoluciferin relative to luciferin

The relative response factor of sulfoluciferin to luciferin was estimated by the relative peak change method, where an enzymatic conversion is sampled at two timepoints and the relative response factor is derived from the difference in the signal of the two compounds. By this method we determined the relative response factor by LC-SRM-MS for sulfoluciferin relative to luciferin to be 1.7 (higher sulfoluciferin response factor). In order to determine the molar ratio of sulfoluciferin to luciferin in vivo, we analyzed the luciferin and sulfoluciferin content in five individual *Photinus pyralis* males by LC-SRM-MS. Correcting for the relative response factor of luciferin and sulfoluciferin, we found the absolute molar ratio of sulfoluciferin to luciferin to be 4.8, 16, 25, 17, and 8 in these specimens respectively (Table S1). The high variability of the sulfoluciferin to luciferin molar ratio is likely due to the diverse life histories of the wild fireflies used for this analysis, as the size, feeding history, and flash history of the sampled fireflies was not controlled. Nonetheless, these results indicate sulfoluciferin is more abundant than luciferin in the firefly lantern, supporting its role as a luciferin storage compound.

Kinetic parameter estimation for LST

A value of k_{cat} for the sulfonation of luciferin by LST was derived from the 15 min LST timepoint of Figure 4. The assumption is made that the assay concentration of 25 μM for the luciferin substrate is >2x over the K_M of LST, based on reported K_M values for other sulfotransferases (Brenda Enzyme Database). The quantity of enzyme in this assay was 0.459 pmol (10 μL of a 2 μg/mL working stock, enzyme M.W. 43539 g/mol). In the case of substrate, the integrated peak area for unconverted luciferin was 6080759 (arb units), whereas the integrated peak area for luciferin at 15 min was 3042290 (arb units), corresponding to a ~50% conversion of the 25 μM luciferin assay concentration. The quantity of substrate converted in 15 min in the assay was 1249 pmol (50% 25 μM substrate concentration in assay, assay volume of 100 μL). Substrate converted per second is then 1.4 pmol/sec. The substrate molecules converted per second per molecule enzyme is then the reported value $3s^{-1}$.

Supporting Tables

Specimen ID	Lantern dry weight (mg)	Extraction solvent	Luciferin SRM area (arb)	Sulfoluciferin SRM area (arb)	Relative response factor corrected molar ratio (sulfoluciferin / luciferin)
Ppyr_1	N.M.	50% MeOH	799331	6599540	4.8
Ppyr_2	2.1	50% ACN	1177492	32785341	16
Ppyr_3	2.1	50% ACN	988742	42682144	25
Ppyr_4	2.9	50% MeOH	1085283	32348061	17
Ppyr_5	2.2	50%	316514	4530908	8

		MeOH			

Table S1: Molar ratio estimate of sulfoluciferin to luciferin from LC-SRM-MS analysis of posterior abdominal (lantern) extracts of five live collected, flash frozen, lyophilized, and -80˚C stored *P. pyralis* male individuals. A relative response factor of 1.7 (sulfoluciferin/luciferin) is used. N.M. = not measured.

Gene	Primer direction	Primer sequence
LST	Forward	5'-GAAAACTTGTACTTCCAGGCCCATGGC atgtttgcatctatcctaggcaa-3'
LST	Reverse	5'-CTCGAATTCGGATCCGCCATGG ttacatttttggaacagatttttga-3'
ST2	Forward	5'-GAAAACTTGTACTTCCAGGCCCATGGC atggaagaaaataactatctccct-3'
ST2	Reverse	5'-CTCGAATTCGGATCCGCCATGG ttataatttataatcagaatgtttaag-3'
ST3	Forward	5'-GAAAACTTGTACTTCCAGGCCCATGGC ATGCCACATAACATTCAAATTGGGG-3'
ST3	Reverse	5'-CTCGAATTCGGATCCGCCATGG TTACATTCGTTCAAACGGTATATTCG-3'

Table S2: PCR cloning primers for candidate firefly sulfotransferases. Red text represents pHis8-4 overlapping sequence for Gibson assembly.

Supporting Figures

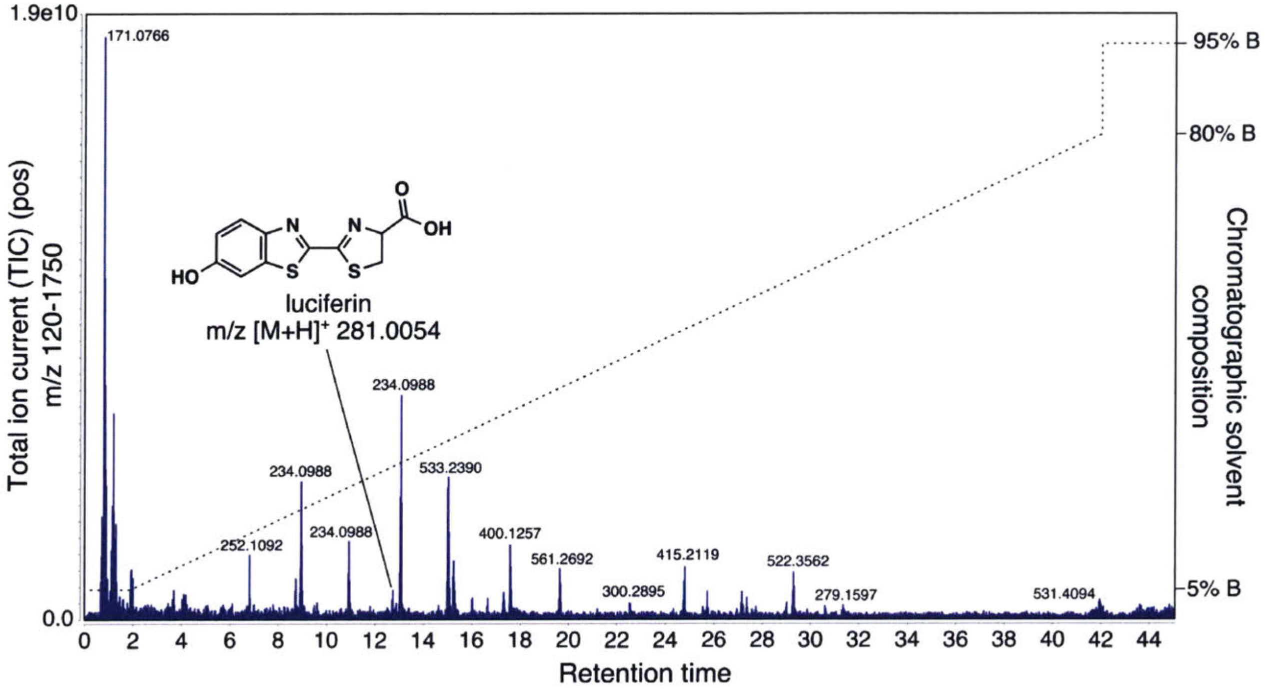


Figure S1. Positive ion-mode total-ion-chromatogram (TIC) from reversed phase C18 chromatography of a *P. pyralis* posterior abdominal (lantern) methanolic extract.

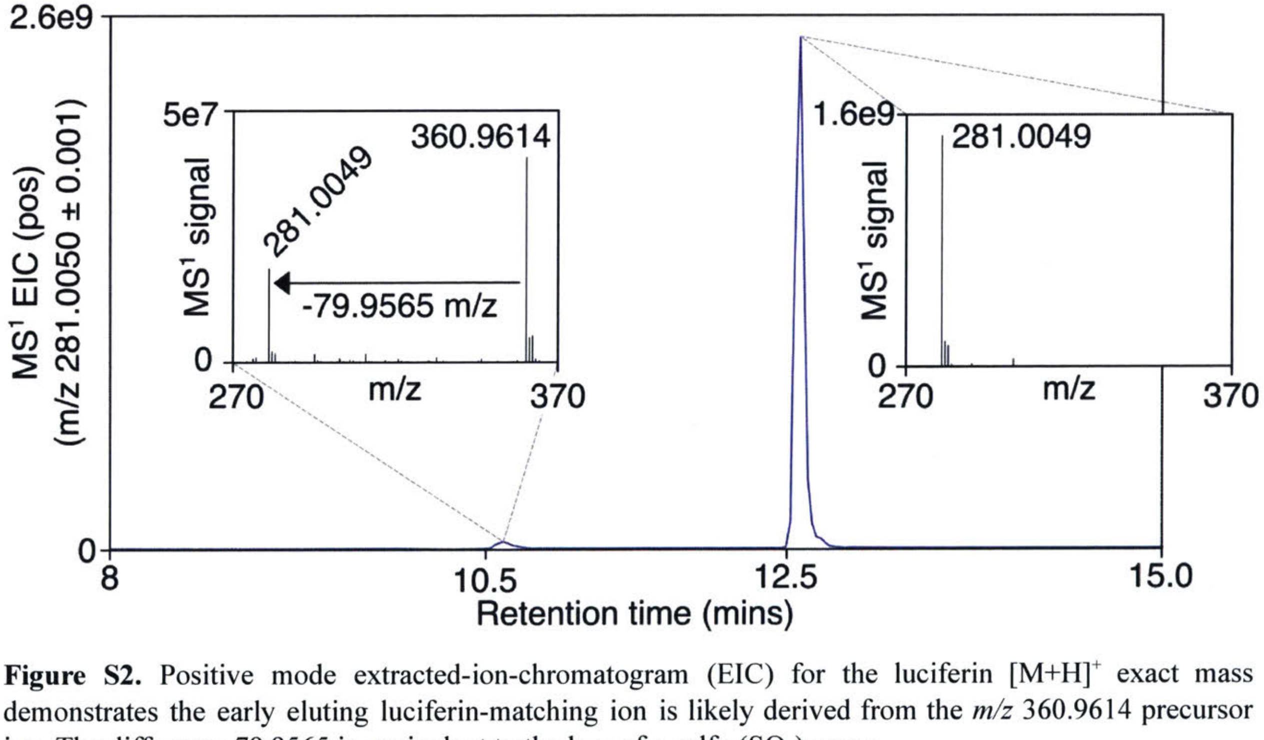


Figure S2. Positive mode extracted-ion-chromatogram (EIC) for the luciferin $[M+H]^+$ exact mass demonstrates the early eluting luciferin-matching ion is likely derived from the *m/z* 360.9614 precursor ion. The difference 79.9565 is equivalent to the loss of a sulfo (SO_3) group.

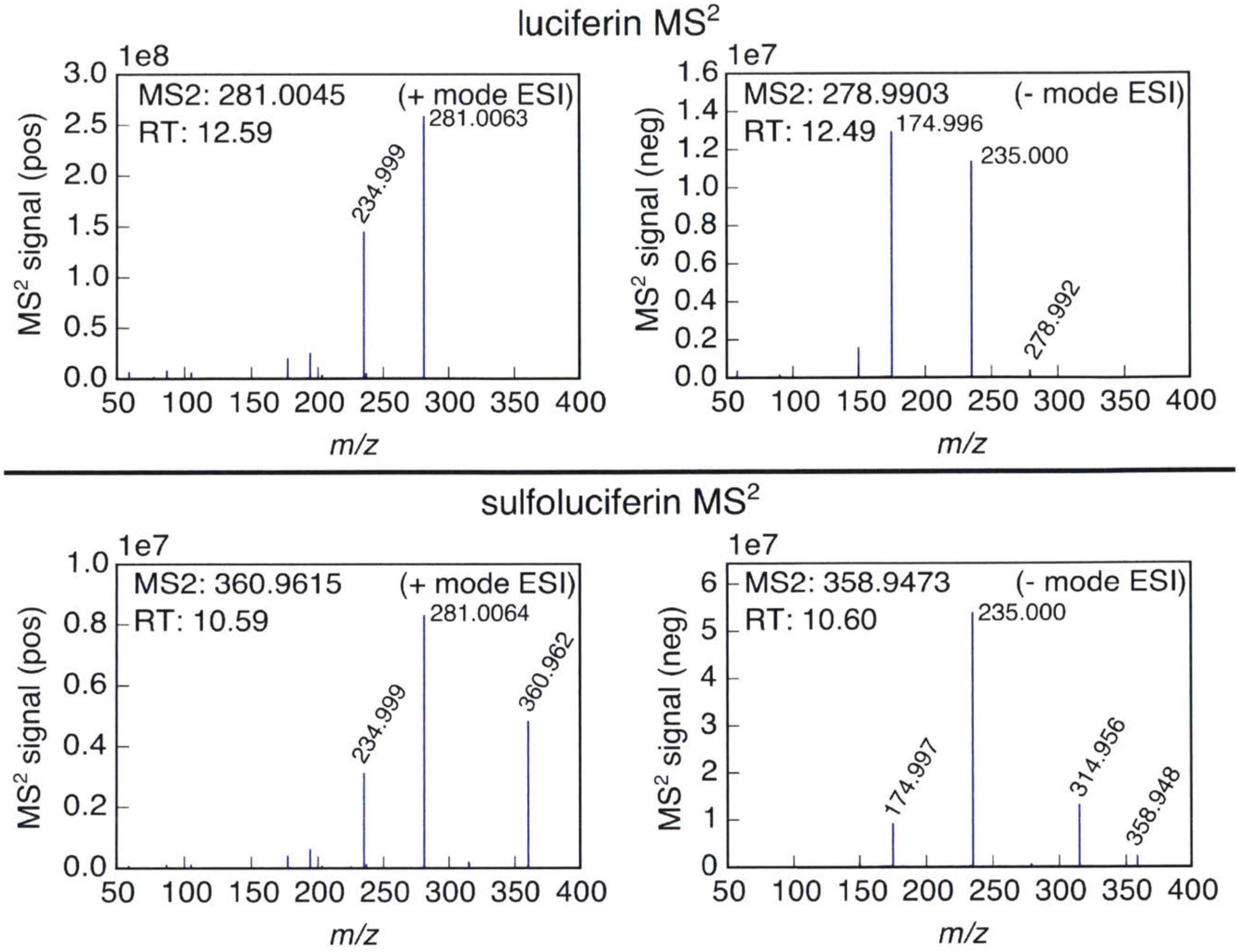


Figure S3. HCD MS^2 fragmentation spectra for luciferin and sulfoluciferin on positive and negative ion mode. Note the degree of similarity between the spectra.

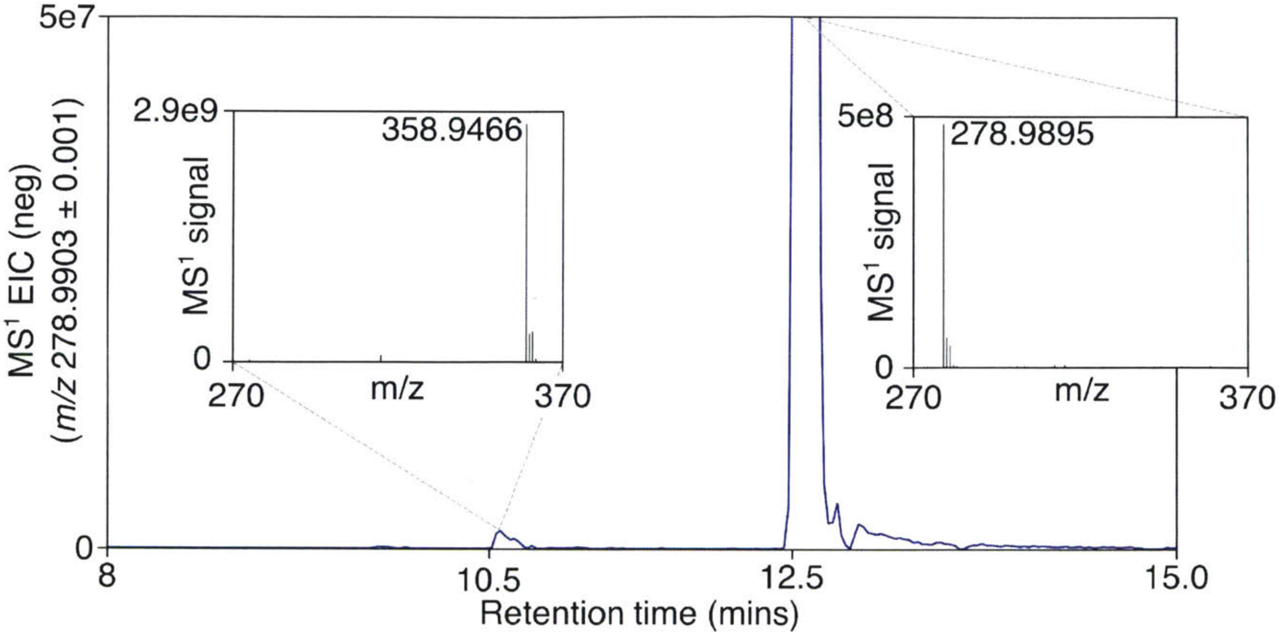
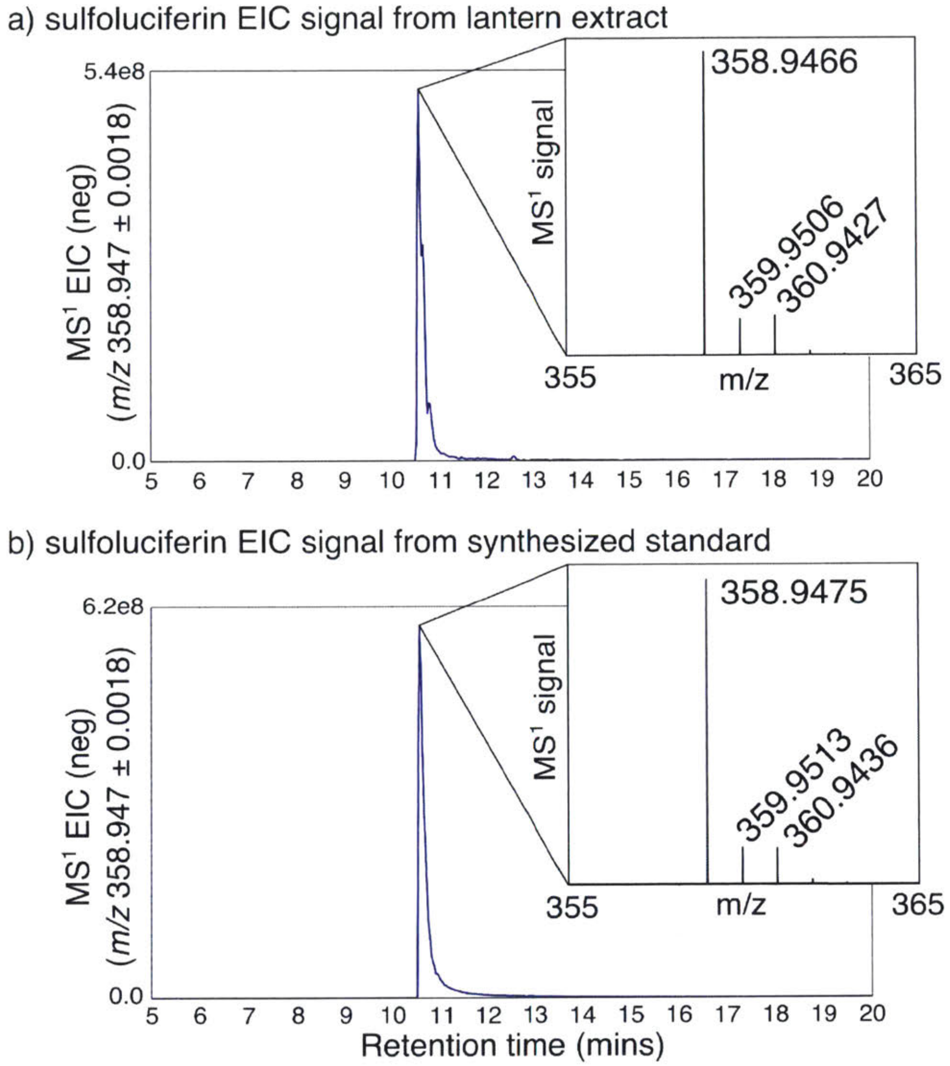


Figure S4. Negative mode EIC for the luciferin [M-H]$^-$ exact mass demonstrates the early eluting luciferin-matching ion is likely derived from the 358.9466 precursor ion.



a) sulfoluciferin EIC signal from lantern extract

b) sulfoluciferin EIC signal from synthesized standard

Figure S5. Comparison of retention time, exact mass, and MS^1 isotopic pattern, of (a) putative firefly sulfoluciferin to (b) synthesized authentic sulfoluciferin standard.

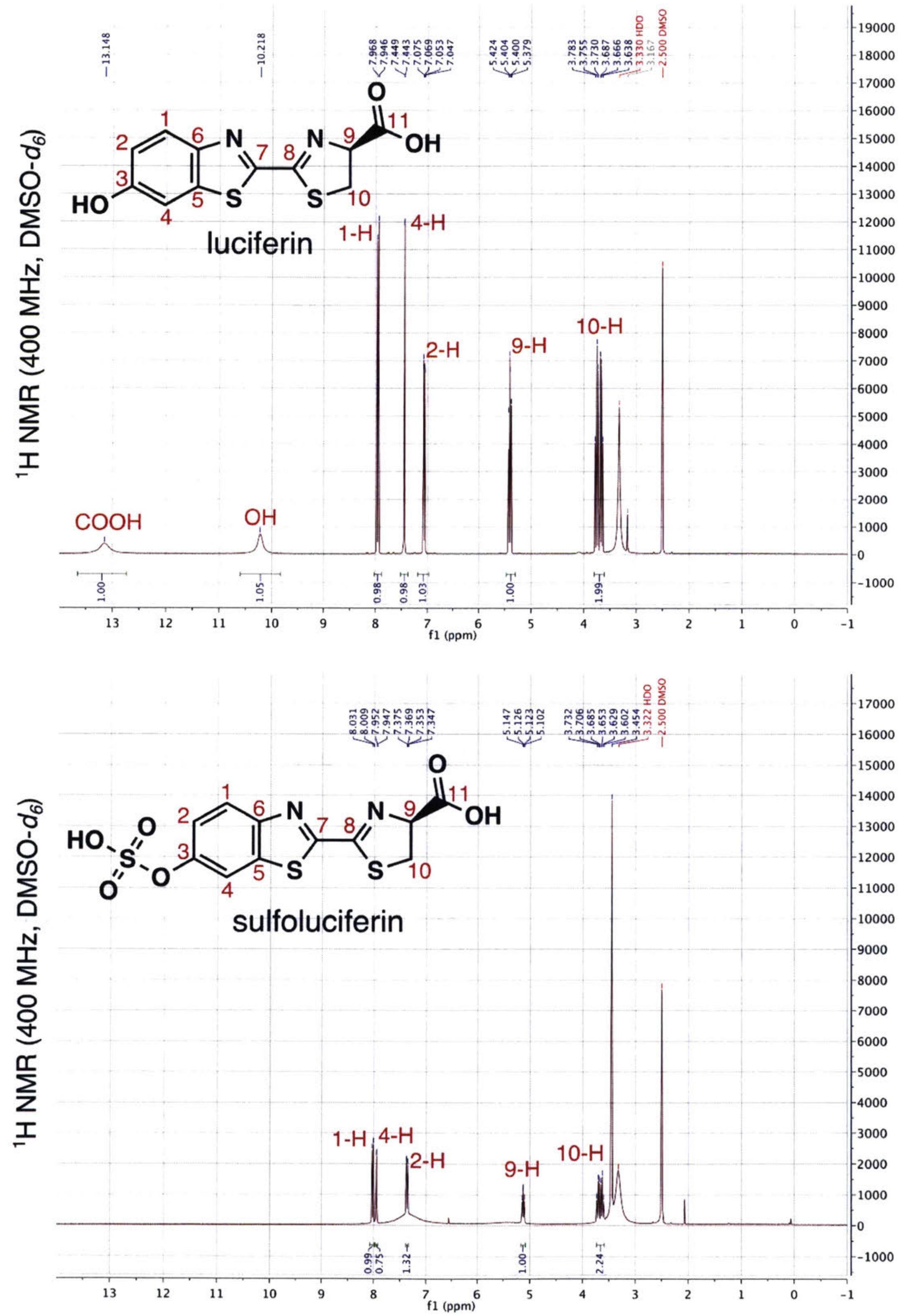


Figure S6. ^{1}H NMR of luciferin and sulfoluciferin in DMSO-d6 on a Bruker Avance III 400 MHz NMR spectrometer.

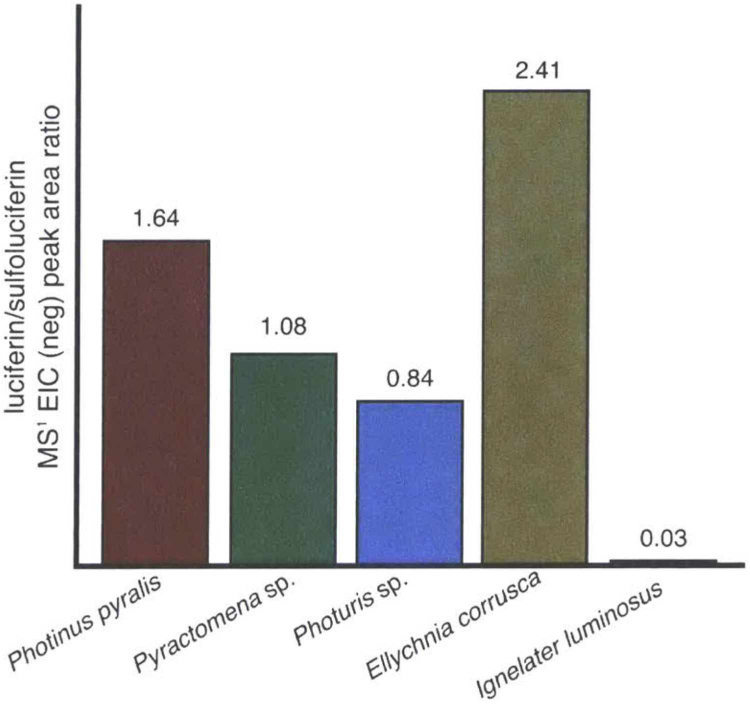


Figure S7. Relative integrated peak area of luciferin to sulfoluciferin from Figure 2B.

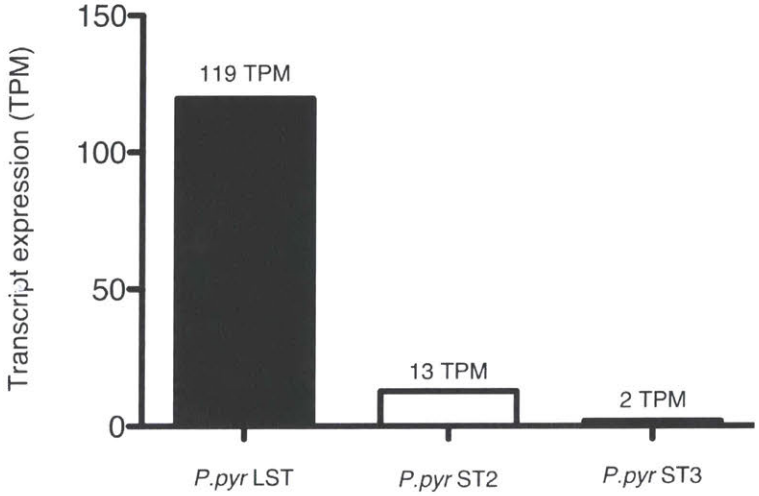


Figure S8. Expression values for the full-length sulfotransferase candidates identified in the de novo *Photinus pyralis* posterior abdominal (lantern) transcriptome. TPM = Transcripts per million.

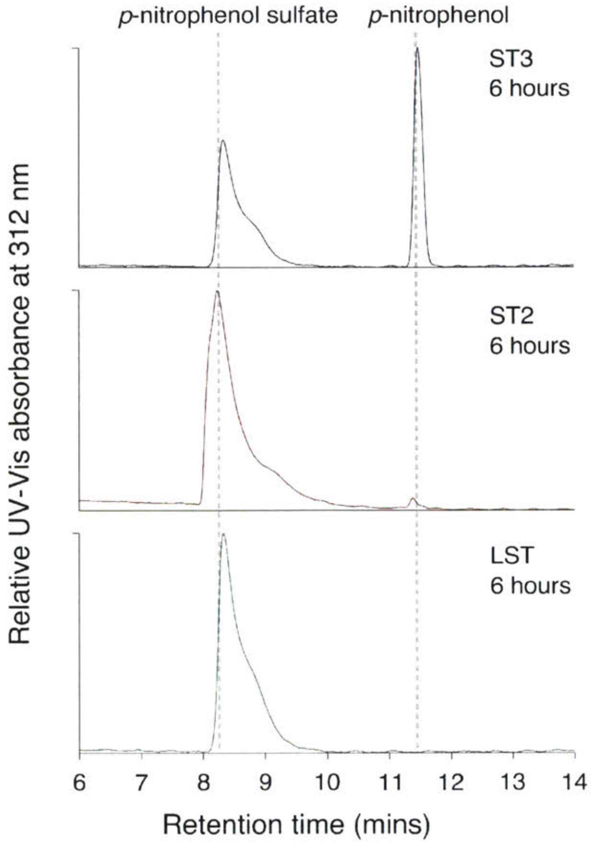


Figure S9. *In vitro* enzymology testing the ability of candidate sulfotransferases to catalyze desulfonation of the model sulfotransferase substrate *p*-nitrophenol sulfate

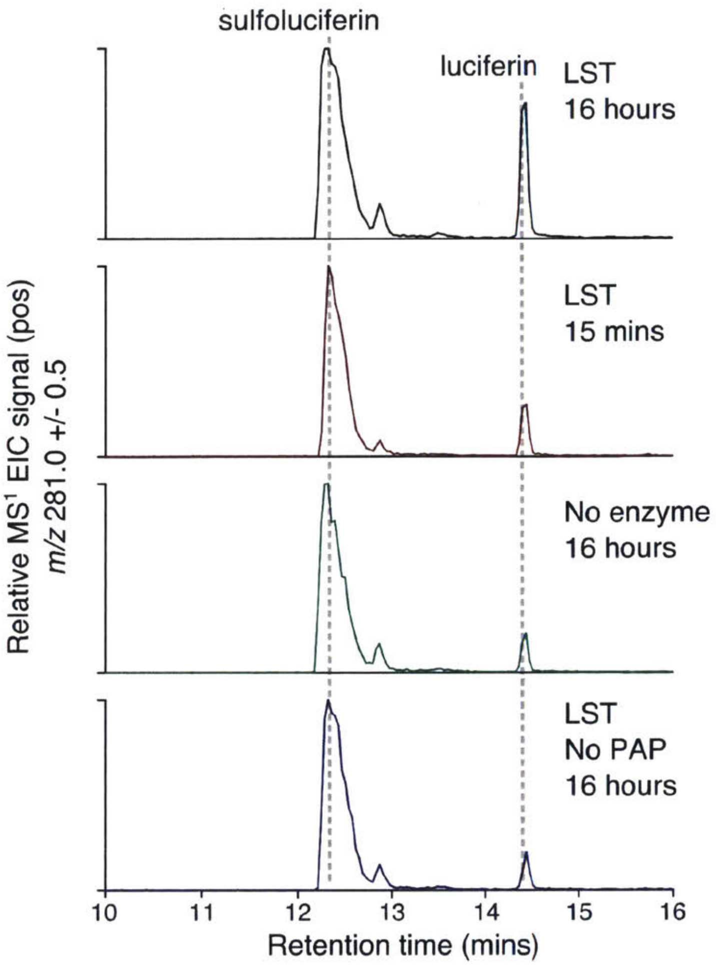


Figure S10. LST catalyzes the PAP-dependent desulfonation of sulfoluciferin.

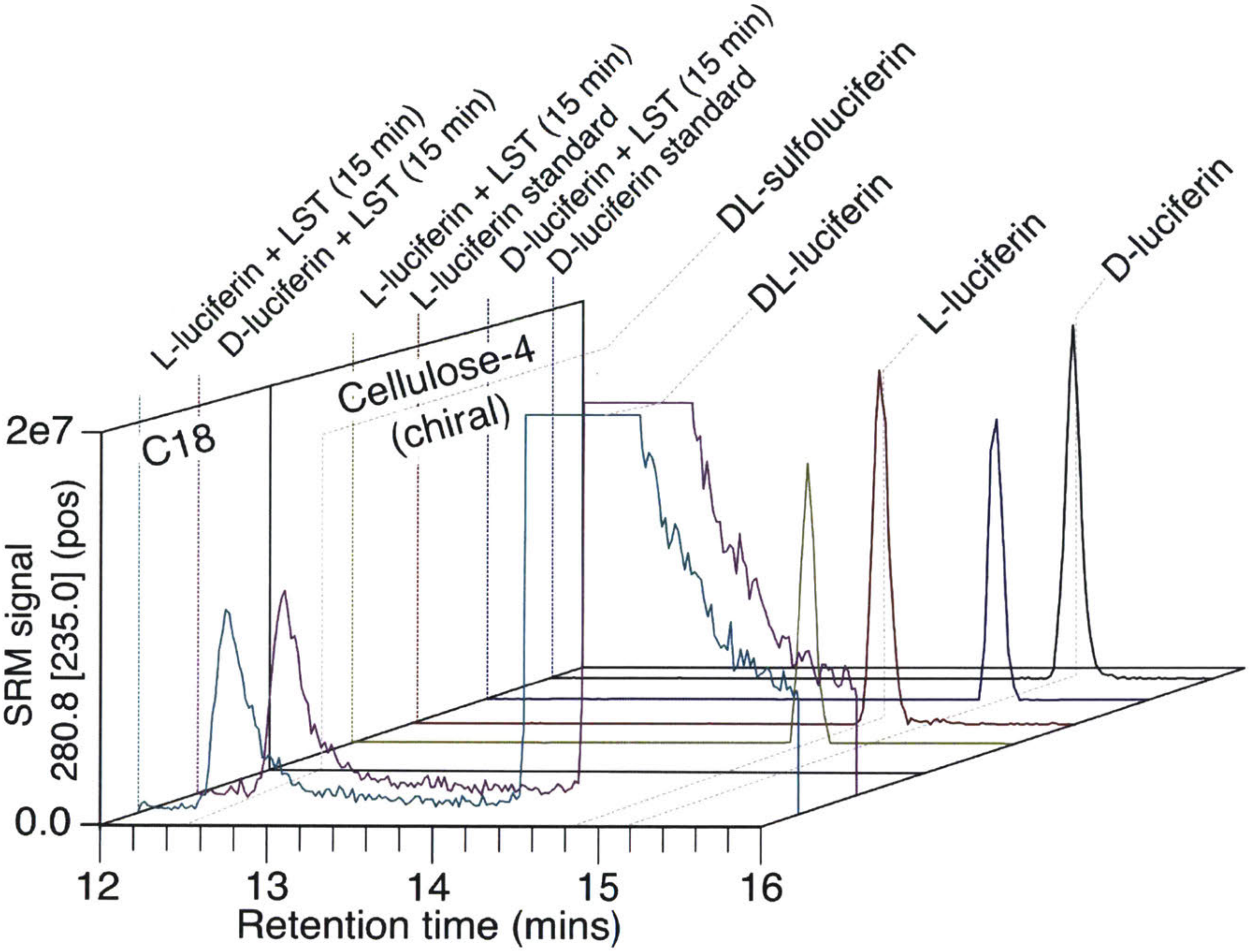


Figure S11. LST does not have a stereochemical preference for sulfonation of either D or L-luciferin. Two reaction mixes were prepared from PAPS (50 μL, 2 mM), D-luciferin (10 μL, 500 μM) or L-luciferin (10 μL, 500 μM), and fresh enzymology buffer (30 μL PBS, 1 mM DTT). The reaction mixes and a 1:250 dilution of LST were equilibrated to 25 ˚C, and 10 μL of the 1:250 LST stock was added to the reaction mix to start the reaction. The final reaction volume was 100 μL, with an assay concentration of 1 mM PAPS and 50 μM D/L-luciferin. The final enzyme concentration was 0.08 μg/mL, 0.15 μg/mL and 0.45 μg/mL, for LST, ST2, and ST3 respectively. Enzymes were added with a 30 second interval between samples to ensure accurate timing, and were incubated at 25 ˚C in the dark. 40 μL aliquots were removed at 15 minutes, and quenched 1:1 with 100% methanol. A low concentration of LST and relatively high concentrations of luciferin are used in the experiment to ensure the reaction was under initial rate conditions at 15 minutes. 10 μL of the quenched samples were analyzed by LC-SRM-MS on a C18 column. 10 μL of the quenched samples were also run on a 250 mm Cellulose-4 column (P/N: 00G-4490-E0 - Lux 3 μm silica - Cellulose-4, Phenomenex), with equivalent gradient chromatography and MS conditions to the reported C18 LC-SRM-MS procedure to confirm luciferin stereochemistry. Sulfoluciferin was not detected under the chiral chromatography conditions.

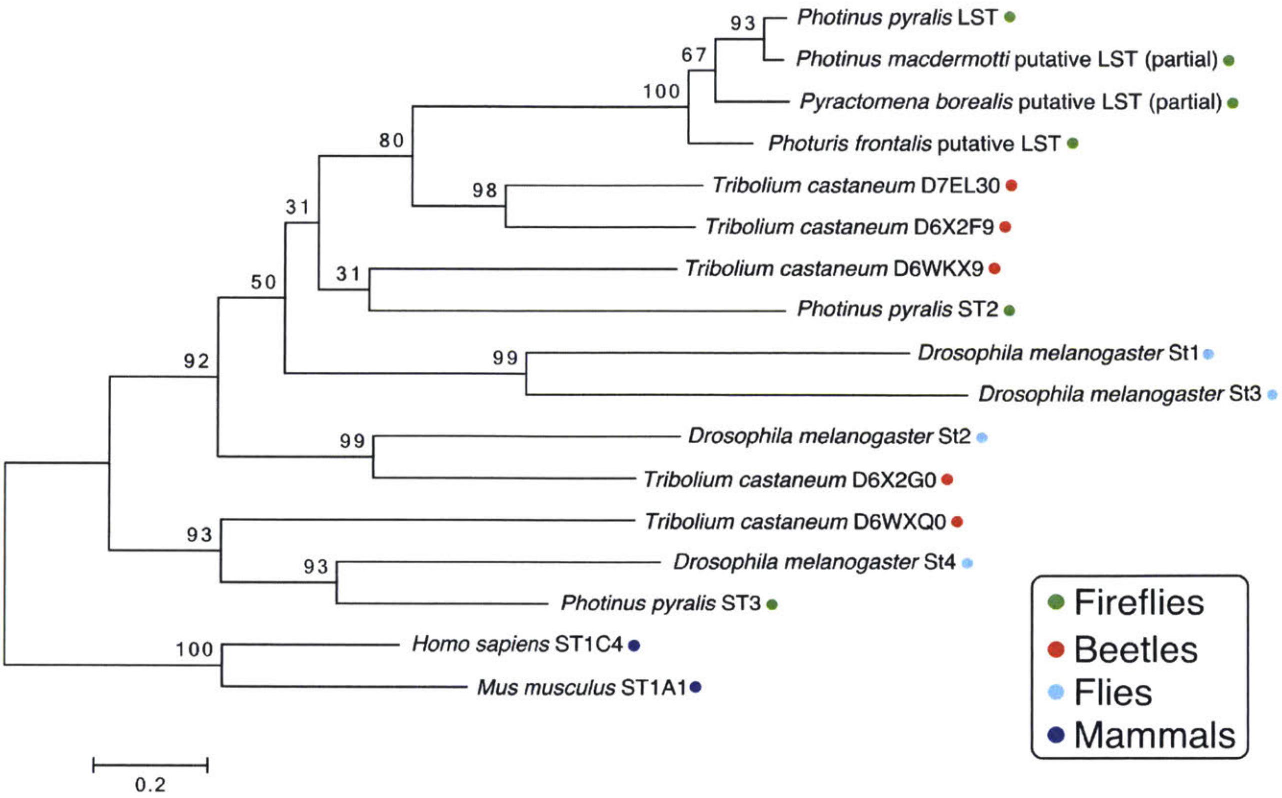

93
67
100
80
98
31
31
50
92
99
99
93
93
100
Photinus pyralis LST ●
Photinus macdermotti putative LST (partial) ●
Pyractomena borealis putative LST (partial) ●
Photuris frontalis putative LST ●
Tribolium castaneum D7EL30 ●
Tribolium castaneum D6X2F9 ●
Tribolium castaneum D6WKX9 ●
Photinus pyralis ST2 ●
Drosophila melanogaster St1 ●
Drosophila melanogaster St3 ●
Drosophila melanogaster St2 ●
Tribolium castaneum D6X2G0 ●
Tribolium castaneum D6WXQ0 ●
Drosophila melanogaster St4 ●
Photinus pyralis ST3 ●
Homo sapiens ST1C4 ●
Mus musculus ST1A1 ●
● Fireflies
● Beetles
● Flies
● Mammals
0.2

Figure S12. Extended maximum likelihood inferred sulfotransferase phylogeny including putative LST ortholog sequences from published firefly lantern transcriptomes (Sander and Hall, 2015), rooted on two mammalian STs as an outgroup. Node labels indicate bootstrap support (5000 replicates). Branch length measures substitutions per site.

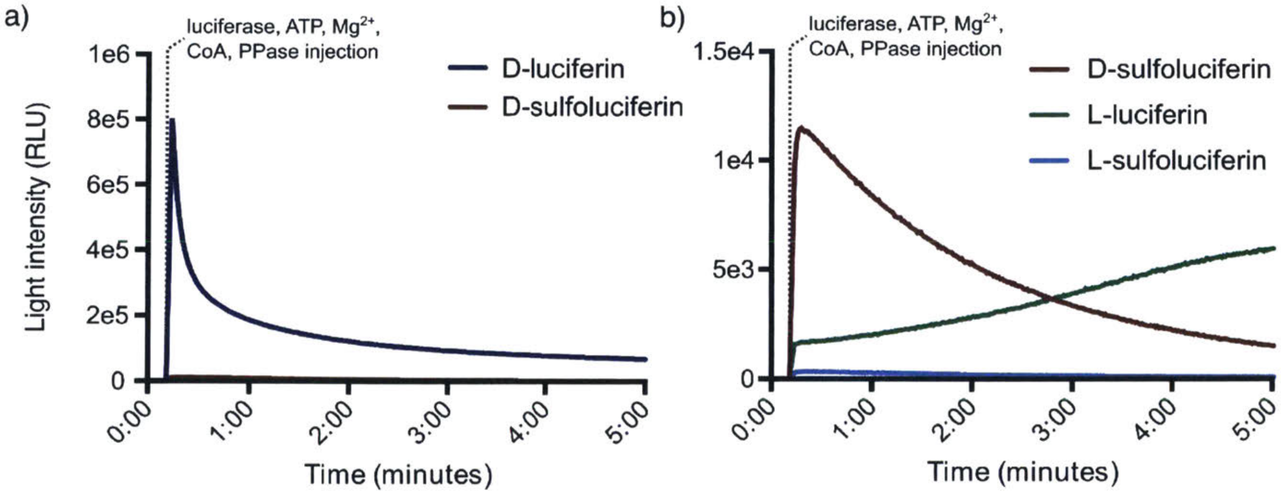


Figure S13. Comparative luminometry of luciferin and sulfoluciferin with *P.pyralis* luciferase indicates sulfoluciferin is not an efficient luminescent substrate for luciferase. Given the difficulty of synthesizing and purifying enantiomerically pure D/L-sulfoluciferin, we utilized LST to synthesize D/L-sulfoluciferin from commercial D/L-luciferin. D-sulfoluciferin and L-sulfoluciferin were synthesized by incubating LST (20 µg/mL), PAPS (500 µM), and D or L-luciferin (100 µM) in luciferase buffer (80 mM HEPES pH

7.3, 150 mM NaCl) for 4 hours. LC-SRM-MS of the sulfoluciferin synthesis reaction after 4 hours indicated near complete conversion of luciferin to sulfoluciferin. D/L-luciferin incubated without LST was included as a control. The reaction mixes were heated to 65˚C for 10 minutes, and filtered through a 4 kDa MWCO ultrafilter to remove LST activity. The reaction mixes were cooled to 25 ˚C, and mixed with a 1:1 injection of luciferase reaction mix consisting of luciferase (50 μg/mL, P/N: SRE0045, Sigma-Aldrich), ATP (2 mM), $MgCl_2$(20 mM), coenzyme A (2 mM, P/N:C4282, Sigma-Aldrich), and pyrophosphatase (0.1 units/μL, P/N: M0361S, NEB) in luciferase buffer. Light output was measured using a Cytation 3 96-well format luminometer with dual reagent injector (BioTek). D-luciferin (Figure S13a), showed approximately 100x the luminescent signal when compared to equimolar D-sulfoluciferin (Figure S13a, S13b). The observed luminescent signal in the presence of D-sulfoluciferin is likely due to residual D-luciferin from the enzymatic synthesis. Oxyluciferin could be detected as a product from these reactions by UV and LC/MS, however an additional peak matching the putative oxidative reaction product sulfooxyluciferin was not detected by either method. Published crystal structures of a luciferyl-AMP analog bound luciferase support our proposal that sulfoluciferin is not a luciferase substrate (Sundlov et al., 2012). In these crystal structures, the hydroxyl group of luciferin is observed to be oriented into the core of luciferase, suggesting the comparatively bulky and charged sulfo group of sulfoluciferin prevents productive enzyme binding. We conclude that sulfoluciferin is not a substrate of *P.pyralis* luciferase.

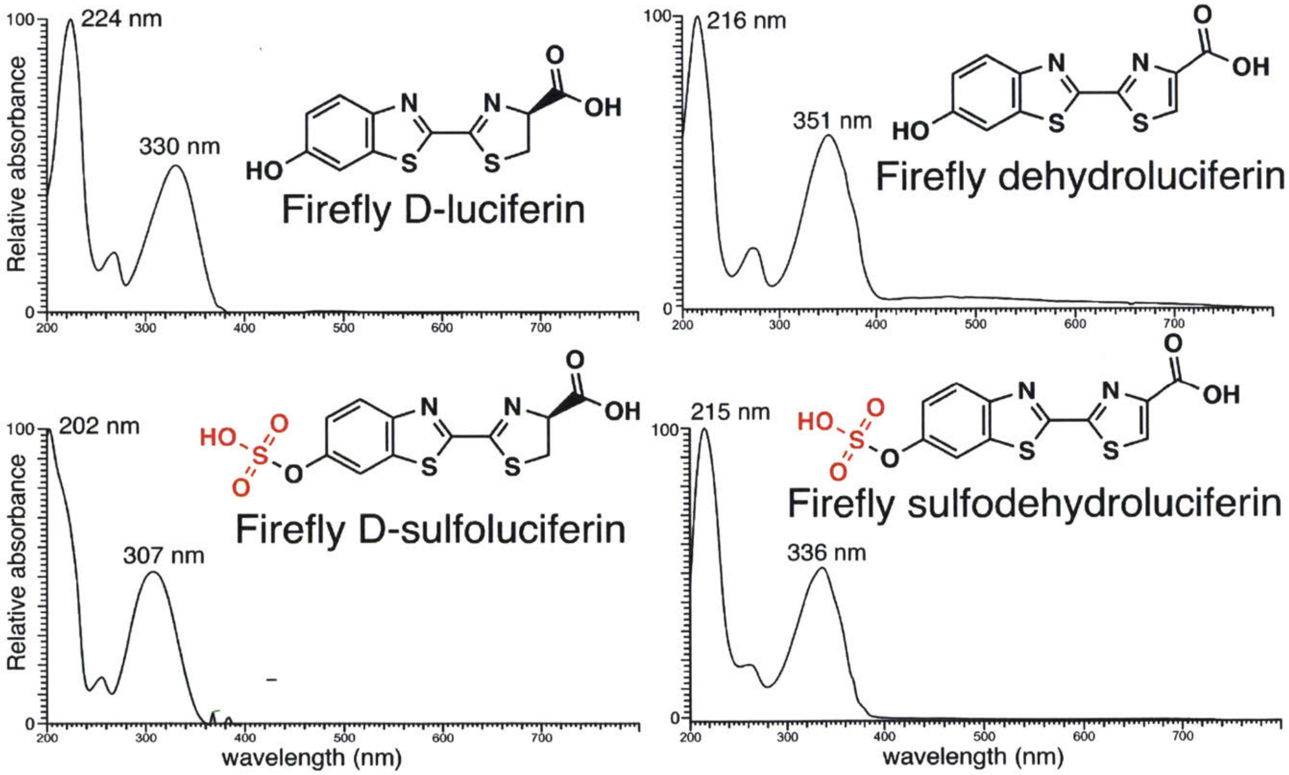


Figure S14. UV-Vis absorption spectra from 200-800 nm for luciferin, dehydroluciferin, sulfoluciferin, and sulfodehydroluciferin.

REFERENCES

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421.

Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak M-Y, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P. 2012. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **30**:918–920.

Cormier MJ, Hori K, Karkhanis YD. 1970. Studies on the bioluminescence of Renilla reniformis. VII. Conversion of luciferin into luciferyl sulfate by luciferin sulfokinase. *Biochemistry* **9**:1184–1189.

de Wet JR, Wood KV, Helinski DR, DeLuca M. 1985. Cloning of firefly luciferase cDNA and the expression of active luciferase in Escherichia coli. *Proc Natl Acad Sci U S A* **82**:7870–7873.

Gomi K, Kajiyama N. 2001. Oxyluciferin, a luminescence product of firefly luciferase, is enzymatically regenerated into luciferin. *J Biol Chem* **276**:36508–36513.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**:644–652.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**:1494–1512.

Hattori K, Motohashi N, Kobayashi I, Tohya T, Oikawa M, Tamura H-O. 2008. Cloning, expression, and characterization of cytosolic sulfotransferase isozymes from Drosophila melanogaster. *Biosci Biotechnol Biochem* **72**:540–547.

Inoue S, Kakoi H, Goto T. 1976. Squid bioluminescence III. Isolation and structure of Watasenia luciferin. *Tetrahedron Lett* **17**:2971–2974.

James MO. 2014. Enzyme kinetics of conjugating enzymes: PAPS sulfotransferase. *Methods Mol Biol* **1113**:187–201.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**:275–282.

Kishi Y, Goto T, Inoue S, Sugiura S, Kishimoto H. 1966. Cypridina bioluminescence III total synthesis of Cypridina luciferin. *Tetrahedron Lett* **7**:3445–3450.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**:R25.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**:323.

Lundin A. 2000. Use of firefly luciferase in ATP-related assays of biomass, enzymes, and metabolites. *Methods Enzymol* **305**:346–370.

Miska W, Geiger R. 1987. Synthesis and Characterization of Luciferin Derivatives for Use in

Bioluminescence Enhanced Enzyme Immunoassays. New Ultrasensitive Detection Systems for Enzyme Immunoassays, I. *Clin Chem Lab Med* **25**:873.

Nakamura H, Kishi Y, Shimomura O, Morse D, Hastings JW. 1989. Structure of dinoflagellate luciferin and its enzymatic and nonenzymatic air-oxidation products. *J Am Chem Soc* **111**:7607–7611.

Nakamura M, Suzuki T, Ishizaka N, Sato J-I, Inouye S. 2014. Identification of 3-enol sulfate of Cypridina luciferin, Cypridina luciferyl sulfate, in the sea-firefly Cypridina (Vargula) hilgendorfii. *Tetrahedron* **70**:2161–2168.

Niwa K, Nakamura M, Ohmiya Y. 2006. Stereoisomeric bio-inversion key to biosynthesis of firefly d-luciferin. *FEBS Lett* **580**:5283–5287.

Oba Y, Yoshida N, Kanie S, Ojika M, Inouye S. 2013. Biosynthesis of firefly luciferin in adult lantern: decarboxylation of L-cysteine is a key step for benzothiazole ring formation in firefly luciferin synthesis. *PLoS One* **8**:e84023.

Ow DW, DE Wet JR, Helinski DR, Howell SH, Wood KV, Deluca M. 1986. Transient and stable expression of the firefly luciferase gene in plant cells and transgenic plants. *Science* **234**:856–859.

Pluskal T, Castillo S, Villar-Briones A, Oresic M. 2010. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**:395.

Priyam A, Woodcroft BJ, Rai V, Munagala A, Moghul I, Ter F, Gibbins MA, Moon H, Leonard G, Rumpf W, Wurm Y. 2015. Sequenceserver: a modern graphical user interface for custom BLAST databases. *bioRxiv*. doi:10.1101/033142

Sander SE, Hall DW. 2015. Variation in opsin genes correlates with signalling ecology in North American fireflies. *Mol Ecol* **24**:4679–4696.

Sekura RD, Jakoby WB. 1979. Phenol sulfotransferases. *J Biol Chem* **254**:5658–5663.

Shimomura O. 2012. Bioluminescence: Chemical Principles and Methods. World Scientific.

Sundlov JA, Fontaine DM, Southworth TL, Branchini BR, Gulick AM. 2012. Crystal structure of firefly luciferase in a second catalytic conformation supports a domain alternation mechanism. *Biochemistry* **51**:6493–6495.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30**:2725–2729.

Weng J-K, Noel JP. 2012. Structure--function analyses of plant type III polyketide synthasesMethods in Enzymology. Elsevier. pp. 317–335.

White EH, McCapra F, Field GF. 1963. The Structure and Synthesis of Firefly Luciferin. *J Am Chem Soc* **85**:337–343.

CHAPTER 3.

Firefly genomes illuminate parallel origins of bioluminescence in beetles

Authors

Timothy R. Fallon[1,2†], Sarah E. Lower[3,4†], Ching-Ho Chang[5], Manabu Bessho-Uehara[6,7,8], Gavin J. Martin[9], Adam J. Bewick[10], Megan Behringer[11], Humberto J. Debat[12], Isaac Wong[5], John C. Day[13], Anton Suvorov[9], Christian J. Silva[5,14], Kathrin F. Stanger-Hall[15], David W. Hall[10], Robert J. Schmitz[10], David R. Nelson[16], Sara M. Lewis[17], Shuji Shigenobu[18], Seth M. Bybee[9], Amanda M. Larracuente[5], Yuichi Oba[6], Jing-Ke Weng[1,2*]

[†]These authors contributed equally to this work

Author Affiliations

[1]Whitehead Institute for Biomedical Research, Cambridge, United States
[2]Department of Biology, Massachusetts Institute of Technology, Cambridge, United States
[3]Department of Molecular Biology and Genetics, Cornell University, Ithaca, United States
[4]Department of Biology, Bucknell University, Lewisburg, United States
[5]Department of Biology, University of Rochester, Rochester, United States
[6]Department of Environmental Biology, Chubu University, Kasugai, Japan
[7]Graduate School of Bioagricultural Sciences, Nagoya University, Nagoya, Japan
[8]Monterey Bay Aquarium Research Institute, Moss Landing, United States
[9]Department of Biology, Brigham Young University, Provo, United States
[10]Department of Genetics, University of Georgia, Athens, United States
[11]Biodesign Center for Mechanisms of Evolution, Arizona State University, Tempe, United States
[12]Center of Agronomic Research, National Institute of Agricultural Technology, Córdoba, Argentina
[13]Centre for Ecology and Hydrology (CEH), Wallingford, United Kingdom
[14]Department of Plant Sciences, University of California Davis, Davis, United States
[15]Department of Plant Biology, University of Georgia, Athens, United States
[16]Department of Microbiology Immunology and Biochemistry, University of Tennessee HSC, Memphis, United States
[17]Department of Biology, Tufts University, Medford, United States
[18]NIBB Core Research Facilities, National Institute for Basic Biology, Okazaki, Japan

Published as

Fallon TR, Lower SE, Chang C-H, Bessho-Uehara M, Martin GJ, Bewick AJ, Behringer M, Debat HJ, Wong I, Day JC, Suvorov A, Silva CJ, Stanger-Hall KF, Hall DW, Schmitz RJ, Nelson DR, Lewis S, Shigenobu S, Bybee SM, Larracuente AM, Oba Y, Weng J-K. 2018. Firefly genomes illuminate parallel origins of bioluminescence in beetles. *eLife*. doi:10.1101/237586

Author contributions:

I produced Figure 1 in collaboration with Sarah Lower, Gavin Martin, Seth Bybee, and Jing-Ke Weng.

I produced Figure 2 in collaboration with Amanda Larracuente, Ivan Liachko (Phase Genomics), Sarah Lower, and Jing-Ke Weng.

I produced Figure 3 in collaboration with Manabu Bessho Uehara, Yuichi Oba, Kathrin Stanger Hall, Sara Lower, and Jing-Ke Weng.

I produced Figure 4 in collaboration with Manabu Bessho Uehara, Yuichi Oba, Kathrin Stanger Hall, Sarah Lower, and Jing-Ke Weng.

I produced Figure 5 in collaboration with Manabu Bessho Uehara, Yuichi Oba, Sarah Lower, and Jing-Ke Weng.

I produced Figure 6 in collaboration with Manabu Bessho Uehara, Yuichi Oba, Sarah Lower, and Jing-Ke Weng.

The supporting information, supporting figures, and supporting tables, were produced with the assistance of the other authors. I edited 100% of the text, and estimate that I wrote and produced about 70% of the text and figures, including re-making certain figures and figures legends provided by collaborators.

ABSTRACT

Fireflies and their luminous courtships have inspired centuries of scientific study. Today firefly luciferase is widely used in biotechnology, but the evolutionary origin of bioluminescence within beetles remains unclear. To shed light on this long-standing question, we sequenced the genomes of two firefly species that diverged over 100 million-years-ago: the North American *Photinus pyralis* and Japanese *Aquatica lateralis*. To compare bioluminescent origins, we also sequenced the genome of a related click beetle, the Caribbean *Ignelater luminosus*, with bioluminescent biochemistry near-identical to fireflies, but anatomically unique light organs, suggesting the intriguing hypothesis of parallel gains of bioluminescence. Our analyses support independent gains of bioluminescence in fireflies and click beetles, and provide new insights into the genes, chemical defenses, and symbionts that evolved alongside their luminous lifestyle.

INTRODUCTION

Fireflies (Coleoptera: Lampyridae) represent the best-studied case of bioluminescence. The coded language of their luminous courtship displays (Figure 1A) has been long studied for its role in mate recognition (Lewis and Cratsley, 2008; Lloyd, 1966; Stanger-Hall and Lloyd, 2015), while non-adult bioluminescence is likely a warning signal of their unpalatable chemical defenses (De Cock and Matthysen, 1999), such as the cardiotoxic lucibufagins of *Photinus* fireflies (Meinwald et al., 1979). The biochemical understanding of firefly luminescence: an ATP, Mg^{2+}, and O_2-dependent luciferase-mediated oxidation of the substrate luciferin (Shimomura, 2012), along with the cloning of the luciferase gene (de Wet et al., 1985; Ow et al., 1986), led to the widespread use of luciferase as a reporter with unique applications in biomedical research and industry (Fraga, 2008). With >2000 species globally, fireflies are undoubtedly the most culturally appreciated bioluminescent group, yet there are at least three other beetle families with bioluminescent species: click beetles (Elateridae), American railroad worms (Phengodidae) and Asian starworms (Rhagophthalmidae) (Martin et al., 2017). These four closely related families (superfamily Elateroidea) have homologous luciferases and structurally identical luciferins (Shimomura, 2012), implying a single origin of beetle bioluminescence. However, as Darwin recognized in his 'Difficulties on Theory' (Charles Darwin, 1872), the light organs amongst the luminous beetle families are clearly distinct (Figure 1B), implying independent origins. Thus, whether beetle bioluminescence is derived from a single or multiple origin(s) remains unresolved.

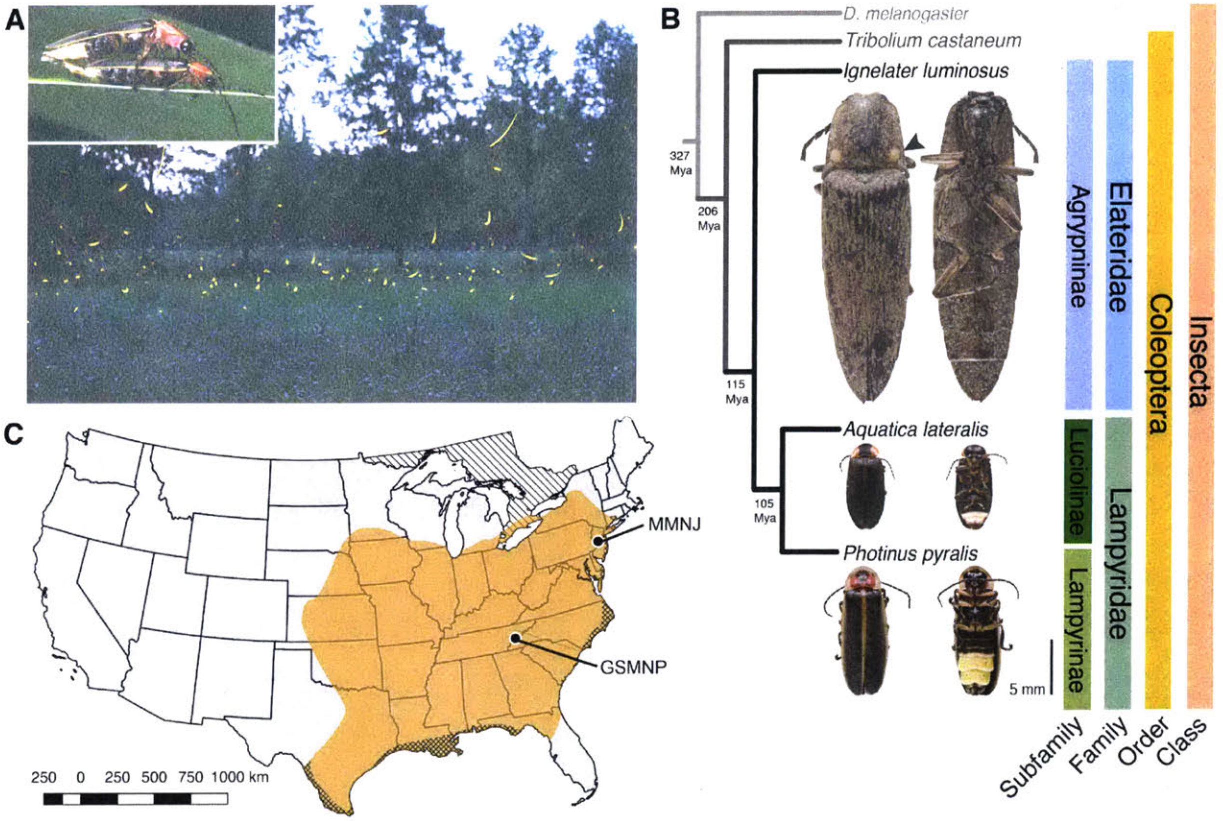


Figure 1: Geographic and phylogenetic context of the Big Dipper firefly, *Photinus pyralis*.
(**A**) *P. pyralis* males emitting their characteristic swooping 'J' patrol flashes over a field in Homer Lake, Illinois. Females cue in on these species-specific flash patterns and respond with their own species-specific flash (Lloyd, 1966). Photo credit: Alex Wild. Inset: male and female *P. pyralis* in early stages of mating. Photo credit: Terry Priest. (**B**) Cladogram depicting the hypothetical phylogenetic relationship between *P. pyralis* and related bioluminescent and non-bioluminescent taxa with *Tribolium castaneum* and *Drosophila melanogaster* as outgroups. Numbers at nodes give approximate dates of divergence in millions of years ago (mya) (McKenna et al., 2015; Misof et al., 2014). Right: Dorsal and ventral photos of adult male specimens. Note the well-developed ventral light organs on the true abdominal segments 6 and 7 of *P. pyralis* and *A. lateralis*. In contrast, the luminescent click beetle, *I. luminosus*, has paired dorsal light organs at the base of its prothorax (arrowhead) and a lantern on the anterior surface of the ventral abdomen (not visible). (**C**) Empirical range of *P. pyralis* in North America, extrapolated from 541 reported sightings (Supporting Information 1.2). Collection sites of individuals used for genome assembly are denoted with circles and location codes. Cross hatches represent areas which likely have *P. pyralis*, but were not sampled. Diagonal hashes represent Ontario, Canada.

To address this long-standing question, we sequenced and analyzed the genomes of three bioluminescent beetle species. To represent the fireflies, we sequenced the widespread North American 'Big Dipper Firefly', *P. pyralis* (Figure 1A,C) and the Japanese 'Heike-botaru' firefly *Aquatica lateralis* (Figure 1B). *P. pyralis* was used in classic studies of firefly bioluminescent biochemistry (Bitler and

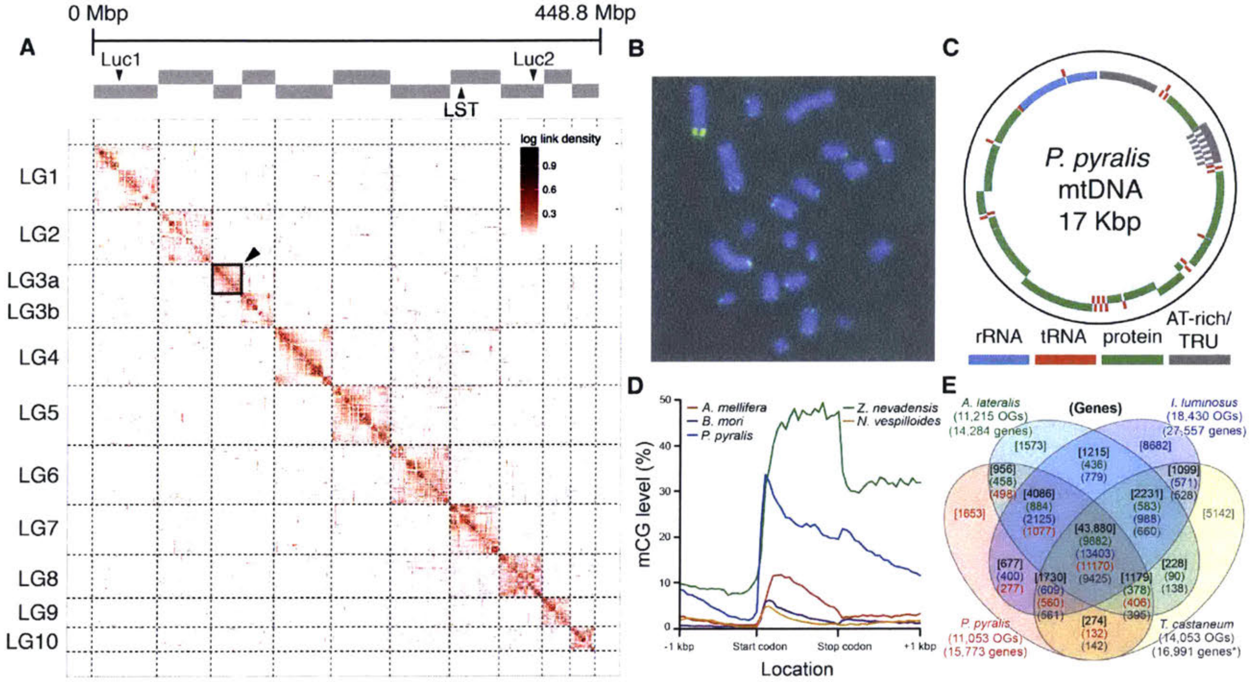
McElroy, 1957), and the cloning of luciferase (de Wet et al., 1985), while *A. lateralis*, a species with specialized aquatic larvae, is one of the few fireflies that can be reliably cultured in the laboratory (Oba et al., 2013a). These two fireflies represent the two major firefly subfamilies, Lampyrinae and Luciolinae, which diverged from a common ancestor over 100 Mya (Figure 1B) (McKenna et al., 2015; Misof et al., 2014). To facilitate evolutionary comparisons, we also sequenced the 'Cucubano', *Ignelater luminosus* (Figure 1B), a Caribbean bioluminescent click beetle, and member of the '*Pyrophorus*' used by Raphaël Dubois (1849-1929) to first establish the enzymatic basis of bioluminescence in the late 1800s (Dubois, 1886, 1885). Comparative analyses of the genomes of these three species allowed us to reconstruct the origin(s) and evolution of beetle bioluminescence.

RESULTS

Sequencing and assembly of firefly and click-beetle genomes

Photinus pyralis adult males were collected from the Great Smoky Mountains National Park, USA (GSMNP) and Mercer Meadows New Jersey, USA (MMNJ) (Figure 1C), and sequenced using short-insert, mate-pair, Hi-C, and long-read Pacific Biosciences (PacBio) approaches (Supporting Information 4—table 1). These datasets were combined in a MaSuRCA (Zimin et al., 2013) hybrid genome assembly (Supporting Information 1.5). The *Aquatica lateralis* genome was derived from an ALL-PATHs (Butler et al., 2008) assembly of short insert and mate-pair reads from a single adult female from a laboratory-reared population, whose lineage, dubbed 'Ikeya-Y90', was first collected 25 years ago from a now extinct population in Yokohama, Japan (Supporting Information 2.5). A single *Ignelater luminosus* adult male, collected in Mayagüez Puerto Rico, USA, was used to produce a high-coverage Supernova (Weisenfeld et al., 2017) linked-read draft genome (Supporting Information 3.5), which was further manually scaffolded using low-coverage long-read Oxford Nanopore MinION sequencing (Supporting Information 3.5.4).

The gene completeness and contiguity statistics of our *P. pyralis* (Ppyr1.3) and *A. lateralis* (Alat1.3) genome assemblies are comparable to the genome of the model beetle *Tribolium castaneum* (Figure 2F; Supporting Information 4.1). The *I. luminosus* genome assembly (Ilumi1.2) is less complete, but is comparable to other published insect genomes (Figure 2F; Supporting Information 4.1). Protein-coding genesets for our study species were produced via an EvidenceModeler-mediated combination of homology alignments, *ab initio* predictions, and *de novo* and reference-guided RNA-seq assemblies followed by manual gene curation for gene families of interest (Supporting Information 1.10; 2.8; 3.8). These coding gene annotation sets for *P. pyralis*, *A. lateralis*, and *I. luminosus* are comprised of 15,773, 14,285, and 27,557 genes containing 94.2%, 90.0%, and 91.8% of the Endopterygota Benchmarking Universal Single-Copy Orthologs (BUSCOs) (Simão et al., 2015), respectively. Protein clustering via predicted orthology indicated 77% of genes were found in orthogroups with at least one other species (Figure 2E; Supporting Information 4—figure 1). We found the greatest orthogroup overlap between the *P. pyralis* and *A. lateralis* genesets, as expected given the more recent phylogenetic divergence of these species. Remaining redundancy in the *P. pyralis* assembly and annotation, as indicated by duplicates of the BUSCOs and the assembly size (Figure 2F; Supporting Information 4—table 2) is likely due to the heterozygosity of the outbred input libraries (Supporting Information 1). The higher BUSCO completeness of the assemblies as compared to the genesets (Supporting Information 4—table 3), suggests that future manual curation efforts will lead to improved annotation completeness.



F

Species	Assembly	Data	Sex	Size** (Mbp)	Assembly size (Mbp)	Scaffolds (#)	NG50 (Mbp)	BUSCO*** C (%)	BUSCO*** D (%)
Photinus pyralis	Ppyr1.3	PacBio + Illumina + Hi-C	M	422 ± 9	471	2160	50.6	97.2	8.4
Aquatica lateralis	Alat1.3	Illumina: short + mate-pair	F	940 ± 1.4	902	7313	0.690	97.4	1.2
Ignelater luminosus	Ilumi1.2	linked-reads + nanopore	M	764 ± 7	845	91324	0.116	94.8	1.4
*Tribolium castaneum**	Tcas5.2	Sanger + BACs + Genetic maps + Illumina + BioNano	M+F	204	166	6580	14.6	98.4	0.5

Figure 2: *Photinus pyralis* genome assembly and analysis.

(**A**) Assembled Ppyr1.3 linkage groups with annotation of the location of known luminescence-related genes, combined with Hi-C linkage density maps. Linkage group 3a (box with black arrow) corresponds to the X chromosome (Supporting Information 1.6.4.1). (**B**) Fluorescence in situ hybridization (FISH) on mitotic chromosomes of a *P. pyralis* larvae. The telomeric repeats TTAGG (green) localize to the ends of chromosomes stained with DAPI (blue). 20 paired chromosomes indicates that this individual was an XX female (Supporting Information 1.13). (**C**) Genome schematic of *P. pyralis* mitochondrial genome (mtDNA). Like other firefly mtDNAs, it has a tandem repetitive unit (TRU) (Supporting Information 1.8). (**D**) mCG is enriched across gene bodies of *P. pyralis* and shows methylation levels that are at least two times higher than other holometabolous insects (Supporting Information 1.12). (**E**) Orthogroup (OGs) clustering analysis of genes with Orthofinder (Emms and Kelly, 2015) shows a high degree of overlap of the *P. pyralis*, *A. lateralis*, and *I. luminosus* genesets with the geneset of *Tribolium castaneum*. Numbers within curved brackets (colored by species) represent gene count from specific species within the shared orthogroups. Numbers with square brackets (black color) represent total gene count amongst shared orthogroups. OGs = orthogroups, *=Not fully filtered to single isoform per gene. See Supporting Information 4.2.1 for more detail. (**F**) Assembly statistics for presented genomes. *=*Tribolium castaneum* model beetle genome assembly (Tribolium Genome Sequencing Consortium et al., 2008) **=Genome size estimated by FC: flow cytometry. *P. pyralis* n = 5 females (SEM) *I. luminosus* n = 5 males (SEM), *A. lateralis* n = 3 technical-replicates of one female (SD). ***=Complete (**C**), and Duplicated (**D**), percentages for the Endopterygota BUSCO (Simão et al., 2015) profile (Supporting Information 1.4, 2.4, 3.4, 4.1).

To enable the characterization of long-range genetic structure, we super-scaffolded the *P. pyralis* genome assembly into 11 pseudo-chromosomal linkage groups using a Hi-C proximity-ligation linkage approach (Figure 2A; Supporting Information 1.5.3). These linkage groups contain 95% of the assembly (448.8 Mbp). Linkage group LG3a corresponds to the X-chromosome based on expected adult XO male read coverage and gene content (Supporting Information 1.6.4.1) and its size (22.2 Mbp) is comparable to the expected X-chromosome size based on sex-specific genome size estimates using flow cytometry (~26 Mbp) (Lower et al., 2017). Homologs to *T. castaneum* X-chromosome genes were enriched on LG3a over every other linkage group, suggesting that the X-chromosomes of these distantly related beetles are homologous, and that their content has been reasonably conserved for >200 MY (Supporting Information 1.6.4.1) (McKenna et al., 2015). We hypothesized that the *P. pyralis* orthologs of known bioluminescence genes, including the canonical luciferase *Luc1* (de Wet et al., 1985) and the specialized luciferin sulfotransferase *LST* (Fallon et al., 2016), would be located on the same linkage group to facilitate chromosomal looping and enhancer assisted co-expression within the light organ. We, however, found these genes on separate linkage groups (Figure 2A).

In addition to nuclear genome assembly and coding gene annotation, we also assembled the complete mitochondrial genomes (mtDNA) of *P. pyralis* (Figure 2C; Supporting Information 1.8) and *I. luminosus* (Supporting Information 3.10), while the mtDNA sequence of *A. lateralis* was recently published (Maeda et al., 2017). These mtDNA assemblies show high conservation of gene content and synteny, with the exception of the variable ~1 Kbp tandem repeat unit (TRU) found in the firefly mtDNAs.

As repetitive elements are common participants and drivers of genome evolution (Feschotte and Pritham, 2007), we next sought to characterize the repeat content of our genome assemblies. Overall, 42.6%, 19.8%, and 34.1% of the *P. pyralis*, *A. lateralis*, and *I. luminosus* assemblies were found to be repetitive, respectively (Supporting Information 1.11; 2.9; 3.9). Of these repeats 66.7%, 39.4%, and 55%

could not be classified as any known repetitive sequence, respectively. Helitrons, DNA transposons that transpose through rolling circle replication (Kapitonov and Jurka, 2001), are among the most abundant individual repeat elements in the *P. pyralis* assembly. Via in situ hybridization, we identified that *P. pyralis* chromosomes have canonical telomeres with telomeric repeats (TTAGG) (Figure 2B; Supporting Information 1.13).

DNA methylation is common in eukaryotes, but varies in degree across insects, especially within Coleoptera (Bewick et al., 2017). Furthermore, the functions of DNA methylation across insects remain obscure (Bewick et al., 2017; Glastad et al., 2017). To examine firefly cytosine methylation, we characterized the methylation status of *P. pyralis* DNA with whole genome bisulfite sequencing (WGBS). Methylation at CpGs (mCG) was unambiguously detected at ~20% within the genic regions of *P. pyralis* and its methylation levels were at least twice those reported from other holometabolous insects (Figure 2D; Supporting Information 1.12). Molecular evolution analyses of the DNA methyltransferases (DNMTs) show that direct orthologs of both DNMT1 and DNMT3 were conserved in *P. pyralis*, *A. lateralis,* and *I. luminosus* (Supporting Information 4—figure 2; Supporting Information 4.2.3), implying that our three study species, and inferentially likely most firefly lineages, possess mCG. Corroborating this claim, *CpG[O/E]* analysis of methylation indicated our three study species had DNA methylation (Supporting Information 4—figure 3).

The genomic context of firefly luciferase evolution

Two luciferase paralogs have been previously described in fireflies (Bessho-Uehara et al., 2017; Oba et al., 2013a). *P. pyralis Luc1* was the first firefly luciferase cloned (de Wet et al., 1985), and its direct orthologs have been widely identified from other fireflies (Oba and Hoffmann, 2014). The luciferase paralog *Luc2* was previously known only from a handful of Asian taxa, including *A. lateralis* (Bessho-Uehara et al., 2017; Oba et al., 2013a). Previous investigations of these Asian taxa have shown that *Luc1* is responsible for light production from the lanterns of adults, larvae, prepupae and pupae, whereas *Luc2* is responsible for the dim glow of eggs, ovaries, prepupae and the whole pupal body (Bessho-Uehara et al., 2017). From our curated genesets (Supporting Information 1.10; 2.8), we unequivocally identified two firefly luciferases, *Luc1* and *Luc2*, in both the *P. pyralis* and *A. lateralis* genomes. Our RNA-Seq data further show that in both *P. pyralis* and *A. lateralis, Luc1* and *Luc2* display expression patterns consistent with previous reports. While *Luc1* is the sole luciferase expressed in the lanterns of both larvae and adults, regardless of sex, *Luc2* is expressed in other tissues and stages, such as eggs (Figure 3C). Notably, *Luc2* expression is detected in RNA libraries derived from adult female bodies (without head or lantern), suggesting detection of ovary expression as described in previous studies (Bessho-Uehara et al., 2017). Together, these results support that since their divergence via gene duplication prior to the divergence of Lampyrinae and Luciolinae, *Luc1* and *Luc2* have established different, but conserved roles in bioluminescence throughout the firefly life cycle.

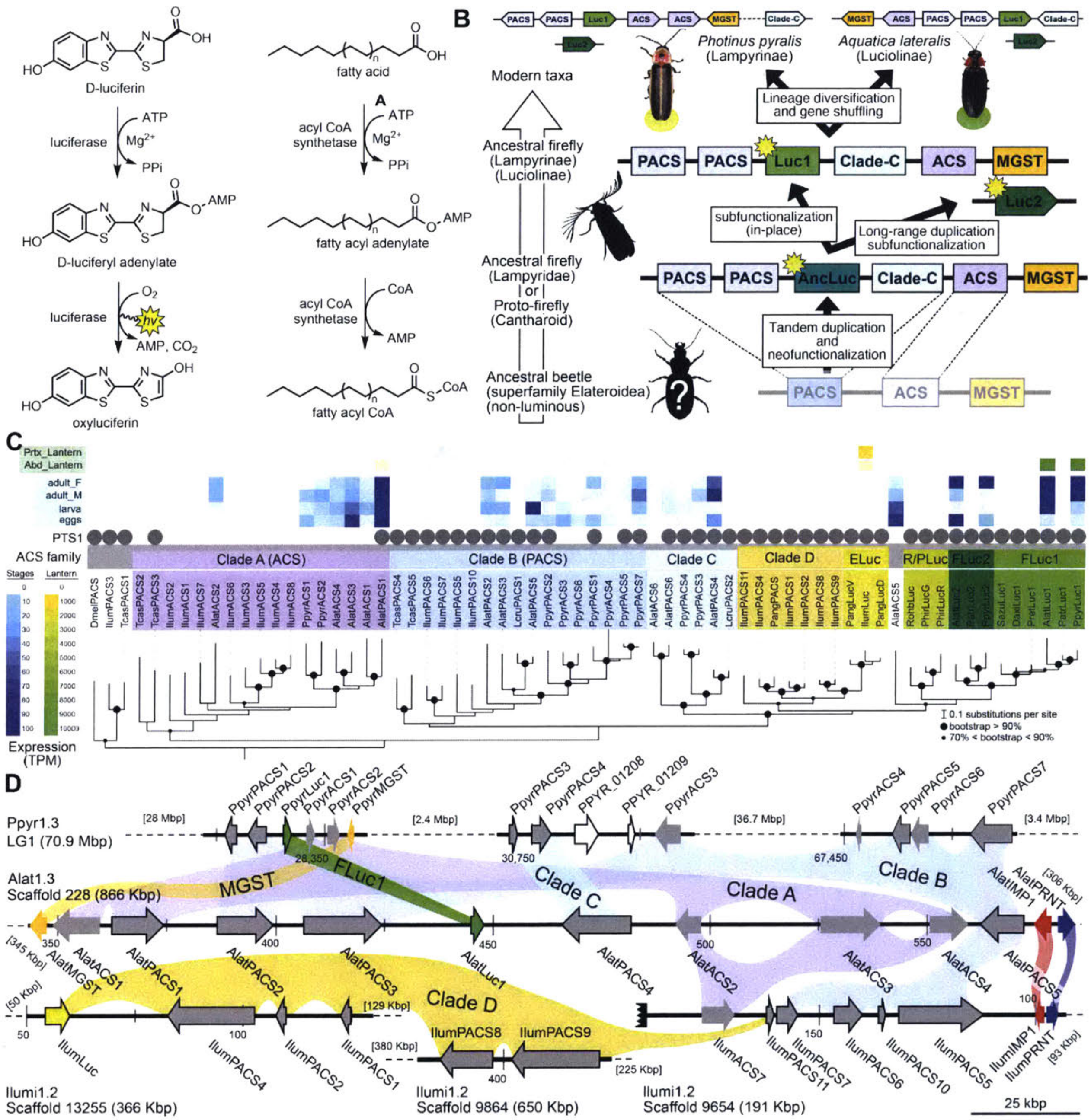


Figure 3: A genomic view of luciferase evolution.

(A) The reaction scheme of firefly luciferase is related to that of fatty acyl-CoA synthetases. **(B)** Model for genomic evolution of firefly luciferases. Ranging from genome structures of luciferase loci in extant fireflies (top), to inferred genomic structures in ancestral species (bottom). Arrow (left) represents ascending time. Not all adjacent genes within the same clade are shown. **(C)** Maximum likelihood tree of luciferase homologs. Grey circles above gene names indicate the presence of peroxisomal targeting signal 1 (PTS1). Color gradients indicate the transcript per million (TPM) values of whole body in each sex/stage (grey to blue) and in the prothorax or abdominal lantern (grey to orange to green). Tree and annotation visualized using iTOL (Letunic and Bork, 2016). Prothorax and abdominal lantern expression values for *I. luminosus* are from whole prothorax plus head, and metathorax plus the two most anterior abdominal segments. Fluc = firefly luciferases, Eluc = elaterid luciferases, R/PLuc =

rhagophthalmid/phengodid luciferases. (Supporting Information 4.3.2) (D) Synteny analysis of beetle luciferase homologs. Nine of the 14 *A. lateralis* PACS/ACS genes closely flank AlatLuc1 on scaffold 228, while 4 of the 13 *P. pyralis* PACS/ACS genes are close neighbors of PpyrLuc1 on LG1, with a further seven genes 2.4 Mbp and 39.1 Mbp away on the same linkage-group. Although the *Luc1* loci in *P. pyralis* and *A. lateralis* are evidently derived from a common ancestor, the relative positions of the most closely related flanking PACS/ACS genes have diverged between the two species. *IlumLuc* was captured on a separate scaffold (Ilumi1.2_Scaffold13255) from its most most closely related PACSs (*IlumPACS8*, *IlumPACS9*) on Ilumi1.2_Scaffold9864, although three more distantly related PACS genes (*IlumiPACS1*, *IlumiPACS2*, *IlumiPACS4*) are co-localized with *IlumLuc*. In contrast, a different scaffold (Ilumi1.2_Scaffold9654) shows orthology to the firefly *Luc1* locus. The full Ilumi1.2_Scaffold13255 was produced by a manual evidence-supported merge of two scaffolds (Supporting Information 3.5.4). Genes with a PTS1 are indicated by a dark outline, except for the genes with white interiors, which instead represent non-PACS/ACS genes without an identified homolog in the other scaffolds. Co-orthologous genes are labeled in the same color in the phylogenetic tree and are connected with corresponding color bands in synteny diagram. Genes and genomic regions are to scale (Scale bar = 25 Kbp). Gaps excluded from the figure are shown with dotted lines and are annotated with their length in square brackets. Scaffold ends are shown with rough black bars. MGST = Microsomal glutathione S-transferase, IMP = Inositol monophosphatase, PRNT = Polyribonucleotide nucleotidyltransferase. Figure produced with GenomeTools 'sketch' (v1.5.9) (Gremme et al., 2013).

Firefly luciferase is hypothesized to be derived from an ancestral peroxisomal fatty acyl-CoA synthetase (PACS) (Figure 3A) (Oba et al., 2006, 2003). We found that, in both firefly species, *Luc1* is genomically clustered with its closely related homologs, including PACSs and non-peroxisomal acyl-CoA synthetases (ACSs), enzymes which can be distinguished by the presence/absence of a C-terminal peroxisomal-targeting-signal-1 (PTS1). We also found nearby microsomal glutathione S-transferase (MGST) family genes (Figure 3D) that are directly orthologous between both species, Genome-wide phylogenetic analysis of the luciferases, PACSs and ACSs genes indicates that *Luc1* and *Luc2* form two orthologous groups, and that the neighboring PACS and ACS genes near *Luc1* form three major clades (Figure 3C): Clade A, whose common ancestor and most extant members are ACSs, and Clades B and C whose common ancestors and most extant members are PACSs. *Luc1* and *Luc2* are highly conserved at the level of gene structure—both are composed of seven exons with completely conserved exon/intron boundaries (Supporting Information 4—figure 4; Supporting Information 4—figure 5), and most members of Clades A, B, and C also have seven exons. The exact syntenic and orthology relationships of the ACS and PACS genes adjacent to the *Luc1* locus remains unclear, likely due to subsequent gene divergence and shuffling (Figure 3C,D).

Luc2 is located on a different linkage-group from *Luc1* in *P. pyralis* and on a different scaffold from *Luc1* in *A. lateralis,* consistent with the interpretation that *Luc1* and *Luc2* lie on different chromosomes in both firefly species. No PACS or ACS genes were found in the vicinity of *Luc2* in either species. These data support that tandem gene duplication in a firefly ancestor gave rise to several ancestral PACS paralogs, one of which neofunctionalized in place to become the ancestral luciferase (*AncLuc*) (Figure 3B). Prior to the divergence of the firefly subfamilies Lampyrinae and Luciolinae around 100 Mya (Supporting Information 4.3), this *AncLuc* duplicated, possibly via a long-range gene duplication event (e.g. transposon mobilization), and then subfunctionalized in its transcript expression pattern to give rise to *Luc2*, while the original *AncLuc* subfunctionalized in place to give rise to Luc1 (Figure 3B). From the shared *Luc* gene clustering in both fireflies, we infer the structure of the pre Luc1/Luc2 duplication *AncLuc* locus contained one or more ACS genes (Clade A), one or more PACS genes (Clade B/C), and one or more MGST family genes (Figure 3B).

Independent origins of firefly and click beetle luciferase

To resolve the number of origins of luciferase activity, and therefore bioluminescence, between fireflies and click beetles, we first identified the luciferase of *I. luminosus* luciferase (*IlumLuc*), and compared its genomic context to the luciferases of *P. pyralis* and *A. lateralis* (Figure 3D). Unlike some other described bioluminescent Elateridae, which have separate luciferases expressed in the dorsal prothorax and ventral abdominal lanterns (Oba et al., 2010a), we identified only a single luciferase in the *I. luminosus* genome which was highly expressed in both of the lanterns (Figure 3C; Supporting Information 3.8). The exon number and exon-intron splice junctions of *IlumLuc* are identical to those of firefly luciferases, but unlike the firefly luciferases which have short introns less than <100 bp long, *IlumLuc* has two long introns (Supporting Information 4—figure 4). We found several PACS genes in the

I. luminosus genome which were related to *IlumLuc* and formed a clade (Clade D) specific to the Elateridae (Figure 3C,D). *IlumLuc* lies on a 366 Kbp scaffold containing 18 other genes, including three related Clade D PACS genes (Scaffold 13255; Figure 3D; Figure 4); however, the Clade D genes that are most closely related to *IlumLuc* are found on a separate 650 Kbp scaffold (Scaffold 9864; Figure 3D). We infer that the *IlumLuc* locus is not orthologous to the extant firefly *Luc1*locus, as *IlumLuc* is not physically clustered with Clade A, B or C ACS or PACS genes (Figure 3C,D). We instead identified a different scaffold in *I. luminosus* that is likely orthologous to the firefly *Luc1* locus (Scaffold 9654; Figure 3D). This assessment is based on the presence of adjacent Clade A and B ACS and PACS genes, as well as orthologous exoribonuclease family (PRNT) and inositol monophosphatase family (IMP) genes, both of which were found adjacent to the *A. lateralis Luc1* locus, but not the *P. pyralis Luc1* locus (Figure 3D). Interestingly, *IlumPACS11*, the most early-diverging member of Clade D, was also found on Scaffold 9654 (Figure 3D). This finding is consistent with an expansion of Clade D following duplication of the *IlumPACS11* syntenic ancestor to a distant site. Overall, these genomic structures are consistent with independent origins of firefly and click beetle luciferases.

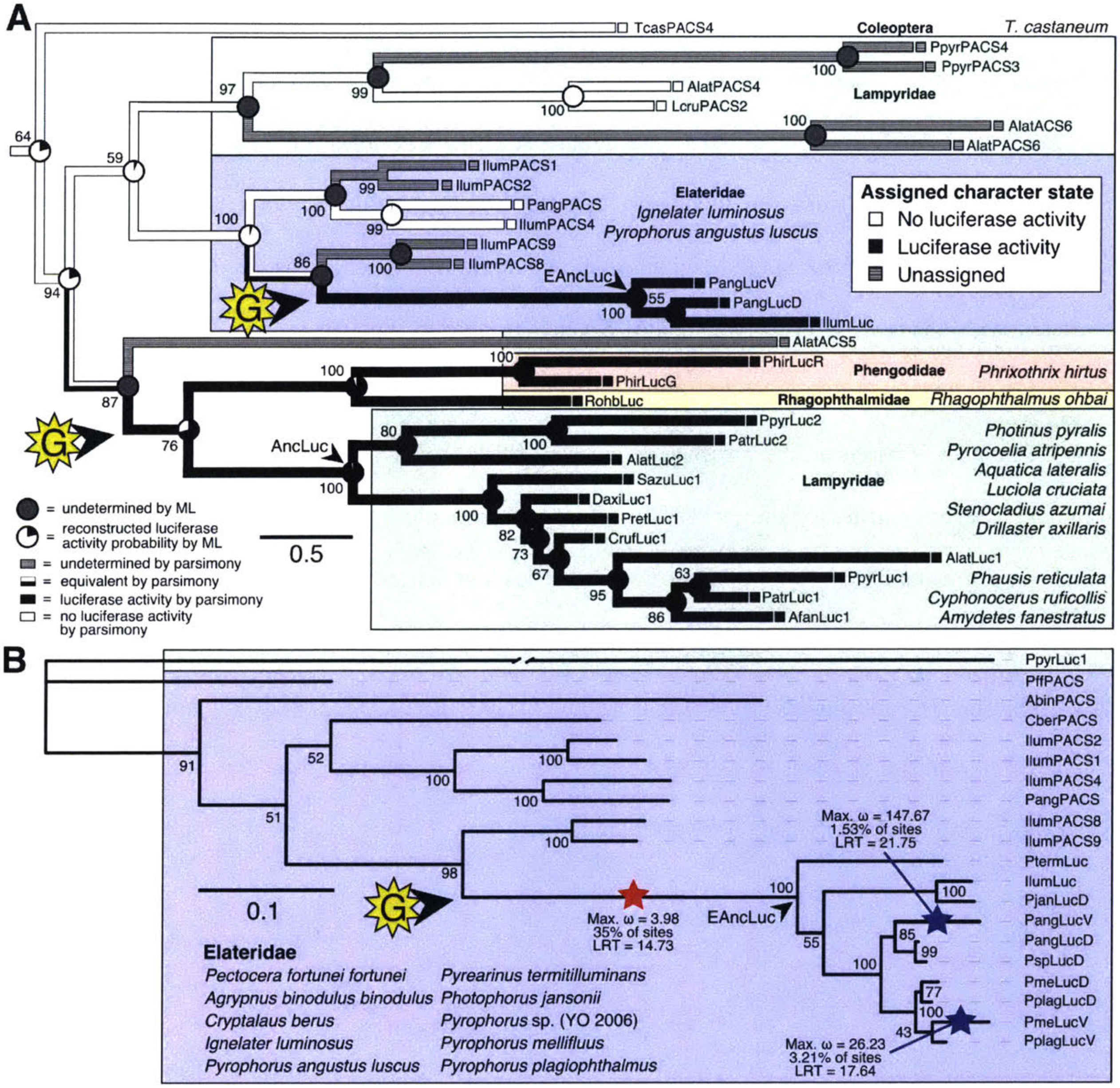


Figure 4: Parallel evolution of elaterid and firefly luciferase.
(**A**) Ancestral state reconstruction recovers at least two gains of luciferase activity in bioluminescent beetles. Luciferase activity (top right figure key; black: luciferase activity, white: no luciferase activity, shaded: undetermined) was annotated on extant firefly luciferase homologs via literature review or inference via direct orthology. The ancestral states of luciferase activity within the putative ancestral nodes were then reconstructed with an unordered parsimony framework and a maximum likelihood (ML) framework (bottom left figure key; Supporting Information 4.3.3). Two gains ('G') of luciferase activity, annotated with black arrows and yellow stars, are hypothesized. These hypothesized gains occurred once in a gene within the common ancestor of fireflies, rhagophthalmid, and phengodid beetles, and once in a gene within the common ancestor of bioluminescent elaterid beetles. Scale bar is substitutions per site. Numbers adjacent to nodes represents node support. (**B**) Molecular adaptation analysis supports independent neofunctionalization of click beetle luciferase. We tested the molecular adaptation of elaterid luciferase using the adaptive branch-site REL test for episodic diversification (aBSREL) method (Smith et

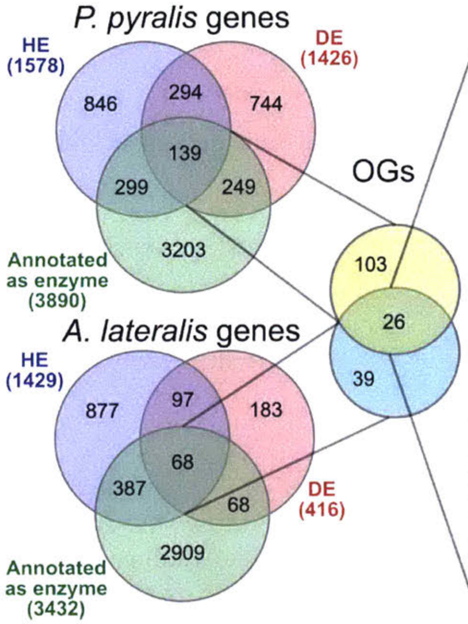
al., 2015) (Supporting Information 4.3.4). The branch leading to the common ancestor of elaterid luciferases (red star) was one of three branches (red and blue stars) recovered with significant ($p<0.01$) evidence of positive selection, with 35% of sites showing strong directional selection (ω or max dN/dS = 3.98), which we interpret as signal of the initial neofunctionalization of elaterid ancestral luciferase (EAncLuc) from an ancestor without luciferase activity. As the selected branches with blue stars are red-shifted elaterid luciferases (Oba et al., 2010a; Stolz et al., 2003), they may represent the post-neofunctionalization selection of a few key sites via sexual selection of emission colors. Specific sites identified as under selection using Mixed Effect Model of Evolution (MEME) and Phylogenetic Analysis by Maximum Likelihood (PAML) methods are described in Supporting Information 4.3.4. The tree and results from the full adaptive model are shown. Branch length, with the exception of the PpyrLuc1 branch which was shortened, reflects the number of substitutions per site. Numbers adjacent to nodes represents node support. Figure was produced with iTOL (Letunic and Bork, 2016).

We then carried out targeted molecular evolution analyses including the known beetle luciferases and their closely related homologs. Ancestral state reconstruction of luminescent activity on the gene tree using Mesquite (Maddison and Maddison, 2017) recovered two independent gains of luminescence as the most parsimonious and likely scenario: once in click beetles, and once in the common ancestor of firefly, phengodid, and rhagophthalmid beetles (Figure 4A; Supporting Information 4.3.3). In an independent molecular adaptation analysis utilizing the coding nucleotide sequence of the elaterid luciferases and their close homologs within Elateridae, 35% of the sites of the branch leading to the ancestral click beetle luciferase showed a statistically significant signal of episodic positive selection with $dN/dS > 1$ (ω or max $dN/dS = 3.98$) as compared to the evolution of its paralogs using the aBSREL branch-site selection test (Smith et al., 2015) (Figure 4B; Supporting Information 4.3.4). This implies that the common ancestor of the click beetle luciferases (*EAncLuc*) underwent a period of accelerated directional evolution. As the branch under selection in the molecular adaptation analysis (Figure 4B) is the same branch of luciferase activity gain via ancestral reconstruction (Figure 4A), we conclude that the identified selection signal represents the relatively recent neofunctionalization of click beetle luciferase from a non-luminous ancestral Clade D PACS gene, distinct from the more ancient neofunctionalization of firefly luciferase. Based on the constraints from our tree, we determine that this neofunctionalization of *EAncLuc* occured after the divergence of the elaterid subfamily Agrypninae. In contrast, we cannot determine if the original neofunctionalization of *AncLuc* occurred in the ancestral firefly, or at some point during the evolution of 'cantharoid' beetles, an unofficial group of beetles including the luminous Rhagophthalmidae, Phengodidae and Lampyridae among other non-luminous groups, but not the Elateridae (Branham and Wenzel, 2003). There is evidence for a subsequent luciferase duplication event in phengodids, but not in rhagophthalmids, that is independent of the duplication event that gave rise to *Luc1* and *Luc2* in fireflies (Figures 3C and 4). Altogether, our results strongly support the independent neofunctionalization of luciferase activity in click beetles and fireflies, and therefore at least two independent gains of luciferin-utilizing luminescence in beetles.

Metabolic adaptation of the firefly lantern

Beyond luciferase, we sought to characterize other metabolic traits which might have co-evolved in fireflies to support bioluminescence. Of particular importance, the enzymes of the *de novo* biosynthetic pathway for firefly luciferin remain unknown (Oba et al., 2013b). We hypothesized that bioluminescent accessory enzymes, either specialized enzymes with unique functions in luciferin metabolism or enzymes with primary metabolic functions relevant to bioluminescence, would be highly expressed (HE: 90th percentile; Supporting Information 4.2.2) in the adult lantern, and would be differentially expressed (DE; Supporting Information 4.2.2) between luminescent and non-luminescent tissues. To determine this, we performed RNA-Seq and expression analysis of the dissected *P. pyralis* and *A. lateralis* adult male lantern tissue compared with a non-luminescent tissue (Supporting Information 4.2.2). We identified a set of predicted orthologous enzyme-encoding genes conserved in both *P. pyralis* and *A. lateralis* that met our HE and DE criteria (Figure 5). Both luciferase and luciferin sulfotransferase (LST), a specialized enzyme recently implicated in luciferin storage in *P. pyralis* (Fallon et al., 2016), were recovered as candidate genes using these four criteria (HE, DE, enzymes, direct orthology across species), confirming the validity of our approach. While a direct ortholog of LST is present in *A. lateralis*, it is absent from *I. luminosus*, suggesting that LST, and the presumed luciferin storage it mediates, is an exclusive ancestral firefly or cantharoid trait. This finding is consistent with previous hypotheses of the absence of LST in

Elateridae (Fallon et al., 2016), and with the overall hypothesis of independent evolution of bioluminescence between the Lampyridae and Elateridae.



P.pyralis ID (OGS1.1)	Predicted function	Ppyr expression rank	Ppyr BSN-TPM	Ppyr PTS1	Orthogroup ID	Alat PTS1	Alat BSN-TPM	Alat expression rank	*A. lateralis* ID (OGS1.0)
PPYR_00001	Luciferase*	2	66743	PTS1	OG0000057	PTS1	36044	1	AQULA_005067
PPYR_11147	Cystathionine gamma-lyase	3	38574		OG0002087		18096	3	AQULA_003032
PPYR_04899	Short chain dehydrogenase	4	28506	PTS1	OG0000476	PTS1	9452	9	AQULA_008573
PPYR_09320	Saccharopine dehydrogenase-like	6	17516	PTS1	OG0000161	PTS1	6355	12	AQULA_012956
PPYR_06194	Alpha/beta hydrolase	7	13554		OG0009024		850	161	AQULA_013805
PPYR_02512	Histidine Triad superfamily	8	11131		OG0005956		5575	13	AQULA_008871
PPYR_00996	Strictosidine synthase-like	12	4870		OG0002066		2529	35	AQULA_002761
PPYR_08432	Adenylate kinase	13	4726		OG0005480	PTS1	3619	22	AQULA_007407
PPYR_08520	Methionine-R-sulfoxide reductase	18	3946		OG0005974		2293	44	AQULA_008914
PPYR_08058	Acetyl-CoA hydrolase/transferase	21	3629		OG0003529		4981	17	AQULA_000701
PPYR_00003	Luciferin sulfotransferase*	25	3167	PTS1	**OG0000054**		2366	43	AQULA_012700
""	""	""	""	""	""		2843	32	AQULA_004004
PPYR_14844	Malic oxidoreductase-like	55	1570	PTS1	**OG0000619**	PTS1	2441	41	AQULA_005495
PPYR_06564	Malic oxidoreductase-like	569	212		""	""	""	""	""
PPYR_04459	ABC transporter	75	1229		**OG0000018**		647	223	AQULA_002548
PPYR_08864	ABC transporter	1119	118		""		""	""	""
PPYR_09240	CoA transferase	76	1210	PTS1	OG0003901	PTS1	690	203	AQULA_001958
PPYR_06879	Metallo-beta-lactamase	79	1200		OG0004565		1880	51	AQULA_004381
PPYR_11151	Enolase	103	926		OG0007981		370	380	AQULA_003033
PPYR_01504	Alpha/beta hydrolase	155	675		OG0000078	PTS1	904	148	AQULA_012908
PPYR_10210	Methionine-S-sulfoxide reductase	174	637		OG0005026		640	227	AQULA_005939
PPYR_14372	Adenylyl-sulfate kinase & sulfate adenylyltransferase	214	537	PTS1	OG0000698	PTS1	4300	19	AQULA_001585
PPYR_05464	Peroxiredoxin	251	474		OG0000556		1434	72	AQULA_013952
PPYR_06980	Cytochrome P450	405	307		OG0000593		251	543	AQULA_002673
PPYR_10578	Short chain dehydrogenase	419	300		OG0004118		412	335	AQULA_002715
PPYR_09779	3'5'-cyclic nucleotide phosphodiesterase	442	286		OG0007963		104	1258	AQULA_002893
PPYR_01821	ABC transporter	478	259		OG0000018		242	566	AQULA_007404
PPYR_12812	Fatty acid hydroxylase	538	228		OG0000864		718	194	AQULA_001837
PPYR_01505	Alpha/Beta hydrolase	664	188		OG0000078		101	1287	AQULA_012915
PPYR_01858	Enoyl-CoA hydratase/isomerase	674	187		OG0002807		652	221	AQULA_010152
PPYR_05219	DD-peptidase superfamily	1526	87		OG0004630		309	448	AQULA_004580

Figure 5: Comparative analyses of firefly lantern expression highlight likely metabolic adaptations to bioluminescence.
Enzymes which are highly expressed (HE), differentially expressed (DE), and annotated as enzymes via InterProScan are shown in the Venn diagrams for their respective species. Those genes in the intersection of the two sets which are within the same orthogroup (OGs) as determined by OrthoFinder are shown in the table. Many-to-one orthology relationships are represented by bold orthogroups and blank cells. See Supporting Information 4.2.2 for more detail. *=genes of previously described function. Underlying expression quantification and Venn analysis available on FigShare: (DOI: 10.6084/m9.figshare.5715151)

Moreover, we identified several additional enzyme-encoding HE and DE lantern genes that are likely important in firefly lantern physiology (Figure 5). For instance, adenylate kinase likely plays a critical role in efficient recycling of AMP post-luminescence, and cystathionine gamma-lyase supports a key role of cysteine in luciferin biosynthesis (Oba et al., 2013b) and recycling (Okada et al., 1974). We also detected a combined adenylyl-sulfate kinase and sulfate adenylyltransferase enzyme (*ASKSA*) among the lantern-enriched gene list (Supporting Information 4—figure 8), implicating active biosynthesis of 3'-phosphoadenosine-5'-phosphosulfate (PAPS), the cofactor of LST, in the lantern. This finding highlights the importance of LST-catalyzed luciferin sulfonation for bioluminescence. These firefly orthologs of *ASKSA* are the only members amongst their paralogs to contain a PTS1 (Supporting Information 4—figure 8), suggesting specialized localization to the peroxisome, the location of the luminescence reaction. This suggests that the levels of sulfoluciferin and luciferin may be actively regulated within the peroxisome of lantern cells in response to luminescence. Overall, our findings of several directly orthologous enzymes that share expression patterns in the light organs of both *P. pyralis* and *A. lateralis* suggests that the enzymatic physiology and/or the gene expression patterns of the photocytes were already fixed in the Luciolinae-Lampyrinae ancestor.

We also performed a similar expression analysis for genes not annotated as enzymes, yielding several genes with predicted lysosomal function (Supporting Information 4—table 6; Supporting Information 4.4). This suggests that the abundant but as yet unidentified 'differentiated zone granule' organelles of the firefly light organ (Ghiradella and Schmidt, 2004) could be lysosomes. Interestingly, we found a HE (TPM value ~300) and DE opsin, *Rh7*, in the light organ of *A. lateralis*, but not *P. pyralis* (Supporting Information 4—figure 9; Supporting Information 4.5), suggesting a potential light perception role for *Rh7* in the *A. lateralis* lantern, akin to the light perception role described for *Drosophila Rh7* (Ni et al., 2017).

Genomic insights into firefly chemical defense

Firefly bioluminescence is postulated to have first evolved as an aposematic warning of larval chemical defenses (Branham and Wenzel, 2003). Lucibufagins are abundant unpalatable defense steroids described from certain North American firefly species, most notably in the genera *Photinus* (Meinwald et al., 1979), *Lucidota* (Gronquist et al., 2005), and *Ellychnia* (Smedley et al., 2017), and hence are candidates for ancestral firefly defense compounds. To test whether lucibufagins are widespread among bioluminescent beetles, we assessed the presence of lucibufagins in *P. pyralis*, *A. lateralis,* and *I. luminosus* by liquid-chromatography high-resolution accurate-mass mass-spectrometry (LC-HRAM-MS). While lucibufagins were found in high abundance in *P. pyralis* adult hemolymph, they were not observed in *A. lateralis* adult hemolymph, nor in *I. luminosus* metathorax extract (Figure 6B; Supporting Information 4.6). Since chemical defense is presumably most critical in the long-lived larval stage, we next tested whether lucibufagins are present in all firefly larvae even if they are not present in the adults of certain species. We found lucibufagins in *P. pyralis* larval extracts; however, they were not observed in *A. lateralis* larval extracts (Figure 6B; Supporting Information 4.6). Together, these results suggest that the lucibufagin biosynthetic pathway is either a derived trait only found in particular firefly taxa (e.g. subfamily: Lampyrinae), or that lucibufagin biosynthesis was an ancestral trait that was lost in *A. lateralis*. Consistent with the former hypothesis, the presence of lucibufagins in non-North-American Lampyrinae has been previously reported (Tyler et al., 2008), but to date there are no reports of lucibufagins in the Luciolinae.

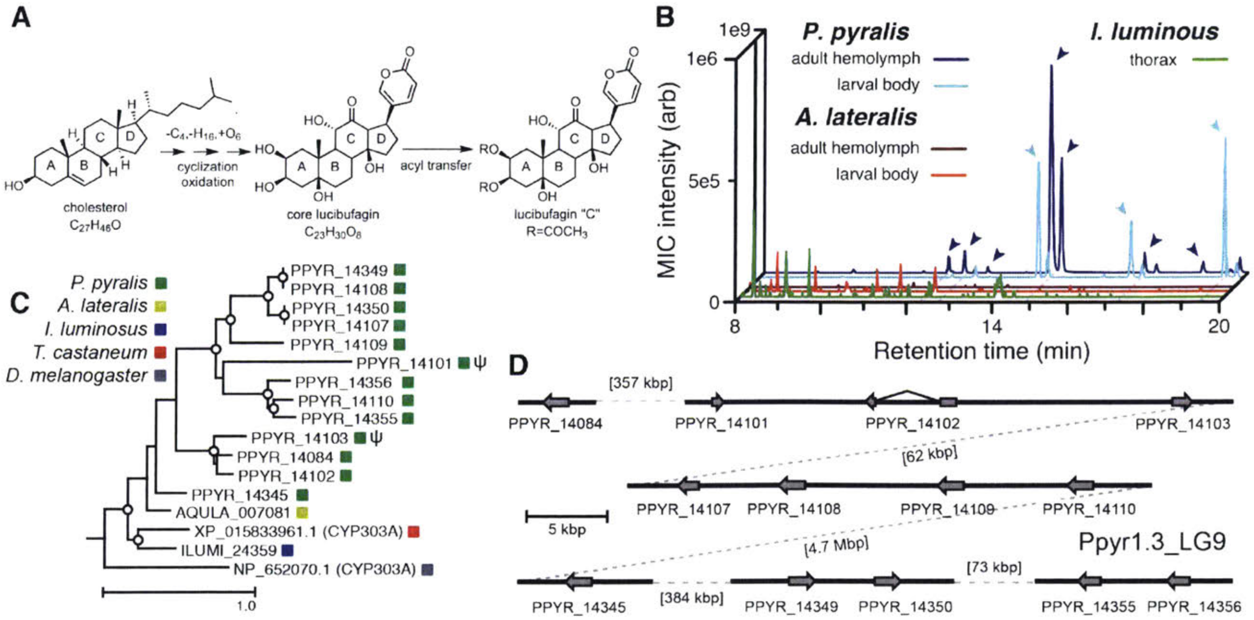


Figure 6: An expansion in the CYP303-P450 family correlates with lucibufagin content.
(**A**) Hypothesized lucibufagin biosynthetic pathway, starting from cholesterol. (**B**) LC-HRAM-MS multi-ion-chromatograms (MIC) showing the summation of exact mass traces for the $[M + H]^+$ of 11 lucibufagin chemical formulas ± 5 ppm, calibrated for run-specific systematic *m/z* error (Supporting Information 4—table 9). Y-axis upper limit for *P. pyralis* adult hemolymph and larval body extract is 1000x larger than other traces. Arrows (blue/teal) indicate features with high MS2 spectral similarity to known lucibufagins. Sporadic peaks in *A. lateralis* body, and *I. luminosus* thorax traces are not abundant, preventing MS2 spectral acquisition and comparison, but do not match the *m/z* and RT of *P. pyralis* lucibufagins (Supporting Information 4.6). (**C**) Maximum likelihood tree of CYP303 family cytochrome P450 enzymes from *P. pyralis*, *A. lateralis*, *T. castaneum*, and *D. melanogaster. P. pyralis* shows a unique CYP303 family expansion, whereas the other species only have a single CYP303. Circles represent node bootstrap support >60%. Branch length measures substitutions per site. Pseudogenes are annotated with the greek letter Ψ (Supporting Information 1.10.1; 4.2.4). (**D**) Genomic loci for *P. pyralis* CYP303 family genes. These genes are found in multiple gene clusters on LG9, supporting origin via tandem duplication. Introns >4 kbp are shown.

The lucibufagin biosynthetic pathway is currently unknown. However, their chemical structure suggests a biosynthetic origin from cholesterol followed by a series of hydroxylations, -OH acetylations, and the side-chain oxidative pyrone formation (Figure 6A) (Meinwald et al., 1979). We hypothesized that cytochrome P450s, an enzyme family widely involved in metabolic diversification of organic substrates (Hamberger Björn and Bak Søren, 2013), could underlie several oxidative reactions in the proposed lucibufagin biosynthetic pathway. We therefore inferred the P450 phylogeny among our three bioluminescent beetle genomes to identify any lineage-specific genes correlated with lucibufagin presence. Our analysis revealed a unique expansion of one P450 family, the CYP303 family, in *P. pyralis*. While 94/97 of currently sequenced winged-insect genomes on OrthoDB (Zdobnov et al., 2017), as well as the *A. lateralis* and *I. luminosus* genomes, contain only a single *CYP303* family gene, the *P. pyralis* genome contains 11 *CYP303* genes and two pseudogenes (Figure 6C), which expanded via tandem duplication on the same linkage group (Figure 6D). The CYP303 ortholog of *D. melanogaster*, CYP303A1, has been shown to play a role in mechanosensory bristle development (Willingham and Keil, 2004). Although the exact biochemical function and substrate of *D. melanogaster* CYP303A1 is unknown, its closely related P450 families operate on an insect steroid hormone ecdysone (Willingham and Keil, 2004). As ecdysone and lucibufagins are structurally similar, CYP303 may operate on steroid-like compounds. Therefore, the lineage-specific expansion of the CYP303 family in *P. pyralis* is a compelling candidate in the metabolic evolution of lucibufagins as chemical defenses associated with the aposematic role of bioluminescence. Alternatively, this CYP303 expansion in *P. pyralis* may be associated with other lineage-specific chemical traits, such as pheromone production.

Symbionts of bioluminescent beetles

Given the increasingly recognized contributions of symbionts to host metabolism (Newman and Cragg, 2015), we characterized the hologenome of all three beetles as potential contributors to metabolic processes related to bioluminescence. Whole genome sequencing of our wild-caught and laboratory reared fireflies revealed a rich microbiome. Amongst our firefly genomes, we found various bacterial genomes, viral genomes, and the complete mtDNA for a phorid parasitoid fly, *Apocephalus antennatus*, the first mtDNA reported for genus *Apocephalus*. This mtDNA was inadvertently included in the *P. pyralis* PacBio library via undetected parasitization of the initial specimens, and was assembled via a metagenomic approach (Supporting Information 5.2). Independent collection of *A. antennatus* which emerged from field-collected *P. pyralis* adults and targeted COI sequencing later confirmed the taxonomic origin of this mtDNA (Supporting Information 5.3). We also sequenced and metagenomically assembled the complete circular genome (1.29 Mbp, GC: 29.7%; ~50x coverage) for a *P. pyralis*-associated mollicute (Phylum: Tenericutes), *Entomoplasma luminosum* subsp. pyralis (Supporting Information 5.1). *Entomoplasma* spp. were first isolated from the guts of North American fireflies (Hackett et al., 1992) and our assembly provides the first complete genomic assembly of any *Entomoplasma* species. Broad read coverage for the *E. luminosus* subsp. pyralis genome was detected in 5/6 of our *P. pyralis* DNA libraries, suggesting that *Entomplasma* is a highly prevalent, possibly vertically inherited, *P. pyralis* symbiont. It has been hypothesized that these *Entomoplasma* mollicutes could play a role in firefly metabolism, specifically via contributing to cholesterol metabolism and lucibufagin biosynthesis (Smedley et al., 2017).

Within our unfiltered *A. lateralis* genomic assembly (Alat1.2), we also found 43 scaffolds (2.3 Mbp; GC:29.8%, ~64x coverage), whose taxonomic annotation corresponded to the Tenericutes (Supporting Information 2.5.2), suggesting that *A. lateralis* may also harbor a mollicute symbiont. Alat1.2 also contains 2119 scaffolds (13.0 Mbp, GC:63.7%, ~25x coverage) annotated as of Proteobacterial origin. Limited Proteobacterial symbionts were detected in the *I. luminosus* assembly (0.4 Mbp; GC:30–65% ~10x coverage) (Supporting Information 3.5.2), suggesting no stable symbiont is present in adult *I. luminosus*. Lastly, we detected two species of novel orthomyxoviridae-like ssRNA viruses, which we dub *Photinus pyralis* orthomyxo-like virus 1 and 2 (PpyrOMLV1/2), that were highly prevalent across our *P. pyralis* RNA-Seq datasets, and showed multi-generational transovarial transmission in the laboratory (Supporting Information 5.4). We also found several endogenous viral elements (EVEs) for PpyrOMLV1/2 in *P. pyralis* (Supporting Information 5.5). These viruses are the first reported in any firefly species, and represent only the second report of transgenerational transfer of any *Orthomyxoviridae* virus (Marshall et al., 2014), and the second report of *Orthomyxoviridae* derived EVEs (Katzourakis and

Gifford, 2010). Together, these genomes from the firefly holobiont provide valuable resources for the continued inquiry of the symbiotic associates of fireflies and their biological and ecological significance.

DISCUSSION

Here, we generated genome assemblies, diverse tissue and life-stage RNA-Seq data, and LC/MS data for three evolutionarily informative and historically well-studied bioluminescent beetles, and used a series of comparative analyses to illuminate long-standing questions on the origins and evolution of beetle bioluminescence. By analyzing the genomic synteny and molecular evolution of the beetle luciferases and their extant and inferred-ancestral homologs, we found strong support for the independent origins of luciferase, and therefore bioluminescence, between fireflies and click beetles. Our approaches and analyses lend molecular evidence to the previous morphology-phylogeny based hypotheses of parallel gain proposed by Darwin and others (Bocakova et al., 2007; Branham and Wenzel, 2003; Charles Darwin, 1872; Costa, 1975; Day, 2013; Oba, 2009; Sagegami-Oba et al., 2007). While our elaterid luciferase selection analysis strongly supports an independent gain, we did not perform an analogous selection analysis of luciferase homologs across all bioluminescent beetles, due to the lack of genomic data from key related beetle families. Additional genomic information from early-diverged firefly lineages, other luminous beetle taxa (e.g. Phengodidae and Rhagophthalmidae), and non-luminous elateroid taxa (e.g. Cantharidae and Lycidae), will be useful to further develop and test models of luciferase evolution, including the hypothesis that bioluminescence also originated independently in the Phengodidae and/or Rhagophthalmidae. As some phylogenetic relationships of fireflies and other lineages of superfamily Elateroidea remain uncertain, continued efforts to produce reference phylogeny for these taxa are required (Bocak et al., 2018; Martin et al., 2017). Toward this goal, the recently published *Pyrocoelia pectoralis*

Lampyrinae firefly genome is an important advance which will contribute to future phylogenetic and evolutionary studies (Fu et al., 2017).

The independent origins of the firefly and click beetle luciferases provide an exemplary natural model system to understand enzyme evolution through parallel mutational trajectories and the evolution of complex metabolic traits generally. The abundance of gene duplication events of PACSs and ACSs at the ancestral luciferase locus in both fireflies and *I. luminosus* suggests that ancestral promiscuous enzymatic activities served as raw materials for the selection of new adaptive catalytic functions (Weng, 2014). But while parallel evolution of luciferase implies evolutionary independence of bioluminescence overall, the reality may be more complex, and the other subtraits of bioluminescence amongst the bioluminescent beetles likely possess different evolutionary histories from luciferase. While subtraits presumably dependent on an efficient luciferase, such as specialized tissues and neural control, almost certainly arose well after luciferase specialization, and thus can be inferred to also have independent origins between fireflies and click beetles, luciferin, which was presumably a prerequisite to luciferase neofunctionalization, may have been present in their common ancestor. Microbial endosymbionts, such as the tenericutes detected in our *P. pyralis* and *A. lateralis* datasets, are intriguing candidate contributors to luciferin metabolism and biosynthesis. Alternatively, recent reports have shown that firefly luciferin is readily produced non-enzymatically by mixing benzoquinone and cysteine (Kanie et al., 2016), and that a compound resulting from the spontaneous coupling of benzoquinone and cysteine acts as a luciferin biosynthetic intermediate in *A. lateralis* (Kanie et al., 2018). Benzoquinone is known to be a defense compound of distantly related beetles (Dettner, 1987) and other arthropods (e.g. millipedes) (Shear, 2015). Therefore, the evolutionary role of sporadic low-level luciferin synthesis through spontaneous chemical reactions, either in the ancestral bioluminescent taxa themselves, or in non-bioluminescent taxa, and dietary acquisition of luciferin by either the ancestral or modern bioluminescent taxa, should be considered. To decipher between these alternative evolutionary possibilities, the discovery of genes involved in luciferin metabolism in fireflies and other bioluminescent beetles will be essential. Here, as a first step toward that goal, we identified conserved, enriched and highly expressed enzymes of the firefly lantern that are strong candidates in luciferin metabolism and the elusive luciferin *de novo* biosynthetic pathway. Ultimately focused experimentation will be needed to decipher the biochemical function of these enzymes.

The early evolution of firefly bioluminescence was likely associated with an aposematic role. The adaptive light production of the primordial firefly (or alternatively, a primordial bioluminescent cantharoid beetle) that enabled the selection and neofunctionalization of luciferase was perhaps linked to a response to predators by a primitive whole-body oxygen-gated luminescence, where a startle-response mediated increase in hemolymph oxygenation through spiracle opening and escape locomotion caused a concomitant increase in luminescence (Buck and Case, 2002; Case, 2004). Alternatively, an early role for firefly luminescence in mate attraction has not been ruled out (Buck and Case, 2002). The presence of particular unpalatable defense compounds in all extant fireflies would be consistent with an ancestral role and the former hypothesis, and the chemical analysis of tissues across species and life stages presented in this work provides new insights into the evolutionary occurrence of lucibufagins, the most well-studied defense compounds associated with fireflies. Our results reject lucibufagins as ancestral defense compounds of fireflies, but rather suggest them as a derived metabolic trait associated with Lampyrinae. Additional chemical analyses across more lineages of fireflies are needed, however, to further support or falsify this hypothesis. Toward this goal, the high sensitivity of our LC-HRAM-MS and MS^2 molecular networking-based lucibufagin identification approach is particularly well suited to broadened sampling in the future, including those of rare taxa and possibly museum specimens. Combined with genomic data showing a concomitant expansion of the CYP303 gene family in *P. pyralis*, we present a promising path toward elucidating the biosynthetic mechanism underlying these potent firefly toxins.

Overall, the resources and analyses generated in this study shed valuable light on the evolutionary questions Darwin first pondered, and will enable future studies of the ecology, behavior, and evolution of bioluminescent beetles. These resources will also accelerate the discovery of new enzymes from bioluminescent beetles that could enhance biotechnological applications of bioluminescence. Finally, we hope that the genomic resources shared here will facilitate the development of effective population genomic tools to monitor and protect wild bioluminescent beetle populations in the face of changing climate and habitats.

MATERIALS AND METHODS

Detailed materials and methods are available in the Supporting Information sections. Methods relating to *P. pyralis* are given in Supporting Information 1, while methods relating to *A. lateralis* and *I. luminosus* are given in Supporting Information 2 and Supporting Information 3, respectively. Methods for comparative genomic analyses are given in Supporting Information 4, while methods for microbiome characterization are given in Supporting Information 5. References to relevant sections of the Supporting Information are placed in-line throughout the maintext.

DATA AND MATERIALS AVAILABILITY

Genomic assemblies (Ppyr1.3, Alat1.3, and Ilumi1.2), associated official geneset data, a SequenceServer (Priyam et al., 2015) BLAST server, and a JBrowse (Skinner et al., 2009) genome browser are available at www.fireflybase.org. Raw genomic and RNA-Seq reads for *P. pyralis*, *A. lateralis*, and *I. luminosus*, are available under the NCBI/EBI/DDBJ BioProjects PRJNA378805, PRJDB6460, and PRJNA418169 respectively. Raw WGBS reads can be found on the NCBI Gene Expression Omnibus (GSE107177). Mitochondrial genomes for *P. pyralis* and *I. luminosus* and *A. antennatus* are available on NCBI GenBank with accessions KY778696, MG242621, and MG546669. The complete genome of *Entomoplasma luminosum* subsp. pyralis is available on NCBI GenBank with accession CP027019. The viral genomes for *Photinus pyralis* orthomyxo-like virus 1 and 2 are available on NCBI Genbank with accessions MG972985-MG972994. LC-MS data is available on MetaboLights (Accession MTBLS698). Other supporting datasets are available on FigShare (Supporting Information 6.1).

Supporting Information 1

Photinus pyralis additional information

1.1 Taxonomy, biology, and life history

Photinus pyralis (Linnaeus, 1767) is amongst the most widespread and abundant of all U.S. fireflies (Lloyd, 2008, 1966). It inspired extensive work on the biochemistry and physiology of firefly bioluminescence in the early 20th century, and the first luciferase gene was cloned from this species (de Wet et al., 1985). A habitat generalist, *P. pyralis* occurs in fields, meadows, suburban lawns, forests, and woodland edges, and even urban environments. For example, the authors have observed *P. pyralis* flashing in urban New York City and Washington D.C. Adults rest on vegetation during the day and signaling begins as early as 20 min before sunset (Lloyd, 1966). Male flashing is cued by ambient light levels, thus shaded or unshaded habitats can show up to a 30 min difference in the initiation of male flashing (Lloyd, 1966). Males can be cued to flash outside of true twilight if exposed to light intensities simulating twilight (Case, 2004). *P. pyralis* were also reported to flash during totality of the total solar eclipse of 2017 (Personal communication: L.F. Faust, M.A. Branham). Courtship activity lasts for 30–45 min and both sexes participate in a bioluminescent flash dialog, as is typical for *Photinus* fireflies.

Males initiate courtship by flying low above the ground while repeating a single ~300 ms patrol flash at ~5–10 s intervals (Case, 2004). Males emit their patrol flash while dipping down and then ascending vertically, creating a distinctive J-shaped flash gesture (Case, 2004; Lloyd, 1966) (Figure 1A). During courtship, females perch on vegetation and respond to a male patrol flash by twisting their abdomen toward the source of the flash and giving a single response flash given after a 2–3 s delay. Receptive females will readily respond to simulated male flashes, such as those produced by an investigator's penlight. Females have fully developed wings and are capable of flight. Both sexes are capable of mating several times during their adult lives. During mating, males transfer to females a fitness-enhancing nuptial gift consisting of a spermatophore manufactured by multiple accessory glands (Reijden et al., 1997); the molecular composition of this nuptial gift has recently been elucidated for *P.*

pyralis (Al-Wathiqui et al., 2016). In other *Photinus* species, male gift size decreases across sequential matings (Cratsley et al., 2003), and multiple matings are associated with increased female fecundity (Rooney and Lewis, 2002).

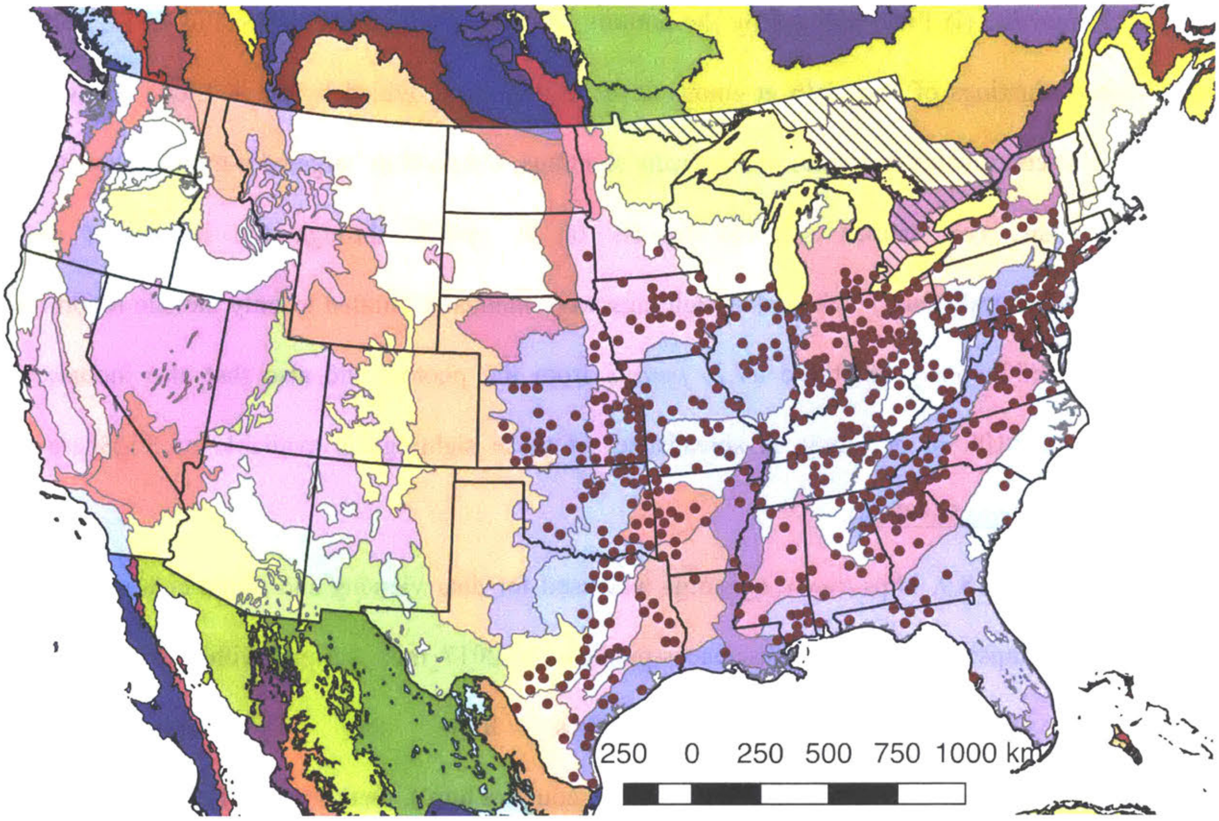
Adult *P. pyralis* live 2–3 weeks, and although these adults are typically considered non-feeding, both sexes have been reported drinking nectar from the flowers of the milkweed *Asclepias syriaca* (Faust and Faust, 2014). Mated females store sperm and lay ~30–50 eggs over the course of a few days on moss or in moist soil. The eggs take 2–3 weeks to hatch. Larval bioluminescence is thought to be universal for the Lampyridae, where it appears to function as an aposematic warning signal. Like other *Photinus*, *P. pyralis* larvae are predatory, live on and beneath the soil, and appear to be earthworm specialists (Hess, 1920). In the northern parts of its range, slower development likely requires *P. pyralis* to overwinter at least twice, most likely as larvae. Farther south, *P. pyralis* may complete development within several months, achieving two generations per year (Faust, 2017), which may be possibly be observed in the South as a 'second wave' of signalling *P. pyralis* in September-October.

Anti-predator chemical defenses of male *P. pyralis* include several bufadienolides, known as lucibufagins, that circulate in the hemolymph (Meinwald et al., 1979). Pterins have also been reported to be abundant in *P. pyralis* (Goetz et al., 1981); however, the potential defense role of these compounds has never been tested (Personal communication: J. Meinwald). When attacked, *P. pyralis* males release copious amounts of rapidly coagulating hemolymph and such 'reflex-bleeding' may also provide physical protection against small predators (Blum and Sannasi, 1974; Faust et al., 2012).

1.2 Species distribution

Although *Photinus pyralis* is widely distributed in the Eastern United States, published descriptions of its range are limited, with the notable exception of Lloyd's 1966 monograph (Lloyd, 1966) which addresses the range of many *Photinus* species. We therefore sought to characterize the current distribution of *P. pyralis* in order to produce an updated map to inform our experimental design and enable future population genetic studies. Four sources of data were used to produce the presented range map of *P. pyralis*: (i) Field surveys by the authors (ii) Published (Lloyd, 1966; Luk et al., 2011) and unpublished sightings of *P. pyralis* at county level resolution, provided by Dr. J. Lloyd (University of Florida), (iii) coordinates and dates of *P. pyralis* sightings, obtained by targeted e-mail surveys to firefly field biologists, (iv) citizen scientist reports of *P. pyralis* through the iNaturalist platform (https://www.inaturalist.org/). iNaturalist sightings were manually curated to only include reports which could be unambiguously identified as *P. pyralis* from the photos, and also that also included GPS geotagging to <100 m accuracy. A spreadsheet of these sightings is available on FigShare (DOI: 10.6084/m9.figshare.5688826).

QGIS (v2.18.9, https://www.qgis.org) was used for data viewing and figure creation. A custom Python script (https://github.com/elifesciences-publications/2017_misc_scripts) within QGIS was used to link *P. pyralis* sightings to counties from the US census shapefile (https://www.census.gov/geo/maps-data/data/cbf/cbf_counties.html). Outlying points that were located in Desert Ecoregions of the World Wildlife Fund (WWF) Terrestrial Ecoregions shapefile (Olson et al., 2001; World Wildlife Fund, 2017) or the westernmost edge of the range were manually removed, as they are likely isolated populations not representative of the contiguous range. For Figure 1B, these points were converted to a polygonal range map using the 'Concave hull' QGIS plugin ('nearest neighbors = 19') followed by smoothing with the Generalizer QGIS plugin with Chaiken's algorithm (Level = 10, and Weight = 3.00). Below (Supporting Information 1—figure 1), red circles indicate county-centroided presence records.



Supporting Information 1—figure 1. Detailed geographic distribution map for *P. pyralis*.
P. pyralis sightings (red circles show county centroided reports) in the United States and Ontario, Canada (diagonal hashes). The World Wildlife Fund Terrestrial Ecoregions (Olson et al., 2001; World Wildlife Fund, 2017) are also shown (colored shapes). The *P. pyralis* sighting dataset shown is identical to that used to prepare Figure 1B.

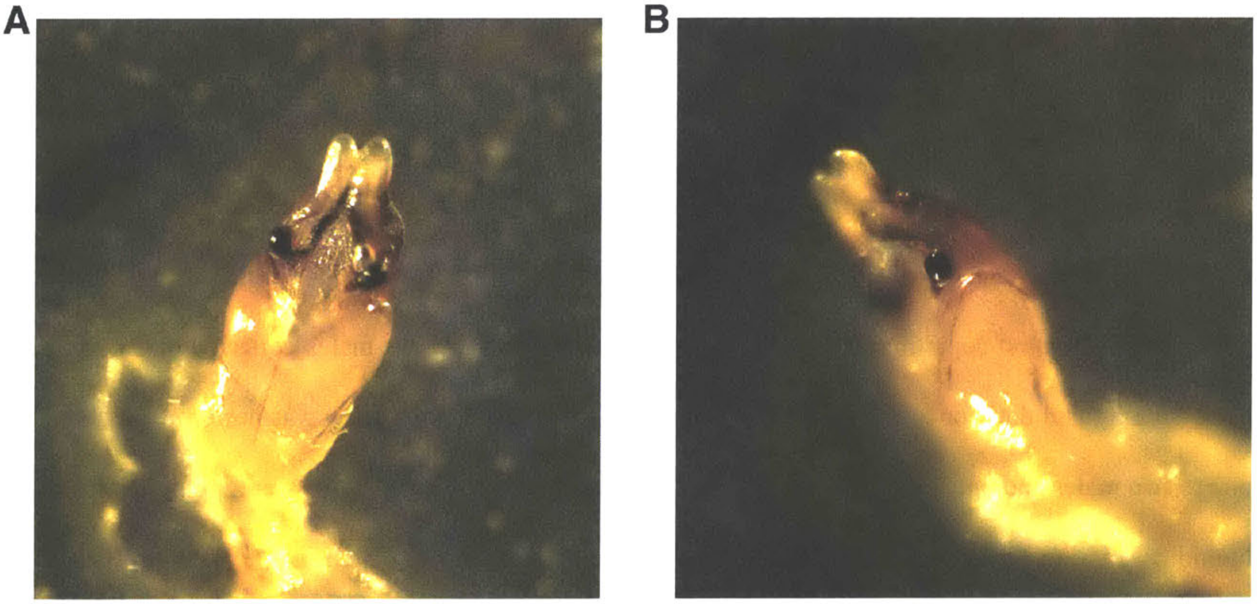
In our field surveys, we found that the range of *P. pyralis* was notably extended from the range reported by Lloyd, specifically we found *P. pyralis* in abundance to the west of the Mill river in Connecticut. *P. pyralis* is found with confidence roughly from Connecticut to Texas, and possibly as far south as Guatemala (Personal communication: A. Catalán). These possible southern populations require further study.

1.3 Specimen collection and identification

Adult male *P. pyralis* specimens for Illumina short-insert and mate-pair sequencing were collected at sunset on June 13th, 2011 near the Visitor's Center at Great Smoky Mountains National Park (permit to Dr. Kathrin Stanger-Hall). Specimens were identified to species and sex via morphology

(Green, 1956), flash pattern and behavior (Lloyd, 1966), and *cytochrome-oxidase I* (*COI*) similarity (partial sequence: primers HCO, LCO (Stanger-Hall and Lloyd, 2015)) when blasted against an in-house database of firefly *COI* nucleotide sequences. Collected fireflies were stored in 95% ethanol at −80°C until DNA extraction.

Adult male *P. pyralis* specimens for Pacific Biosciences (PacBio) RSII sequencing were captured during flight at sunset on June 9th, 2016, from Mercer Meadows in Lawrenceville, NJ (40.3065 N 74.74831 W), on the basis of the characteristic 'rising J' flash pattern of *P. pyralis* (permit to TRF via Mercer County Parks Commission). Collected fireflies were sorted, briefly checked to be likely *P. pyralis* by the presence of the margin of ventral unpigmented abdominal tissue anterior to the lanterns, flash frozen with liquid N2, lyophilized, and stored at −80°C until DNA extraction. A single aedeagus (male genitalia) was dissected from the stored specimens and confirmed to match the *P. pyralis* taxonomic key (Green, 1956) (Supporting Information 1—figure 2).



Supporting Information 1—figure 2. *P. pyralis* aedeagus (male genitalia).
(**A**) Ventral and (**B**) side view of a *P. pyralis* aedeagus dissected from specimens collected on the same date and locality as those used for PacBio sequencing. Note the strongly sclerotized paired ventro-basal processes ('mickey mouse ears') emerging from the median process, characteristic of *P. pyralis* (Green, 1956).

1.3.2 Collection and rearing of P. pyralis larvae

We intended to survey the lucibufagin content of *P. pyralis* larvae (Figure 6B; Supporting Information 4.6), and as well as the transovarial transmission of Photinus pyralis orthomyxo-like viruses from parent to larvae (Supporting Information 5.4), but as *P. pyralis*larvae are subterranean and extremely difficult to collect from the wild, we reared *P. pyralis* larvae from eggs laid from mated pairs. It is important to note that these *P. pyralis* larval rearing experiments were unexpectedly successful. Although there has been some success in laboratory rearing and domestication of Asian *Aquatica* spp. (Ho et al., 2010), including the *A. lateralis* Ikeya-Y90 strain described in this manuscript, rearing of North American fireflies is considered extremely difficult with numerous unpublished failures for unclear reasons (Lloyd, 1996), and limited reports of successful rearing of mostly non-Photinus genera, including *Photuris* sp. (McLean et al., 1972), *Pyractomena* angulata (Buschman, 1988), and *Pyractomena borealis* (Personal communication: Scott Smedley). The below protocol for *P. pyralis* larval rearing is presented in the context of disclosure of the methods of this manuscript, and should be considered a preliminary, unoptimized rearing protocol. A full description of the *P. pyralis* larvae and it's life history and behavior will be presented in a separate manuscript.

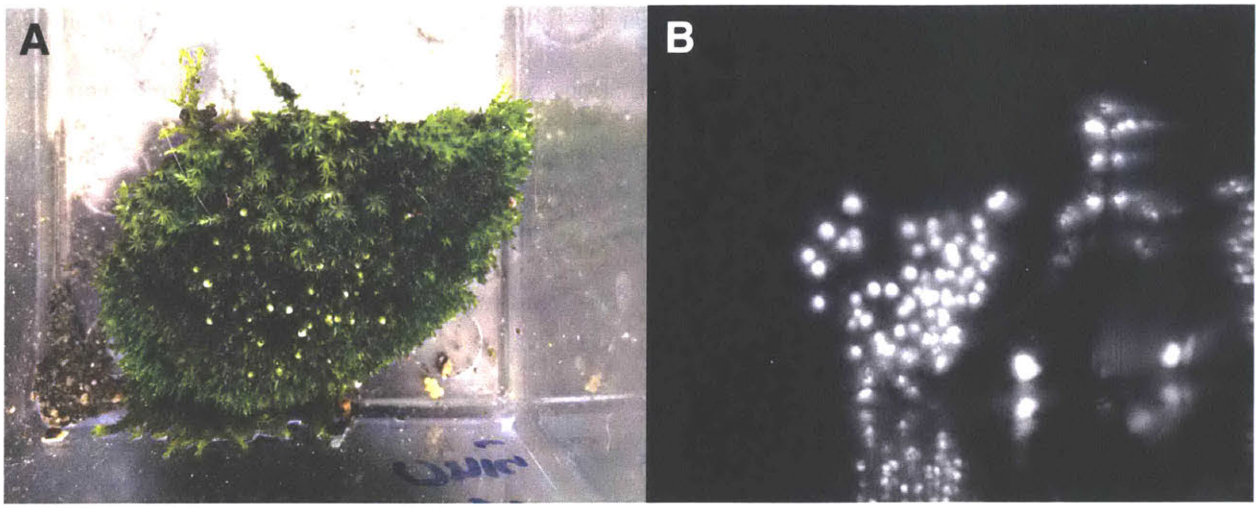
Four adult female *P. pyralis* were collected from the Bluemont Junction Trail in Arlington, VA from June 12th through June 18th 2017 (collection permission obtained by TRF from Arlington County Parks and Recreation department). The females were mated to *P. pyralis* males collected either from the same locality and date, or to males collected from Kansas in late June. Mating was performed by housing one to two males and one female in small plastic containers for ~1–3 days with a wet kimwipe to maintain humidity. Mating pairs were periodically checked for active mating, which in *Photinus* fireflies takes several hours. Successfully mated females were transferred to Magenta GA-7 plastic boxes (Sigma-Aldrich, USA), and provided a ~4 cm x 4 cm piece of locally collected moss (species diverse and unknown) as egg deposition substrate, and allowed to deposit eggs until their death in ~1–4 days. Deceased females were removed, artificial freshwater (AFW; 1:1000 diluted 32 PSU artificial seawater) was sprayed into the box to maintain high humidity, and eggs were kept for 2–3 weeks at room temperature and periodically checked until hatching. Like other firefly eggs, the eggs of *P. pyralis*were observed to be faintly luminescent imaging using a cooled CCD camera (Supporting Information 1—figure 3); however, this luminescence was not visible to the dark-adapted eye, indicating that this luminescence is less intense than other firefly species such as *Luciola cruciata* (Harvey, 1952).

Upon hatching, first instar larvae were mainly fed ~1 cm cut pieces of Canadian Nightcrawler earthworms (*Lumbricus terrestris*; Windsor Wholesale Bait, Ontario, Canada), and occasional live White Worms (*Enchytraeus albidus;* Angels Plus, Olean, NY). Although *P. pyralis* first instar larvae were observed to attack live *Enchytraeus albidus,* an experiment to determine if this would be suitable as a single food source was not performed. Uneaten and putrefying earthworm pieces were removed after 1 day, and the container cleaned. Once the larvae had been manually fed for ~2 weeks and deemed sufficiently strong, they were transferred to plastic shoeboxes (P/N: S-15402, ULINE, USA) which were intended to mimic a soil ecosystem. In personal discussions of unpublished firefly rearing attempts by various firefly researchers, we noted that a common theme was the difficulty of preventing the uneaten prey of these predatory larvae from putrifying. Therefore, we sought to create ecologically inspired 'eco-shoeboxes', where fireflies would prey on live organisms, and other organisms would assist in cleanup of uneaten or partially eaten prey that had been fed to the firefly larvae, to prevent the growth of pathogenic microorganisms on uneaten prey.

First, these shoeboxes were filled with 1L of mixed 50% (v/v) potting soil, and 50% coarse sand (Quikrete, USA) that had been washed several times with distilled water to remove silt and dust. The soil-sand mix was wet well with AFW, and live *Enchytraeus albidus* (50+), temperate springtails (50+; *Folsomia candida*; Ready Reptile Feeders, USA), and dwarf isopods (50+; *Trichorhina tomentosa*; Ready Reptile Feeders, USA) were added to the box, and several types of moss, coconut husk, and decaying leaves were sparingly added to the corners of the box. The non-firefly organisms were included to mimic a primitive detritivore (*Enchytraeus albidus and Trichorhina tomentosa*) and fungivore (*Folsomia candida*) system. About 50 firefly larvae were included per box. No interactions between the *P. pyralis* larvae and the additional organisms were observed. Predation on *Enchytraeus albidus* seems likely, but careful observations were not made. Distilled water was sprayed into the box every ~2 days to maintain a high humidity. Throughout this period, live *Lumbricus terrestris* (~10–15 cm) were added to the box every 2–3 days as food. These earthworms were first prepared by washing with distilled water several times to remove attached soil, weakened and stimulated to secrete coelomic fluid and gut contents by spraying with 95% ethanol, washed several times in distilled water, and left overnight in ~2 cm depth distilled water at 4°C. Anecdotally this pre-cleaning and preparation process reduced the rate and degree that dead earthworms putrefied. Young *P. pyralis* larvae were observed to successfully kill and gregariously feed on these live earthworms (Supporting Information 1—figure 4). The possibility that firefly larvae possess a paralytic venom used to stun or kill prey has been noted by other researchers (Hess, 1920; Williams, 1917). In our observations, an earthworm would immediately react to the bite from a single *P. pyralis* larvae, thrashing about for several minutes, but would then become seemingly paralyzed over time, supporting the role of a potent, possibly neurotoxic, firefly venom. The *P. pyralis* larvae would then begin extra-oral digestion and gregarious feeding on the liquified earthworm. Once the earthworm had been killed and broken apart by firefly larvae, *Enchytraeus albidus* would enter through gaps in the cuticle and begin to feed in large numbers throughout the interior of the earthworm. The other detritivores were observed at later stages of feeding. Between the combined action of the *P. pyralis* larvae, and the other detritivores, the live earthworm was completely consumed within 1–2 days, and no manual cleanup was required.

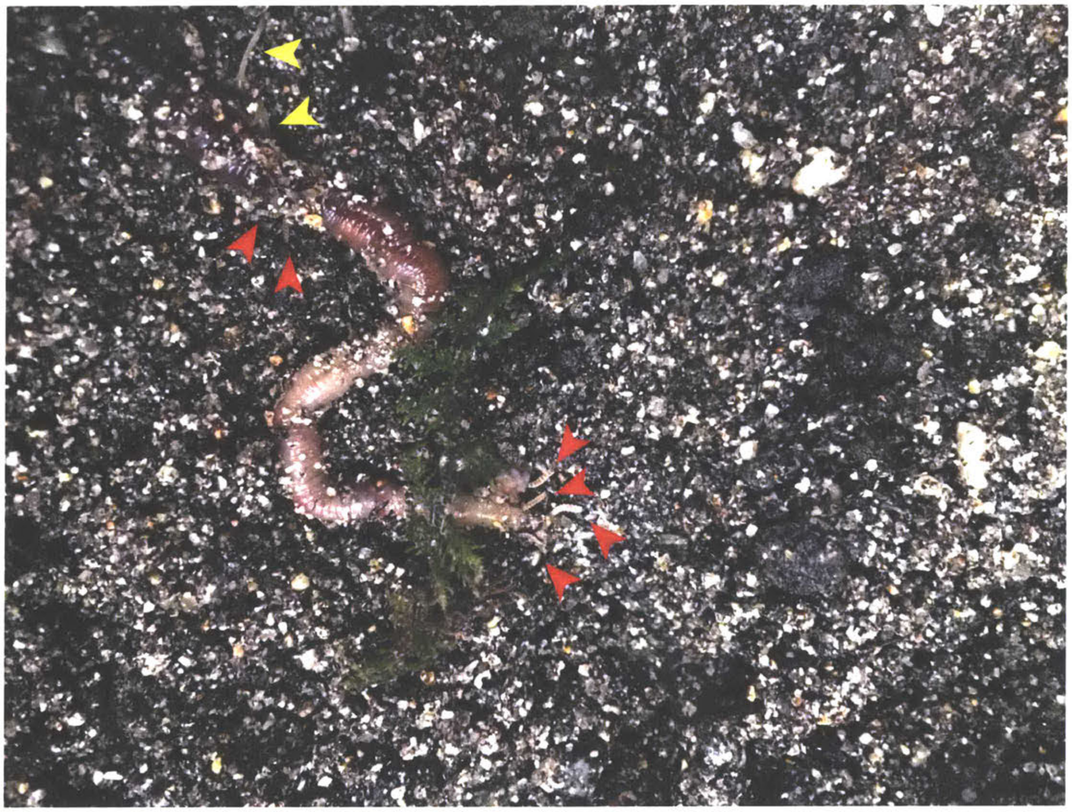
Compared to the initial manual feeding and cleaning protocol for *P. pyralis* 1st instar larvae, the ‘eco-shoebox’ rearing method was low-input and convenient for large numbers of larvae. The feeding and cleanup process was efficient for ~2 months (July through September), leading to a large number of healthy 3-4th instar larvae (Supporting Information 1—figure 5). However, after that point, *P. pyralis* larvae, possibly in preparation for a winter hibernation, seemingly became quiescent, and were less frequently seen patrolling throughout the box. At the same time, the *Enchytraeus albidus* earthworms were observed to become less abundant, either due to continual predation by *P. pyralis*, or due to population collapse from insufficient fulfillment of nutritional requirements from feeding of *Enchytraeus albidus* on *Lumbricus terrestris* alone.

At this point, earthworms were not consumed within 1–2 days, and became putrid, and *P. pyralis* which had been feeding on these earthworms were frequently found dead nearby, and themselves quickly putrefied. Generally after this point *P. pyralis* larvae were more frequently found dead and partially decayed, indicating the possibility of pathogenesis from microorganisms from putrefying earthworms. At this stage, it was observed that mites (Acari), probably from the soil contained in the guts of the fed earthworms, became abundant, and were observed to act as ectoparasitic on *P. pyralis* larvae. An attempt to simulate hibernation of *P. pyralis larvae* was made by storing them at 4°C for ~3 weeks, however a large proportion (~30%) of larvae died during this hibernation to a seeming fungal infection. Other larvae revived quickly when returned to room temperature, but all *Trichorhina tomentosa* were killed by even transient exposure to 4°C. To date, a smaller number of fifth and sixth instar *P. larvae* have been obtained, but pupation in the laboratory has not occured. The lack of pupation is unsurprising as it is likely occurs in the wild after 1–2 years of growth, is likely under temperature and photoperiodic control, and may require a licensing stage of cold temperature hibernation for several weeks. Overall, manual feeding of first1 st instar larvae followed by the 'eco-shoebox' method was unexpectedly successful approach for the maintenance and growth of *P. pyralis* larvae.



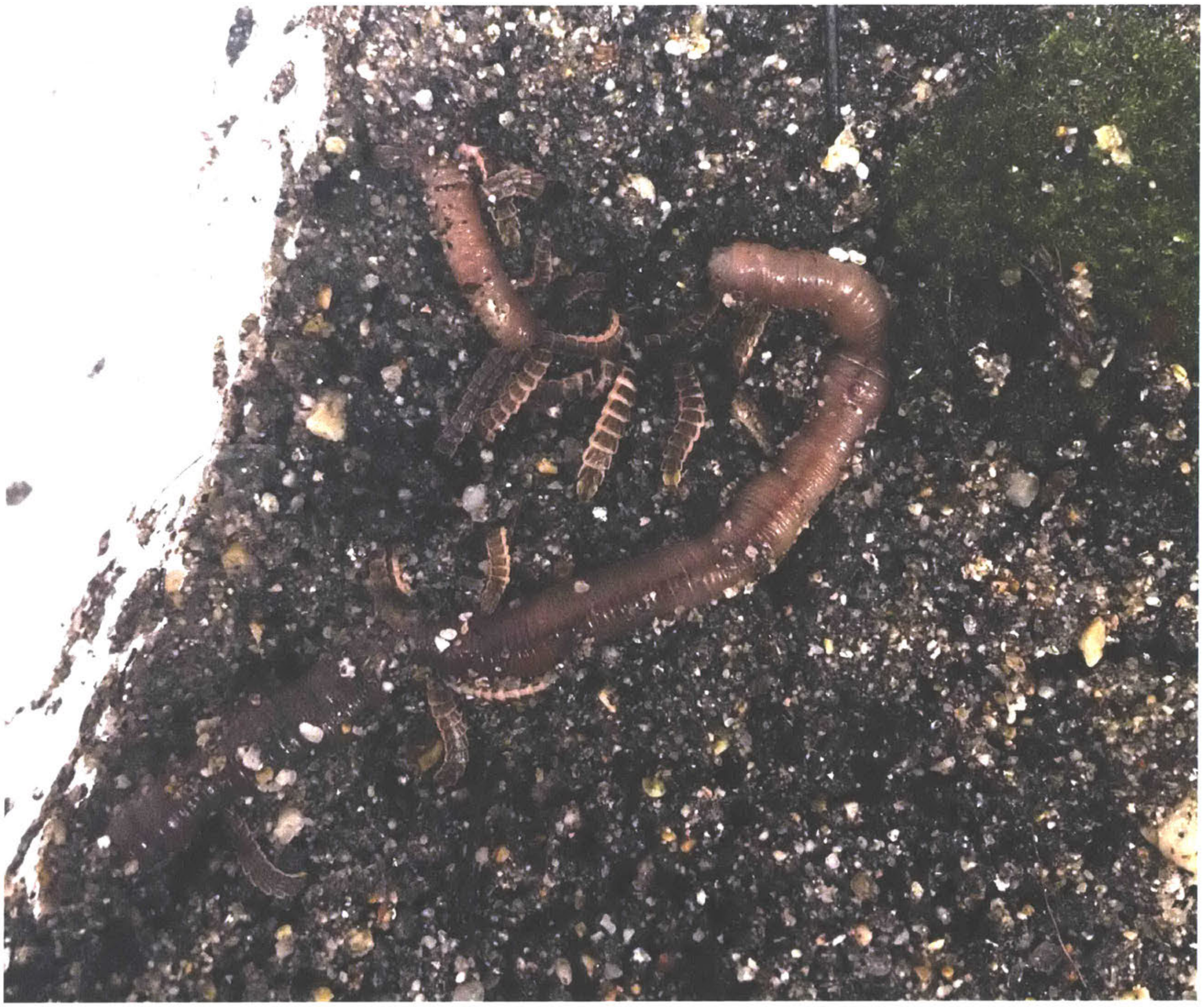
Supporting Information 1—figure 3. Luminescence of P. pyralis eggs.

(**A**) Photograph under ambient light of ~1 day post-deposition *P. pyralis* eggs. (**B**) Photograph of self-luminescence of ~1 day post-deposition *P. pyralis* eggs. Both photographs taken with a NightOwl LB98 cooled CCD luminescence imager (Berthold Technologies, USA). Luminescence was not visible to the dark-adapted eye.



Supporting Information 1—figure 4. Gregarious predation of young P. pyralis larvae on a live Lumbricus terrestris.

Both *P. pyralis* larvae (red arrows), and *Enchytraeus albidus* (yellow arrows), were observed to feed on the paralyzed earthworms.



Supporting Information 1—figure 5. Gregarious predation of 3rd-4th instar *P. pyralis* larvae on a live Lumbricus terrestris.

1.4 Karyotype and genome size

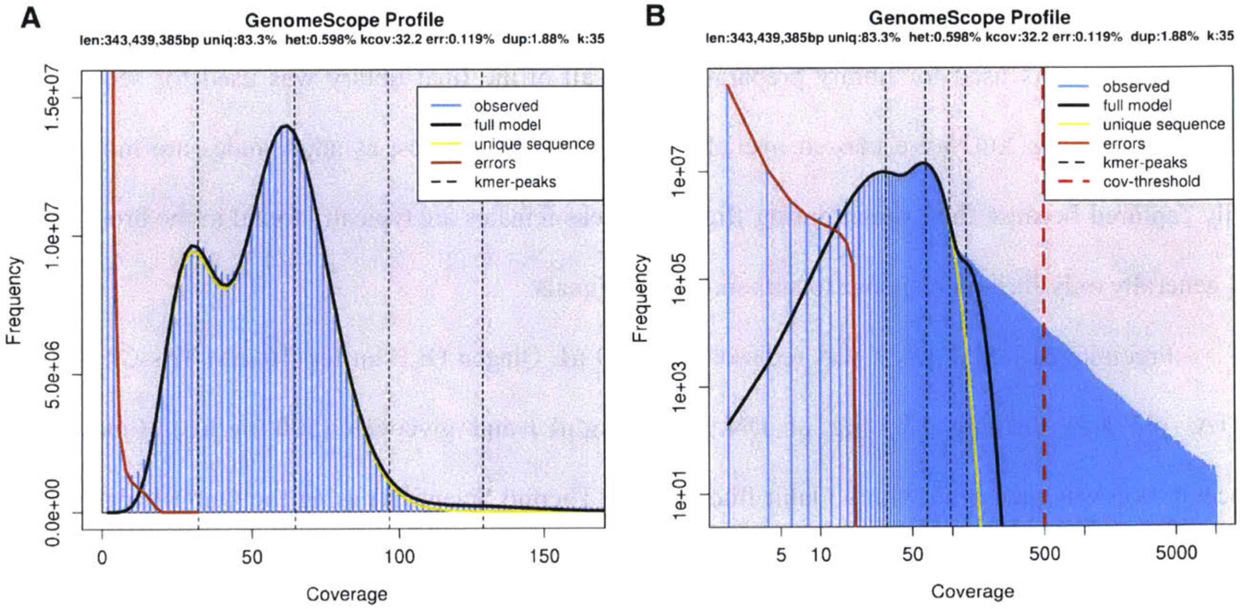
The karyotype of *P. pyralis* was previously reported to be 2n = 20 with XO sex determination (male, 18A + XO; female, 18A + XX) (Wasserman and Ehrman, 1986). The genome sizes of four *P. pyralis* adult males were previously determined to be 422 ± 9 Mbp (SEM, n = 4), whereas the genome sizes of five *P. pyralis* adult females were determined to be 448 ± 7 (SEM, n = 5) by nuclear flow cytometry analysis (Lower et al., 2017). From these analyses, the size of the X-chromosome is inferred to be ~26 Mbp. Genome size inference via kmer spectral analysis of the *P. pyralis* short-insert Illumina data from a single adult *P. pyralis* male estimated a genome size of 343 Mbp (Supporting Information 1—figure 6).

1.5 Library preparation and sequencing

See Supporting Information 4—table 1 for a overview of all sequence libraries. Library specific construction methods are detailed below.

1.5.1 Illumina

DNA was extracted from sterile-water-washed thorax of Great Smoky Mountains National Park collected specimens using phenol-chloroform extraction with RNAse digestion, checked for quality via gel electrophoresis, and quantified by Nanodrop or Qubit (Thermo Scientific, USA). To obtain sufficient DNA for both short insert and mate-pair library construction, libraries were constructed separately from DNA from each of two individual males and pooled DNA of three males, all from the same population. Males were selected for sequencing as they are more easily found in the field than females. In addition, as *P. pyralis* males are XO (Dias et al., 2007), differences in sequencing coverage could inform localization of scaffolds to the X chromosome. Illumina TruSeq short insert (average insert size: 300 bp) and Nextera mate-pair libraries (insert size: 3 Kbp, 6 Kbp) were constructed at the Georgia Genomics Facility (Athens, GA) and subsequently sequenced on two lanes of Illumina HiSeq 2000 100 × 100 bp PE reads (University of Texas; Supporting Information 4—table 1).



Supporting Information 1—figure 6. Genome scope kmer analysis of the P. pyralis short read library.

(**A**) Linear and (**B**) log plot of a kmer spectral genome composition analysis of the '8369' *P. pyralis* Illumina short-read library from a single *P. pyralis* XO adult male (Supporting Information 1.5.1; Supporting Information 4—table 1) with jellyfish (v2.2.9; parameters: -C -k 35) (Marçais and Kingsford, 2011) and GenomeScope (v1.0; parameters: Kmer length = 35, Read length = 100, Max kmer coverage = 1000) (Vurture et al., 2017). len = inferred haploid genome length, uniq = percentage non-repetitive sequence, het = overall rate of genome heterozygosity, kcov = mean kmer coverage for heterozygous bases, err = error rate of the reads, dup: average rate of read duplications. These results are consistent with the genome size of a XO male, when possible systematic error of kmer spectral analysis and flow cytometry genome size estimates is considered. The heterozygosity is somewhat low when compared to some other arthropods.

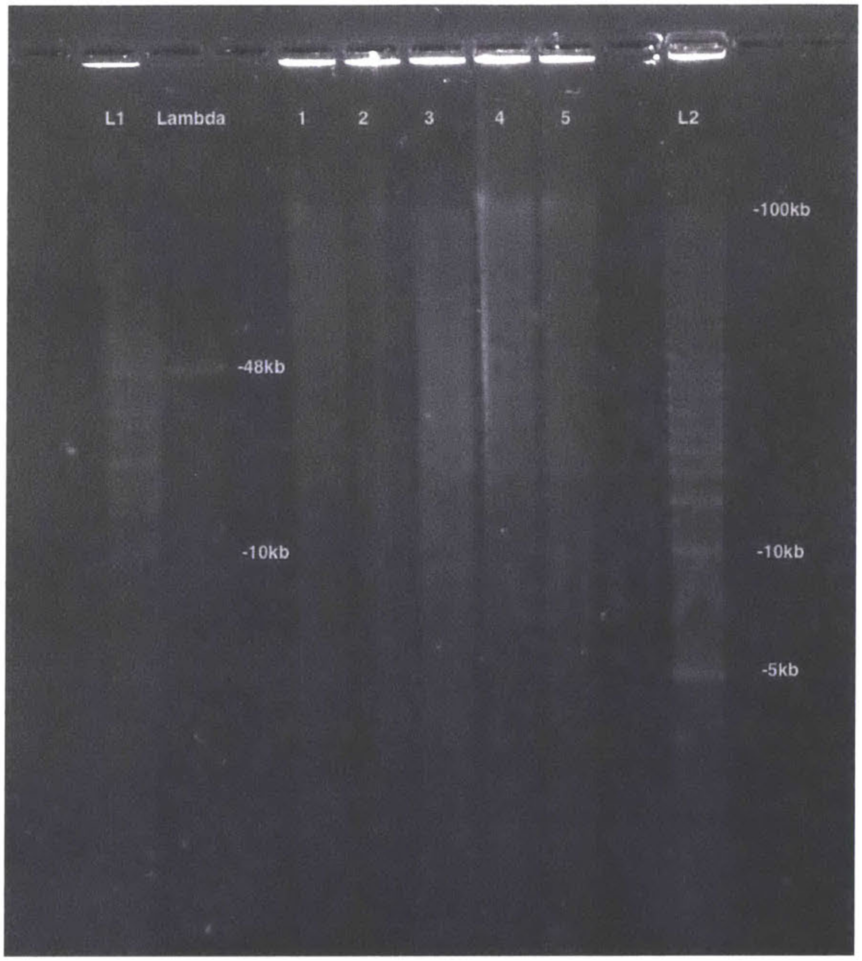
1.5.2 PacBio

High-molecular-weight DNA (HMW DNA) was extracted from four pooled lyophilized adult male *P. pyralis* (dry mass 90.8 mg) from the MMNJ field site. These specimens were first externally washed using 95% ethanol, after which DNA extraction proceeded with a 100/G Genomic Tip plus Genomic Buffers kit (Qiagen, USA). DNA extraction followed the manufacturer's protocol, with the exception of the final precipitation step, where HMW DNA was pelleted with 40 µg RNA grade glycogen (Thermo Scientific, USA) and centrifugation (3000 x g, 30 min, 4 ˚C) instead of spooling on a glass rod. Although increased genomic heterozygosity from four pooled males and a resulting more complicated genome assembly was a concern for a wild population like *P. pyralis*, four males were used in order to extract enough DNA for workable coverage using 15 Kbp+ size selected PacBio RSII sequencing. All extracted DNA was used for library preparation, and all of the final library was used for sequencing. Adult males, being XO, were chosen over the preferable XX females, as adult males are much more easily captured because they signal during flight, whereas females are typically found in the brush below and generally only flash in response to authentic male signals.

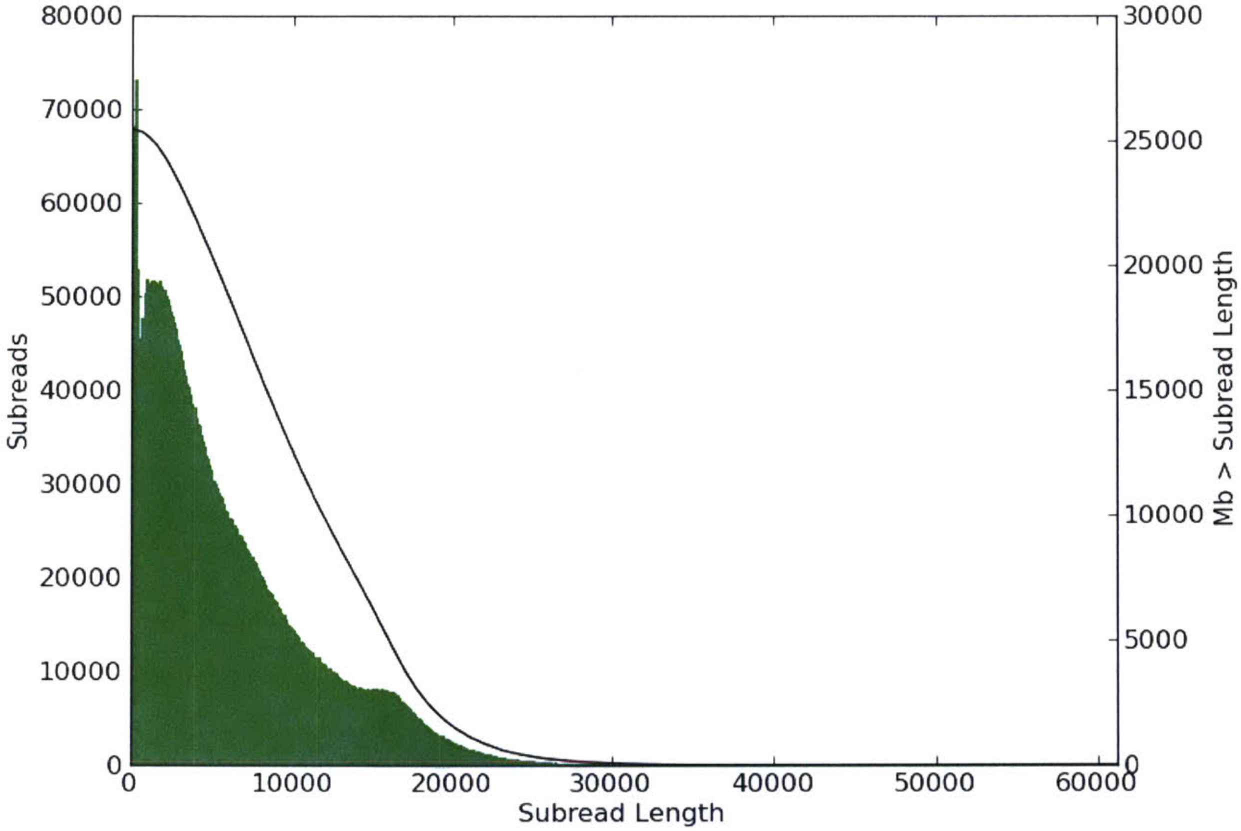
Precipitated HMW DNA was redissolved in 80 µL Qiagen QLE buffer (10 mM Tris-Cl, 0.1 mM EDTA, pH 8.5) yielding 17.1 µg of DNA (214 ng/µL) and glycogen (500 ng/µL). Final DNA concentration was measured with a Qubit fluorometer (Thermo Scientific) using the Qubit Broad Range kit. Manipulations hereafter, including HMW DNA size QC, fragmentation, size selection, library construction, and PacBio RSII sequencing, were performed by the Broad Technology Labs of the Broad Institute (Cambridge, MA).

First, the size distribution of the HMW DNA was confirmed by pulsed-field-gel-electrophoresis (PFGE). In brief, 100 ng of HMW DNA was run on a 1% agarose gel (in 0.5x TBE) with the Bio-Rad CHEF-DR III system. The sample was run out for 16 hr at six volts/cm with an angle of 120 degrees with a running temperature of 14 °C. The gel was stained with SYBR Green dye (Thermo Scientific - Part No. S75683). 1 µg of 5 Kbp ladder (Bio-Rad, part no 170–3624) was used as a standard. These results demonstrated the HMW DNA had a mean size of >48 Kbp (Supporting Information 1—figure 7). This pool of HMW DNA is designated 1611_PpyrPB1 (NCBI BioSample SAMN08132578).

Next, HMW DNA (17.1 µg) was sheared to a targeted average size of 20–30 Kbp by centrifugation in a Covaris g-Tube (part no. 520079) at 2500 x g for 2 min. SMRTbell libraries for sequencing on the PacBio platform were constructed according to the manufacturer's recommended protocol for 20 Kbp inserts, which includes size selection of library constructs larger than 15 Kbp using the BluePippin system (Sage Science, Beverly, MA). Two separate cassettes were run. In each cassette, two lanes were used in which there was 1362 ng/lane (PAC20kb kit). Constructs 15 Kbp and above were eluted over a period of 4 hr. An additional damage repair step was carried out post size-selection. Insert size range for the final library was determined using the Fragment Analyzer System (Advanced Analytical, Ankeney, IA). The size-selected SMRTbell library was then sequenced over 61 SMRT cells on a PacBio RSII instrument of the Broad Technology Labs (Cambridge, MA), using the P6 v.2 polymerase and the v.4 DNA Sequencing Reagent (P6-C4 chemistry; part numbers 100-372-700, 100-612-400). PacBio sequencing data is available on the NCBI Sequence Read Archive (Bioproject PRJNA378805).



Supporting Information 1—figure 7. PFGE of P. pyralis HMW DNA used for PacBio sequencing. Lane 1 was used for further library prep and sequencing, Lanes 2–5 represent separate batches of *P. pyralis* HMW DNA that was not used for PacBio sequencing. Lane 1 was used as it had the highest DNA yield, and an equivalent DNA size distribution to the other samples.



Supporting Information 1—figure 8. Subread length distribution for P. pyralis PacBio RSII sequencing.

Figure produced with SMRTPortal (v2.3.0.140936, Pacific Biosciences) by aligning all PacBio reads from data from the 61 SMRT cells against Ppyr1.3 using the RS_Resequencing.1 protocol with default parameters. Subread length unit is basepair (bp).

1.5.3 Hi-C library preparation

Two adult *P. pyralis* MMNJ males were flash frozen in liquid nitrogen, stored at −80°C, and shipped on dry-ice to Phase Genomics (Seattle, WA). Manipulations hereafter occurred at Phase Genomics, following previously published protocols (Bickhart et al., 2017; Burton et al., 2013; Lieberman-Aiden et al., 2009). Briefly, a streamlined version of the standard Hi-C protocol (Lieberman-Aiden et al., 2009) was used to perform a series of steps resulting in proximity-ligated DNA fragments, in which physically proximate sequence fragments are joined into linear chimeric molecules. First, in vivo chromatin was cross-linked with formaldehyde, fixing physically proximate loci to each other. Chromatin was then extracted from cellular material and digested with the *Sau*3AI restriction enzyme, which cuts at the GATC motif. The resulting fragments were proximity ligated with biotinylated nucleotides and pulled down with streptavidin beads. These chimeric sequences were then sequenced with 80 bp PE sequencing on the Illumina NextSeq platform, resulting in Hi-C read pairs.

1.6 Genome assembly

The *P. pyralis* genome assembly followed three stages: (1) a hybrid assembly using Illumina and PacBio reads, producing assembly Ppyr1.1 (Supporting Information 1.6.2), (2) Ppyr1.1 scaffolded using Hi-C data, producing assembly Ppyr1.2 (Supporting Information 1.6.3), and (3) Ppyr1.2 manually curation for proper X-chromosome assembly and removal of putative non-firefly sequences, producing Ppyr1.3 (1.6.4).

1.6.2 Ppyr1.1: MaSuRCA hybrid assembly

Several genome assembly approaches were evaluated with the general goal of maximizing conserved gene content and contiguity. The highest quality *P. pyralis assembly* was generated by a hybrid assembly approach using a customized MaSuRCA (v3.2.1_01032017) (Zimin et al., 2017, 2013) pipeline that combined both Illumina-corrected PacBio reads (Mega-reads) and synthetic long reads constructed from short-insert reads alone (Super-reads) using a custom small overlap length (59 bp).

We first applied MaSuRCA (v3.2.1_01032017) (Zimin et al., 2017, 2013) to correct our long reads (38x coverage; Library ID 1611_PpyrPB1; Supporting Information 4—table 1) using our short-insert and mate-pair reads (Libraries: 8369, 375_3K, 8375_6K, 83_3K, 83_6K; Supporting Information 4—table 1). No pre-filtering of reads was performed, as Illumina adaptors are automatically removed within the MaSuRCA pipeline. We modified the pipeline to assemble the genome using both corrected long reads (Mega-reads) and synthetic long reads (Super-reads) with a custom smaller overlap length (59 bp). All reads (short-insert, mate-pair and PacBio) were then used within the MaSuRCA pipeline to call a genomic consensus.

To scaffold the contigs, we first filtered Illumina short-reads from the mate-pair libraries (Libraries 8375_3K, 8375_6K, 83_3K, 83_6K) with Nxtrim (v0.4.1) (O'Connell et al., 2015) with parameters '--separate --rf --justmp'. We then manually integrated the MaSuRCA assembly by replacing the incomplete mitochondrial contigs with complete mitochondrial assemblies from *P. pyralis* and *Apocephalus antennatus* (Supporting Information 5.2). We scaffolded and gap-filled the assembly using the Illumina short-insert and filtered mate-pair reads (Libraries: 8369, 8375_3K, 8375_6K, 83_3K, 83_6K) via Redundans (v0.13a) (Pryszcz and Gabaldón, 2016) with default settings. After scaffolding with our Illumina data, redundant sequences were removed by the MaSuRCA 'deduplicate_contigs.sh' script. We then applied PBjelly (v15.8.24) (English et al., 2012) and PacBio reads to scaffold and gap-fill the assembly, and redundancy reduction with 'deduplicate_contigs.sh' script was run again. Finally, we replaced mitochondrial sequences which had been artificially extended by the scaffolding, gap-filling and sequence extension process with the proper sequences. The resultant assembly was dubbed Ppyr1.1.

1.6.3 Ppyr1.2: Scaffolding with Hi-C

The Hi-C read pairs were applied in a manner similar to that originally described here (Burton et al., 2013) and later expanded upon (Bickhart et al., 2017). Briefly, Hi-C reads were mapped to Ppyr1.1 with BWA (v1.7.13) (Li and Durbin, 2009), requiring perfect, unique mapping locations for a read pair to be considered usable. The number of read pairs joining a given pair of contigs is referred to as the 'link frequency' between those contigs, and when normalized by the number of restriction sites in the pair of contigs, is referred to as the 'link density' between those contigs.

A three-stage scaffolding process was used to create the final scaffolds, with each stage based upon previously described analysis of link density (Bickhart et al., 2017; Burton et al., 2013). First, contigs were placed into chromosomal groups. Second, contigs within each chromosomal group were placed into a linear order. Third, the orientation of each contig is determined. Each scaffolding stage was performed many times in order to optimize the scaffolds relative to expected Hi-C linkage characteristics.

In keeping with previously described methods (Bickhart et al., 2017; Burton et al., 2013), the number of chromosomal scaffolds to create–10–was an *a priori* input to the scaffolding process derived from the previously published chromosome count of *P. pyralis* (Wasserman and Ehrman, 1986). However, to verify the correctness of this assumption, scaffolds were created for haploid chromosome numbers ranging from 5 to 15. A scaffold number of 10 was found to be optimal for containing the largest proportion of Hi-C linkages within scaffolds, which is an expected characteristic of actual Hi-C data.

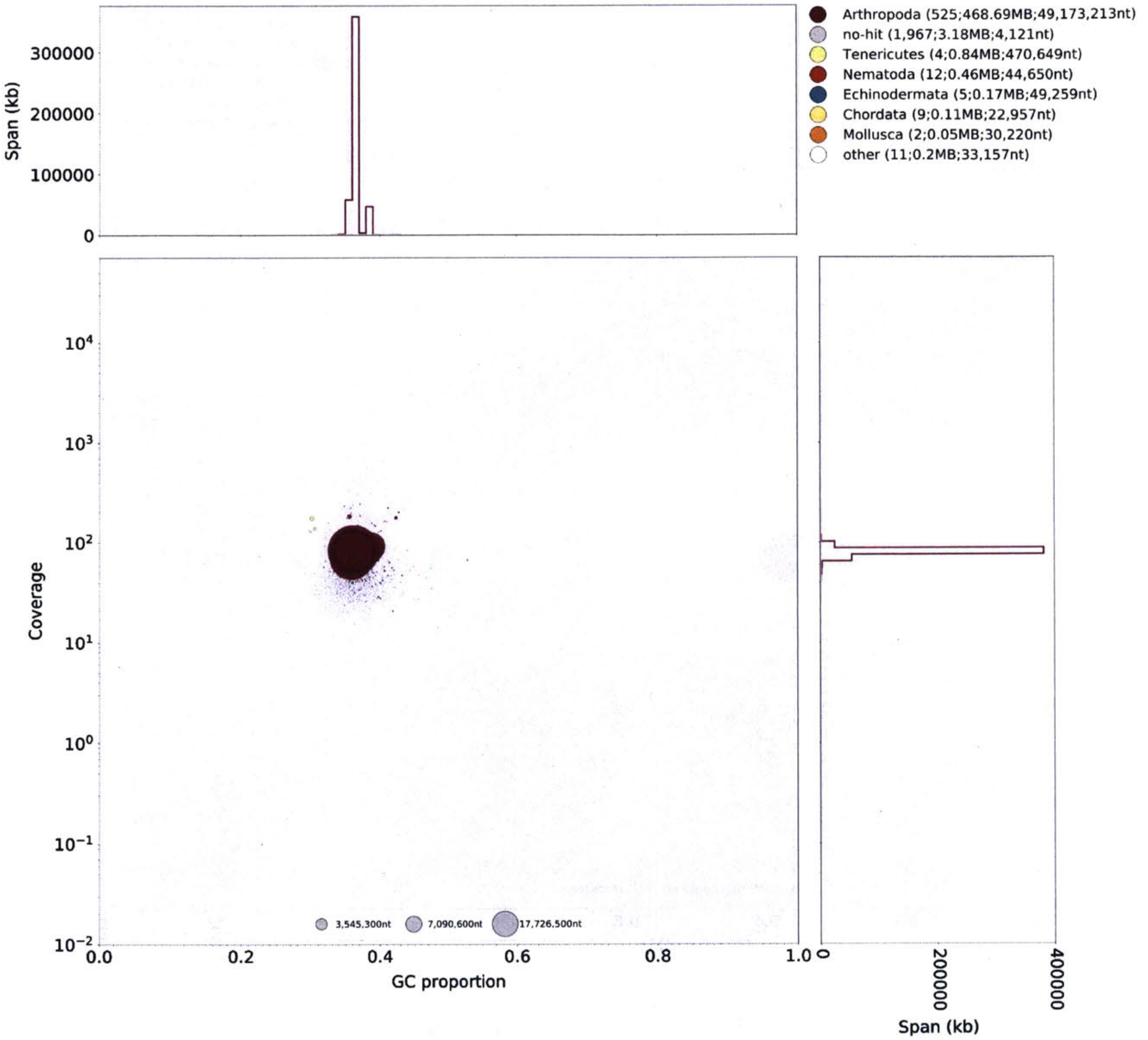
1.6.4 Ppyr1.3: Manual curation and taxonomic annotation filtering

1.6.4.1 Defining the X chromosome

Hi-C data was mapped and converted to the 'hic' file format with the juicer pipeline (v1.5.6) (Durand et al., 2016b), and then visualized using juicebox (v1.5.2) (Durand et al., 2016a). This visualization revealed a clear breakpoint in Hi-C linkage density on LG3 at ~22,220,000 bp. Mapping of Illumina short-insert and PacBio reads with Bowtie2 (v2.3.1) (Langmead and Salzberg, 2012) and SMRTPortal (v2.3.0.140893) with the 'RS_Resequencing.1' protocol, followed by visualization with Qualimap (v2.2.1) (Okonechnikov et al., 2016), revealed that the first section of LG3 (1–22,220,000 bp), here termed LG3a, was present at roughly half the coverage of LG3b (22,220,001–50,884,892 bp) in both the Illumina and PacBio libraries. Mapping of *Tribolium castaneum* X chromosome proteins (NCBI Tcas 5.2) to the Ppyr1.2 assembly using both tblastn (v2.6.0) (Camacho et al., 2009) and Exonerate (v2.2.0) (Slater and Birney, 2005) based 'protein2genome' alignment through the MAKER pipeline revealed a relative enrichment on LG3a only. Taken together, this data suggested that the half-coverage section of LG3 (LG3a) corresponded to the X-chromosome of *P. pyralis*, and that it was misassembled onto an autosome. Therefore, we manually split LG3 into LG3a and LG3b in the final assembly.

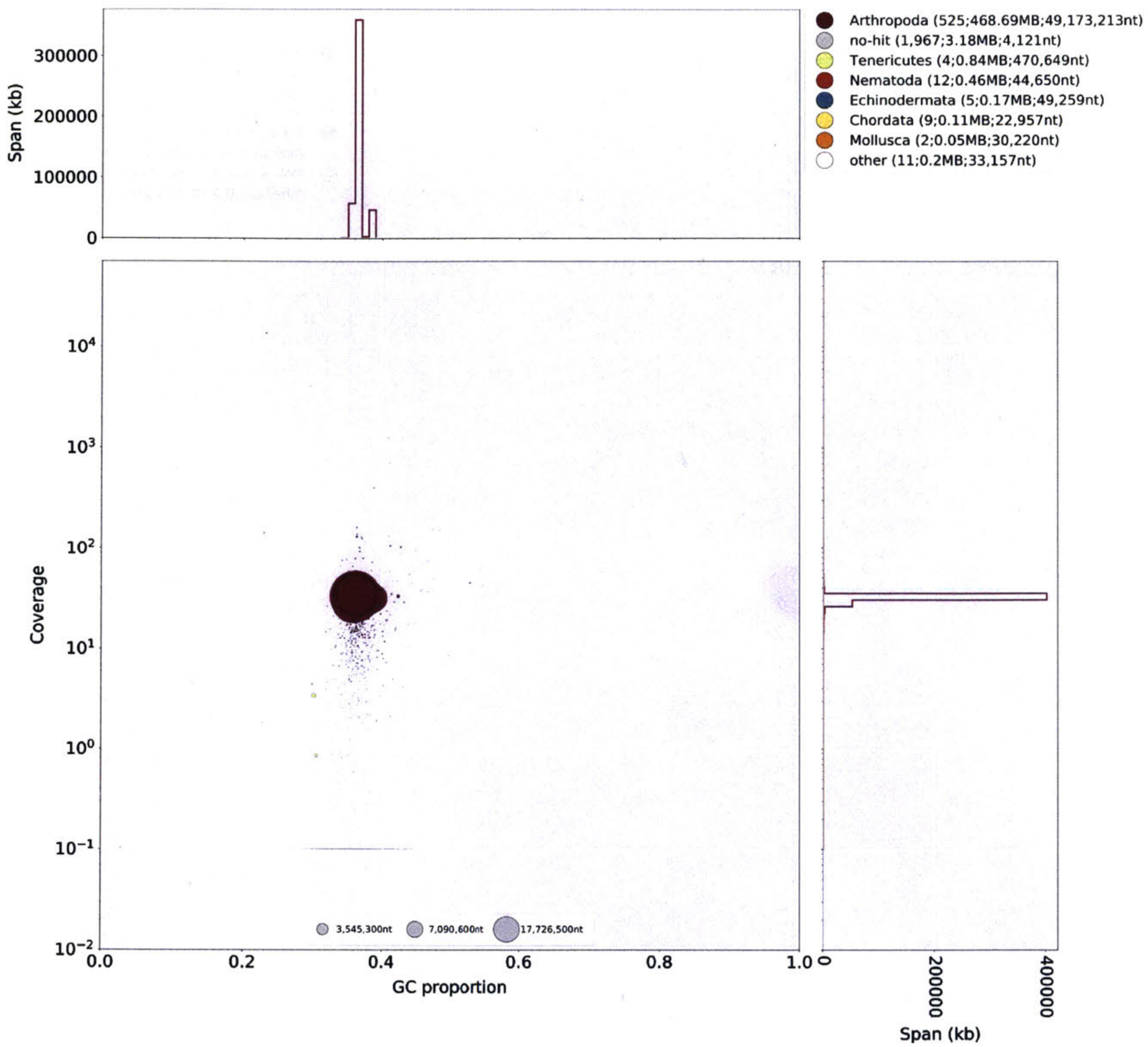
1.6.4.2 Taxonomic annotation filtering

Given the recognized importance of filtering genome assemblies to avoid misinterpretation of the data (Koutsovoulos et al., 2016), we sought to systematically remove assembled non-firefly contaminant sequence from Ppyr1.2. Using the blobtools toolset (v1.0.1) (Laetsch and Blaxter, 2017), we taxonomically annotated our scaffolds by performing a blastn (v2.6.0+) nucleotide sequence similarity search against the NCBI nt database, and a diamond (v0.9.10.111) (Buchfink et al., 2015) translated nucleotide sequence similarity search against the of Uniprot reference proteomes (July 2017). Using this similarity information, we taxonomically annotated the scaffolds with blobtools using parameters '-x bestsumorder --rank phylum'. A tab delimited text file containing the results of this blobtools annotation are available on FigShare (DOI: 10.6084/m9.figshare.5688982). We then generated the final genome assembly by retaining scaffolds that either contained annotated features (genes or non-simple/low-complexity repeats), had coverage >10.0 in both the Illumina (Supporting Information 1—figure 9) and PacBio libraries (Supporting Information 1—figure 10), and if the taxonomic phylum was annotated as 'Arthropod' or 'no-hit' by the blobtools pipeline (Supporting Information 1—figure 11). This approach removed 374 scaffolds (2.1 Mbp), representing 15% of the scaffold number and 0.4% of the nucleotides of Ppyr1.2. Notably, four tenericute scaffolds, likely corresponding to a partially assembled *Entomoplasma sp.* genome, distinct from the *Entomoplasma luminosus var. pyralis assembled* from the PacBio library (Supporting Information 5) were removed. Furthermore, we removed two contigs representing the mitochondrial genome of *P. pyralis* (complete mtDNA available via Genbank: KY778696). The final filtered assembly, Ppyr1.3, is available at www.fireflybase.org.

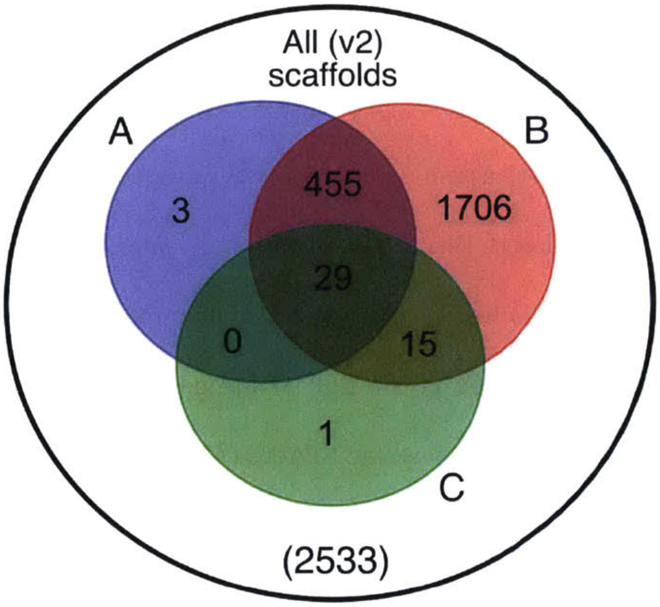


Supporting Information 1—figure 9. BlobPlot of Illumina short-insert reads aligned against the Ppyr1.2 reference.

Coverage shown represents mean coverage of reads from the Illumina short-insert library (Sample name 8369; Supporting Information 4—table 1), aligned against Ppyr1.2 using Bowtie2 with parameters (--local). Scaffolds were taxonomically annotated as described in Supporting Information 1.6.4.2.



Supporting Information 1—figure 10. BlobPlot of *P. pyralis* PacBio reads aligned against Ppyr1.2. Coverage shows represents mean coverage of reads from the PacBio library (Sample name 1611; Supporting Information 4—table 1). The reads were aligned using SMRTPortal v2.3.0.140893 with the 'RS_Resequencing.1' protocol with default parameters. Scaffolds were taxonomically annotated as described in Supporting Information 1.6.4.2.



Supporting Information 1—figure 11. Venn diagram representation of blobtools taxonomic annotation filtering approach for Ppyr1.2 scaffolds.

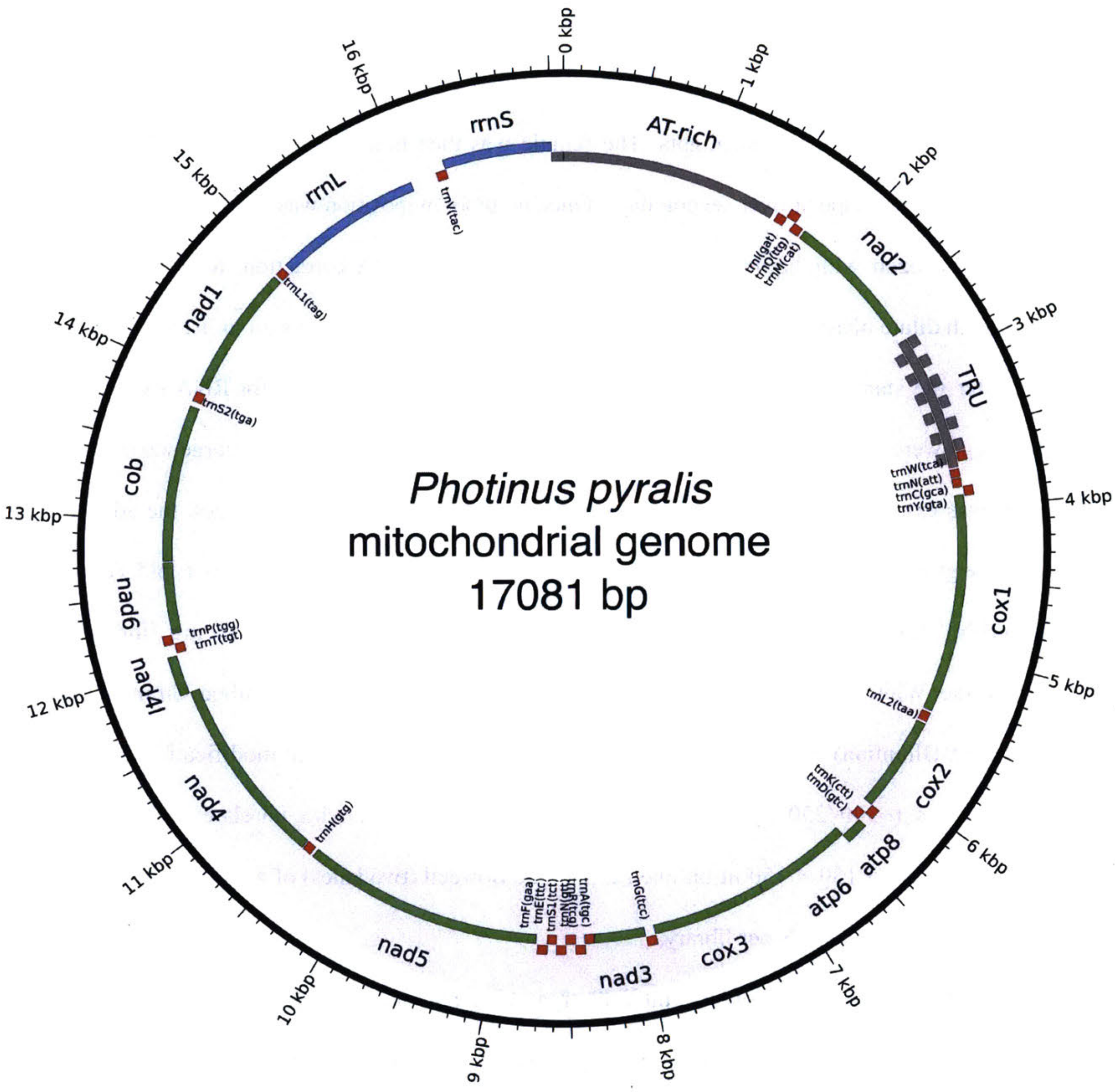
(**A**) The blue set represents scaffolds which have >10.0 coverage in both Illumina and PacBio libraries. (**B**) The red set represents scaffolds which had either genes on repeats (non simple or low-complexity) annotated. (**C**) The green set represents scaffolds with suspicious taxonomic assignment (Non 'Arthropod' or 'no-hit'). Outside A, B, and C, represents low-coverage, unannotated scaffolds. Ppyr1.3 consists of the intersection of A and B, minus the intersection of C. All linkage groups (LG1-LG10) were annotated as 'Arthropod' by blobtools, and captured in the intersection between A and B but not set C.

1.7 Ppyr0.1-PB: PacBio only genome assembly

In addition to our finalized genome assembly (Ppyr1.3), we sought to better understand the symbiont composition that varied between our *P. pyralis* PacBio and Illumina libraries. Therefore, we produced a long-read only assembly of our PacBio data to assemble the sequence that might be unique to this library. To achieve this, we first filtered the HDF5 data from the 61 sequence SMRT cells to. FASTQ format subreads using the SMRTPortal data processing software package (v2.3.0.140893) (http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/) with the 'RS_Subreads.1' protocol with default parameters. These subreads were then input into Canu (Github commit 28ecea5/v1.6) (Koren et al., 2017) with parameters 'genomeSize = 450 m corOutCoverage = 200 ovlErrorRate = 0.15 obtErrorRate = 0.15 -pacbio-raw'. The unpolished contigs from this produced genome assembly are dubbed Ppyr0.1-PB.

1.8 Mitochondrial genome assembly and annotation

To achieve a full length mitochondrial genome (mtDNA) assembly of *P. pyralis*, sequences were assembled separately from the nuclear genome. Short insert Illumina reads from a single GSMNP individual (Sample 8369; Supporting Information 4—table 1) were mapped to the known mtDNA of the closest available relative, *Pyrocoelia rufa* (NC_003970.1 (Bae et al., 2004)) using bowtie2 v2.3.1 (parameters: --very-sensitive-local). All concordant read pairs were input to SPAdes (v3.8.0) (Nurk et al., 2013) (parameters: --plasmid --only-assembler -k35,55,77,90) for assembly. The resulting contigs were then combined with the *P. rufa* mitochondrial reference genome for a second round of read mapping and assembly. The longest resulting contig aligned well to the *P. rufa* mitochondrial genome; however, it was ~1 Kbp shorter than expected, with the unresolved region appearing to be the tandem repetitive region (TRU) (Bae et al., 2004), previously described in the *P. rufa* mitochondrial genome. To resolve this, all PacBio reads were mapped to the draft mitochondrial genome, and a single high-quality PacBio circular-consensus-sequencing (CCS) read that spanned the unresolved region was selected using manual inspection and manually assembled with the contiguous sequence from the Illumina sequencing to produce a complete circular assembly. The full assembly was confirmed by re-mapping the Illumina short-read data using bowtie2 followed by consensus calling with Pilon v1.21 (Walker et al., 2014). Re-mapped PacBio long-read data also confirmed the structure of the mtDNA, and indicated variability in the repeat unit copy number of the TRU amongst the four sequenced *P. pyralis* individuals (Sample 1611_PpyrPB1; Supporting Information 4—table 1). The *P. pyralis* mtDNA was then 'restarted' using seqkit (Shen et al., 2016), such that the FASTA record break occurred in the AT-rich region, and annotated using the MITOS2 annotation server (http://mitos2.bioinf.uni-leipzig.de/). Low confidence and duplicate gene predictions were manually removed from the MITOS2 annotation. The final *P. pyralis* mtDNA with annotations is available on GenBank (KY778696).



Supporting Information 1—figure 12. Mitochondrial genome of P. pyralis.
The mitochondrial genome of *P. pyralis* was assembled and annotated as described. Note the firefly specific tandem-repeat-unit (TRU) region. Figure produced with Circos (Krzywinski et al., 2009).

1.9 Transcriptome analysis

1.9.1 RNA-extraction, library preparation and sequencing

In order to capture expression from diverse life stages, stranded RNA-Seq libraries were prepared from whole bodies of four life stages/sexes (eggs, 1 st instar larvae, adult male, and adult female; Supporting Information 1—table 1). Eggs and larvae were derived from a laboratory mating of *P. pyralis*

(Collected MMNJ, July 2016). Briefly, live adult *P. pyralis* were transported to the lab and allowed to mate in a plastic container over several days. The female, later sequenced, was observed mating with two independent males on two separate nights. The female was then transferred to a plastic container with moss, and allowed to oviposit over several days. Once no more oviposition was observed, the female was removed, flash frozen with liquid N2, and stored at −80°C for RNA extraction. Resulting eggs were washed 3x with dilute bleach/ H2O and reared in aggregate in plastic containers on moist Whatman paper. ~13 days after the start of egg oviposition, a subset of eggs were flash frozen for RNA extraction. The remaining eggs were allowed to hatch and larvae were flash frozen the day after emergence (first instar). Total RNA was extracted from a single stored adult male (non-paternal to eggs/larvae), the adult female (maternal to eggs/larvae), seven pooled eggs, and four pooled larvae using the RNeasy Lipid Tissue Mini Kit (QIAGEN) with the optional on-column DNase treatment. Illumina sequencing libraries were prepared by the Whitehead Genome Technology Core (WI-GTC) using the TruSeq Stranded mRNA library prep kit (Illumina) and following the manufacturer's instructions with modification to select for larger insert sizes (~300–350 bp). These samples were multiplexed with unrelated plant RNA-Seq samples and sequenced 150 × 150 nt on one rapid mode flowcell (two lanes) of a HiSeq2500 (WI-GTC), to a depth of ~30M paired reads per library.

To examine gene expression in adult light organs, we generated non-strand specific sequencing of polyA pulldown enriched mRNA from dissected photophore tissue (Supporting Information 1—table 1). Photophores were dissected from the abdomens of adult *P. pyralis* males (Collected MMNJ, July 2015) by Dr. Adam South (Harvard School of Public Health), using three individuals per biological replicate. These tissues and libraries were co-prepared and sequenced with other previously published libraries (full library preparation and sequencing details available in (Al-Wathiqui et al., 2016)) at a depth of ~10M paired reads per library.

To examine gene expression in larval light organs, we performed RNA-seq on dissected larval light organs. We first extracted total RNA from a pool of six dissected larval photophores from three individuals using the RNeasy Lipid Tissue Mini Kit (QIAGEN) with the optional on-column DNase treatment. The larvae were the same larvae described in Supporting Information 1.3.2. The total RNA was enriched to mRNA via polyA pulldown and prepared into a paired unstranded Illumina sequencing using the Kapa HyperPrep kit (Kapa Biosystems, USA), and sequenced to a depth of 43M 100 × 100 paired reads on a HiSeq2500 sequencer (Illumina, USA).

All these data were combined with previously published tissue, sex, and stage-specific libraries (Supporting Information 1—table 1) for reference-guided transcriptome assembly (Supporting Information 1.9.3). Strand-specific data was used for *de novo* transcriptome assembly (Supporting Information 1.9.2).

Supporting Information 1—table 1. P. pyralis RNA sequencing libraries.

N: number of individuals pooled for sequencing; **Sex/stage**: M = male, F = female, A = adult, L = larva, L1 = larva 1 st instar, L4 = larvae fourth instar, E13 = 13 days post fertilization eggs; **Tissue**: H = head, PA = lantern abdominal segments, FB = abdominal fat body, T = thorax, OAG = other accessory glands, SD = spermatophore digesting gland/bursa, SG = spiral gland, SC = spermatheca, p=dissected photophore, E = egg, WB = whole body.

Library name	Source*	SRA ID	N	Sex/stage	Tissue	Library type
8175 *Photinus pyralis* male head (adult) transcriptome	SRA1	SRR2103848	1	M/A	H	
8176 *Photinus pyralis* male light organ (adult) transcriptome	SRA1	SRR2103849	1	M/A	PA	
8819 *Photinus pyralis* light organ (larval) transcriptome	SRA1	SRR2103867	1	L	PA	
9_Photinus_sp_1_lantern	SRA2	SRR3521424	1	M/A	PA	Strand-specific. Ribo-zero
Ppyr_FatBody_1	SRA3	SRR3883756	6	M/A	FB	
Ppyr_FatBody_2	SRA3	SRR3883757	6	M/A	FB	
Ppyr_FatBody_3	SRA3	SRR3883766	6	M/A	FB	
Ppyr_FatBody_Mated	SRA3	SRR3883767	4	M/A	FB	
Ppyr_FThorax	SRA3	SRR3883768	3	F/A	T	

Ppyr_MThorax_1	SRA3	SRR3883769	6	M/A	T	
Ppyr_MThorax_2	SRA3	SRR3883770	6	M/A	T	
Ppyr_MThorax_3	SRA3	SRR3883771	6	M/A	T	
Ppyr_OAG_1A	SRA3	SRR3883772	6	M/A	AG	
Ppyr_OAG_1B	SRA3	SRR3883773	6	M/A	AG	
Ppyr_OAG_2	SRA3	SRR3883758	6	M/A	AG	
Ppyr_OAG_Mated	SRA3	SRR3883759	4	M/A	AG	
Ppyr_SDGBursa	SRA3	SRR3883760	3	F/A	SD	
Ppyr_SG_Mated	SRA3	SRR3883761	4	M/A	SG	
Ppyr_Spermatheca	SRA3	SRR3883762	3	F/A	SC	
Ppyr_SpiralGland_1	SRA3	SRR3883763	6	M/A	SG	
Ppyr_SpiralGland_2	SRA3	SRR3883764	6	M/A	SG	
Ppyr_SpiralGland_3	SRA3	SRR3883765	6	M/A	SG	
Ppyr_Lantern_1A	‡	SRR6345453	6	M/A	P	
Ppyr_Lantern_2	‡	SRR6345454	6	M/A	P	
Ppyr_Lantern_3	‡	SRR6345446	6	M/A	P	
Ppyr_Eggs	‡	SRR6345447	7	E13	E	Strand-specific
Ppyr_Larvae	‡	SRR6345445	4	L1	WB	Strand-specific
Ppyr_wholeFemale†	‡	SRR6345449	1	F/A	WB	Strand-specific
Ppyr_wholeMale	‡	SRR6345452	1	M/A	WB	Strand-specific
TF_VA2017_3pooled_larval_lantern	‡	SRR7345580	3	L4	P	

*SRA1 = NCBI BioProject PRJNA289908 (Sander and Hall, 2015); SRA2 = NCBI BioProject PRJNA321737 (Fallon et al., 2016); SRA3 = NCBI BioProject PRJNA328865 (Al-Wathiqui et al., 2016).
†Parent of eggs and larvae with data from this study.
‡This study.

1.9.2 De novo transcriptome assembly and genome alignment

One strand-specific *de novo* transcriptome was produced from all available MMNJ strand-specific reads (WholeMale, WholeFemale, eggs, larvae) and strand-specific reads from SRA (SRR3521424) (Supporting Information 1—table 1). Reads from these five libraries were pooled (158.6M paired-reads) as input for *de novo* transcriptome assembly. Transcripts were assembled using Trinity (v2.4.0) (Grabherr et al., 2011) with default parameters except the following: (--SS_lib_type RF --trimmomatic --min_glue 2 min_kmer_cov 2 --jaccard_clip --no_normalize_reads). Gene structures were then predicted from alignment of the *de novo* transcripts to the Ppyr1.3 genome using the PASA pipeline (v2.1.0) (Haas et al., 2008) with the following steps: first, poly-A tails were trimmed from transcripts using the internal seqclean component; next, transcript accessions were extracted using the accession_extractor.pl component; finally, the trimmed transcripts were aligned to the genome with modified parameters (--aligners blat,gmap --ALT_SPLICE --transcribed_is_aligned_orient --tdn tdn.accs). Using both the blat (v. 36 × 2) (Kent, 2002) and gmap (v2017-09-11) (Wu and Watanabe, 2005) aligners was required, as an appropriate gene model for Luc2 was not correctly produced using only a single aligner. Importantly, it was also necessary to set (--NUM_BP_PERFECT_SPLICE_BOUNDARY = 0) for the validate_alignments_in_db.dbi step, to ensure transcripts with natural variation near the splice sites were not discarded. Post alignment, potentially spurious transcripts were filtered out using a custom script (https://github.com/elifesciences-publications/PASA_expression_filter_2017) that removed extremely lowly-expressed transcripts (<1% of the expression of a given PASA assembly cluster). Expression values used for filtering were calculated from the WholeMale library reads using the Trinity align_and_estimate_abundance.pl utility script. The WholeMale library was selected because it was the highest quality library - strand-specific, low contamination, and many reads - thereby increasing the reliability of the transcript quantification. Finally, the PASA pipeline was run again with this filtered transcript set to generate reliable transcript structures. Peptides were predicted from the final transcript structures using Transdecoder (v.5.0.2) (https://github.com/TransDecoder/TransDecoder) with default parameters. Direct coding gene models (DCGMs) were then produced with the Transdecoder 'cdna_alignment_orf_to_genome_orf.pl' utility script with the PASA assembly GFF and transdecoder predicted peptide GFF as input. The unaligned *de novo* transcriptome assembly is dubbed

'PPYR_Trinity_stranded', whereas the aligned direct coding gene models are dubbed 'Ppyr1.3_Trinity-PASA_stranded-DCGM'.

1.9.3 Reference guided transcriptome assembly

Two reference guided transcriptomes, one strand-specific and one non-strand-specific, were produced from all available *P. pyralis* RNA-Seq reads (Supporting Information 1—table 1) using HISAT2 (v2.0.5) (Kim et al., 2015) and StringTie (v1.3.3b) (Pertea et al., 2015). For each library, reads were first mapped to the Ppyr1.3 genome assembly with HISAT2 (parameters: -X 2000 --dta --fr) and then assembled using StringTie with default parameters except use of '--rf' for the strand-specific libraries. The resulting library-specific assemblies were then merged into a final assembly using StringTie (--merge), one for the strand-specific and one for the non-strand specific libraries, producing two final assemblies. For each final assembly, a transcript fasta file was produced and peptides predicted using Transdecoder with default parameters. Then, the StringTie. GTFs were converted to GFF format with the Transdecoder 'gtf_to_alignment_gff3.pl' utility script and direct coding gene models (DCGMs) were produced with the Transdecoder 'cdna_alignment_orf_to_genome_orf.pl' utility script, with the StringTie GFF and transdecoder predicted peptide GFF as input. The final GFFs were validated and sorted with genometools (v1.5.9) with parameters (parameters: gff3 -tidy -sort -retainids), and then sorted again for IGV format with igvtools (parameters: sort). The aligned direct coding gene models for the stranded and unstranded reference guided transcriptomes are dubbed 'Ppyr1.3_Stringtie_stranded-DCGM' and 'Ppyr1.3_Stringtie_unstranded-DCGM'.

1.9.4 Transcript expression analysis

P. pyralis RNA-Seq reads (Supporting Information 1—table 1) were pseudoaligned to the PPYR_OGS1.1 geneset CDS sequences using Kallisto (v0.44.0) (Bray et al., 2016) with 100 bootstraps (-b 100), producing transcripts-per-million reads (TPM). Kallisto expression quantification analysis results are available on FigShare (DOI: 10.6084/m9.figshare.5715139).

1.10 Official coding geneset annotation (PPYR_OGS1.1)

We annotated the coding gene structure of *P. pyralis* by integrating direct coding gene models produced from the *de novo* transcriptome (Supporting Information 1.9.2) and reference guided transcriptome (Supporting Information 1.9.3), with a lower weighted contribution of *ab initio* gene predictions, using the Evidence Modeler (EVM) algorithm (v1.1.1) (Haas et al., 2008). First, Augustus (v3.2.2) (Stanke et al., 2006) was trained against Ppyr1.2 with BUSCO (parameters: -l endopterygota_odb9 --long --species tribolium2012). Next, preliminary gene models for prediction training were produced by the alignment of the *P. pyralis de novo* transcriptome to Ppyr1.2 with the MAKER pipeline (v3.0.0β) (Holt and Yandell, 2011) in 'est2genome' mode. Preliminary gene models were used to train SNAP (v2006-07-28) (Korf, 2004) following the MAKER instructions (http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial_for_GMOD_Online_Training_2014). Augustus and SNAP gene predictions of Ppyr1.3 were then produced through the MAKER pipeline, with hints derived from MAKER blastx/exonerate mediated protein alignments of peptides from *Drosophila melanogaster* (NCBI GCF_000001215.4_Release_6_plus_ISO1_MT_protein.faa), *Tribolium castaneum*(NCBI GCF_000002335.3_Tcas5.2_protein), and *Aquatica lateralis* (AlatOGS1.0; this report), and MAKER blastn/exonerate transcript alignments of the *P. pyralis de novo* transcriptome. These *ab initio* coding gene models are dubbed 'Ppyr1.3_abinitio_Augustus-SNAP-MAKER-GMs.gff3'

We then integrated the *ab initio* predictions with our *de novo* and reference guided direct coding gene models, using EVM. A variety of evidence sources, and EVM evidence weights were empirically tested and evaluated using a combination of inspection of known gene models (e.g. Luc1/Luc2), and the BUSCO score of the geneset. In the final version, six sources of evidence were used for EVM: *de novo transcriptome* direct coding gene models (Ppyr1.3_Trinity-PASA_stranded-DCGM; weight = 11), protein alignments (*D. melanogaster*, *T. castaneum*, *A. lateralis;*weight = 8), GMAP and BLAT alignments of *de novo* transcriptome (via PASA; weight = 5), reference guided transcriptome direct coding gene models

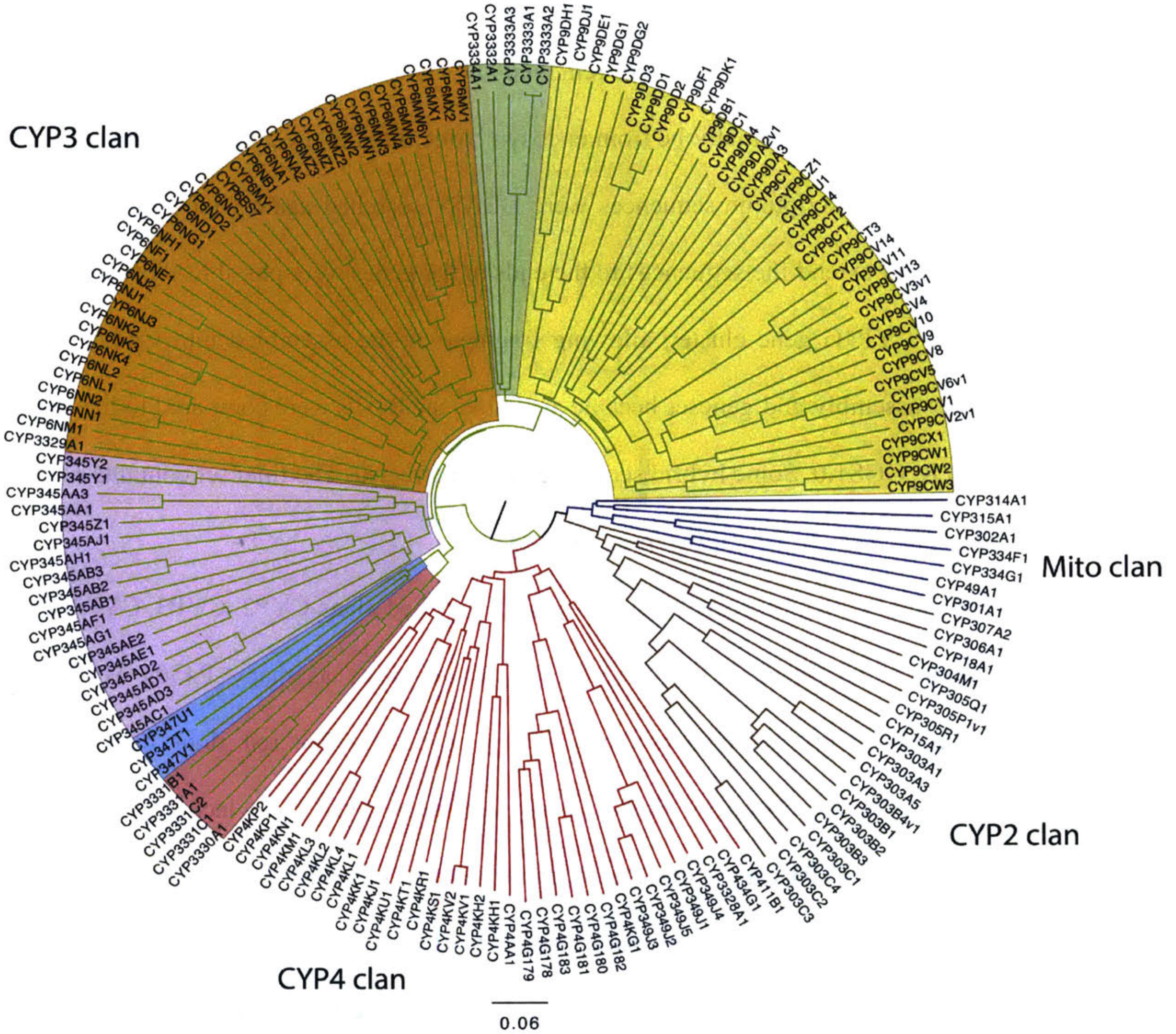
(Ppyr1.3_Stringtie_stranded-DCGM; weight = 3), Augustus and SNAP *ab initio* gene models (via MAKER; weight = 2). A custom script (https://github.com/elifesciences-publications/maker_gff_to_evm_gff_2017) was necessary to convert MAKER GFF format to an EVM compatible GFF format.

Lastly, gene models for luciferase homologs, P450s (Supporting Information 1.10.1), and *de novo* methyltransferases (DNMTs) which were fragmented or were incorrect (e.g. fusions of adjacent genes) were manually corrected based on the evidence of the *de novo* and reference guided direct coding gene models. Manual correction was performed by performing TBLASTN searches with known good genes from these gene families within SequenceServer(v1.10.11) (Priyam et al., 2015), converting the TBLASTN results to gff3 format with a custom script (https://github.com/elifesciences-publications/firefly_genomes_general_scripts), and viewing these alignments alongside the alternative direct coding gene models (Supporting Information 1.9.2; 1.9.3) in Integrative Genomics Viewer(v2.4.8) (Thorvaldsdóttir et al., 2013). The official gene set models gff3 file was manually modified in accordance with the evidence from the direct gene models. Different revision numbers of the official geneset (e.g. PPYR_OGS1.0, PPYR_OGS1.1) represent the improvement of the geneset over time due to these continuing manual gene annotations.

1.10.1 P450 annotation

Translated *de novo* transcripts were formatted to be BLAST searchable with NCBI's standalone software. The peptides were searched with 58 representative insect P450s in a batch BLAST (evalue = 10). The query set was chosen to cover the diversity of insect P450s. The top 100 hits from each search were retained. The resulting 5837 hit IDs were filtered to remove duplicates, leaving 472 unique hits. To reduce redundancy due to different isoforms, the Trinity transcript IDs (style DNXXX_cX_gX_iX) were filtered down to the 'DN' level, resulting in 136 unique IDs. All peptides with these IDs were retrieved and clustered with CD-Hit (v4.5.4) (Li and Godzik, 2006) to 99% identity to remove short overlapping peptides. These 535 protein sequences were batch BLAST compared to a database of all named insect P450s to identify best hits. False positives were removed and about 30 fungal sequences were removed. These fungal sequences could potentially be from endosymbiotic fungi in the gut. Overlapping sequences were combined and the transcriptome sequences were BLAST searched against the *P. pyralis* genome assembly to fill gaps and extend the sequences to the ends of the genes were possible. This approach was very helpful with the CYP4G gene cluster, allowing fragments to be assembled into whole sequences. When a new genome assembly and geneset became available, the P450s were compared to the integrated gene models in PPYR_OGS1.0. Some hybrid sequences were corrected. The final set contains 170 named cytochrome P450 sequences (166 genes, two pseudogenes).

The cytochrome P450s in insects belong to four established clans CYP2, CYP3, CYP4 and Mito (Supporting Information 1—figure 13). *P. pyralis* has about twice as many P450s as *Drosophila melanogaster* (86 genes, four pseudogenes) and slightly more than the red flour beetle *Tribolium castaneum* (137 genes, 10 pseudogenes). Pseudogenes were determined by a lack of conserved sites common to all P450s. The CYP3 clan is the largest, mostly due to three families: CYP9 (40 sequences), CYP6 (36 sequences) and CYP345 (18 sequences). Insects have few conserved sequences across species. These include the halloween genes for 20-hydroxyecdysone synthesis and metabolism CYP302A1, CYP306A1, CYP307A2, CYP314A1 and CYP315A1 (Rewitz et al., 2007) in the CYP2 and Mito clans. The CYP4G subfamily makes a hydrocarbon waterproof coating for the exoskeleton (Helvig et al., 2004). Additional conserved P450s are CYP15A1 (juvenile hormone (Helvig et al., 2004)) and CYP18A1 (20-hydroxyecdysone degradation (Guittard et al., 2011)) in the CYP2 clan. Most of the other P450s are limited to a narrower phylogenetic range. Many are unique to a single genus, although this may change as more sampling is done. It is common for P450s to expand into gene blooms (Sezutsu Hideki et al., 2013).



Supporting Information 1—figure 13. P. pyralis P450 gene phylogenetic tree.
Neighbor-joining phylogenetic tree of 165 cytochrome P450s from *P. pyralis*. Four pseudogenes and one short sequence were removed. The P450 clans have colored spokes (CYP2 clan brown, CYP3 clan green, CYP4 clan red, Mito clan blue). Shading highlights different families and family clusters within the CYP3 clan. The tree was made using Clustal Omega at EBI (https://www.ebi.ac.uk/Tools/msa/clustalo/) with default settings. The resulting multiple sequence alignment is available on FigShare (DOI: 10.6084/m9.figshare.5697643). The tree was drawn with FigTree v1.3.1 using midpoint rooting.

1.10.2 Virus annotation and analysis

Viruses were discovered from analysis of published *P. pyralis* RNA sequencing libraries (NCBI TSA: GEZM00000000.1) and the Ppyr1.2 genome assembly (Supporting Information 5.4). 24 *P. pyralis*

RNA sequencing libraries were downloaded from SRA (taxid: 7054, date accessed: 15th June 2017). RNA sequence reads were first *de novo* assembled using Trinity v2.4.0 (Grabherr et al., 2011) with default parameters. Resulting transcriptomes were assessed for similarity to known viral sequences by TBLASTN searches (max e-value = 1 × 10−5) using as a probe the complete predicted non redundant viral Refseq proteins retrieved from NCBI (date accessed: 15th June 2017). Significant hits were explored manually and redundant contigs discarded. False-positives were eliminated by comparing candidate viral contigs to the entire non-redundant nucleotide (nt) and protein (nr) database to remove false-positives.

Candidate virus genome segment sequences were curated by iterative mapping of reads using Bowtie 2 (v2.3.2) (Langmead and Salzberg, 2012). Special attention was taken with the segments' terminis -- an arbitrary cut off of 10x coverage was used as threshold to support terminal base calls. The complementarity and folded structure of untranslated ends, as would be expected for members of the Orthomyxoviridae, was assessed by Mfold 2.3 (Zuker, 2003). Further, conserved UTR sequences were identified using ClustalW2 (Larkin et al., 2007) (support of>65% required to call a base). To identify/rule out additional segments of no homology to the closely associated viruses we used diverse *in silico* approaches based on RNA levels including: the sequencing depth of the transcript, predicted gene product structure, or conserved genome termini, and significant co-expression with the remaining viral segments. After these filtering steps, putative viral sequences were annotated manually. First, potential open-reading frames (ORF) were predicted by ORFfinder (Wheeler et al., 2003) and manually inspected by comparing predicted ORFS to those from the closest-related reference virus genome sequence. Then, translated ORFs were blasted against the non-redundant protein sequences NR database and best hits were retrieved. Predicted ORF protein sequences were also subjected to a domain-based Blast search against the Conserved Domain Database (CDD) (v3.16) (Marchler-Bauer et al., 2017) and integrated with SMART (Letunic and Bork, 2018), Pfam (Finn et al., 2016), and PROSITE (Sigrist et al., 2002) results to characterize the functional domains. Secondary structure was predicted with Garnier as implemented in

EMBOSS (v6.6) (Rice et al., 2000), signal and membrane cues were assessed with SignalP (v4.1) (Petersen et al., 2011), and transmembrane topology and signal peptides were predicted by Phobius (Käll et al., 2004). Finally, the potential functions of predicted ORF products were explored using these annotations as well as similarity to viral proteins of known function.

To characterize *Orthomyxoviridae* viral diversity in *P. pyralis* in relation to known viruses, predicted *P. pyralis* viral proteins were used as probes in TBLASTN (max e-value = 1×10−5) searches of the complete 2754 Transcriptome Shotgun Assembly (TSA) projects on NCBI (date accessed: 15th June 2017). Significant hits were retrieved and the target TSA projects further explored with the complete *Orthomyxoviridae* refseq collection to assess the presence of additional similar viral segments. Obtained transcripts were extended/curated using the SRA associated libraries for each TSA hit and then the curated virus sequences were characterized and annotated as described above.

To identify *P. pyralis* viruses to family/genus/species, amino acid sequences of the predicted viral polymerases, specifically the PB1 subunit, were used for phylogenetic analyses with viruses of known taxonomy. To do this, multiple sequence alignment were generated using MAFFT (v7.310) (Katoh and Standley, 2013) and unrooted maximum-likelihood phylogenetic trees were constructed using FastTree (Price et al., 2010) with standard parameters. FastTree accounted for variable rates of evolution across sites by assigning each site to one of 20 categories, with the rates geometrically spaced from 0.05 to 20, and set each site to its most likely rate category using a Bayesian approach with a gamma prior. Support for individual nodes was assessed using an approximate likelihood ratio test with the Shimodaira-Hasegawa-like procedure. Tree topology, support values and substitutions per site were based on 1000 tree resamples.

To facilitate taxonomic identification, we complemented BLASTP data with two levels of phylogenetic insights: (i) Trees based on the complete refseq collection of ssRNA (-) viruses which permitted a conclusive assignment at the virus family level. (ii) Phylogenetic trees based on reported, proposed, and discovered *Orthomyxoviridae* viruses that allowed tentative species demarcation and genera postulation. PB1-based trees were complemented independently with phylogenetic studies derived from amino acids of predicted nucleoproteins, hemagglutinin protein, PB2 protein, and PA protein which supported species, genera and family demarcation based on solely on PB1, the standard in *Orthomyxoviridae*. In addition, sequence similarity of concatenated gene products of International Committee on Taxonomy of Viruses (ICTV) allowed demarcation to species and firefly viruses were assessed by Circoletto diagrams (Darzentas, 2010) (e-value = 1e-2). Where definitive identification was not easily assessed, protein Motif signatures were determined by identification of region of high identity between divergent virus species, visualized by Sequence Logo (Crooks et al., 2004), and contrasted with related literature. Heterotrimeric viral polymerase 3D structure prediction was generated with the SWISS-MODEL automated protein structure homology-modeling server (Biasini et al., 2014) with the best fit template 4WSB: the crystal structure of Influenza A virus 4WSB. Predicted structures were visualized in UCSF Chimera (Pettersen et al., 2004) and Needleman-Wunsch sequence alignments from structural superposition of proteins were generated by MatchMaker and the Match->Align Chimera tool. Alternatively, 3D structures were visualized in PyMOL (v1.8.6.0; Schrodinger).

Viral RNA levels in the transcriptome sequences were also examined. Virus transcripts RNA levels were obtained by mapping the corresponding raw SRA FASTQ read pairs using either Bowtie2 (Langmead and Salzberg, 2012) or the reference mapping tool of the Geneious 8.1.9 suite (Biomatters, Ltd.) with standard parameters. Using the mapping results and retrieving library data, absolute levels, TPMs and FPKM were calculated for each virus RNA segment. Curated genome segments and coding annotation of the identified PpyrOMLV1 and 2 are available on FigShare at (DOI: 10.6084/m9.figshare.5714806) and (DOI: 10.6084/m9.figshare.5714812) respectively, and NCBI Genbank (accessions MG972985 through MG972994)

All curation, phylogeny construction, and visualization were conducted in Geneious 8.1.9 (Biomatters, Ltd.). Animal silhouettes in Supporting Information 5—figure 2 were developed based on non-copyrighted public domain images. Figure compositions were assembled using Photoshop CS5 (Adobe). Bar graphs were generated with Excel 2007 software (Microsoft). RNA levels normalized as mapped transcripts per million per library were visualized using Shinyheatmap (Khomtchouk et al., 2017).

Finally, to identify endogenous viral-like elements, tentative virus detections and the viral refseq collection were contrasted to the *P. pyralis* genome assembly Ppyr1.2 by BLASTX searches (e-value = 1e-6) and inspected by hand. Then 15 Kbp genome flanking regions were retrieved and annotated. Lastly, transposable elements (TEs) were determined by the presence of characteristic conserved domains (e.g. RNASE_H, RETROTRANSPOSON, INTEGRASE) on predicted gene products and/or significant best BLASTP hits to reported TEs (e-value <1e-10).

1.11 Repeat annotation

Repeat prediction for *P. pyralis* was performed *de novo* using RepeatModeler (v1.0.9) (http://www.repeatmasker.org/) and MITE-Hunter (v11-2011) (Han and Wessler, 2010). RepeatModeler uses RECON (Bao and Eddy, 2002) and RepeatScout (Price et al., 2005) to predict interspersed repeats, and then refines and classifies the consensus repeat models to build a repeat library. MITE-Hunter detects candidate MITEs (miniature inverted-repeat transposable elements) by scanning the assembly for terminal inverted repeats and target site duplications < 2 kb apart. To identify tandem repeats, we also ran Tandem Repeat Finder (v4.09; parameters: 2 7 7 80 10) (Benson, 1999), and added repeats whose repeat block length was >5 kb to the repeat library annotated as 'complex tandem repeat'. The RepeatModeler and MITE-Hunter libraries were combined and classified using RepeatClassifier (RepeatModeler 1.0.9 distribution) (http://www.repeatmasker.org/). The complex repeats identified by Tandem Repeat Finder were added to this classified list to create the final library of 3118 repeats. This repeat library is dubbed the *P. pyralis* Official Repeat Library 1.0 (PPYR_ORL1.0).

Supporting Information 1—table 2. Annotated repetitive elements in *P. pyralis*.

Repeat class	Family	Counts	Bases	% of assembly
DNA	All	122551	38364685	8.14
	Helitrons	35068	9308100	1.97
LTR	All	28860	11401648	2.42
Non-LTR	All	52107	17744320	3.76
	LINE	48983	16763499	3.56
	SINE	1241	139637	0.03
Unknown interspersed		696511	141970977	30.1
Complex tandem repeats		10395	2352796	0.50
Simple repeat		48224	2372183	0.50
rRNA		449	161517	0.034

1.12 *P. pyralis* methylation analysis

MethylC-seq libraries were prepared from HMW DNA prepared from four *P. pyralis* MMNJ males using a previously published protocol (Urich et al., 2015), and sequenced to ~36x expected depth on an Illumina NextSeq 500. Methylation analysis was performed using methylpy (Schultz et al., 2015) Methylpy calls programs for read processing and aligning: (i) reads were trimmed of sequencing adapters using Cutadapt (Martin, 2011), (ii) processed reads were mapped to both a converted forward strand (cytosines to thymines) and converted reverse strand (guanines to adenines) using bowtie (flags: -S, -k 1, -m 1, --chunkmbs 3072, --best, --strata, -o 4, -e 80, -l 20, -n 0 (Langmead et al., 2009)), and (iii) PCR duplicates were removed using Picard (http://broadinstitute.github.io/picard). In total, 49.4M reads were mapped corresponding to an actual sequencing depth of ~16x. A sodium bisulfite non-conversion rate of 0.17% was estimated from Lambda phage genomic DNA. Raw WGBS data can be found on the NCBI

Gene Expression Omnibus (GSE107177). Previously published whole genome bisulfite sequencing (WGBS)/MethylC-seq libraries for *Apis mellifera* (Herb et al., 2012), *Bombyx mori* (Xiang et al., 2010), *Nicrophorus vespilloides* (Cunningham et al., 2015), and *Zootermopsis nevadensis (Glastad et al., 2016)* were downloaded from the Short Read Archive (SRA) using accessions SRR445803–4, SRR027157–9, SRR2017555, and SRR3139749, respectively. Libraries were subjected to identical methylation analysis as *P. pyralis*.

Weighted DNA methylation was calculated for CG sites by dividing the total number of aligned methylated reads by the total number of methylated plus un-methylated reads (Schultz et al., 2012). For genic metaplots, the gene body (start to stop codon), 1000 base pairs (bp) upstream, and 1000 bp downstream was divided into 20 windows proportional windows based on sequence length (bp). Weighted DNA methylation was calculated for each window and then plotted in R (v3.2.4) (Team and Others, 2013).

1.13 Telomere FISH analysis

We synthesized a 5' fluorescein-tagged (TTAGG)5 oligo probe (FAM; Integrated DNA Technologies) for fluorescence in situ hybridization (FISH). We conducted FISH on squashed larval tissues according to previously published methods (Larracuente and Ferree, 2015), with some modification. Briefly, we dissected larvae in 1X PBS and treated tissues with a hypotonic solution (0.5% Sodium citrate) for 7 min. We transferred treated larval tissues to 45% acetic acid for 30 s, fixed in 2.5% paraformaldehyde in 45% acetic acid for 10 min, squashed, and dehydrated in 100% ethanol. We treated dehydrated slides with detergent (1% SDS), dehydrated again in ethanol, and then stored until hybridization. We hybridized slides with probe overnight at 30°C, washed in 4X SSCT and 0.1X SSC at 30°C for 15 min per wash. Slides were mounted in VectaShield with DAPI (Vector Laboratories), visualized on a Leica DM5500 upright fluorescence microscope at 100X, imaged with a Hamamatsu Orca R2 CCD camera. Images were captured and analyzed using Leica's LAX software.

Supporting Information 2

Aquatica lateralis additional information

2.1 Taxonomy, biology, and life history

Aquatica lateralis (Motschulsky, 1860) (Japanese name, Heike-botaru / ヘイケボタル) is one of the most common and popular luminous insects in mainland Japan. This species is a member of the subfamily Luciolinae and had long belonged in the genus *Luciola*, but was recently moved to the new genus *Aquatica* with some other Asian aquatic fireflies (Fu et al., 2010).

The life cycle of *A. lateralis* is usually 1 year. Aquatic larva possesses a pair of outer gills on each abdominal segment and live in still or slow streams near rice paddies, wetlands and ponds. Larvae mainly feed on freshwater snails. They pupate in a mud cocoon under the soil near the water. Adults emerge in early to end of summer. While both males and females are full-winged and can fly, there is sexual dimorphism in adult size: the body length is about 9 mm in males and 12 mm in females (Ohba, 2004).

Like other firefly larvae, *A. lateralis* larvae are bioluminescent. Larvae possess a pair of lanterns at the dorsal margin of the abdominal segment 8. Adults are also luminescent and possess lanterns at true abdominal segments 6 and 7 in males and at segment six in females (Branham and Wenzel, 2003; Kanda, 1935; Ohba, 2004). The adult is dusk active. Male adults flash yellow-green for about 1.0 s in duration every 0.5–1.0 s while flying ~1 m above the ground. Female adults, located on low grass, respond to the male signal with flashes of 1–2 s in duration every 3–6 s. Males immediately approach females and copulate on the grass (Ohba, 2004, 1983). Like many other fireflies, *A. lateralis* is likely toxic: both adults and larvae emit an unpleasant smell when disturbed and both invertebrate (dragonfly) and vertebrate (goby) predators vomit up the larva after biting (Ohba and Hidaka, 2002). *A. lateralis* larvae have eversible glands on each of the eight abdominal segments (Fu et al., 2010). The contents of the eversible glands is perhaps similar to that reported for *A. leii* (Fu et al., 2007).

2.2 Species distribution

The geographical range of *A. lateralis* includes Siberia, Northeast China, Kuril Islands, Korea, and Japan (Hokkaido, Honshu, Shikoku, Kyushu, Tsushima Isls.) (Kawashima et al., 2003). Natural habitats of these Japanese fireflies have been gradually destroyed through human activity, and currently these species can be regarded as 'flagship species' for conservation (Higuchi, 1996). For example, in 2017, Japanese Ministry of Environment began efforts to protect the population of *A. lateralis* in the Imperial Palace, Tokyo, where 3000 larvae cultured in an aquarium were released in the pond beside the Palace (Imperial Palace Outer Garden Management Office, 2017).

2.3 Specimen collection

Individuals used for genome sequencing, RNA sequencing, and LC-HRAM-MS were derived from a small population of laboratory-reared fireflies. This population was established from a few individuals collected from rice paddy in Kanagawa Prefecture of Japan in 1989 and 1990 (Ikeya, 2016) by Mr. Haruyoshi Ikeya, a highschool teacher in Yokohama, Japan. Mr. Ikeya collected adult *A. lateralis* specimens from their natural habitat in Yokohama and has propagated them for over 25 years (~25 generations) in a laboratory aquarium without any addition of wild individuals. This population has since been propagated in the laboratory of YO and JKW, and is dubbed the 'Ikeya-Y90' cultivar. Because of the small number of individuals used to establish the population and the number of generations of propagation, this population likely represents a partially inbred strain. Larvae were kept in aquarium at 19–21°C and fed using freshwater snails (*Physella acuta and Indoplanorbis exustus).* Under laboratory rearing conditions, the life cycle is reduced to 7–8 months. The original habitat of this strain has been destroyed and the wild population which led to the laboratory strain is now extinct.

2.4 Karyotype and genome size

Unlike *P. pyralis*, the karyotype of *A. lateralis* is reported to be 2n = 16 with XY sex determination (male, 14A + XY; female, 14A + XX) (Inoue and Yamamoto, 1987). The Y chromosome is much smaller than X chromosome, and the typical behaviors of XY chromosomes, such as partial conjugation of X/Y at the first meiotic metaphase and a separation delay of X/Y at the first meiotic anaphase, were observed in testis cells (Inoue and Yamamoto, 1987).

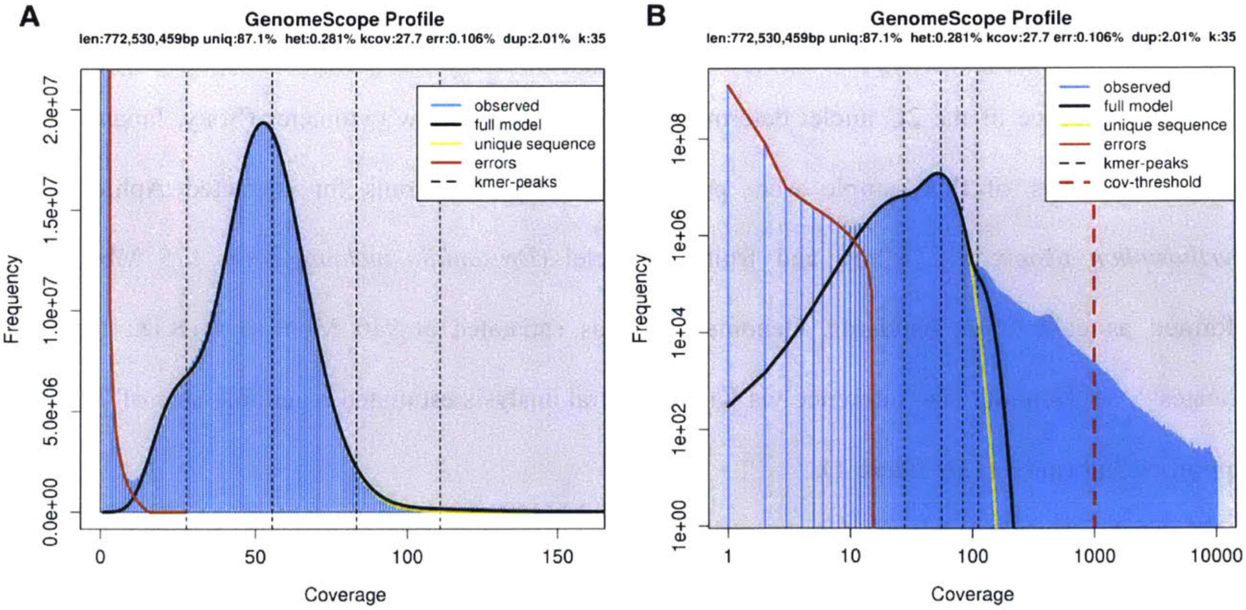
We determined the genome size of *A. lateralis* using flow cytometry-mediated calibrated-fluorimetry of DNA content with propidium iodide stained nuclei. First, the head+prothorax of a single pupal female (gender identified by morphological differences in abdominal segment VIII) was homogenized in 100 µL PBS. These tissues were chosen to avoid the ovary tissue. Once homogenized, 900 µL PBS, 1 µL Triton X-100 (Sigma-Aldrich), and 4 µL 100 mg/mL RNase A (QIAGEN) were added. The homogenate was incubated at 4°C for 15 min, filtered with a 30 µm Cell Tries filter (Sysmex), and further diluted with 1 mL PBS. 20 µL of 0.5 mg/mL propidium iodide was added to the mixture and then average fluorescence of the 2C nuclei determined with a SH-800 flow cytometer (Sony, Japan). Three technical replicates of this sample were performed. Independent runs for extracted Aphid nuclei (*Acyrthosiphon pisum*; 517 Mbp), and fruit fly nuclei (*Drosophila melanogaster*; 175 Mbp) were performed as calibration standards. Genome size was estimated at 940 Mbp ±1.4 (S.D.; technical replicates = 3). Genome size inference via Kmer spectral analysis estimated a genome size of 772 Mbp (Supporting Information 2—figure 1).

2.5 Genomic sequencing and assembly

Genomic DNA was extracted from the whole body of a single laboratory-reared *A. lateralis* adult female (c.v. Ikeya-Y90) using the QIAamp Kit (Qiagen). Purified DNA was fragmented with a Covaris S2 sonicator (Covaris, Woburn, MA), size-selected with a Pippin Prep (Sage Science, Beverly, MA), and then used to create two paired-end libraries using the TruSeq Nano Sample Preparation Kit (Illumina) with insert sizes of ~200 and~800 bp. These libraries were sequenced on an Illumina HiSeq 1500 using a

125 × 125 paired-end sequencing protocol. Mate-pair libraries of 2–20 Kb with a peak at ~5 Kb were created from the same genomic DNA using the Nextera Mate Pair Sample Preparation Kit (FC-132–1001, Illumina), and sequenced on HiSeq 1500 using a 100 × 100 paired-end sequencing protocol at the NIBB Functional Genomics Facility (Aichi, Japan). In total, 133.3 Gb of sequence (159x) was generated.

Reads were assembled using ALLPATHS-LG (build# 48546) (Gnerre et al., 2011), with default parameters and the 'HAPLOIDIFY = True' option. Scaffolds were filtered to remove non-firefly contaminant sequences using blobtools (Laetsch and Blaxter, 2017), resulting in the final assembly (Alat1.3). The final assembly (Alat1.3) consists of 5388 scaffolds totaling 908.5 Gbp with an N50 length of 693.0 Kbp, corresponding to 96.6% of the predicted genome size of 940 Mbp based on flow cytometry (Supporting Information 2.4). Genome sequencing library statistics are available in Supporting Information 4—table 1.



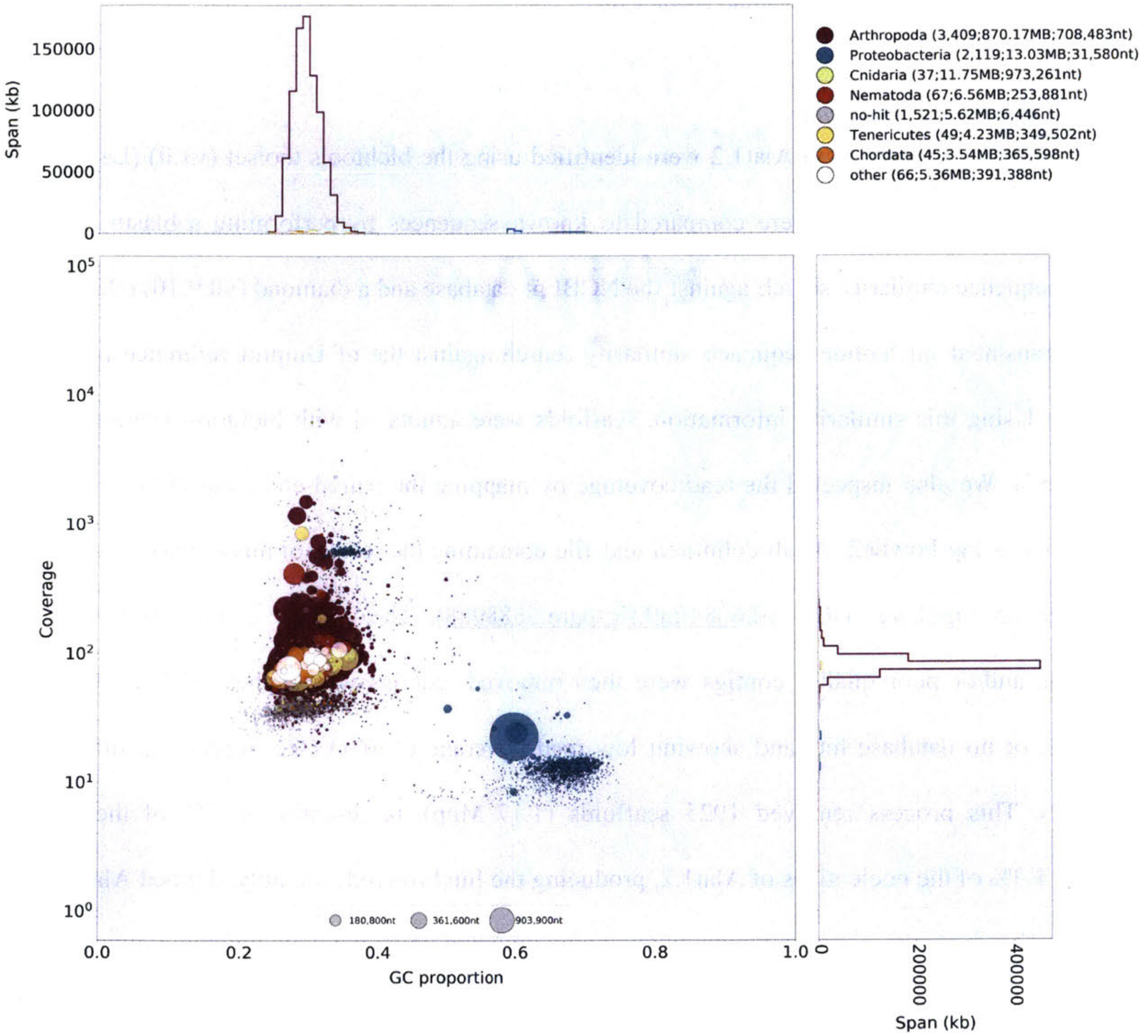
Supporting Information 2—figure 1. Genome scope kmer analysis of the A. lateralis short-insert genomic library.

(**A**) Linear and (**B**) log plot of a kmer spectral genome composition analysis of the 'FFGPE_PE200' *A. lateralis* Illumina short-insert library (Supporting Information 2.5; Supporting Information 4—table 1) with jellyfish (v2.2.9; parameters: -C -k 35) (Marçais and Kingsford, 2011) and GenomeScope (v1.0; parameters: Kmer length = 35, Read length = 100, Max kmer coverage = 1000) (Vurture et al., 2017). len = inferred haploid genome length, uniq = percentage non-repetitive sequence, het = overall rate of genome heterozygosity, kcov = mean kmer coverage for heterozygous bases, err = error rate of the reads, dup: average rate of read duplications. These results are consistent when considering the possible systematic error of kmer spectral analysis and flow cytometry genome size estimates.

The heterozygosity is lower than that measured for *P. pyralis*, possibly reflecting the long-term laboratory rearing in reduced population sizes of *A. lateralis* strain Ikeya-Y90.

2.5.2 Taxonomic annotation filtering

Potential contaminants in Alat1.2 were identified using the blobtools toolset (v1.0) (Laetsch and Blaxter, 2017). First, scaffolds were compared to known sequences by performing a blastn (v2.5.0+) nucleotide sequence similarity search against the NCBI nt database and a diamond (v0.9.10) (Buchfink et al., 2015) translated nucleotide sequence similarity search against the of Uniprot reference proteomes (July 2017). Using this similarity information, scaffolds were annotated with blobtools (parameters '-x bestsumorder'). We also inspected the read coverage by mapping the paired-end reads (FFGPE_PE200) on the genome using bowtie2. A tab delimited text file containing the results of this blobtools annotation are available on FigShare (DOI: 10.6084/m9.figshare.5688928). The contigs derived from potential contaminants and/or poor quality contigs were then removed: contigs with higher %GC (>50%) with bacterial hits or no database hits and showing low read coverage (<30 x) (see Supporting Information 2—figure 2). This process removed 1925 scaffolds (1.17 Mbp), representing 26.3% of the scaffold number and 1.3% of the nucleotides of Alat1.2, producing the final filtered assembly, dubbed Alat1.3.



Supporting Information 2—figure 2. BlobPlot of A. lateralis Illumina reads aligned against Alat1.2. Coverage shown represents mean coverage of reads from the Illumina short-insert library (Sample name FFGPE_PE200; Supporting Information 4—table 1), aligned against Alat1.2 using Bowtie2. Scaffolds were taxonomically annotated as described in Supporting Information 2.5.2.

2.6 RNA-extraction, library preparation and sequencing

In order to capture transcripts from diverse life-stages and tissues, non-stranded RNA-Seq libraries were prepared from fresh specimens of nine life stages/sexes/tissues (eggs, fifth (the last) instar larvae, both sex of pupae, adult male head, male abdomen (prothorax-to-fifth segment), male lantern, adult female head, and female lantern (Supporting Information 2—table 1). Live specimens were anesthetized on ice and dissected during the day. The lantern tissue was dissected from the abdomen and contains the cuticle, photocyte layer and reflector layer. For eggs, larvae, and pupae, total RNA was extracted using the RNeasy Mini Kit (QIAGEN) with the optional on-column DNase treatment. For adult specimens, total RNA was extracted using TRIzol reagent (Invitrogen) to avoid contamination of pigments and uric acid. These were then treated with DNase in solution and then cleaned using a RNeasy Mini kit.

cDNA libraries were generated from purified Total RNA (500 ng from each sample) using a TruSeq RNA Sample Preparation Kit v2 (Illumina) according to the manufacturer's protocol (Low-Throughput Protocol), except that all reactions were carried out at half scale. The fragmentation of mRNA was performed for 4 min. The enrichment PCR was done using six cycles. A subset of nine libraries (BdM1, HeF1, HeM1, LtF1, LtM1, Egg1, Lrv1, PpEF, PpLM; Supporting Information 2—table 1) were multiplexed and sequenced in a single lane of Hiseq 1500 101 × 101 bp paired-end reads. The remaining 23 libraries (BdM2, BdM3, HeF2, HeF3, HeM2, HeM3, LtF2, LtF3, LtM2, LtM3, WAF1, WAF2, WAF3, WAM1, WAM2, WAM3, Egg2, Lrv2, Lrv3, PpEM, PpLF, PpMF, PpMM) were multiplexed and sequenced in two lanes of Hiseq 1500 66 bp single-end reads. Sequence quality was inspected with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Supporting Information 2—table 1. ***Aquatica lateralis*** **RNA sequencing.**
N: number of individuals pooled for sequencing; **Sex/stage**: M = male, F = female, A = adult, L = larva, L = larvae, E = Eggs, p=Pupae, P-E = Pupae early, P-M = Pupae middle, P-L = Pupae late; **Tissue**: H = head, La = dissected lantern containing cuticle, photocyte layer and reflector layer, H = head, B = Thorax, plus abdomen excluding lantern containing segments. W = whole specimen. AEL = After egg laying.

Library name	Label	SRA ID	N	Sex/ Stage	Tissue	Library type
R102L6_idx13	BdM1	DRR119264	1	M/A	B	Illumina paired-end, non-stranded specific, PolyA
R128L1_idx25	BdM2	DRR119265	1	M/A	B	Illumina single-end, non-stranded specific, PolyA
R128L2_idx27	BdM3	DRR119266	1	M/A	B	Illumina single-end, non-stranded specific, PolyA

R102L6_idx15	HeF1	DRR119267	3	F/A	H	Illumina paired-end, non-stranded specific, PolyA
R128L1_idx22	HeF2	DRR119268	3	F/A	H	Illumina single-end, non-stranded specific, PolyA
R128L2_idx23	HeF3	DRR119269	3	F/A	H	Illumina single-end, non-stranded specific, PolyA
R102L6_idx12	HeM1	DRR119270	2	M/A	H	Illumina paired-end, non-stranded specific, PolyA
R128L1_idx20	HeM2	DRR119271	2	M/A	H	Illumina single-end, non-stranded specific, PolyA
R128L2_idx21	HeM3	DRR119272	2	M/A	H	Illumina single-end, non-stranded specific, PolyA
R102L6_idx16	LtF1	DRR119273	5	F/A	La	Illumina paired-end, non-stranded specific, PolyA
R128L1_idx06	LtF2	DRR119274	5	F/A	La	Illumina single-end, non-stranded specific, PolyA
R128L2_idx12	LtF3	DRR119275	5	F/A	La	Illumina single-end, non-stranded specific, PolyA
R102L6_idx14	LtM1	DRR119276	5	M/A	La	Illumina paired-end, non-stranded specific, PolyA
R128L1_idx05	LtM2	DRR119277	5	M/A	La	Illumina single-end, non-stranded specific, PolyA
R128L2_idx19	LtM3	DRR119278	5	M/A	La	Illumina single-end, non-stranded specific, PolyA
R128L2_idx15	WAF1	DRR119279	1	F/A	W	Illumina single-end, non-stranded specific, PolyA
R128L1_idx16	WAF2	DRR119280	1	F/A	W	Illumina single-end, non-stranded specific, PolyA
R128L2_idx18	WAF3	DRR119281	1	F/A	W	Illumina single-end, non-stranded specific, PolyA

R128L1_idx11	WAM1	DRR119282	1	M/A	W	Illumina single-end, non-stranded specific, PolyA
R128L2_idx13	WAM2	DRR119283	1	M/A	W	Illumina single-end, non-stranded specific, PolyA
R128L1_idx14	WAM3	DRR119284	1	M/A	W	Illumina single-end, non-stranded specific, PolyA
R102L6_idx4	Egg1	DRR119285	19.6 mg (~30–50)	E ~6 hr AEL	W	Illumina paired-end, non-stranded specific, PolyA
R128L1_idx01	Egg2	DRR119286	21.6 mg (~30–50)	E ~7 d AEL	W	Illumina single-end, non-stranded specific, PolyA
R102L6_idx5	Lrv1	DRR119287	1	L	W	Illumina paired-end, non-stranded specific, PolyA
R128L1_idx03	Lrv2	DRR119288	1	L	W	Illumina single-end, non-stranded specific, PolyA
R128L2_idx04	Lrv3	DRR119289	1	L	W	Illumina single-end, non-stranded specific, PolyA
R128L1_idx07	PpEM	DRR119290	1	M/P-E	W	Illumina single-end, non-stranded specific, PolyA
R128L2_idx10	PpLF	DRR119291	1	F/P-L	W	Illumina single-end, non-stranded specific, PolyA
R128L1_idx09	PpMF	DRR119292	1	F/P-M	W	Illumina single-end, non-stranded specific, PolyA
R128L2_idx08	PpMM	DRR119293	1	M/P-M	W	Illumina single-end, non-stranded specific, PolyA
R102L6_idx7	PpEF	DRR119294	1	F/P-E	W	Illumina paired-end, non-stranded specific, PolyA
R102L6_idx6	PpLM	DRR119295	1	M/P-L	W	Illumina paired-end, non-stranded specific, PolyA

2.7 Transcriptome analysis

2.7.1 De novo transcriptome assembly and alignment

To build a comprehensive set of reference transcript sequences, reads derived from the pool of nine libraries (BdM1, HeF1, HeM1, LtF1, LtM1, Egg1, Lrv1, PpEF, PpLM; Supporting Information 2—table 1) were pooled. These represent RNA prepared from various tissues (head, thorax + abdomen, lantern) and stages (egg, pupae, adult) of both sexes. A non strand-specific *de novo* transcriptome assembly was produced with Trinity (v2.6.6) (Grabherr et al., 2011) using default parameters exception the following: (--min_glue 2 min_kmer_cov 2 --jaccard_clip --no_normalize_reads --trimmomatic). Peptides were predicted from the *de novo* transcripts via Transdecoder (v5.3.0; default parameters). *De novo* transcripts were then aligned to the *A. lateralis* genome (Alat1.3) using the PASA pipeline with blat (v36 × 2) and gmap (v2018-05-03) (--aligners blat,gmap), parameters for alternative splice analysis and strand specificity (--ALT_SPLICE --transcribed_is_aligned_orient), and input of the previously extracted Trinity accessions (--tdn tdn.accs). Importantly, it was necessary to set (--NUM_BP_PERFECT_SPLICE_BOUNDARY = 0) for the validate_alignments_in_db.dbi step, to ensure transcripts with natural variation near the splice sites were not discarded. Direct coding gene models (DCGMs) were then produced with the Transdecoder 'cdna_alignment_orf_to_genome_orf.pl' utility script, with the PASA assembly GFF and transdecoder predicted peptide GFF as input. The unaligned *de novo* transcriptome assembly is dubbed 'AQULA_Trinity_unstranded', whereas the aligned direct coding gene models are dubbed 'Alat1.3_Trinity_unstranded-DCGM'.

2.7.2 Reference guided transcriptome alignment and assembly

A reference guided transcriptome was produced from all available *A.lateralis* RNA-seq reads (Supporting Information 2—table 1) using HISAT2 (v2.1.0) (Kim et al., 2015) and StringTie (v1.3.3b) (Pertea et al., 2015). Reads were first mapped to the *A. lateralis genome* (Alat1.3) with HISAT2 (parameters: -X 2000 --dta --fr). Then StringTie assemblies were performed on each separate bam file corresponding to the original libraries using default parameters. Finally, the produced. GTF files were merged using StringTie (--merge). A transcript fasta file was produced from the StringTie GTF file with the transdecoder 'gtf_genome_to_cdna_fasta.pl' utility script, and peptides were predicted for these transcripts using Transdecoder (v5.3.0) with default parameters. The StringTie GTF was converted to GFF format with the Transdecoder 'gtf_to_alignment_gff3.pl' utility script, and direct coding gene models (DCGMs) were then produced with the Transdecoder 'cdna_alignment_orf_to_genome_orf.pl' utility script, with the StringTie-provided GFF and transdecoder predicted peptide GFF as input. The reference guided transcriptome assembly was dubbed 'AQULA_Stringtie_unstranded', whereas the aligned direct coding gene models were dubbed 'Alat1.3_Stringtie_unstranded-DCGM'.

2.7.3 Transcript expression analysis

A. lateralis RNA-Seq reads (Supporting Information 2—table 1) were pseudoaligned to the AQULA_OGS1.0 geneset mRNAs using Kallisto (v0.43.1) (Bray et al., 2016) with 100 bootstraps (-b 100), producing transcripts-per-million reads (TPM). Kallisto expression quantification analysis results are available on FigShare (DOI: 10.6084/m9.figshare.5715139).

2.8 Official coding geneset annotation (AQULA_OGS1.0)

A protein-coding gene reference set for *A. lateralis* was generated by Evidence Modeler (v1.1.1) using both aligned transcripts and aligned proteins. For transcripts, we combined reference guided and *de novo* transcriptome assembly approaches. Notably, these reference guided and *de novo* transcriptome assembly approaches differed from the current *de novo* (Supporting Information 2.7.1) and reference guided (Supporting Information 2.7.2) transcriptome assembly approaches. In the reference-guided approach applied here, RNA-Seq reads were mapped to the genome assembly with TopHat and assembled into transcripts with Cufflinks (parameters: --min-intron-length 30) (Trapnell et al., 2010). The Cufflinks transcripts were subjected to the TransDecoder program to extract ORFs. In the *de novo transcriptome* approach applied here, RNA-seq reads were assembled *de novo* by Trinity and ORFs were predicted using TransDecoder. We used CD-HIT-EST (Li and Godzik, 2006) to reduce the redundancy of the predicted

ORFs. The ORF sequences were mapped to the genome using Exonerate in est2genome mode for splice-aware alignment. We processed homology evidence at the protein level using the reference proteomes of *D. melanogaster* and *T. castaneum*. These reference proteins were split-mapped to the *A. lateralis* genome in two steps: first with BLASTX to find approximate loci, and then with Exonerate in protein2genome mode to obtain more refined alignments. These gene models derived from multiple evidence were merged by the EVM program to obtain the reference annotation for the genomes. We also predicted *ab initio* gene models using Augustus, but we didn't include Augustus models for the EVM integration because our preliminary analysis showed the *ab initio* gene models had no positive impact on gene prediction.

Lastly, gene models for luciferase homologs, P450s, and *de novo* methyltransferases (DNMTs) which were fragmented or were incorrect (e.g. fusions of adjacent genes) were manually corrected based on the evidence of the *de novo* and reference guided direct coding gene models. Manual correction was performed by performing TBLASTN searches with known good genes from these gene families within SequenceServer(v1.10.11) (Priyam et al., 2015), converting the TBLASTN results to gff3 format with a custom script (https://github.com/photocyte/general_scripts/blob/master/blastxml2gff.py), and viewing these alignments alongside the alternative direct coding gene models (Supporting Information 2.7.1; 2.7.2) in Integrative Genomics Viewer(v2.4.8) (Thorvaldsdóttir et al., 2013). The official gene set. gff3 file was manually modified in accordance with the alternative gene models. Different revision numbers of the official geneset (e.g. AQULA_OGS1.0, AQULA_OGS1.1) represent the improvement of the geneset over time due to these continuing manual gene annotations.

2.9 Repeat annotation

A *de novo* species-specific repeat library for *A. lateralis* was constructed using RepeatModeler (v1.0.9), and Tandem Repeat Finder (v4.09; settings: 2 7 7 80 10) (Benson, 1999). Only tandem repeats from Tandem Repeat Finder with a repeat block length >5 kb (annotated as 'complex tandem repeat') were added to the RepeatModeler library. This process yielded a final library of 1695 interspersed repeats. We then used this library and RepeatMasker (v4.0.5) (http://repeatmasker.org/) to identify and mask interspersed and tandem repeats in the genome assembly. This repeat library is dubbed the *Aquatica lateralis* Official Repeat Library 1.0 (AQULA_ORL1.0).

Supporting Information 2—table 2. Annotated repetitive elements in *A. lateralis*.

Repeat class	Family	Counts	Bases	% of assembly
DNA	All	229064	73263593	8.06
	Helitrons	930	466679	0.051
LTR	All	59499	23391956	2.57
Non-LTR	All	151788	50394853	5.55
	LINE	151788	50394853	5.55
	SINE	0	0	0
Unknown interspersed		450934	99998958	11.01
Complex tandem repeats		295	33237	0.004
Simple repeat		155265	6656757	0.73
rRNA		0	0	0

Supporting Information 3

Ignelater luminosus additional information

3.1 Taxonomy, biology, and life history

Ignelater luminosus is a member of the beetle family Elateridae ('click beetles'), related to Lampyridae within the superfamily Elateroidea. The Elateridae includes about 10,000 species (Slipinski, S. A., Leschen, R. A. B. & Lawrence, J. F., 2011) (17 subfamilies) (Costa, C., Lawrence, J. F. & Rosa, S. P., 2010), which are widespread throughout the globe. Unlike in fireflies, where bioluminescence is universal, only ~200 described elaterid species are luminous. These luminous species are recorded only from tropical and subtropical regions of Americas and some small Melanesian islands, such as Fiji and Vanuatu (Costa, 1975; Costa, C., Lawrence, J. F. & Rosa, S. P., 2010). For instance, the tropical American *Pyrophorus noctilucus is* considered the largest (~30 mm) and brightest bioluminescent insect (Harvey and Stevens, 1928; Levy, 1998); Levy, 1998). All luminous species are closely related - luminous click beetles belong to the tribes Pyrophorini and Euplinthini (Arias-Bohart, 2015; Costa, 1975) of the subfamily Agrypninae, with the single exception of *Campyloxenus pyrothorax* (Chile) in the related subfamily Campyloxeninae (Stibick, 1979). The luminescence of a pair of pronotal 'light organs' of the adult *Balgus schnusei* (Costa, 1984), a species that has now been assigned to the Thylacosterninae of the Elateridae (Costa, C., Lawrence, J. F. & Rosa, S. P., 2010), has not been confirmed by later observation. This near-monophyly of bioluminescent elaterid taxa is supported by both morphological (Douglas, 2011) and molecular phylogenetic analysis (Kundrata and Bocak, 2011; Sagegami-Oba et al., 2007), although early morphological phylogenies were inconsistent (Dolin, 1978; Hyslop, 1917; Ôhira, 2013, 1962; Stibick, 1979). This suggests a single origin of bioluminescence in this family.

The genus *Ignelater* was established by Costa in 1975 and *I. luminosus* was included in this genus (Costa, 1975). Often this species is called *Pyrophorus luminosus* as an 'auctorum', a name used to describe a variety of taxa (Johnson, 2002). This use of 'Pyrophorus' as an auctorum may be due to the heightened difficulty of classifying Elateridae (Costa, 1975). The genus *Ignelater* is characterized by the presence of both dorsal and ventral photophores (Costa, 1975; Rosa, 2007). An unreviewed report suggested that the adult *I. luminosus* has a ventral light organ only in males (Reyes and Lee, 2010). Phylogenetic analyses based on the morphological characters suggested that the genera *Ignelater* and *Photophorus* (which contain only two species from Fiji and Vanuatu) are the most closely related genera in the tribe Pyrophorini (Rosa, 2007). The earliest fossil of an Elateridae species was recorded from the Middle Jurassic of Inner Mongolia, China (Chang et al., 2009). McKenna and Farrell suggested that, based on molecular analyses, the family Elateridae originated in the Early Cretaceous (130 Mya) (McKenna and Farrell, 2009). It is expected that many recent genera in Elateroidea were established by the Early Tertiary (<65 Mya) (Grimaldi and Engel, 2005).

The exact function of bioluminescence across different life stages remains unknown for many luminous elaterid species. Bioluminescent elaterid beetles typically have two paired lanterns on the dorsal surface of the prothorax, and a single lantern on the ventral abdomen, which is only exposed during flight. Several bioluminescent Elateridae produce different colored luminescence from their prothorax and abdominal lanterns (Feder and Velez, 2009; Oba et al., 2010a). Harvey reported that there was not a marked difference in the luminescence color of the dorsal and ventral lanterns of Puerto Rican *I. luminosus* (Harvey, 1952). Like fireflies, elaterid larvae often produce light, with the glowing termite mounds of Brazil that contain the predatory larvae of *Pyrearinus termitilluminans* being a striking example (Costa and Vanin, 2010). A description of the anatomy of the larval light organ of *Pyrophorus* is provided by (Harvey, 1952), and a more modern photograph of the larval light organ is provided by (Bechara and Stevani, 2018). Like other bioluminescent elaterid larvae, *I. luminosus* larvae produce a diffuse light from their prothorax, however they are only luminous when disturbed (Wolcott, 1948). *I. luminosus* larvae are subterranean predators and are an enthusiastic predator of the white grub (*Ancylonycha* spp.), reportedly consuming 50 + to reach full size (Wolcott, 1950). Adult *I. luminosus* are luminescent and a bioluminescent courtship behavior was described in an unreviewed study (Kretsch, 2000). Reportedly, males search during flight with their prothorax lanterns illuminated steadily, while females stay on the ground modulating the intensity of their prothorax lanterns in ~2 s intervals. Once a female is observed, the prothorax lanterns of the male go dark, the ventral lantern becomes illuminated, and the male approaches the female via a circular search pattern. Mating is brief, reportedly taking only 5 seconds.

Unlike fireflies, bioluminescent elaterid species are not known to have potent chemical defenses. For example, the Jamaican bioluminescent elaterid beetle *Pyrophorus plagiophthalmus*, does not appear to be strongly unpalatable, as bats were observed to regularly capture the beetles during their flying bioluminescent displays (Vélez, 2006). A defense role for *I. luminosus* luminescence to startle predators is possible.

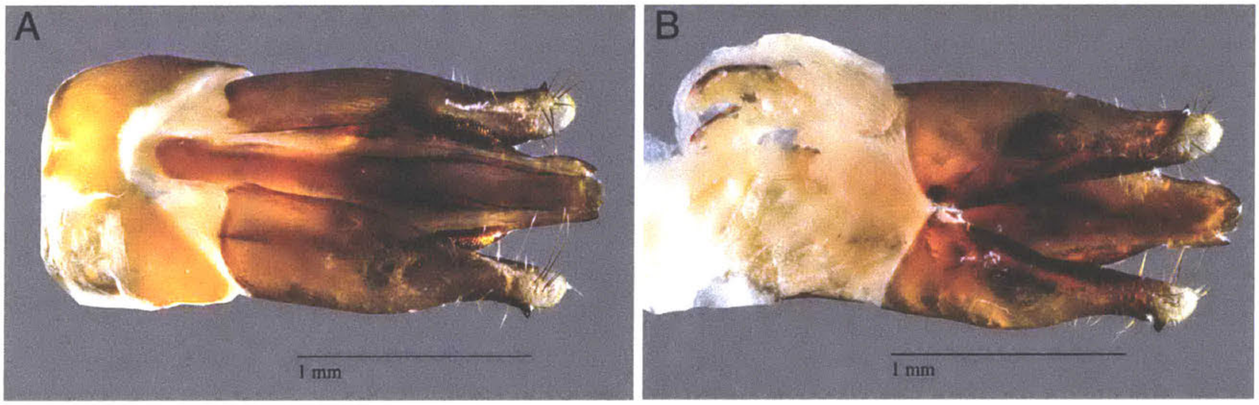
3.2 Species distribution

I. luminosus is often considered to be endemic to Puerto Rico (Virkki et al., 1984); however, the genus *Ignelater* is reported in Florida (USA), Vera Cruz (Mexico), the Bahamas, Cuba, Isla de la Juventud, Hispaniola (Haiti + Dominican Republic), Puerto Rico, and the Lesser Antilles (Costa, 1975). Similarly, *I. luminosus* itself has been reported on the island of Hispaniola (Kretsch, 2000; Perez-Gelabert, 2008), indicating *I. luminosus* is not restricted to Puerto Rico. This geographic distribution of *Ignelater* suggests that Puerto Rico may contain multiple *Ignelater* species and, given the difficulty of distinguishing different species of bioluminescent Elateridae by morphological characters, a definitive species distribution for *I. luminous* cannot be stated, other than this species is seemingly not strictly endemic to Puerto Rico.

3.3 Collection

I. luminosus (Illiger, 1807) adult specimens were collected from private land in Mayagüez, Puerto Rico (18° 13' 12.1974' N, 67° 6' 31.6866' W) with permission of the landowner by Dr. David Jenkins

(USDA-ARS). Individuals were captured at night on April 20th and April 28th 2015 during flight on the basis of light production. The *I. luminosus* specimens were frozen in a −80°C freezer, lyophilized, shipped to the laboratory (MIT) on dry ice, and stored at −80 °C. Full collection metadata is available from the NCBI BioSample records of these specimens (NCBI Bioproject PRJNA418169). Identification to species was performed by comparing antenna and dissected genitalia morphology to published keys (Costa, 1975; Rosa, 2010, 2007) (Supporting Information 3—figure 1). All inspected specimens were male (3/3). Specimens collected at the same time, but not those used for genitalial dissection, were used for sequencing. Although the genitalia morphology of the sequenced specimens was not inspected to confirm their sex, sequenced specimens were inferred to be male, based on the fact that female bioluminescent elaterid beetles are rarely seen in flight (Personal communication: S. Velez) and the dissected specimens collected in the same batch as the sequenced specimens were confirmed to be male.



Supporting Information 3—figure 1. Ignelater luminosus aedeagus (male genitalia).
(**A**) Dorsal and (**B**) ventral view of an *Ignelater luminosus* aedeagus, dissected from the same batch of specimens used for linked-read sequencing and genome assembly. The species identity of this specimen was confirmed as *I. luminosus* by comparison of the aedeagus to the keys of Costa and Rosa (Costa, 1975; Rosa, 2010, 2007).

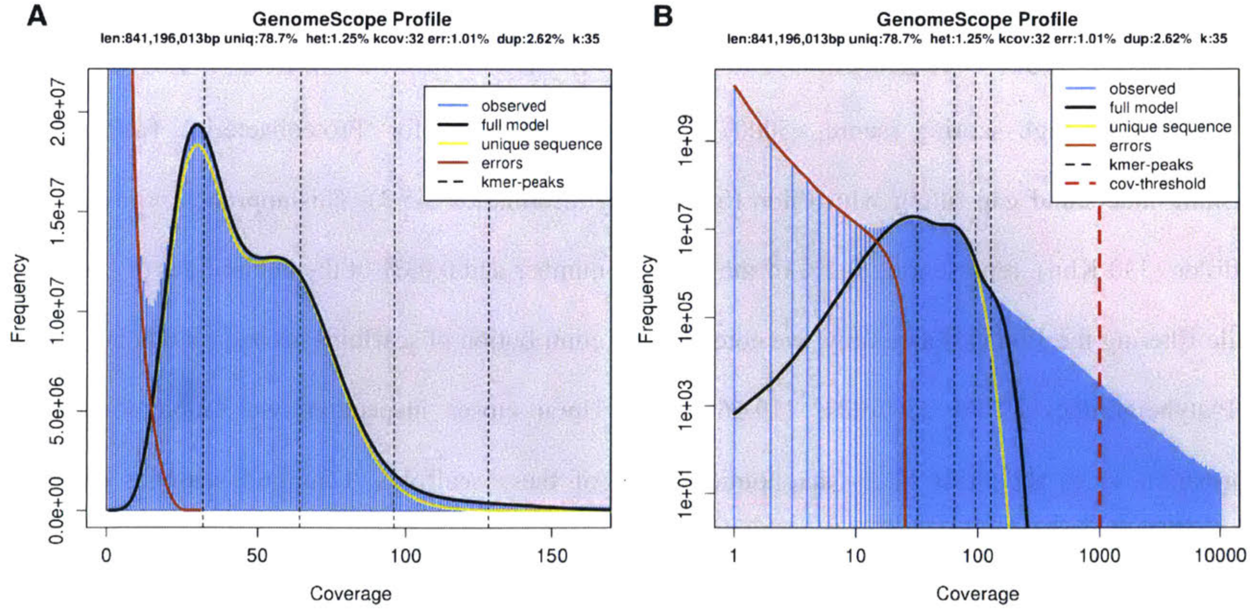
3.4 Karyotype and genome size

The karyotype of male Puerto Rican *I. luminosus* (as *Pyrophorus luminosus*) was reported as 2n = 14A + X1X2Y (Virkki et al., 1984). The genome sizes of 5 male *I. luminosus* were determined by flow cytometry-mediated calibrated-fluorimetry of DNA content with propidium iodide stained nuclei by Dr. J.

Spencer Johnston (Texas A&M University). The frozen head of each individual was placed into 1 mL of cold Galbraith buffer in a 1 mL Kontes Dounce Tissue Grinder along with the head of a female *Drosophila virilis* standard (1C = 328 Mbp). The nuclei from the sample and standard were released with 15 strokes of the 'B' (loose) pestle, filtered through 40 μm Nylon mesh, and stained with 25 mg/mL Propidium Iodide (PI). After a minimum of 30 min staining in the dark and cold, the average fluorescence channel number for the PI (red) fluorescence of the 2C (diploid) nuclei of the sample and standard were determined using a CytoFlex Flow Cytometer (Beckman-Coulter). The 1C amount of DNA in each sample was determined as the ratio of the 2C channel number of the sample and standard times 328 Mbp. The genome size of these *I. luminosus* males was determined to be 764 ± 7 Mbp (SEM, n = 5). Genome size inference via Kmer spectral analysis of the *I. luminosus* linked-read data estimated a genome size of 841 Mbp (Supporting Information 3—figure 2).

3.5 Genomic sequencing and assembly

HMW DNA (25 μg) was extracted from a single male specimen of *I. luminosus*using a 100/G Genomic Tip with the Genomic buffers kit (Qiagen, USA). The *I. luminosus* specimen was first washed with 95% ethanol, and DNA was extracted following the manufacturer's protocol, with the exception of the final precipitation step, where HMW DNA was pelleted with 40 μg RNA grade glycogen (Thermo Scientific, USA) and centrifugation (3000 x g, 30 min, 4°C) instead of spooling on a glass rod. HMW DNA was sent on dry-ice to the Hudson Alpha Institute of Biotechnology Genomic Services Lab (HAIB-GSL), where pulsed-field-gel-electrophoresis (PFGE) quality control and 10x Genomics Chromium Genome v1 library construction was performed. PFGE quality control indicated the mean size of the input DNA was >35 kbp+. The resulting library was then sequenced on one HiSeqX lane. 408,838,927 paired reads (150 × 150 PE) were produced, corresponding to a genomic coverage of 153x. To evaluate the effect of different Illumina instruments on data and assembly quality, the library was also sequenced on one HiSeq2500 lane, where 145,250,480 reads (150 × 150 PE) were produced, corresponding to a genomic coverage of 54x. A summary of the library statistics for the genomic sequencing is available in Supporting Information 4—table 1. The draft genome of *I. luminosus* (Ilumi1.0) was assembled from the obtained HiSeqX genomic sequencing reads using the Supernova assembler (v1.1.1) (Weisenfeld et al., 2017), on a 40 core 1 TB RAM server at the Whitehead Institute for Biomedical Research. The reported mean molecule size was 12.23 kbp. The assembly was exported to FASTA format using Supernova mkoutput (parameters: --style=pseudohap), and modified by taxonomic annotation filtering (Supporting Information 3.5.2) and polishing (Supporting Information 3.5.3) to form Ilumi1.1. A Supernova (v2.0.0) assembly was also produced from combined HiSeqX and HiSeq2500 reads, but on a brief inspection the quality was equivalent to Ilumi1.1, so the new assembly was not used for further analyses. Manual long-read based scaffolding was then applied to produce a final assembly Ilumi1.2 (Supporting Information 3.5.4).



Supporting Information 3—figure 2. Genome scope kmer analysis of the I. luminosus linked-read genomic library.

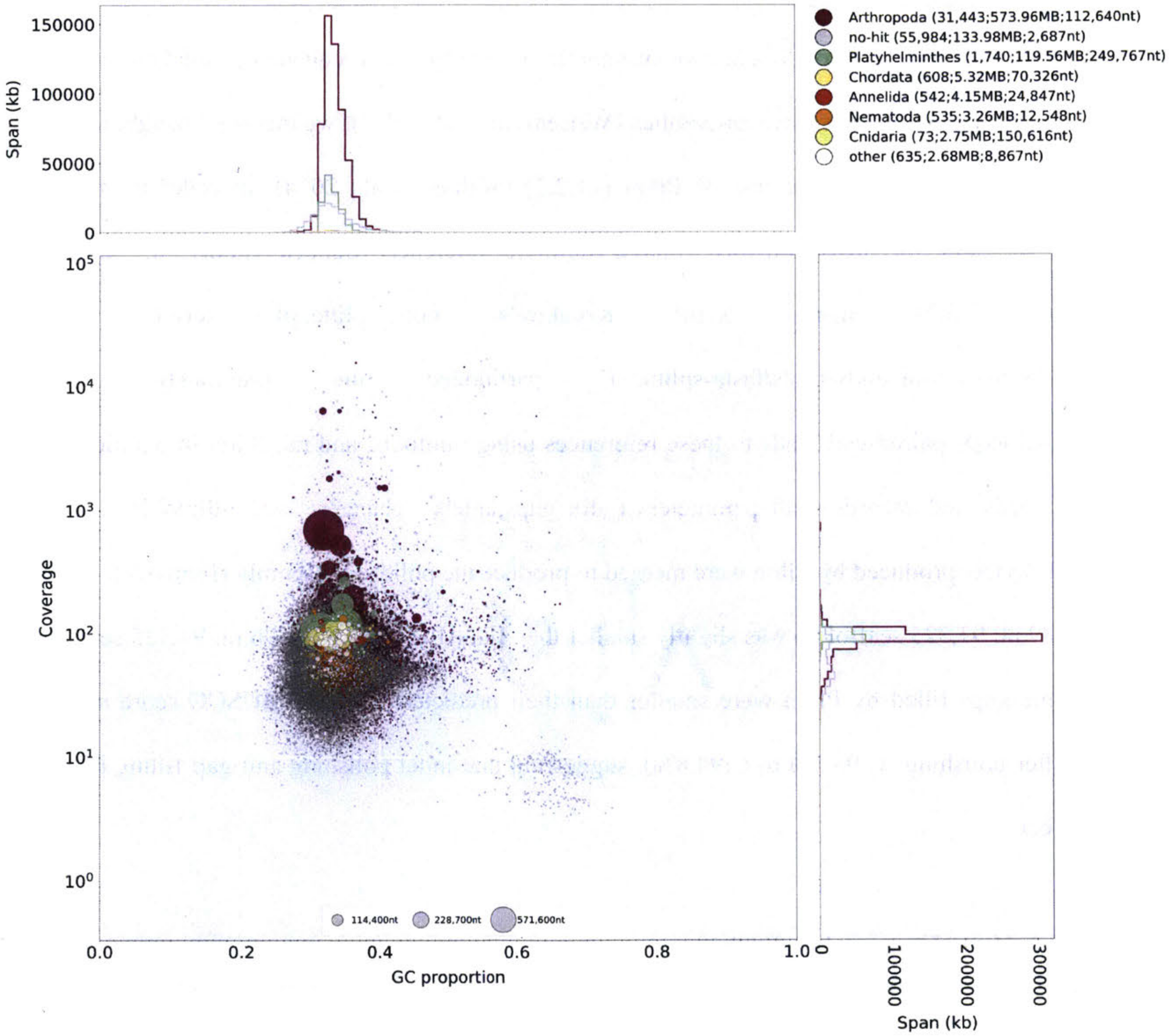
(**A**) Linear and (**B**) log plot of a kmer spectral genome composition analysis of the '1610_IlumiHiSeqX' *I. luminosus* Illumina linked-read library (Supporting Information 2.5; Supporting Information 4—table 1) with jellyfish (v2.2.9; parameters: -C -k 35) (Marçais and Kingsford, 2011) and GenomeScope (v1.0; parameters: Kmer length = 35, Read length = 138, Max kmer coverage = 1000) (Vurture et al., 2017). Before analysis, 10x Chromium barcodes were trimmed off Read1 using cutadapt (v1.8; parameters: -u 23) (Martin, 2011). vlen = inferred haploid genome length, uniq = percentage non-repetitive sequence, het = overall rate of genome heterozygosity, kcov = mean kmer coverage for heterozygous bases, err = error rate of the reads, dup: average rate of read duplications.

These results are consistent when considering the possible systematic error of kmer spectral analysis and flow cytometry genome size estimates. The heterozygosity is higher than that measured for *P. pyralis* and *A. lateralis*. The read error rate for this library is also significantly higher than the *P. pyralis* and *A. lateralis* results, possibly highlighting the difference in raw read error rate between HiSeq2500 and HiSeqX sequencing, or is possibly an artifact of the Chromium library.

3.5.2 Taxonomic annotation filtering

We sought to systematically remove assembled non-elaterid contaminant sequence from Ilumi1.0. Using the blobtools toolset (v1.0.1), (Laetsch and Blaxter, 2017), we taxonomically annotated our scaffolds by performing a blastn (v2.6.0+) nucleotide sequence similarity search against the NCBI nt database, and a diamond (v0.9.10.111) (Buchfink et al., 2015) translated nucleotide sequence similarity search against the of Uniprot reference proteomes (July 2017). Using this similarity information, we taxonomically annotated the scaffolds with blobtools using parameters '-x bestsumorder --rank phylum' (Supporting Information 3—figure 3). A tab delimited text file containing the results of this blobtools annotation is available on FigShare (DOI: 10.6084/m9.figshare.5688952). We then generated the final genome assembly by retaining scaffolds that had coverage >10.0 in the 1610_IlumiHiSeqX library, and did not have a high scoring (score >5000) taxonomic assignment for 'Proteobacteria', followed by polishing indels and gap-filling with Pilon (Supporting Information 3.5.3). This approach removed 235 scaffolds (330 Kbp), representing 0.2% of the scaffold number and 0.03% of the nucleotides of Ilumi1.0. While filtering the Ilumi1.0 assembly, we noted a large contribution of scaffolds taxonomically annotated as Platyhelminthes (1740 scaffolds; 119.56 Mbp). Upon closer inspection, we found conflicting information as to the most likely taxonomic source of these scaffolds. Diamond searches of these scaffolds had hits in Coleoptera, whereas blastn searches showed these scaffold had confident hits (nucleotide identity >90%, evalue = 0) against the Rat Tapeworm *Hymenolepis diminuta* genome (NCBI BioProject PRJEB507). Removal of these scaffolds decreased the endopterygota BUSCO score, from C:97% D:1.3% to C:76.0% D:1.1%. This loss of the endopterygota BUSCOs led us to conclude that the Platyhelminthes annotated scaffolds were authentic scaffolds of *I. luminosus*, but sequences of *Hymenolepis* sp. may have been transferred into the *I. luminosus* genome via horizontal-gene-transfer

(HGT). Although *Hymenolepis diminuta* infects mammals, it also spends a period of its life cycle in intermediate insect hosts, including beetles, as cysticercoids (Sheiman et al., 2006). For a beetle like *I. luminosus*, which has a extended predatory larval stage, the accidental ingestion and harboring of a *Hymenolepis* sp. is plausible, potentially enabling HGT between *Hymenolepis* sp. and *I. luminosus* over evolutionary timescales.



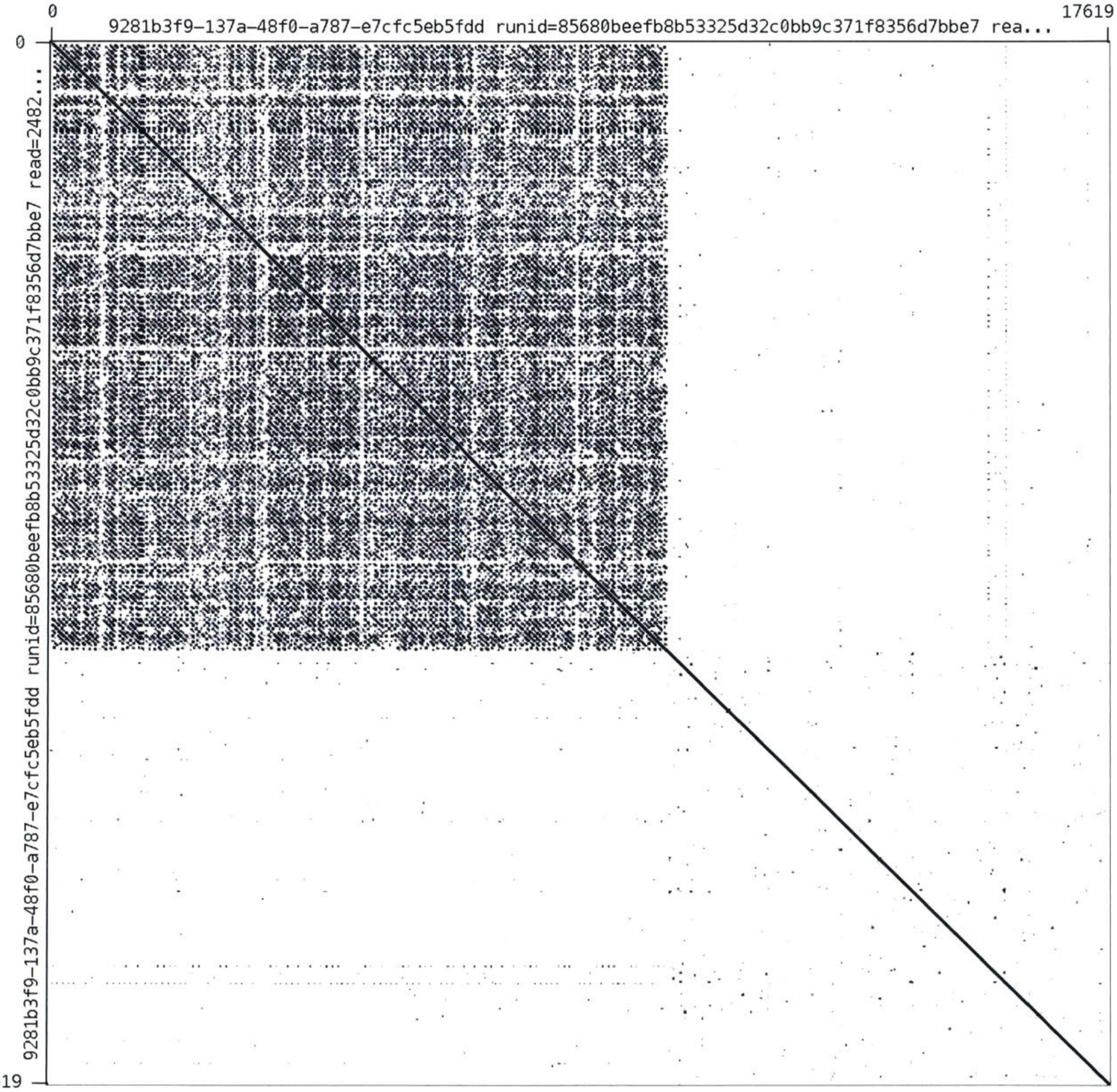
Supporting Information 3—figure 3. BlobTools plot of Ilumi1.0.
Coverage shown represents mean coverage of reads from the HiSeqX Chromium library sequencing (Sample name 1610_IlumiHiSeqX; Supporting Information 4—table 1), aligned against Ilumi1.0 using Bowtie2 with parameters (--local). Scaffolds were taxonomically annotated as described in Supporting Information 3.5.2.

3.5.3 Ilumi1.1: Indel polishing

Manual inspection of the initial gene-models for Ilumi1.0 revealed a key luciferase homolog had an unlikely frameshift occurring after a polynucleotide run. Mapping of the 1610_IlumiHiSeqX and 1706_IlumiHiSeq2500 reads (Supporting Information 4—table 1) with Bowtie2 using parameters (--local), revealed that this indel was not supported by the majority of the data, and that indels were present at a notable frequency after polynucleotide runs. As a greatly increased indel rate after polynucleotide runs (~10% error) is a known systematic error of Illumina sequencing, and has been noted as the major error type in Supernova assemblies (Weisenfeld et al., 2017), we therefore sought to correct these errors globally through the use of Pilon (v1.2.2) (Walker et al., 2014). In order to run Pilon efficiently, we split the taxonomically filtered Ilumi1.0 reference (dubbed Ilumi1.0b; Supporting Information 3.5.2) using Kirill Kryukov's fasta_splitter.pl script (v0.2.6) http://kirill-kryukov.com/study/tools/fasta-splitter/), partitioned the previously mapped 1610_IlumiHiSeqX paired-end reads to these references using samtools, and ran Pilon in parallel on the partitioned reads and records with parameters (--fix gaps,indels --changes --vcf --diploid). The final consensus FASTAs produced by Pilon were merged to produce the polished assembly (Ilumi1.1). Ilumi1.1 (842,900,589 nt; 91,325 scaffolds) was slightly smaller than Ilumi1.0b (845,332,796 nt; 91,325 scaffolds), indicating the gaps filled by Pilon were smaller than their predicted size. The BUSCO score increased modestly after polishing (C:93.3% to C:94.8%), suggesting that indel polishing and gap filling had a net positive effect.

3.5.4 Ilumi1.2: Manual long-read scaffolding

We determined via manual gene-model annotation of Ilumi1.1 (Supporting Information 3.8), that the second through seventh exon of IlumPACS4 (ILUMI_06433 PA) were present on Ilumi1.1_Scaffold13255, but that the first exon was missing from this scaffold. Targeted tblastn using PangPACS (AB479114.1) (Oba et al., 2010a), the most closely related gene sequence to IlumPACS4, indicated that the most similar region in the *I. luminosus* genome to the predicted PangPACS first exon was a right-pointing region on Ilumi1.1_Scaffold11560, not captured in any gene model, but downstream of the existing luciferase homolog genes IlumPACS1 and IlumPACS2. We surmised that this region was the correct first exon for IlumPACS4, and that the IlumPACS4 gene model spanned Ilumi1.1_Scaffold13255 and Ilumi1.1_Scaffold11560, and thus that the right edge of Ilumi1.1_Scaffold13255 and the left edge of the reverse complement of Ilumi1.1_Scaffold11560 should be joined. To substantiate this, we performed long-read Oxford Nanopore MinION sequencing at the MIT BioMicroCenter. The HMW DNA used was the same DNA used for Chromium library prep, and had been stored at −80°C since extraction. Thawing of DNA and size distribution QC on a FEMTO Pulse capillary electrophoresis instrument (Advanced Analytical Technologies Inc, USA) indicated the DNA had a mean size distribution peak of ~17 kbp. A 1D Nanopore library was prepared from this DNA using the standard kit and protocol (Part #: SQK-LSK108). The resulting library was sequenced for 48 hr on a MinION sequencer using a R9.4 flow cell (Part #:FLO-MIN106). Raw trace data was basecalled live within the MinKNOW software (v18.01.6). 824,248 reads (2.4 Gbp; ~1–2x of the *I. luminosus genome*) were obtained. Reads were mapped to Ilumi1.1 with minimap2 (v2.8-r686-dirty) (Li, 2018) using parameters (-ax map-ont). Inspection of mapped reads with Integrative Genomics Viewer (v2.4.8) (Thorvaldsdóttir et al., 2013) revealed a 17.6 kbp read with seven kbp antiparallel alignment to the right edge of Scaffold13255. Inspection of the extension of this read off Scaffold13255 revealed it contained 10 Kbp+ of a non-palindromic complex tandem repeat DNA with an ~100 bp repeat unit (Supporting Information 3—figure 4). The repeat unit of this complex tandem repeat DNA (Supporting Information 3—table 1) is annotated in our *de novo* repeat library construction as 'Ilumi.complex.repeat.1' (Supporting Information 3.9), and via blastn is clearly interspersed at low copy numbers throughout the Ilumi1.1 genome assembly. Notably, this repeat unit was present the right edge of Ilumi1.1_Scaffold13255, while the reverse complement of this repeat unit was present on the right edge of Ilumi1.1_Scaffold11560, supporting that these scaffolds were adjacent to one another, but the assembly had been broken by this large stretch of tandem repetitive DNA. Although our Nanopore sequencing did not unambiguously span this repetitive element and bridge the two scaffolds, we surmised that this information was sufficient to manually merge these scaffolds (Supporting Information 3—figure 5). The long Ilumi1.1_Scaffold13255 extending read was adaptor trimmed with porechop (v0.2.3) (https://github.com/rrwick/Porechop), removing 35 bp from the start of the read. Next, the 3' end of the read which aligned up to the last nucleotide of Ilumi1.1_Scaffold13255 was trimmed. Finally, the remaining read was reverse complemented, and concatenated to the right edge of Ilumi1.1_Scaffold13255. 1337 Ns were concatenated to the right edge of the extended Ilumi1.1_Scaffold13255 to indicate an uncertainty in the repeat copy number, and Ilumi1.1_Scaffold11560 was reverse complemented and concatenated to Ilumi1.1_Scaffold13255 to produce the final version of Ilumi1.2_Scaffold13255 (Supporting Information 3—figure 5). Further whole genome scaffolding using this Nanopore data and the LINKS pipeline (v1.8.5) (Warren et al., 2015) with parameters (-d 4000,8000,10000,14000,16000,20000 t 2,3,5,9 l 2 -a 0.75) was attempted, but only a single additional pair of scaffolds was merged, so this whole-genome scaffolding was not used further.

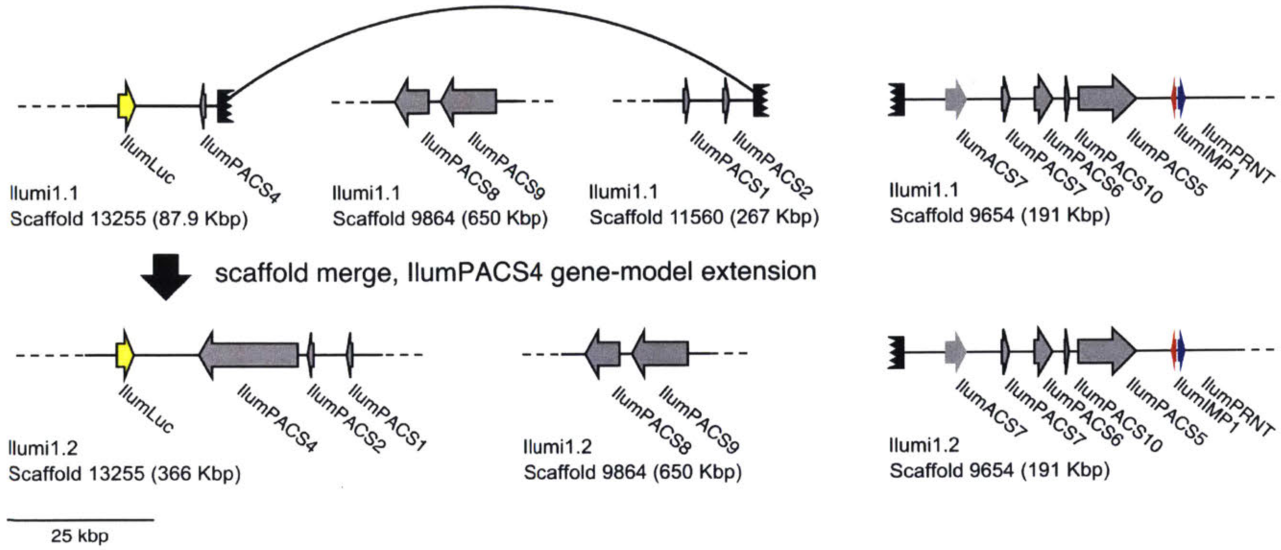


Supporting Information 3—figure 4. Self alignment of the Ilumi1.1_Scaffold13255 right-edge extending long MinION read.

Alignment performed in in Gepard (Krumsiek et al., 2007). Note the large (10 kbp+) tandem repetitive region.

Supporting Information 3—table 1. Sequence of the *I. luminosus* luciferase cluster splitting complex tandem repeat.

Repeat name	Repeat unit length	Repeat unit sequence
Ilumi.complex.repeat.1	~100 bp	TGGTACGAACTATACACGTATACTCAAATCTAAT TGTGATACAGCAAAGTAATAATGCAGCATTGTTT GCCGCTCTATACTGCGATTTTATAGTGGT



Supporting Information 3—figure 5. Diagram of manual scaffold merges between Ilumi1.1 and Ilumi1.2.

Diagram of the manual merge of Ilumi1.1_Scaffold13255 with Ilumi1.1_Scaffold11560 between *I. luminosus* genome assembly versions Ilumi1.1 and Ilumi1.2. This merge was supported by: (1) The putative missing first exon of IlumPACS4 being present on the right edge of Ilumi1.2_Scaffold11560. (2) The right edge of Ilumi1.1_Scaffold13255, and the right edge of Ilumi1.1_Scaffold11560, having anti-parallel versions of a homologous complex tandem repeat. See Figure 3 in the maintext for explanation of presented genes.

3.6 RNA extraction, library prep, and sequencing

3.6.1 HiSeq2500

Total RNA was extracted from the head + prothorax of an *I. luminosus* presumed male using the RNeasy Lipid Tissue Mini Kit (Qiagen, USA). Illumina sequencing libraries were prepared from total RNA enriched to mRNA with a polyA pulldown using the TruSeq RNA Library Prep Kit v2 (Illumina, San Diego, CA). The library was sequenced at the Whitehead Institute Genome Technology Core (Cambridge, MA) on two lanes of an Illumina HiSeq 2500 using rapid mode 100 × 100 bp PE. This library was multiplexed with the *P. pyralis* RNA-Seq libraries of Al-Wathiqui and colleagues (Al-Wathiqui et al., 2016), and thus, *P. pyralis* reads arising from index misassignment were present in this library which necessitated downstream filtering to avoid misinterpretation.

3.6.2 BGISEQ-500

Total RNA was extracted from the head + prothorax, mesothorax + metathorax, and abdomen of adult presumed *I. luminosus* males using the RNeasy Lipid Tissue Mini Kit (Qiagen, USA), and sent on dry-ice to Beijing Genomics Institute (BGI, China). Transcriptome libraries for RNA each sample were prepared from total RNA using the BGISEQ-500 (BGI, China) RNA sample prep protocol. Briefly, poly-A mRNA was purified using oligo (dT) primed magnetic beads and chemically fragmented into smaller pieces. Cleaved fragments were converted to double-stranded cDNA by using N6 primers. After gel purification and end-repair, an 'A' base was added at the 3'-end of each strand. The Ad153-2B adapters with barcode was ligated to both ends of the end repaired/dA tailed DNA fragments, then amplification by ligation-mediated PCR. Following this, a single strand DNA was separated at a high temperature and then a Splint oligo sequence was used as bridge for DNA cyclization to obtain the final library. Then rolling circle amplification (RCA) was performed to produce DNA Nanoballs (DNBs). The qualified DNBs were loaded into the patterned nanoarrays and the libraries were sequenced as 50 × 50 bp (PE-50) read through on the BGISEQ-500 platform. Sequencing-derived raw image files were processed by BGISEQ-500 base-calling software with the default parameters, generating the 'raw data' for each sample stored in FASTQ format. This library preparation and sequencing was provided free of charge as an evaluation of the BGISEQ-500 platform.

Supporting Information 3—table 2. *Ignelater luminosus* RNA-Seq libraries.

Library name	SRA ID	N	Sex	Tissue	Notes
Pyrophorus_luminosus_head	SRR6339835	1	M*	Prothorax and head (lantern containing)	Illumina RNA-Seq
Prothorax_A3	SRR6339834	1	M*	Prothorax and head (lantern containing)	BGISEQ-500 RNA-Seq
Thorax_A3	SRR6339833	1	M*	Mesothorax and metathorax	BGISEQ-500 RNA-Seq

Abdomen_A3	SRR6339832	1	M*	Abdomen (lantern containing)	BGISEQ-500 RNA-Seq
Prothorax_A4	SRR6339831	1	M*	Prothorax and head (lantern containing)	BGISEQ-500 RNA-Seq
Thorax_A4	SRR6339830	1	M*	Mesothorax and metathorax	BGISEQ-500 RNA-Seq
Abdomen_A4	SRR6339838	1	M*	Abdomen (lantern containing)	BGISEQ-500 RNA-Seq

*Sex inferred. See Supporting Information 3.3 for a discussion on this inference.

3.7 Transcriptome analysis

Both *de novo* (Supporting Information 3.7.1) and reference guided (Supporting Information 3.7.2) transcriptome assembly approaches using Trinity and Stringtie were used, respectively.

3.7.1 De novo transcriptome assembly and alignment

For the *de novo* transcriptome approach, all available *I. luminosus* RNA-Seq reads (head + prothorax,metathorax + mesothorax, abdomen - both Illumina and BGISEQ-500) were pooled and input into Trinity. A non-strand-specific *de novo transcriptome* assembly was produced with Trinity (v2.4.0) (Grabherr et al., 2011) using default parameters exception the following: (--min_glue 2 min_kmer_cov 2 --jaccard_clip --no_normalize_reads --trimmomatic). Peptides were predicted from the *de novo* transcripts via Transdecoder (v5.0.2; default parameters). *De novo transcripts* were then aligned to the *I. luminosus* genome (Ilumi1.1) using the PASA pipeline with blat (v36 × 2) and gmap (v2017-09-11) (--aligners blat,gmap), parameters for alternative splice analysis and strand specificity (--ALT_SPLICE --transcribed_is_aligned_orient), and input of the previously extracted Trinity accessions (--tdn tdn.accs). Importantly, it was necessary to set (--NUM_BP_PERFECT_SPLICE_BOUNDARY = 0) for the validate_alignments_in_db.dbi step, to ensure transcripts with natural variation near the splice sites were not discarded. Direct coding gene models (DCGMs) were then produced with the Transdecoder 'cdna_alignment_orf_to_genome_orf.pl' utility script, with the PASA assembly GFF and transdecoder predicted peptide GFF as input. The resulting DCGM GFF3 file was manually lifted over to the Ilumi1.2 assembly. The unaligned *de novo* transcriptome assembly is dubbed 'ILUMI_Trinity_unstranded', whereas the aligned direct coding gene models are dubbed 'Ilumi1.2_Trinity_unstranded-DCGM'.

3.7.2 Reference guided transcriptome alignment and assembly

A reference guided transcriptome was produced from all available *I. luminosus* RNA-seq reads (head + prothorax, mesothorax + metathorax, abdomen - both Illumina and BGISEQ-500) using HISAT2 (v2.0.5) (Kim et al., 2015) and StringTie (v1.3.3b) (Pertea et al., 2015). Reads were first mapped to the *I. luminosus* draft genome with HISAT2 (parameters: -X 2000 --dta --fr). Then StringTie assemblies were performed on each separate bam file corresponding to the original libraries using default parameters. Finally, the produced GTF files were merged using StringTie (--merge). A transcript fasta file was produced from the StringTie GTF file with the transdecoder 'gtf_genome_to_cdna_fasta.pl' utility script, and peptides were predicted for these transcripts using Transdecoder (v5.0.2) with default parameters. The StringTie GTF was converted to GFF format with the Transdecoder 'gtf_to_alignment_gff3.pl' utility script, and direct coding gene models (DCGMs) were then produced with the Transdecoder 'cdna_alignment_orf_to_genome_orf.pl' utility script, with the StringTie-provided GFF and transdecoder predicted peptide GFF as input. The resulting DCGM GFF3 file was manually lifted over to the Ilumi1.2 assembly. The reference guided transcriptome assembly was dubbed 'ILUMI_Stringtie_unstranded', whereas the aligned direct coding gene models were dubbed 'Ilumi1.2_Stringtie_unstranded-DCGM'

3.7.3 Transcript expression analysis

I. luminosus RNA-Seq reads (Supporting Information 3—table 2) were pseudoaligned to the ILUMI_OGS1.2 geneset CDS sequences using Kallisto (v0.44.0) (Bray et al., 2016) with 100 bootstraps

(-b 100), producing transcripts-per-million reads (TPM). Kallisto expression quantification analysis results are available on FigShare (DOI: 10.6084/m9.figshare.5715139).

3.8 Official coding geneset annotation (ILUMI_OGS1.2)

We annotated the coding gene structure of *I. luminosus* by integrating direct coding gene models produced from the *de novo* transcriptome (Supporting Information 3.7.1) and reference guided transcriptome (Supporting Information 3.7.2), with a lower weighted contribution of *ab initio* gene predictions, using the Evidence Modeler (EVM) algorithm (v1.1.1) (Haas et al., 2008). First, Augustus (v3.2.2) (Stanke et al., 2006) was trained against Ilumi1.0 with BUSCO (parameters: -l endopterygota_odb9
--long --species tribolium2012). Augustus predictions of Ilumi1.0 were then produced through the MAKER pipeline, with hints derived from MAKER blastx/exonerate mediated protein alignments of peptides from *Drosophila melanogaster* (NCBI GCF_000001215.4_Release_6_plus_ISO1_MT_protein.faa), *Tribolium castaneum* (NCBI GCF_000002335.3_Tcas5.2_protein), *Photinus pyralis*(PPYR_OGS1.0; this report), *Aquatica lateralis* (AlatOGS1.0; this report), the *I. luminosus de novo* transcriptome translated peptides, and MAKER blastn/exonerate transcript alignments of the *I. luminosus de novo* transcriptome transcripts.

We then integrated the *ab initio* predictions with our *de novo* and reference guided direct coding gene models, using EVM. In the final version, eight sources of evidence were used for EVM: *de novo* transcriptome direct coding gene models (Ilumi1.1_Trinity_unstranded-DCGM; weight = 8), reference guided transcriptome direct coding gene models (Ilumi1.1_Stringtie_unstranded-DCGM; weight = 4), MAKER/Augustus *ab initio* predictions (Ilumi1.1_maker_augustus_ab-initio; weight = 1), protein alignments (*P. pyralis*, *A. lateralis*, *D. melanogaster*, *T. castaneum, I. luminosus;* weight = 1 each). A custom script (https://github.com/photocyte/maker_gff_to_evm_gff_2017) was used to convert the input MAKER GFF to an EVM compatible GFF format.

Lastly, gene models for luciferase homologs, P450s, and *de novo methyltransferases* (DNMTs) which were fragmented or were incorrectly assembled (e.g. adjacent gene fusions) were manually corrected based on the evidence of the *de novo* and reference guided direct coding gene models (Supporting Information 3.7.1; 3.7.2). Manual correction was performed by performing TBLASTN searches with known good genes from these gene families within SequenceServer(v1.10.11) (Priyam et al., 2015), converting the TBLASTN results to gff3 format with a custom script (https://github.com/photocyte/general_scripts/blob/master/blastxml2gff.py), and viewing these TBLASTN alignments alongside the alternative direct coding gene models and the official geneset in Integrative Genomics Viewer (v2.4.8) (Thorvaldsdóttir et al., 2013). The official gene set models gff3 file was then manually modified based on the observed evidence. Different revision numbers of the official geneset (e.g. ILUMI_OGS1.0, ILUMI_OGS1.1) represent the improvement of the geneset over time due to these continuing manual gene annotations.

3.9 Repeat annotation

A *de novo* species-specific repeat library for *I. luminosus* was constructed using RepeatModeler (v1.0.9), and Tandem Repeat Finder (v4.09; settings: 2 7 7 80 10) (Benson, 1999). Only tandem repeats from Tandem Repeat Finder with a repeat block length >5 kb (annotated as 'complex tandem repeat') were added to the RepeatModeler library. This process yielded a final library of 2259 interspersed repeats. We then used this library and RepeatMasker (v4.0.5) (http://www.repeatmasker.org/) to identify and mask interspersed and tandem repeats in the genome assembly. This repeat library is dubbed the *Ignelater luminosus* Official Repeat Library 1.0 (ILUMI_ORL1.0).

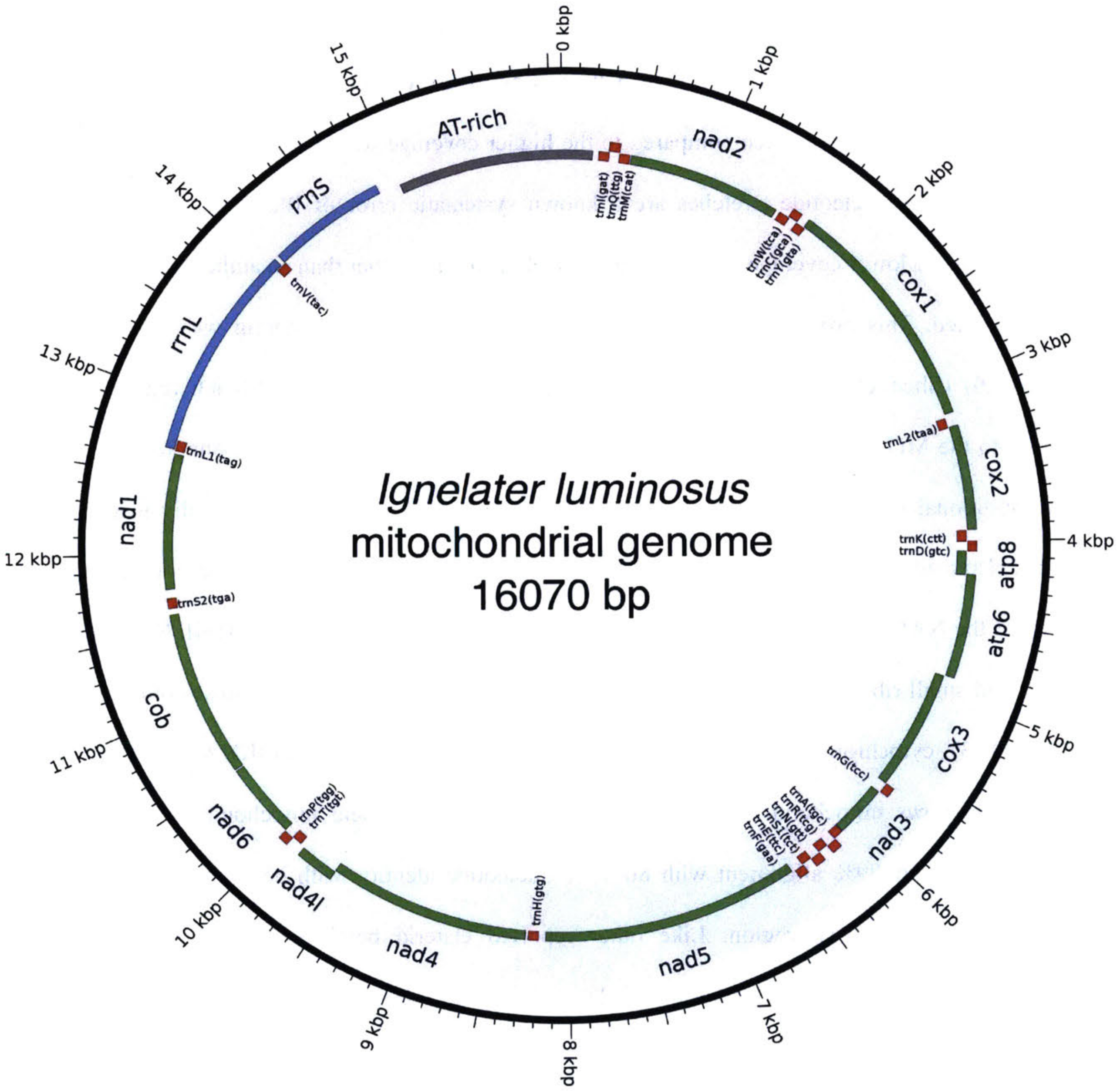
Supporting Information 3—table 3. Annotated repetitive elements in *I. luminosus*.

Repeat class	Family	Counts	Bases	% of assembly
DNA	All	158853	71221843	8.45
	Helitrons	344	139863	0.016

LTR	All	23433	11341577	1.35
Non-LTR	All	151788	50394853	4.75
	LINE	97703	40052840	4.75
	SINE	0	0	0
Unknown interspersed		757206	159587269	18.93
Complex tandem repeats		4976	848992	0.1
Simple repeat		108914	4439967	0.52
rRNA		0	0	0

3.10 Mitochondrial genome assembly and annotation

The mitochondrial genome sequence of *I. luminosus* was assembled by a targeted sub-assembly approach. First, Chromium linked-reads were mapped to the previously sequenced mitochondrial genome of the Brazilian elaterid beetle *Pyrophorus divergens* (NCBI ID: NC_009964.1) (Arnoldi et al., 2007), using Bowtie2 (v2.3.1; parameters: --very-sensitive-local) (Langmead et al., 2009). Although these reads still contain the 16 bp Chromium library barcode on read 1 (R1), Bowtie2 in local mapping mode can accurately map these reads. Mitochondrial mapping R1 reads with a mapping read 2 (R2) pair were extracted with 'samtools view -bh -F 4 f 8', whereas mapping R2 reads with a mapping R1 pair were extracted with 'samtools view -bh -F 8 f 4'. R1 and R2 singleton mapping reads were extracted with 'samtools view -bh -F 12' for diagnostic purposes, but were not used further in the assembly. The R1, R2, and singleton reads in. BAM format were merged, sorted, and converted to FASTQ format with samtools and 'bedtools bamtofastq', respectively. The resultant R1 and R2 FASTQ files containing only the paired mapped reads (995523 pairs, 298 Mbp) were assembled with SPAdes (Nurk et al., 2013) without error correction and with the plasmidSPAdes module (Antipov et al., 2016) enabled (parameters: -t 16 --plasmid -k55,127 --cov-cutoff 1000 --only-assembler). The resulting 'assembly_graph.fastg' file was viewed in Bandage (Wick et al., 2015), revealing a 16,088 bp node with 1119x average coverage that circularized through two possible paths: a 246 bp node with 252x average coverage, or a 245 bp node with 1690x coverage. The lower coverage path was observed to differ only in a 'T' insertion after a 10-nucleotide poly-T stretch when compared to the higher coverage path. Given that increased levels of insertions after polynucleotide stretches are a known systematic error of Illumina sequencing, it was concluded that the lower coverage path represented technical error rather than an authentic genetic variant and was deleted. This produced a single 16,070 bp circular contig. This contig was 'restarted' with seqkit(v0.7.0) (Shen et al., 2016) to place the FASTA record break in the AT-rich region, and was submitted to the MITOSv2 mitochondrial genome annotation web server. Small mis-annotations (e.g. low scoring additional predictions of already annotated mitochondrial genes) were manually inspected and removed. This annotation indicated that all expected features were present on the contig, including subunits of the NAD+ dehydrogenase complex (NAD1, NAD2, NAD3, NAD4, NAD4l, NAD5, NAD6), the large and small ribosomal RNAs (rrnL, rrnS), subunits of the cytochrome c oxidase complex (COX1, COX2, COX3), cytochrome b oxidase (COB), ATP synthase (atp6, atp8), and tRNAs. BLASTN of the *Ignelater luminosus* mitochondrial genome against published complete mitochondrial genomes from beetles indicated 96–89% alignment with 86–73% nucleotide identity, with poor or no sequence level alignment in the A-T rich region. Like other reported elaterid beetle genomes, the *I. luminosus* mitochondrial genome does not contain the tandem repeat unit (TRU) previously reported in Lampyridae (Bae et al., 2004).



Supporting Information 3—figure 6. Mitochondrial genome of *I. luminosus*.
The mitochondrial genome of *I. luminosus* was assembled and annotated as described. in the Supporting Information 3.10. Figure produced with Circos (Krzywinski et al., 2009).

Supporting Information 4

Interspecies and comparative analyses

4.1 Assembly statistics and comparisons

The level of non-eukaryote contamination of the raw read data for each *P. pyralis library* was assessed using kraken v1.0 (Wood and Salzberg, 2014) using a dust-masked minikraken database to eliminate comparison with repetitive sequences. Overall contamination levels were low (Supporting Information 4—table 1), in agreement with a low level of contamination in our final assembly (Supporting Information 1—figure 9, Supporting Information 2—figure 2, Supporting Information 3—figure 3). On average, contamination was 3.5% in the PacBio reads (whole body) and 1.6% in the Illumina reads (only thorax) (Supporting Information 4—table 1). There was no support for Wolbachia in any of the *P. pyralis* libraries, with the exception of a single read from a single library which had a kraken hit to Wolbachia. QUAST version 4.3 (Gurevich et al., 2013) was used to calculate genome quality statistics for comparison and optimization of assembly methods (Supporting Information 4—table 2). BUSCO (v3.0.2) (Simão et al., 2015) was used to estimate the percentage of expected single copy conserved orthologs captured in our assemblies and a subset of previously published beetle genome assemblies (Supporting Information 4—table 3). The endopterygota_odb9 (metamorphosing insects) BUSCO set was used. The bacteria_odb9 gene set was used to identify potential contaminants by screening contigs and scaffolds for conserved bacterial genes. For genome predictions from beetles, the parameter '--species tribolium2012' was used to improve the BUSCO internal Augustus gene predictions. For *Drosophila melanogaster* BUSCO genome predictions (Supporting Information 4—table 3) '--species=fly' was used.

Supporting Information 4—table 1. Genomic sequencing library statistics.
ID: NCBI BioProject or Gene Expression Omnibus (GEO) ID. **N**: Number of individuals used for sequencing. **Date**: collection date for wild-caught individuals. **Locality**: GSMNP: Great Smoky Mountains National Park, TN; MMNJ: Mercer Meadows, Lawrenceville, NJ; IY90: laboratory strain Ikeya-Y90; MAPR: Mayagüez, Puerto Rico. **Tissue**: Thr: thorax; WB: whole-body; **Type**: SI: Illumina short insert; MP: Illumina mate pair; PB: Pacific Biosciences, RSII P6-C4; HC: Hi-C; BS: Bisulfite; CH: 10x Chromium; ONT: Oxford Nanopore MinION R9.4. Reads: PE: paired-end, CLR: continuous long read. **Number**: number of reads. **Cov**: Mode of autosomal coverage (mode of putative X chromosome, LG3a, coverage), determined from mapped reads with QualiMap (v2.2). ND: Not

Determined. **Insert size**: Mode of insert size after alignment (orientation: FR: forward, RF: reverse), determined from mapped reads with QualiMap.

Library	**SRA ID**	**N**	**Sex**	**Tissue**	**Type**	**Reads**	**Number**	**Cov**	**Insert size (Ori)**
Photinus pyralis									
8369*	SRR6345451 / SRR2127932	1	M	Thr	SI	101 × 101 PE	203,074,230	98 (49)	354 bp (FR)
8375_3 K†	SRR6345448	1	M	Thr	MP	101 × 101 PE	101,624,630	21	2155 bp (RF)
8375_6 K†	SRR6345457	1	M	Thr	MP	101 × 101 PE	23,564,456	5	4889 bp (RF)
83_3 K†	SRR6345450	3	M	Thr	MP	101 × 101 PE	121,757,858	13	2247 bp (RF)
83_6 K†	SRR6345455	3	M	Thr	MP	101 × 101 PE	17,905,700	1	4877 bp (RF)
1611_PpyrPB1	SRX3444870	4	M	WB	PB	CLR-PB	3,558,201	38 (21)	7 Kbp‡
1704	SRR6345456	2	M	WB	HC	80 × 80 PE	93,850,923	ND	ND
1705	GSE107177	1	M	WB	BS	150 SE	113,761,746	~16x§	ND
Aquatica lateralis									
FFGPE_PE200	DRR119296	1	F	WB	SI	126 × 126 PE	561,450,686	72	180 bp (FR)
FFGPE_PE800	DRR119297			WB	SI	126 × 126 PE	218,830,950	20	476 bp (FR)
FFGMP_MPGF	DRR119298			WB	MP	101 × 101 PE	358,601,808	31	2300 bp (RF)
Ignelater luminosus									
1610_Ilumi HiSeqX#	SRR6339837	1	M¶	WB	CH	151 × 151 PE	408,838,927	99	339 bp (FR)

1706_Ilumi HiSeq2500#	SRR6339836	WB	CH	150 × 150 PE	145,250,480	48	334 bp (FR)
18_lib1	SRR6760567		ONT	CLR	824,248	~2x	2984‡

*Mean of three sequencing lanes
†Mean of two sequencing lanes
‡Mean subread (PacBio) or read (Oxford Nanopore) length after alignment
§Estimate from quantity of mapped reads
#Same library, different instruments
¶Inferred from specimens collected at the same time and locality

Supporting Information 4—table 2. Assembly statistics

Assembly	Libraries	Assembly scheme	Assembly* /measured** genome size (Gbp)	Scaffold/ Contig (#)	Contig NG50*** (Kbp)	Scaffold NG50*** (Kbp)	BUSCO statistics
Ppyr0.1-PB	PacBio (61 RSII SMRT cells)	Canu (no polishing)	721/422	25986/ 25986	86	86	C:93.8% [S:65.2%, D:28.6%], F:3.3%, M:2.9%
Ppyr1.1	Short read Mate Pair PacBio	MaSuRCA + redundancy reduction	473/422	8065/ 8285	193.4	202	C:97.2% [S:88.8%, D:8.4%], F:1.9%, M:0.9%
Ppyr1.2	Short PacBio Hi-C	Ppyr1.1+ Phase Genomics scaffolder (in-house)	473/422	2535/ 7823	193.4	50,607	C:97.2% [S:88.8% ,D:8.4%], F:1.9%, M:0.9%
Ppyr1.3	Short read Mate Pair PacBio	Ppyr1.2 +Blobtools + manual filtering	472/422	2160/ 7533	192.5	49,173	C:97.2% [S:88.8%, D:8.4%], F:1.9%, M:0.9%

Alat1.2	Short read Mate Pair	ALLPATHS-LG	920/940	7313/ 36467	38	673	C:97.4% [S:96.2%, D:1.2%], F:1.8%, M:0.8%
Alat1.3	Short read Mate Pair	Alat1.2+ Blobtools + manual filtering	909/940	5388/ 34298	38	670	C:97.4% [S:96.2%, D:1.2%], F:1.8%, M:0.8%
Ilumi1.0	Linked-read	Supernova	845/764	91560/ 105589	31.6	116.5	C:93.7% [S:92.3%, D:1.4%], F:4.3%, M:2.0%,
Ilumi1.2	Linked read+ nanopore	Ilumi1.0+ BlobTools+ Pilon indel and gap polishing. Manual scaffolding	842/764	91305/ 105262	34.5	115.8	C:94.8% [S:93.4%, D:1.4%], F:3.5%, M:1.7%

*Calculated from genome assembly file with 'seqkit stat'

**Measured via flow cytometry of propidium iodide stained nuclei. See Supporting Information 1.4, 2.4, 3.4.

***Calculated with QUAST (v4.5) (Gurevich et al., 2013), parameters '-e --scaffolds --est-ref-size X --min-contig 0' and the measured genome size for 'est-ref-size'

Supporting Information 4—table 3. Comparison of BUSCO conserved gene content with other insect genome assemblies

Species	Genome version (NCBI assemblies)	Note	Genome BUSCO (endopterygota_odb9)	Protein geneset BUSCO (endopterygota_odb9)**
Drosophila melanogaster	GCA_000001215.4 Release 6	Model insect	C:99.4%[S:98.7%,D:0.7%], F:0.4%,M:0.2%,n:2442	C:99.6%[S:92.8%,D:6.8%], F:0.3%,M:0.1%,n:2442
Tribolium castaneum	GCF_000002335.3 Release 5.2	Model beetle	C:98.4%[S:97.9%,D:0.5%], F:1.2%,M:0.4%,n:2442	C:98.0%[S:95.8%,D:2.2%], F:1.6%,M:0.4%,n:2442

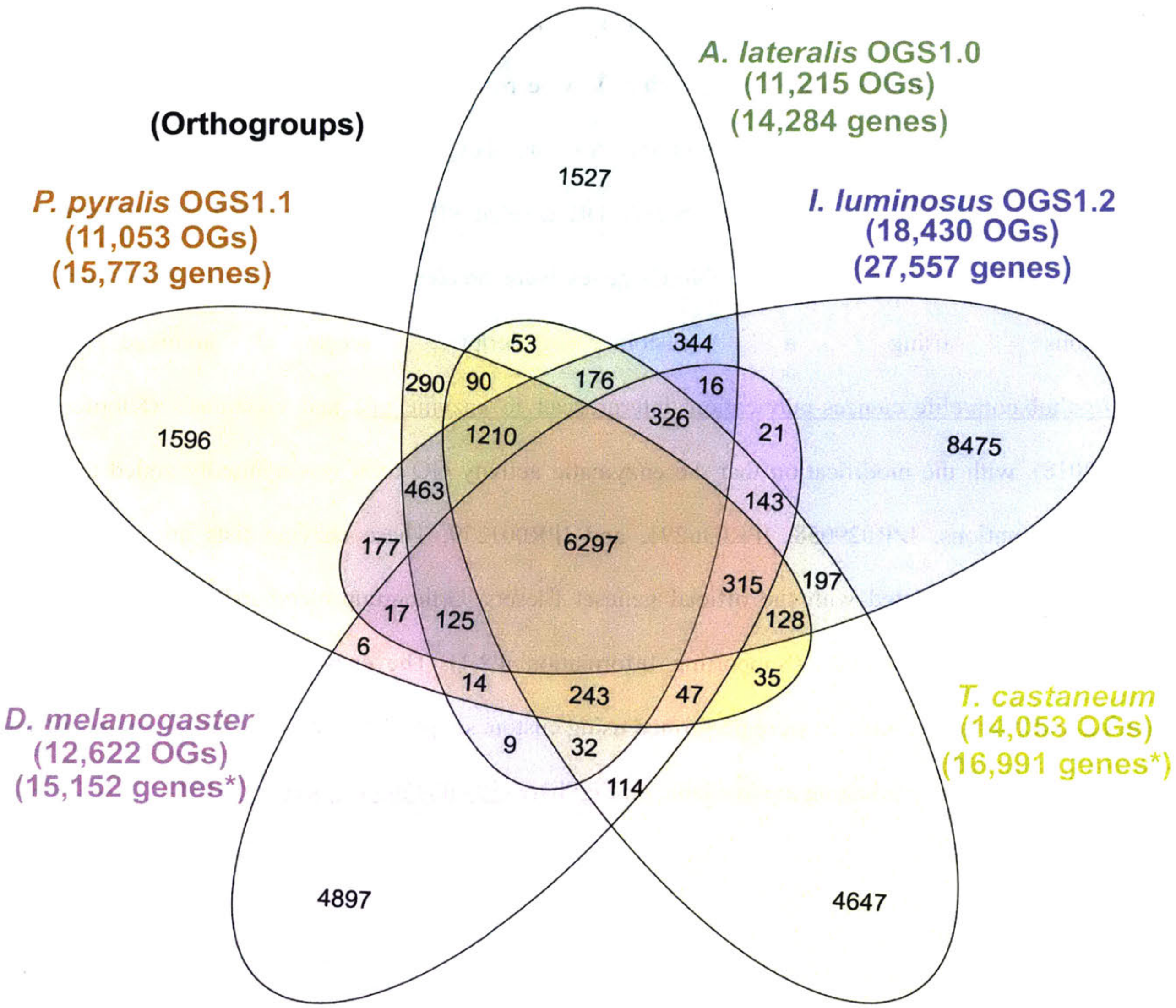
*Photinus pyralis**	Ppyr1.3*	North American firefly	C:97.2%[S:88.8%,D:8.4%], F:1.8%,M:1.0%,n:2442	C:94.2%[S:84.0%,D:10.2%], F:1.2%,M:4.6%, n:2442
*Aquatica lateralis**	Alat1.3*	Japanese firefly	C:97.4%[S:96.2%,D:1.2%], F:1.8%,M:0.8%,n:2442	C:90.0%[S:89.1%,D:0.9%], F:3.2%,M:6.8%,n:2442
Nicrophorus vespilloides (Cunningham et al., 2015)	GCF_001412225.1 Release 1.0	Burying beetle	C:96.8%[S:95.3%,D:1.5%], F:2.1%,M:1.1%, n:2442	C:98.7%[S:69.4%,D:29.3%], F:0.8%,M:0.5%,n:2442
Agrilus planipennis (Poelchau et al., 2015)	GCF_000699045.1 Release 1.0	Emerald Ash Borer beetle	C:92.7%[S:91.8%,D:0.9%], F:4.6%,M:2.7%,n:2442	C:92.1%[S:64.1%,D:28.0%], F:4.5%,M:3.4%,n:2442
*Ignelater luminosus**	Ilumi1.2	Puerto Rican bioluminescent click beetle	C:94.8%[S:93.4%,D:1.4%], F:3.5%,M:1.7%,n:2442	C:91.8%[S:89.8%,D:2.0%], F:4.4%,M:3.8%, n:2442

*=This report, **=Protein genesets downloaded from the NCBI Genome resource associated with the mentioned assembly in the 2nd column, or in the case of *D. melanogaster*, and *T. castaneum*, protein genesets were produced from Uniprot Reference Proteomes which had been heuristically filtered down to 'canonical' isoforms with a custom script and BLASTP against the *D. melanogaster*, *T. castaneum, Apis mellifera*, *Bombyx mori*, *Caenorhabditis elegans*, and *Anopheles gambiae* protein genesets associated with their more recent genome assembly on NCBI. See Supporting Information 4.2.1 for more detail.

4.2 Comparative analyses

4.2.1 Protein orthogroup clustering

Orthologs were identified by clustering the *P. pyralis*, *A. lateralis*, and *I. luminous genset* peptides with the *D. melanogaste*r (UP000007266) and *T. castaneum* (UP000000803) reference Uniprot protein genesets using the OrthoFinder (v2.2.6) (Emms and Kelly, 2015) pipeline with parameters '-M msa -A mafft -T fasttree -I 1.5'. The pipeline was executed with NCBI blastp + v0.2.7.1, mafft 7.313, and FastTree v2.1.10 with Double precision (No SSE3). The Uniprot reference proteomes were first filtered using a custom script to remove multiple isoforms-per-gene using a custom script (copy archived at https://github.com/elifesciences-publications/filter_uniprot_to_best_isoform), which utilized blastp evidence against either the *Drosophila melanogaster* or *Tribolium castaneum* NCBI datasets (whichever species was not being filtered), and the *Apis mellifera*, *Bombyx mori*, *Caenorhabditis elegans*, *Anopheles gambiae* NCBI peptide genesets. Not all redundant isoforms are removed as there may not have been sufficient evidence to support a particular isoform as the canonical isoform, or there were unusual annotation situations (alternative splice variants annotated as separate genes). OrthoFinder clustering results are available on FigShare (DOI: 10.6084/m9.figshare.5715136). Overlaps of number of shared orthogroups across species are shown in Supporting Information 4—figure 1. Overlaps on a gene-basis (only *P. pyralis*, *A. lateralis*, *I. luminosus*, and *T. castaneum*) are shown in Figure 2E.



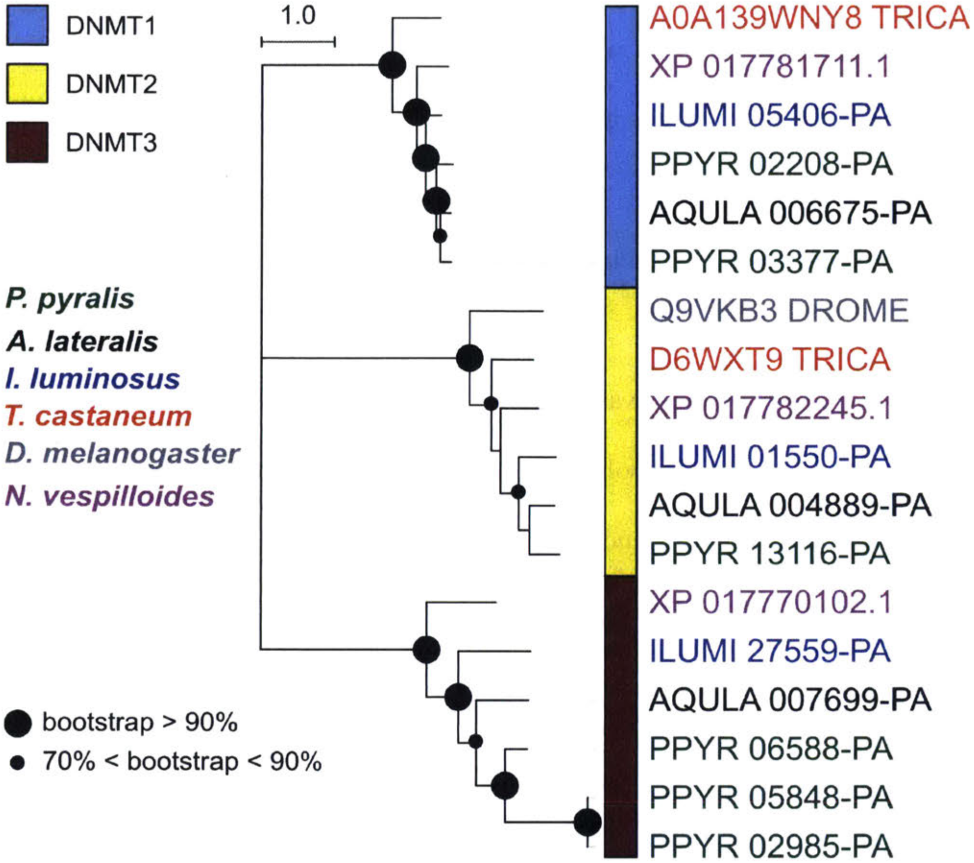
Supporting Information 4—figure 1. Venn diagram of P. pyralis, A. lateralis, I. luminosus, T. castaneum, and D. melanogaster orthogroup relationships.

Orthogroups were calculated between the PPYR_OGS1.1, AQULA_OGS1.0, ILUMI_OGS1.2, genesets, and the *T. casteneum* and *D. melanogaster* filtered Uniprot reference proteomes using OrthoFinder(Emms and Kelly, 2015). See Supporting Information 4.2.1 for description of clustering method. OGs = Orthogroups, OGS = Official gene set, *=Not completely filtered to single peptide per gene. Figure produced with InteractiVenn (Heberle et al., 2015).

4.2.2 Comparative RNA-Seq differential expression analysis (Figure 5)

For differential expression testing, Kallisto transcript expression results for *P. pyralis* (Supporting Information 1.9.4) and *A. lateralis* (Supporting Information 2.7.3) were independently between-sample normalized using Sleuth (v0.30.0) (Pimentel et al., 2017) with default parameters, producing between-sample-normalized transcripts-per-million reads (BSN-TPM). Differential expression (DE) tests for *P. pyralis* (adult male dissected fatbody vs. adult male dissected lantern - three biological replicates per condition), and for *A. lateralis* (adult male thorax + abdominal segments 1–5 vs. adult male dissected lantern - three biological replicates per condition), were performed using the Wald test within Sleuth. Genes whose mean BSN-TPM across bio-replicates was above the 90th percentile were annotated as ‘highly expressed’ (HE). Genes with a Sleuth DE q-value <0.05 were annotated as ‘differentially expressed.’ (DE). Enzyme encoding (E/NotE) genes were predicted from the InterProScan functional annotations using a custom script (copy archived at https://github.com/elifesciences-publications/interproscan_to_enzyme_go) and GOAtools (Klopfenstein et al., 2018), with the modification that the enzymatic activity GO term was manually added to select InterPro annotations: IPR029058, IPR036291, and IPR001279. These enzyme lists are available as supporting files associated with the official geneset filesets. Orthogroup membership was determined from the OrthoFinder analysis (Supporting Information 4.2.1). The enzyme HE/DE/E + NotE gene filtering and overlaps (Figure 5) were performed using custom scripts. These custom scripts and results of the differential expression testing are available on FigShare (10.6084/m9.figshare.5715151).

4.2.3 Comparative methylation analyses



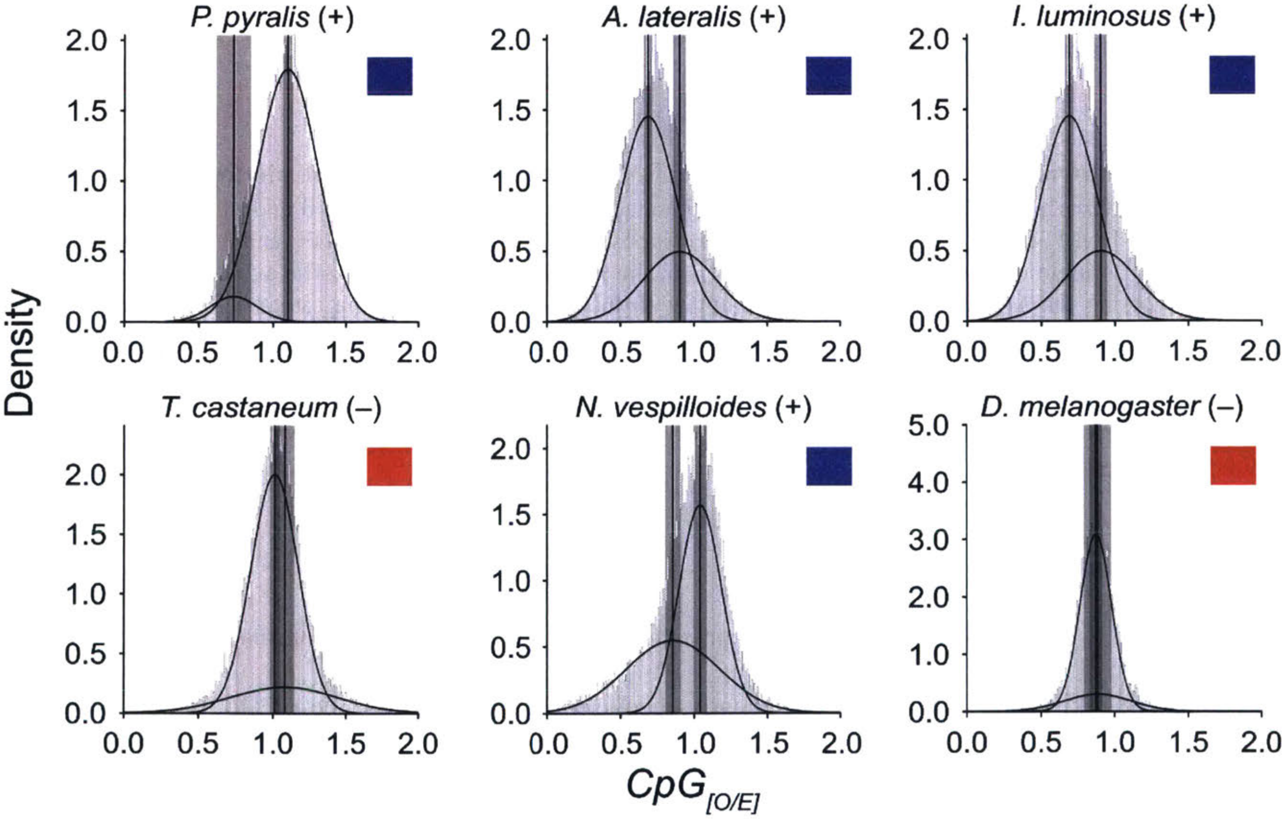
Supporting Information 4—figure 2. DNA and tRNA methyltransferase gene phylogeny. Levels and patterns of mCG in *P. pyralis* are corroborated by the presence of *de novo* and maintenance DNMTs (DNMT3 and DNMT1, respectively). Notably, *P. pyralis* possesses two copies of DNMT1, and 3 copies of DNMT3, in contrast to a single copy of DNMT1 and DNMT3 in the firefly *Aquatica lateralis*. The evolutionary history was inferred by using the Maximum Likelihood method with the LG + G (five gamma categories) (Le and Gascuel, 2008). Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016). Size of circles at nodes corresponds to bootstrap support (100 bootstrap replicates). Branch lengths are in amino acid substitutions per site. *T. castaneum* = *Tribolium castaneum*, *D. melanogaster* = *Drosophila melanogaster*, *N. vespilloides* = *Nicrophorus vespilloides*. The multiple sequence alignment and phylogenetic topology are available on FigShare (10.6084/m9.figshare.6531311).

4.2.3.2 *CpG[O/E]* methylation analysis

CpG[O/E] is a non-bisulfite sequencing metric that captures spontaneous deamination of methylated cytosines (Suzuki et al., 2007), and confidently recovers the presence/absence of DNA methylation in insects (Bewick et al., 2017). In a mixture of loci that are DNA methylated and low to un-methylated, a bimodal distribution of *CpG[O/E]* values is expected. Conversely, a unimodal distribution is suggestive of a set of loci that are mostly low to un-methylated.

CpG[O/E] was estimated for each annotated gene in the official gene set of *A. lateralis*, *I. luminosus*, and *P. pyralis*. Additionally, *CpG[O/E]* was estimated for each annotated gene for a true positive and negative coleopteran (*Nicrophorus vespilloides* [https://i5k.nal.usda.gov/nicrophorus-vespilloides] and *Tribolium castaneum* [https://i5k.nal.usda.gov/tribolium-castaneum], respectively), and a true negative dipteran (*Drosophila melanogaster* [http://flybase.org/]).

The modality of *CpG[O/E]* distributions was tested using Gaussian mixture modeling in R (https://www.r-project.org/: mclust v5.4 and mixtools v1.0.4). Two modes were modeled for each *CpG[O/E]* distribution, and the subsequent means and 95% confidence interval (CI) of the means were compared with overlapping or non overlapping CI's signifying unimodality or bimodality, respectively.



Supporting Information 4—figure 3. Detection of DNA methylation using *CpG$_{[O/E]}$*
Distributions of *CpG$_{[O/E]}$*(*CpG$_{[O/E]}$* methylation analysis) within sequenced species (*P. pyralis*, *A. lateralis*, and *I. luminosus*), other coleopterans (*N. vespilloides* and *T. castaneum*), and the dipteran *D. melanogaster*. Curves represent two independently modeled Gaussian distributions, and the solid vertical

lines and shaded areas represent the mean and 95% confidence interval (CI) of the mean of each distribution. Modality of the distributions accurately predicts presence (+)/blue square or absence (–)/red square of DNA methylation in each species.

4.2.4 CYP303 evolutionary analysis (Figure 6C)

Candidate P450s were identified using BLASTP (e-value: 1 × 10−20) of a *P. pyralis* CYP303 family member (PPYR_OGS1.0: PPYR_14345-PA) against the *P. pyralis*, *A. lateralis*, and *I. luminosus* reference set of peptides, and the *D. melanogaste*r (NCBI GCF_000001215.4) and *T. castaneum* (NCBI GCF_000002335.3) geneset peptides. Resulting hits were merged, aligned with MAFFT E-INS-i (v7.243) (Katoh and Standley, 2013), and a preliminary neighbor-joining (NJ) tree was generated using MEGA7 (Kumar et al., 2016). Genes descending from the common ancestor of the *CYP303* and *CYP304* genes were selected from this NJ tree, and the peptides within this subset re-aligned with MAFFT using the L-INS-i algorithm. Then the maximum likelihood evolutionary history of these genes was inferred within MEGA7 using the LG + G model (five gamma categories (+G, parameter = 2.4805). Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with the best log likelihood value. The resulting tree was rooted using *D. melanogaster Cyp6a17* (NP_652018.1). The tree shown in Figure 6C was truncated in Dendroscope (v3.5.9) (Huson and Scornavacca, 2012) to display only the *CYP303* clade.

4.3 Luciferase evolution analyses

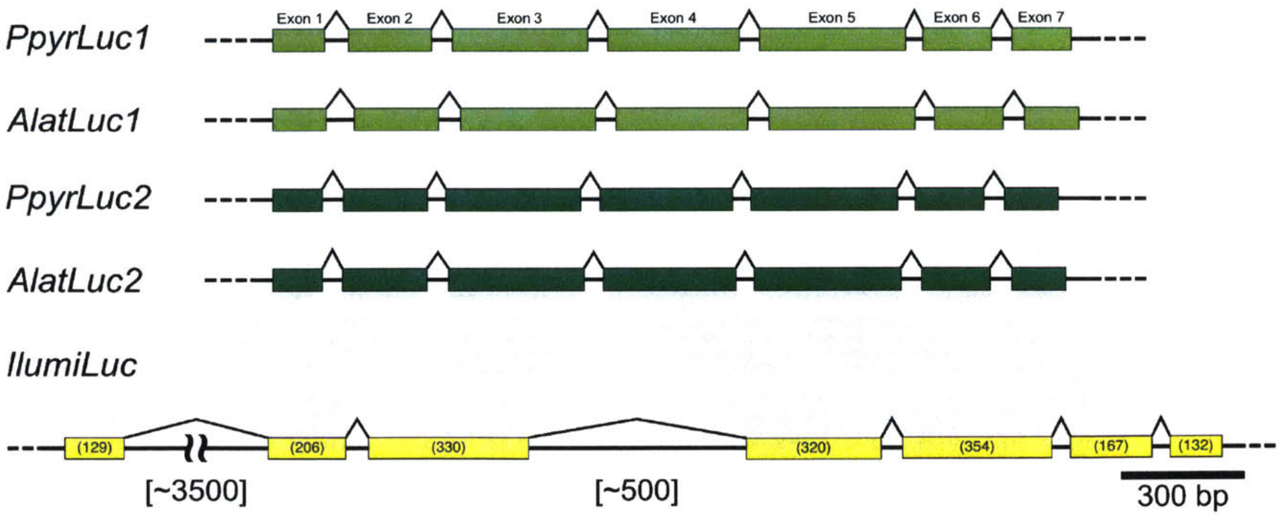
4.3.1 Luciferase genetics overview

The gene for firefly luciferase was first isolated from the North American firefly *P. pyralis* ((De Wet et al., 1987; de Wet et al., 1985; Wood et al., 1984) and then identified from the Japanese fireflies *Luciola cruciata* (Masuda et al., 1989) and *Aquatica lateralis* (Tatsumi et al., 1992). To date, firefly luciferase genes have been isolated from more than 30 lampyrid species in the world. Two different types of luciferase genes, *Luc1* and *Luc2*, have been reported from *Photuris pennsylvanica (Ye et al., 1997)*

(Photurinae), *L. cruciata* (Oba et al., 2010b) (Luciolinae), *A. lateralis (Oba et al., 2013a)* (Luciolinae), *Luciola parvula* (Bessho-Uehara and Oba, 2017) (Luciolinae), and *Pyrocoelia atripennis* (Bessho-Uehara et al., 2017) (Lampyrinae).

Luciferase genes have also been isolated from members of the other luminous beetles families: Phengodidae, Rhagophthalmidae, and Elateridae (Ohmiya et al., 2000; Viviani et al., 1999a, 1999b; Wood et al., 1989) with amino acid identities to firefly luciferases at >48% (Oba and Hoffmann, 2014). The chemical structures of the substrates for these enzymes are identical to firefly luciferin. These results that the bioluminescence systems of luminous beetles are essentially the same, supports a single origin of the bioluminescence in elateroid beetles. Recent molecular analyses based on the mitochondrial genome sequences strongly support a sister relationship between the three luminous families: Lampyridae, Phengodidae, and Rhagophthalmidae (Timmermans et al., 2010; Timmermans and Vogler, 2012), suggesting the monophyly of Elateroidea and a single origin of the luminescence in the ancestor of these three lineages (Oba and Hoffmann, 2014). However, ambiguity in the evolutionary relationships among luminous beetles, including luminous elaterids, does not yet exclude multiple origins.

Molecular analyses have suggested that the origin of Lampyridae was dated back to late Jurassic (McKenna and Farrell, 2009) or mid-Cretaceous periods (McKenna et al., 2015). Luciolinae and Lampyrinae was diverged at the basal position of the Lampyridae (Martin et al., 2017) and the fossil of the Luciolinae firefly dated at Cretaceous period was discovered in Burmese amber (Kazantsev, 2015; Shi et al., 2012). Taken together, the divergence of Luciola and Lampyridae is dated back at least 100 Mya.



Supporting Information 4—figure 4. Intron-exon structure of beetle luciferases.
(**A**) Intron-exon structure of *P. pyralis* and *A. lateralis Luc1* and *Luc2* from Ppyr1.3 and Alat1.3, and *IlumLuc* from Ilumi1.2. Between fireflies and click-beetles, the structure of the luciferase genes are globally similar, with seven exons, similar intron lengths, and identical splice junction locations (Supporting Information 4—figure 5). The intron-exon structure of *IlumLuc* is consistent with the reported intron-exon structure of *Pyrophorus plagiophthalamus* luciferase (Velez and Feder, 2006).

```
PpyrLuc1    ATG---------GAAGACGCCAAAAACATAAAGAAAGGCCCGGCGCCATTCTATCCTCTAGAGGATGGAACCGCTGGAGAGCAACTGCATAAGGCTATGAAGAGATACGCCCTGGTTCCT
AlatLuc1    ATGGAAAACATGGAGAACGATGAAAATATTGTATATGGTCCTGAACCATTTTACCCTATTGAAGAGGGATCTGCTGGAGCACAATTGCGCAAGTATATGGATCGATATGC---AAAACTT
PpyrLuc2    ATG------------GAAAATAAGAATATCTTGTATGGACCTAAACCATTTTATCCTGTTTCGGATGGTACGGCAGGCGAGGAGATATTTAGGGCACTTAAAAAGTATGCAAGGATACCA
AlatLuc2    ATG---------------AACAAGAATATATTATACGGTCCACCACCGGTACACCCTCTTGACGATGGGACGGGTGGTGAACAATTGTACAAATGTATTTTAAAATACGCTCAAATTCCC

PpyrLuc1    GGAACAATTGCTTTTgtgagt---------atttctgtc---tgatttctttcgagttaacgaaatgttcttaatgtttctttagACAGATGCACATATCGAGGTGAACATCACGTACGC
AlatLuc1    GGAGCAATTGCTTTTgtaagttcgaaattaattttataaaaaaattcttctaaactcaatttttgtattaaactaaaatttagACTAACGCACTTACCGGTGTCGATTATACGTACGC
PpyrLuc2    GGTTGTATTGCTATGgtaagc-----ttgtacctatgca-------------cattgcttgcagcttgttcaaacatttttagACGAACGCGCATACTAAAGAAAATCTGCTGTATGA
AlatLuc2    GGATGCATTGCTTTGgtaagtacc--ttttatttttata----------------ttaagtcgttagctttttttatactttagACAAGTGCGCATACTAAAGAAAATATGCTATATAA

PpyrLuc1    GGAATACTTCGAAATGTCCGTTCGGTTGGCAGAAGCTATGAAACGATATGGGCTGAATACAAATCACAGAATCGTCGTATGCAGTGAAAACTCTCTTCAATTCTTTATGCCGGTGTTGGG
AlatLuc1    CGAATACTTAGAAAAATCATGCTGTCTAGGAGAGGCTTTAAAGAATTATGGTTTGGTTGTTGATGGAAGAATTGCGTTATGCAGTGAAAATTGTGAAGAATTCTTTATTCCTGTATTAGC
PpyrLuc2    AGACGTACTGACATTAACCACTCGATTGGCGGTTGCTTACAAAAACTACGGTCTCGACATTAACAGCACAATTGCGGTGTGCAGCGAAAACAGCTTGCAATTCTTTCTACCAGTGATCGC
AlatLuc2    AGACTTATTACAATCAACATGCCGATTAGCCGAAAGTTTAAAAAAATATGGAATTACAACAAATAGCACAATTGCCGTGTGCAGTGAAAATAACTTACAGTACTTTATTCCTGTTATTGC

PpyrLuc1    CGCGTTATTTATCGGAGTTGCAGTTGCGCCCGCGAACGACATTTATAATGAACgtaagcaccctcgccatcagacccaaagg--gaatgacgtatttaat--ttttaagGTGAATTGCTC
AlatLuc1    CGGTTTATTTATAGGTGTCGGTGTGGCTCCAACTAATGAGATTTACACTCTACgtaagcacctaaacgtttagtagaacgtagtatttacagtaaacaaa--tttttagGTGAATTGGTT
PpyrLuc2    CGCCTTATACCTCGGAGTGACCGTTGCGTCCATAAATGACAAGTACACCGAGCgtaagta-------aagtgctcggtattg--ctgaaaagaaaacaat--atttagGTGAACTACTT
AlatLuc2    AGCTTTATACATCGGAGCTGCTACCGCAGCTGTTAACGACAAATACAATGAACgtaagaacgtaagaatgtaatagaaactg--actagctttataaaataattttagGAGAGTTAATT

PpyrLuc1    AACAGTATGAACATTTCGCAGCCTACCGTAGTGTTTGTTTCCAAAAAGGGGTTGCAAAAAATTTTGAACGTGCAAAAAAAATTACCAATAATCCAGAAAATTATTATCATGGATTCTAAA
AlatLuc1    CACAGTTTAGGCATCTCTAAGCCAACAATTGTATTTAGTTCTAAAAAAGGATTAGATAAAGTTATAACTGTACAAAAAACGGTAACTGCTATTAAAACCATTGTTATATTGGACAGCAAA
PpyrLuc2    CATAACTTTGAGATAACGAAACCTAGCGTGGTTTTCTGTTCCAAAAGGGCCGTAAAGAACATTCAGACAGTGAAGCACCGGCTAACTTACATTAATACAGTGGTCATATTGGATGACATC
AlatLuc2    AATTGTTTAAATTTATCAAAACCGACTTTTTTATTCTGTTCAAAAGAAACTTGGCCAAAAATACGTCAAGCTAAAAAAAAACTAGATTTTATTAAAAAAAATAATTATTCTTGATAATAAA

PpyrLuc1    ACGGATTACCAGGGATTTCAGTCGATGTACACGTTCGTCACATCTCATCTACCTCCCGGTTTTAATGAATACGATTTTGTACCAGAGTCCTTTGATCGTGACAAAACAATTGCACTGATA
AlatLuc1    GTGGATTATAGAGGTTATCAATCCATGGACAACTTTATTAAAAAAAACACTCCACCAGGTTTCAAAGGATCAAGTTTTAAAACTGTAGAAGTTAACCGCAAAGAACAAGTTGCGCTTATA
PpyrLuc2    ACCGACTGGCAAGATTTCCCTTGCCTAAACAACTTCATTTGAAGTTTTGCGATCCAAATTTAAATATTGGAGATTTCAAGCCCAATTCGTTCGATCGTGATAACCAAGTTGCACTTGTT
AlatLuc2    AACGACAGTGATTCACCACAATCCTTAGAAAATTTTATTTTTCAAAATTGTGACAAAGATTTTAACGTAAGTCAATTTAAACCAAATATATTTAACCGCGATGAGCACGTTGCATTGATA

PpyrLuc1    ATGAATTCCTCTGGATCTACTGGGTTACCTAAGGGTGTGGCCCTTCCGCATAGAACTGCCTGCGTCAGATTCTCGCATGCCAGgtat------gtcgta-taacaagagattaagtaatg
AlatLuc1    ATGAACTCTTCGGGTTCTACCGGTTTGCCAAAAGGTGTGCAACTTACTCATGAAAATGCAGTCACTAGATTTTCTCACGCTAGgtacatattagttata-tagtaaaagtctatattta
PpyrLuc2    ATGTACTCATCTGGCACAACAGGCGTGTCTAAAGGTGTCATGATAACCCATAAGAACATCATTGCTCGATTTTCGCACTGCAAgtcc------gtaatactcgcatcgcgcttgttaacc
AlatLuc2    TTAAATTCGTCGGGGTCGAGTGGATTGCCTAAAGGTGTAATGTTAACACATAAAAACTTAGCGGTGAGATTTTGTCATTGCAAgtaa------gtaaaa-aaattacacatgctttttct

PpyrLuc1    ttgctacacacattgtagAGATCCTATTTTTGGCAATCAAATCATTCCGGATACTGCGATTTTAAGTGTTGTTCCATTCCATCACGGTTTTGGAATGTTTACTACACTCGGATATTTGAT
AlatLuc1    taatttc-----tattagAGATCCAATTTATGGAAACCAAGTTTCACCAGGCACGGCTATTTTAACTGTAGTACCATTCCATCATGGTTTTGGTATGTTTACTACTTTAGGCTATCTAAC
PpyrLuc2    acgctat-aattttcagAGATCCGACTTTTGGGAACCAAATCAATCCGACCACTGTCATTTTAACGGTGGTACCATTCCAACACAGCTTTGGTATGTTTACAAGTCTAGGATACATGAC
AlatLuc2    ttacgtttaacacttaagGGATCCCATTTTTGGTAATCAAATAAGTCCGGGTACTGCAATTTTAACAGTTATACCATTTCACCATGGATTTGGAATGTTCACTACTTTGGGATATTTTAC

PpyrLuc1    ATGTGGATTTCGAGTCGTCTTAATGTATAGATTTGAAGAAGAGCTGTTTTTACGATCCCTTCAGGATTACAAAATTCAAAGTGCGTTGCTAGTACCAACCCTATTTTCATTCTTCGCCAA
AlatLuc1    TTGTGGTTTTCGTATTGTCATGTTAACAAAATTTGACGAAGAAACTTTTTTAAAAACACTGCAAGATTACAAATGTTCAAGCGTTATTCTTGTACCGACTTTGTTTGCAATTCTTAATAG
PpyrLuc2    CTGCGGATTTCGAATCGTCGTATTAACCACGTTTGATGAAAAGCTCTTTTTGCAATCCCTTCAAGATTATAAAGTGGCAAGCACTTTACTAGTGCCTACCCTGATGTCCTTGTTCGCAAA
AlatLuc2    ATGCGGGTTTCGAATTGTTTTAATGCATACATTTGAAGAACATTTGTTTTTACAATCATTACAAGATTATAAAGTTAAAAGTACTTTGTTGGTACCTACGTTAATGACTTTTTTTGCCAA

PpyrLuc1    AAGCACTCTGATTGACAAATACGATTTATCTAATTTACACGAAATTGCTTCTGGGGGCGCACCTCTTTCGAAAGAAGTCGGGGAAGCGGTTGCAAAACGgtgagttaagcgcattgctag
AlatLuc1    AAGTGAATTACTCGATAAATATGATTTATCAAATTTAGTTGAAATTGCATCTGGCGGAGCACCTTTATCTAAAGAAATTGGTGAAGCTGTTGCTAGACGgtaatttttgtttataaattt
PpyrLuc2    AAGCGCAATCGTCGAGAACTACGATCTGTCGCACTTGGAAGAGATCGCCTCGGGTGGAGCACCTTTATCCAAGCAAATCAGCGATGCGGTTAGGAAACGgtgagtctgcggcgtttttttg
AlatLuc2    AAGTCCATTAGTAGACAAATTTCATTTGCCTTATTTACACGAAATTGCGTCGGGAGGTGCACCTCTGTCAAAAGAAATTGGTGAAGCTGTTGCACTAAGgtaatatttttttgaattattt

PpyrLuc1    tatttcaa--ggctctaaaacggcgcgtagCTTCCATCTTCCAGGGATACGACAAGGATATGGGCTCACTGAGACTACATCAGCTATTCTGATTACACCCGAGGGGGATGATAAACCGGG
AlatLuc1    ttaatcaaatactttataaatctgttgcagTTTTAATTTACCGGGTGTTCGTCAAGGCTATGGTTTAACAGAAACAACCTCTGCAATTATTATCACACCGGAAGGCGATGATAAACCAGG
PpyrLuc2    accat-----cctcttatcttccagtacagATTTAAGCTAAACCAGATCAGGCAAGGATACGGGCTCACCGAAACTACCTCGGCAGTGTTAATTACGCCAGATACCGGCGTCATACCGGG
AlatLuc2    tcaat-----attaattacgtaaagtttagATTTAAATTGAAATCAATTAGACAAGGTTATGGTTTAACCGAAACAACTTCGGCTATTTTATTAACACCTGAAGGAGAAATAGTACCTGG

PpyrLuc1    CGCGGTCGGTAAAGTTGTTCCATTTTTTGAAGCGAAGGTTGTGGATCTGGATACCGGGAAAACGCTGGGCGTTAATCAGAGAGGCGAATTATGTGTCAGAGGACCTATGATTATGTCCGG
AlatLuc1    TGCTTCTGGCAAAGTTGTGCCATTATTTAAAGCAAAAGTTATCGATCTTGATACTAAAAAAACTTTGGGCCCGAACAGACGTGGAGAAGTTTGTGTAAAGGGTCCTATGCTTATGAAAGG
PpyrLuc2    CTCTACCGGAAAAATTGTCCCCTTTCACGCCGTAAAAGTTGTCGATACAGCTACTGGAGAAAACTTGGGGCCCAATCGAACTGGCGAATTGTATTTCAAAGGTGACATGATAATGAAGGG
AlatLuc2    ATCGACAGGAAAAGTAGTACCCTTTTTTGCAGCTAAAGTTGTAGATAACGACACTGGTAGAATACTAGGACCAAATGAAGTTGGAGAATTGTGCTTTAAAGGAGATATGAATATGAAAGG

PpyrLuc1    TTATGTAAACAATCCGGAAGCGACCAACGCCTTGATTGACAAGGATGGATGGCTACATTCTGGAGACATAGCTTACTGGGACGAAGACGAACACTTCTTCATAGTTGACCGCTTGAAGTC
AlatLuc1    TTATGTAGATAATCCAGAAGCAACAAGAGAAATCATAGATGAAGAAGGTTGGTTGCACACAGGAGATATTGGGTATTACGATGAAGAAAAACATTTCTTTATCGTGGATCGTTTGAAGTC
PpyrLuc2    CTACTGTAACAACGCCCCAGCTACCGACGCAATTATTGACCCAAATGGGTGGTTGCGATCCGGCGACATCGGCTATTACGATGGGAATGGAAATTTTTTCATCGTGGACAGAATTAAATC
AlatLuc2    TTACTGTAATGATATCAAAGCTACCAACGCTATTATTGATAAAGAAGGATGGTTACATTCAGGTGATCTCGGATATTATGACGAAAACGAACATTTTTTTATTGTTGATCGACTAAAATC

PpyrLuc1    TTTAATTAAATACAAAGGATATCAGgtaatgaagattttacatgcacacacgctacaatacc------tgtagGTGGCCCCCGCTGAATTGGAATCGATATTGTTACAACACCCCAACA
AlatLuc1    TTTAATCAAATACAAAGGATATCAAgtaatatttttttaaccgataaaaataattctaaatatt---taatttagGTACCACCTGCTGAATTAGAATCTGTTCTTTTGCAACATCCAAATA
PpyrLuc2    ACTAATAAAGTACAAGGGCTTCCAGgcaggttttcctacagttttggtcgattttaaaatg-----tattgtagGTTGCACCCGCCGAAATTGAAGCAGTACTACTGCAACACCCGGACA
AlatLuc2    TTTAATCAAATACAAAGGATACCAGgtacgtttttttaaagtcatttctttgtgttattttgtccgatgctttagGTTGCTCCTGCCGAATTGGAAGGAATATTATTAACTCATCCAAGTA

PpyrLuc1    TCTTCGACGCGGGCGTGGCAGGTCTTCCCGACGATGACGCCGGTGAACTTCCCGCCGCCGTTGTTGTTTTGGAGCACGGAAAGACGATGACGGAAAAAGAGATCGTGGATTACGTCGCCA
AlatLuc1    TTTTTGATGCCGGCGTTGCTGGCGTTCCAGATCCTATAGCTGGTGAGCTTCCGGGAGCTGTTGTTGTACTTGAAAAAGGAAAATCTATGACTGAAAAAGAAGTAATGGATTACGTTGCAA
PpyrLuc2    TTCTCGACGCGGGCGTTACGGGTATTAAAGACGACGAAGCGGGCGAAATACCGGCGGCGGCTATAGTCATAAAGAAAGGCGCACATTTAGACGAAGAAGACGTGAAGAAATACGTTGAAA
AlatLuc2    TCATGGACGCGGGTGTTACTGGTATACCGGATGAACACGCTGGTGAACTTCCAGCAGCATGTGTCGTAGTTAAACCAGGGCGAAACCTCACTGAAGAAAATGTCATAAATTACGTCTCAA

PpyrLuc1    gtaaatgaat-------tcgttttacgttactcgtactaca-attcttttcatagGTCAAGTAACAACCGCGAAAAAGTTGCGCGGAGGAGTTGTGTTTGTGGACGAAGTACCGAAAGGT
AlatLuc1    gtaactattattcaacactagttaaagtaaatactactaca---tttttgtgtagGTCAAGTTTCAAATGCAAAACGTTTGCGTGGTGGTGTCCGTTTTGTGGACGAAGTGCCTAAAGGT
PpyrLuc2    gtaagtgtcg-gcatcaagaggccgacgaactaatttt------tcggttttcagGCCAAATGTCTTCGACAAGGTGGTTACGGGGCGGTGTGCGCTTTTTGGATGAAATCCCAAAAGGT
AlatLuc2    gtaattcttt-tttatattggtatttttaatatttatatataatttctattagGCCAGGTATCTTCTTCGAAGAGATTGCGTGGAGGTGTTCGTTTTATAGATAACATTCCAAAAGGA

PpyrLuc1    CTTACCGGAAAACTCGACGCAAGAAAAATCAGAGAGATCCTCATAAAGGCCAAGAAGGGCGGAAAGTCCAAATTGTAA
AlatLuc1    CTTACTGGTAAAATTGACGGTAAAGCAATTAGAGAAATACTGAAGAAA------------CCAGTTGCTAAGATGTAA
PpyrLuc2    CCGACCGGTAAAATTGATGGAAAAGCCATACGGGAAATATTTGAGAAG------------CAAAAATCTAAGCTGTAA
AlatLuc2    TCTACCGGCAAAATTGACACAAAAGCTTTAAAACAAATTTTACAAAAA------------CAAAAATCCAAGTTATAA
```

Supporting Information 4—figure 5. Multiple sequence alignment of firefly luciferase genes.

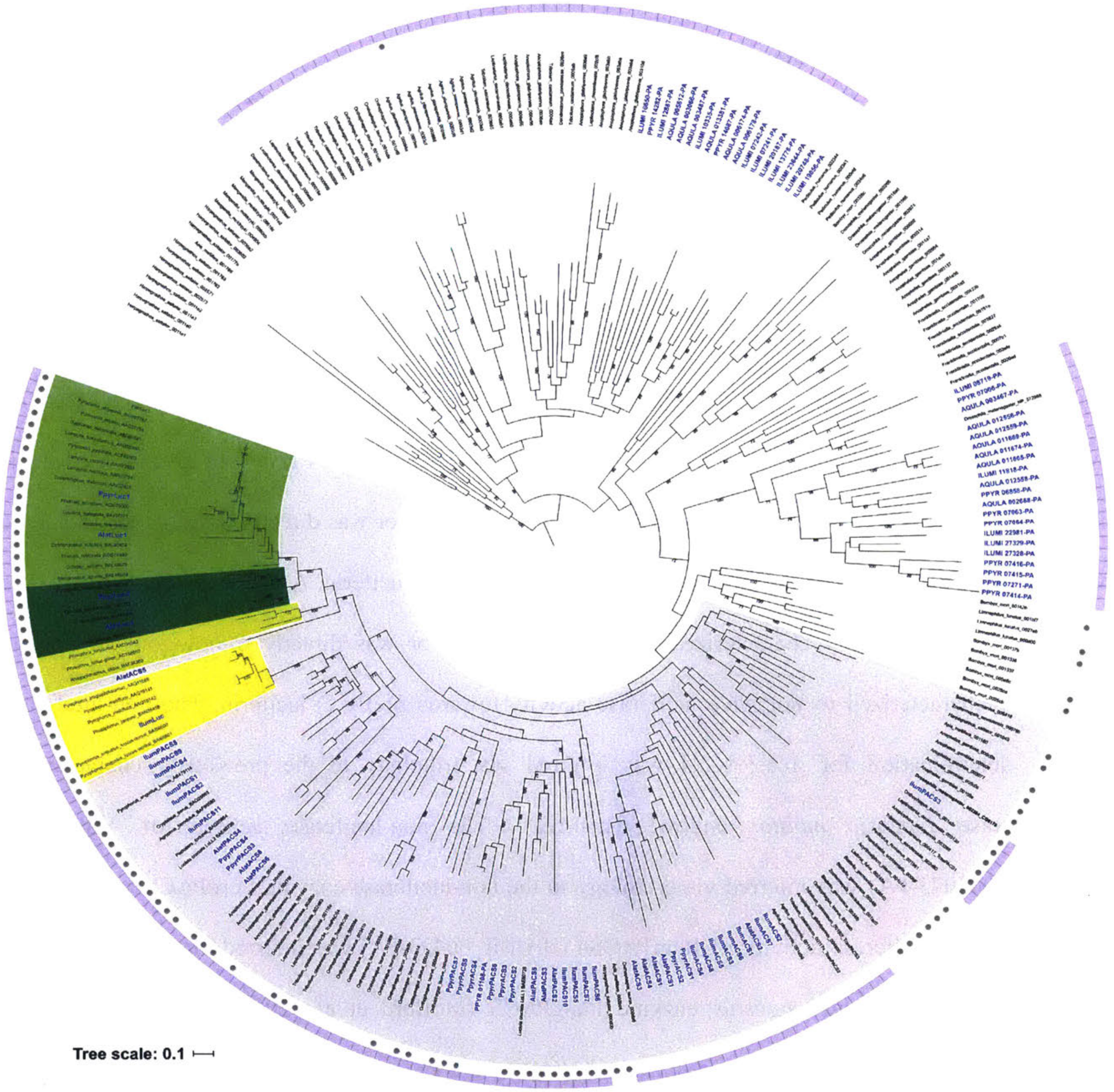
MAFFT (Katoh and Standley, 2013) L-INS-i multiple sequence alignment of luciferase gene nucleotide sequences from PpyrOGS1.1 and AlatOGS1.0 demonstrates the location of intron-exon junctions (bolded blue text) is completely conserved amongst the four luciferases. Exonic sequence is capitalized, whereas intronic sequence is lowercase.

4.3.2 Luciferase homolog gene tree (Figure 3C)

From our reference genesets, a protein BLAST search detected 24, 20, 32, and two luciferase homologs (E-value <1 × 10−60) to *P. pyralis* luciferase (PpyrLuc1; Genbank accession AAA29795) from the *P. pyralis, A. lateralis, I. luminosus* genesets, and *Drosophila melanogaster,* respectively. We defined the luciferase co-orthology as followings: (1) shows an BLASTP E-value lower than 1.0 × 10−60 toward *DmelPACS*(CG6178), (2) phylogenetically sister to *DmelPACS*, which is the most similar gene to firefly luciferase in *D. melanogaster,* based on a preliminary maximum likelihood (ML) phylogenetic reconstruction (Supporting Information 4—figure 6). Preliminary ML phylogenetic reconstruction was performed as follows: The sequences of luciferase homologs from *Mengenilla moldrzyki, Pediculus humanus, Limnephilus lunatus, Ladona fulva, Frankliniella occidentalis, Zootermopsis nevadensis, Onthophagus taurus, Anoplophora glabripennis, Agrilus planipennis, Harpegnathos saltator, Blattella germanica, Acyrthosiphon pisum, Tribolium castaneum, Bombyx mori, Anopheles gambiae, Apis mellifera, Leptinotarsa decemlineata,* and *Dendroctonus ponderosae* were obtained from OrthoDB (https://www.orthodb.org) (Zdobnov et al., 2017). The sequences which show 99% similarity were filtered by CD-HIT (v4.7) (Fu et al., 2012). The resulting sequences and beetle luciferases were aligned using (MAFFT v7.309) (Katoh and Standley, 2013) using the BLOSUM62 matrix and filtered for spurious sequences and poorly aligned regions using trimAl (v.1.2rev59) (Capella-Gutiérrez et al., 2009) (parameters: -strict). The final alignment was 385 blocks and 264 sequences. Then, the best fit amino acid substitution model, LG + F Gamma, was estimated by Aminosan (v1.0.2016.11.07) (Tanabe, 2011) using the Akaike Information Criterion. Finally, a maximum likelihood gene phylogeny was estimated using RAxML (v8.2.9; 100 bootstrap replicates) (Stamatakis, 2006). Supporting files such as multiple sequence alignment, gene accession numbers, and other annotations are available on FigShare (DOI: 10.6084/m9.figshare.6687086).

To more closely examine luciferase evolution, an independent maximum likelihood gene tree was constructed for luciferase co-orthologous genes defined above (highlighted clade as grey in Supporting Information 4—figure 6) with well important genes: non-luminescent luciferase homolog from two model insect *D. melanogaster*(DmelPACS and DmelACS as outgroup) and *T. castaneum* (TcasPACSs and

TcasACSs), biochemically characterized non-luminescent PACS (LcruPACS1 and LcruPACS2 from *Luciola cruciata,* DmelPACS, and PangPACS from *Pyrophorus angustus*) and biochemically characterized luciferases from Lampyrinae (PatrLuc1 and 2: *Pyrocoelia atripennis*), Ototretinae (DaxiLuc1 and SazuLuc1: *Drilaster axillaris* and *Stenocladius azumai), Phausis* (PretLuc1: *Phausis reticulata*) from Lampyridae, Rhagophthalmidae (RohbLuc: *Rhagophthalmus ohbai*), Phengodidae (PhirLucG and R: *Phrixothrix hirtus*), and Elateridae (PangLucD and V: *P. angustus*). Then co-orthologous genes were confirmed to be phylogenetically sister to *DmelPACS* (CG6178) and their evolution examined using a maximum likelihood (ML) gene phylogeny approach. First, amino acid sequences were aligned using (MAFFT v7.308) (Katoh and Standley, 2013) using the BLOSUM62 matrix (parameters: gap open penalty = 1.53, offset value = 0.123) and filtered for spurious sequences and poorly aligned regions using trimAl (Capella-Gutiérrez et al., 2009) (parameters: gt = 0.8). The final alignment was 533 blocks and 67 sequences. Then, the best fit amino acid substitution model, LG + F Gamma, was estimated by Aminosan (v1.0.2016.11.07) (Tanabe, 2011) using the Akaike Information Criterion. Finally, a maximum likelihood gene phylogeny was estimated using RAxML (v8.2.9; 100 bootstrap replicates) (Stamatakis, 2006). The tree was rooted using *DmelACS* as an outgroup. The peroxisomal targeting signal 1 (PST1) was predicted using the regular expressions provided by the Eukaryotic Linear Motif database (Dinkel et al., 2012) and verified using the mendel PTS1 prediction server (http://mendel.imp.ac.at/pts1/). Supporting files such as multiple sequence alignment, gene accession numbers, and other annotation and expression values are available as Figure 3—source data 1.



Supporting Information 4—figure 6. Preliminary maximum likelihood phylogeny of luciferase homologs.

A preliminary maximum likelihood tree was reconstructed from a 385 amino acid multiple sequence alignment, generated via a BLASTP and orthoDB search using *P. pyralis* luciferase as query (e-value: $1.0 \times 10_{-60}$). Members of the clade that includes both known firefly luciferase and CG6178 of *D. melanogaster* (bold) are defined as luciferase co-orthologous genes (highlighted in gray), and were selected and used for the independent maximum likelihood analysis in Figure 3C (Supporting Information 4.3.2). Branch length represents substitutions per site. Genes found from this study are indicated in blue. Lampyridae Luc1-type and Luc2-type luciferases are highlighted in yellow-green and green. Rhagophthalmidae and Phengodidae luciferases are highlighted in lime-green. Elateridae luciferases are highlighted in yellow. Genbank accession numbers of luciferase orthologs genes are indicated after the species name. OrthoDB taxon and protein IDs of luciferase co-orthologs are indicated after species name. Bootstrap values are indicated on the nodes. The genes from Coleoptera are indicated as purple strip. Grey closed circles indicate genes that have PTS1.

4.3.3 Ancestral state reconstruction of luciferase activity (Figure 4A)

We performed an ancestral character state reconstruction of luciferase activity on the luciferase homolog gene tree within Mesquite (v3.31) (Maddison and Maddison, 2017), using an unordered parsimony analysis, and maximum likelihood (ML) analyses. First, the gene tree from Figure 3C in Newick format was filtered using Dendroscope(v3.5.9) (Huson and Scornavacca, 2012) to include only the clade descending from the common ancestor of TcasPACS4 and PpyrLuc1. TcasPACS4 was set as the rooting outgroup. Luciferase activity of these extant genes was coded as a character state within Mesquite with: (0 = no luciferase activity, 1 = luciferase activity, ?=undetermined). A gene was given the 1-state if it had been previously characterized as having luciferase activity, or was directly orthologous to a gene with previously characterized luciferase activity against firefly D-luciferin. A gene was given the 0-state if it had been previously characterized as a non-luciferase, or was directly orthologous to a gene previously characterized to not have luciferase activity towards firefly D-luciferin. The non-luciferase activity determination for TcasPACS4 was inferred via orthology to the previously characterized non-luciferase *Tenebrio molitor* enzyme Tm-LL2 . The non-luciferase activity of AlatPACS4 (AQULA_005073-PA) was inferred via orthology to the non-luciferase enzyme LcruPACS2 (Oba et al., 2006). The non-luciferase activity of IlumPACS4 (ILUMI_06433-PA) was inferred via orthology to the non-luciferase *Pyrophorus angustus* enzyme PangPACS (Mofford et al., 2017; Oba et al., 2010a). IlumLuc luciferase activity was inferred via orthology to the *P. angustus* dorsal and ventral luciferases (Oba et al., 2010a). The luciferase activity of PpyrLuc2 (PPYR_00002-PA) was inferred via orthology to other Luc2s, e.g. *A. lateralis* Luc2 (Oba et al., 2013a). The luciferase activity of the included phengodid (Amaral et al., 2017; Arnoldi et al., 2010; Viviani et al., 1999a), rhagophthalmid (Ohmiya et al., 2000; Viviani et al., 1999a), and firefly luciferases (Branchini et al., 2017; Oba et al., 2012; Viviani et al., 2011) were annotated from the literature. We then reconstructed the ancestral luciferase activity character state over the tree, using an unordered parsimony model, and a maximum likelihood (ML) model. ML analyses were performed under the AsymmMk model with default parameters (i.e. Root State Frequencies Same as Equilibrium). NEXUS files with presented parsimony and ML reconstructions are available as Figure 4—source data 1.

4.3.4 Testing for ancestral selection of elaterid ancestral luciferase (Figure 4B)

Selection of peptide sequences

Peptide sequences for elaterid luciferase homologs descending from the putative common ancestor of firefly and elaterid luciferase as determined by a preliminary maximum likelihood molecular evolution analysis of luciferase homologs (not shown), were selected from Uniprot, whereas their respective CDS sequences were selected from the European Nucleotide Archive (ENA) or National Center for Biotechnology Information (NCBI). These sequences include: The dorsal (PangLucD; ENA ID = BAI66600.1) and ventral (PangLucV; ENA ID = BAI66601.1) luciferases, and a luciferase-like homolog without luciferase-activity (PangPACS; ENA ID = BAI66602.1) from *Pyrophorus angustus* (Oba et al., 2010a), and two unpublished but database deposited luciferase homologs without luciferase-activity (data not shown) from *Cryptalaus berus* (CberPACS; ENA ID = BAQ25863.1) and *Pectocera fortunei fortunei* (PffPACS; ENA ID = BAQ25864.1). The peptide and CDS sequence of the *Pyrearinus termitilluminans* luciferase (PtermLuc) were manually transcribed from the literature (Viviani et al., 1999b), as these sequences were seemingly never deposited in a publically accessible sequence database. The dorsal (PmeLucD; NCBI ID = AF545854.1) and ventral (PmeLucV; NCBI ID = AF545853.1) luciferases of *Pyrophorus mellifluus* (Stolz et al., 2003). The dorsal (AF543412.1) and ventral (AF543401.1) luciferase alleles of *Pyrophorus plagiophthalmus (Stolz et al., 2003)*, which were most similar to that of *Pyrophorus mellifluus* in a maximum likelihood analysis (data not shown). The CDS sequence of the complete *I. luminosus* luciferase (IlumLuc; ILUMI_00001-PA), two closely related paralogs (IlumPACS9: ILUMI_26849-PA, IlumPACS8: ILUMI_26848-PA), and two other paralogs (IlumPACS2:

ILUMI_02534-PA; IlumPACS1: ILUMI_06433-PA), and the CDS for *Photinus pyralis* luciferase (PpyrLuc1: PPYR_00001-PA) were added as an outgroup sequence.

Alignment and Gene Phylogeny

The 20 merged CDS sequences were multiple-sequenced-aligned with MUSCLE (Edgar, 2004) in 'codon' mode within MEGA7 (Kumar et al., 2016), using parameters (Gap Open = −0.2.9; Gap Extend = 0; Hydrophobicity Multiplier 1.2, Clustering Method = UPGMB, Min Diag Length (lambda) = 24, Genetic Code = Standard), producing a nucleotide multiple-sequence-alignment (MSA). A maximum likelihood gene tree was produced from the nucleotide MSA within MEGA7 using the General Time Reversible model (Nei and Kumar, 2000), with five gamma categories (+G, parameter = 0.8692). The analysis involved 20 nucleotide sequences. Codon positions included were 1st + 2nd + 3rd + Noncoding. There were a total of 1659 positions in the final dataset. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with the superior log likelihood value. The tree with the highest log likelihood (−16392.22) was selected. 1000 bootstrap replicates were performed to evaluate the topology, and the percentage of trees in which the associated taxa clustered together is shown next to the branches in Figure 4B.

Tests of selection: aBSREL

An adaptive branch-site REL test for episodic diversification was performed on the previously mentioned gene-tree and nucleotide MSA using the adaptive branch-site REL test for episodic diversification (aBSREL) method (Smith et al., 2015) within the HyPhy program (v2.3.11) (Pond et al., 2005). The input MSA contained 20 sequences with 553 sites (codons). All 37 branches of the gene phylogeny were formally tested for diversifying selection. The aBSREL analysis found evidence of episodic diversifying selection on 3 out of 37 branches in the phylogeny. Significance was assessed using the Likelihood Ratio

Test at a threshold of $p \leq 0.01$, after the Holm-Bonferroni correction for multiple hypothesis testing. The intermediate files and results of this analysis, including the nucleotide MSA, GTR based gene-tree, and aBSREL produced adaptive rate class model gene tree are available as Figure 4—source data 2.

Tests of selection: MEME

After identification of the selected branch via the aBSREL method, we turned to the MEME method within the HyPhy program (v2.3.11) (Pond et al., 2005), to identify those sites which may have adaptively evolved. We tested the branch leading to EAncLuc, which was previously identified as under selection in the aBSREL analysis. A single partition was recovered with 28 sites under episodic diversifying positive selection at $p<=0.1$ (Supporting Information 4—table 5). Input files and full results are available on FigShare (10.6084/m9.figshare.6626651).

Tests of selection: PAML-BEB

To validate our findings from aBSREL and MEME using a different method, we applied Phylogenetic Analysis by Maximum Likelihood (PAML) branch by site analysis to the luciferase sequences. We tested the alternative hypothesis, that there is a class of sites under selection ($\omega > 1$) on the EAncLuc ancestral branch identified as under selection in the aBSREL analysis, against the null hypotheses, that all classes of sites on all branches are evolving either under constraint ($\omega < 1$) or neutrality ($\omega = 1$). A likelihood ratio test supported the alternative hypothesis, that 13% of sites in lucifcrase were in a positively selected class ($\omega = 3.25$). Subsequent Bayes Empirical Bayes (BEB) estimation identified 31 sites with evidence of selection on these branches, 5 of which were significant. Full results are available on FigShare (10.6084/m9.figshare.6725081).

Tests of selection: Overlap

Nineteen of the overall sites were shared between the MEME analysis, and are shown in Supporting Information 4—table 5. The frequency of extant amino acids at these sites are shown in Supporting Information 4—figure 7.

Supporting Information 4—table 4. Results of PAML branch x sites analysis.
Proportion indicates the proportion of sites in each site class (0, 1, 2a, 2b). Site classes 0 and 1 are those in the constrained and neutral classes, respectively. 2a are sites that were constrained on the background branches, but are either neutral (H0) or in the selective class (HA) on the foreground branches. 2b are sites that were neutral on the background branches, but are either neutral (H0) or in the selective class (HA) on the foreground branches.

Hypothesis	Site class:	0	1	2a	2b	lnL
H0: no selection	proportion	0.62	0.14	0.18	0.04	−15888.16
	background ω	0.12	1	0.12	1	
	foreground ω	0.12	1	1	1	
HA: selection	proportion	0.71	0.15	0.11	0.02	−15833.50*
	background ω	0.12	1	0.12	1	
	foreground ω	0.12	1	3.25	3.25	

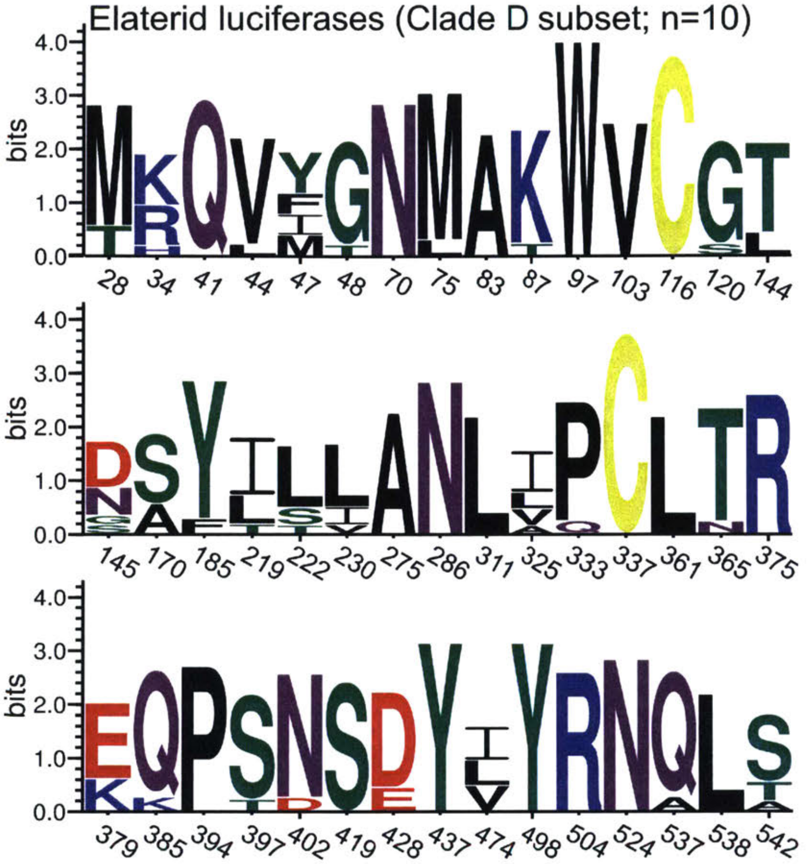
*significant (LRT: 9.32, df = 1)

Supporting Information 4—table 5. Sites identified as under selection on foreground branches using both Bayes Empirical Bayes (BEB) and Mixed Effects Model of Evolution (MEME).

Site numbering			MEME[2]					PAML-BEB	
MSA	*IlumLuc*	*IlumLuc site AA*[1]	α	β+	LRT	Episodic selection p-value	# branches	BEB site class probability	BEB significance
28	28	M						0.986	*
34	34	K	0.47	23.5	4.1	0.0603	0		
41	41	Q						0.5	
46	44	V	0	3	4.5	0.0485	0		
49	47	I	0.93	792.4	3.8	0.0692	0		
50	48	G	0.57	3332.3	4.8	0.0427	0	0.836	
72	70	N	0.55	3333.1	3.1	0.0998	0	0.776	
77	75	M						0.964	*
85	83	A						0.962	*
89	87	K						0.958	*

99	97	W						0.598	
105	103	V	0.44	6.8	4.3	0.0549	0	0.768	
118	116	C	0.3	3333.1	7.4	0.0109	1		
122	120	G						0.82	
146	144	L	0.34	12.8	4.9	0.039	0		
147	145	G	0.75	3333.6	5.9	0.0236	0		
172	170	A						0.698	
189	185	F						0.534	
223	219	L						0.507	
226	222	T	1.44	29.6	4.8	0.0427	0	0.889	
234	230	I	1.13	9.6	3.1	0.0991	0	0.613	
279	275	A						0.559	
290	286	N	0.92	3333	4	0.064	0		
315	311	L	0.69	29.5	5.1	0.0362	0	0.884	
329	325	L						0.766	
337	333	P	0.26	13.3	6.3	0.0198	0		
341	337	C						0.812	
365	361	L	0.58	7.6	4.4	0.052	0	0.912	
369	365	T	0.21	6.8	6.6	0.0169	0	0.843	
379	375	R						0.932	
383	379	E	0	2.8	4.1	0.0594	0		
389	385	Q						0.792	
398	394	P	0.96	1999.2	4.5	0.05	0	0.951	*
401	397	S						0.617	
406	402	N	0.58	5.5	3.7	0.0745	0	0.949	
423	419	S	0.67	1574.6	4.7	0.043	0	0.569	
432	428	E	0	2.9	3.1	0.0999	1		
441	437	Y	1.43	39.3	4.2	0.0573	0	0.912	
478	474	V	0	10.3	6.9	0.0139	1	0.646	
502	498	Y	0.5	1790.4	4.9	0.0393	0	0.583	
508	504	R						0.519	
528	524	N	0	2.2	3.6	0.0772	0		
541	537	Q	0	1999.2	10.4	0.0024	1		
542	538	L	0.56	68	6.3	0.0197	0		
550	542	T	0.74	3332.9	4.3	0.0541	0		

1 = amino acid. 2 = All recovered sites in a single partition with a p+ value of 1.000.



Supporting Information 4—figure 7. Amino acid variation at sites recovered in selection analysis. Amino acid variation of extant Elaterid luciferases (Clade D 'Eluc' subset; Figure 3) at all sites recovered via both the MEME and PAML-BEB selection analysis (Supporting Information 4—table 5). Site numbering relative to *IlumLuc*. Figure produced with seqkit (Shen et al., 2016) and WebLogo(v3.6.0) (Crooks et al., 2004).

4.4 Non-enzyme highly and differentially expressed genes of the firefly lantern

PPYR_04589, a predicted fatty acid binding protein is almost certainly orthologous to the light organ fatty acid binding protein reported from *Luciola cerata* (Goh and Li, 2011). This fatty acid binding protein was previously reported to bind strongly to fatty acids, and weakly to luciferin. Notably, PPYR_04589 is the most highly expressed gene in the *P. pyralis* adult lantern, ahead of firefly luciferase. Three G-protein coupled receptors (GPCRs) with similarity to annotated octopamine/tyramine receptors were also detected to be highly and differentially expressed in the *P. pyralis* light organ (PPYR_11673-PA, PPYR_11364-PA, PPYR_12266-PA). Octopamine is known to be the key effector neurotransmitter of the adult and larval firefly lantern and this identified GPCR likely serves as the upstream receptor of octopamine activated adenylate cyclase, previously reported as abundant in *P. pyralis* lanterns (Nathanson et al., 1989).

The neurobiology of flash control, including regulation of flash pattern and intensity, is a fascinating area of behavioral research. Our data generate new hypotheses regarding the molecular players in flash control. A particularly interesting highly and differentially expressed gene in both *P. pyralis* and *A. lateralis* is the full length 'octopamine binding secreted hemocyanin'(PPYR_14966; AQULA_008529; Supporting Information 4—table 6) previously identified from *P. pyralis* light organ extracts via photoaffinity labeling with an octopamine analog and partial N-terminal Edman degradation (Nathanson et al., 1989). This protein is intriguing as hemocyanins are typically thought to be oxygen binding. We speculate that this octopamine binding secreted hemocyanin, previous demonstrated to be abundant, octopamine binding, and secreted from the lantern (presumably into the hemolymph of the light organ), could be triggered to release oxygen upon octopamine binding, thereby providing a triggerable O_2 store within the light organ under control of neurotransmitter involved in flash control. As O_2 is believed to be limiting in the adult light reaction, such a release of O_2 could enhance flash intensity or accelerate flash kinetics. Further research is required to test this hypothesis.

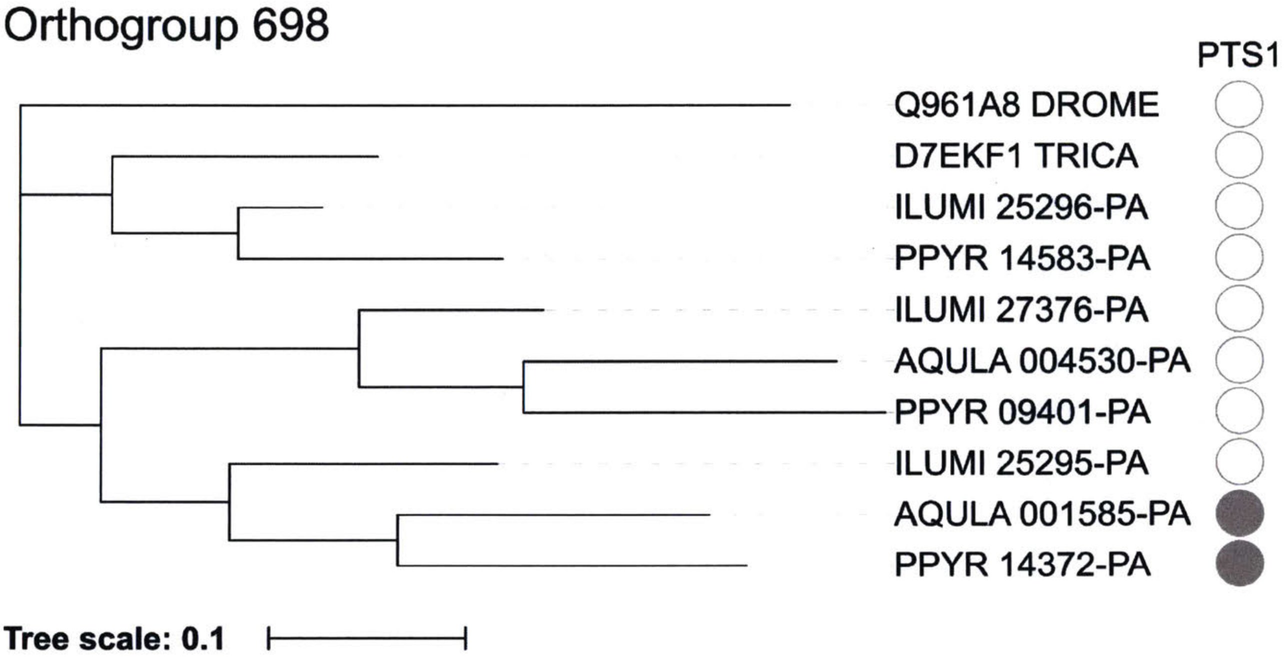
Supporting Information 4—table 6. Non-enzyme genes of the firefly lantern.

Highly expressed (HE), differentially expressed (DE), non-enzyme annotated (NotE), lantern genes whose closest relative in the opposite species is also HE, DE, NotE. BSN-TPM = between sample normalized TPM.

P. pyralis ID (OGS1.1)	Predicted function	Ppyr expression rank	Ppyr BSN-TPM	Orthogroup	Alat expression rank	Alat BSN-TPM	*A. lateralis* ID (OGS1.0)
PPYR_04589	Fatty-acid binding protein	1	70912	OG0000524	2	31943	AQULA_005253
PPYR_04589	Fatty-acid binding protein	1	70912	OG0000524	8	10464	AQULA_005257
PPYR_04589	Fatty-acid binding protein	1	70912	OG0000524	10	8520	AQULA_005259
PPYR_05098	Peroxisomal biogenesis factor 11 (PEX11)	15	4005	OG0001490	26	3294	AQULA_005466
PPYR_14966	Octopamine binding secreted hemocyanin	34	2353	OG0000369	21	3658	AQULA_008529
PPYR_11733	MFS transporter superfamily	42	1853	OG0000980	84	1335	AQULA_012209
PPYR_07633	Reticulon	56	1556	OG0004764	109	1123	AQULA_005090
PPYR_09394	lysosomal Cystine Transporter	87	1098	OG0000847	69	1494	AQULA_009474
PPYR_08979	PF03670 Uncharacterised protein family	114	860	OG0003009	340	411	AQULA_012099
PPYR_05852	Vacuolar ATP synthase 16 kDa subunit	118	836	OG0001039	287	475	AQULA_001418
PPYR_11443	RNA-binding domain superfamily	134	782	OG0004268	1221	108	AQULA_003174
PPYR_02465	Peroxin 13	189	581	OG0001667	196	710	AQULA_010288
PPYR_06160	V-type ATPase, V0 complex	209	543	OG0000381	541	251	AQULA_000400
PPYR_11300	Mitochondrial outer membrane	232	509	OG0004557	402	349	AQULA_004355

	translocase complex						
PPYR_08174	PF03650 Uncharacterised protein family	249	475	OG0000647	163	836	AQULA_009867
PPYR_04602	Leucine-rich repeat domain superfamily	262	459	OG0004508	378	373	AQULA_004134
PPYR_01678	MFS transporter superfamily	264	458	OG0000347	455	302	AQULA_002485
PPYR_08192	PF03650 Uncharacterised protein family	271	453	OG0000647	163	836	AQULA_009867
PPYR_13497	Mitochondrial substrate/solute carrier	285	438	OG0004402	379	372	AQULA_003680
PPYR_08917	LysM domain superfamily	315	398	OG0002035	483	278	AQULA_002396
PPYR_04424	Domain of unknown function (DUF4782)	332	379	OG0007447	1296	101	AQULA_013946
PPYR_08278	Protein of unknown function DUF1151	348	365	OG0001306	430	325	AQULA_000628
PPYR_13261	Major facilitator superfamily	404	309	OG0000410	158	862	AQULA_007558
PPYR_14848	Homeobox-like domain superfamily - Abdominal-B-like	413	304	OG0001849	737	186	AQULA_000483
PPYR_11623	GNS1/SUR4 family	446	281	OG0008603	308	449	AQULA_009341
PPYR_01828	TLDc domain	490	250	OG0002035	483	278	AQULA_002396
PPYR_03449	Innexin	533	230	OG0000992	619	219	AQULA_013430
PPYR_05702	Sulfate permease family	543	225	OG0007205	396	357	AQULA_013064

PPYR_05993	V-type ATPase, V0 complex, 116 kDa subunit family	579	210	OG0000381	541	251	AQULA_000400
PPYR_04179	Haemolymph juvenile hormone binding protein	606	202	OG0002916	879	152	AQULA_011187
PPYR_08298	Peroxisomal membrane protein (Pex16)	623	198	OG0007339	395	358	AQULA_013536
PPYR_06294	Homeobox-like domain superfamily - Abdominal-B-like	627	197	OG0001849	737	186	AQULA_000483
PPYR_05397	PDZ superfamily	773	164	OG0006975	367	379	AQULA_012321
PPYR_12625	Homeobox domain	796	160	OG0002661	1395	95	AQULA_008665
PPYR_08494	Armadillo-type fold	846	152	OG0001600	986	133	AQULA_008183
PPYR_09217	Haemolymph juvenile hormone binding protein	853	151	OG0001089	441	316	AQULA_003304
PPYR_01677	MFS transporter superfamily	1234	108	OG0000347	455	302	AQULA_002485



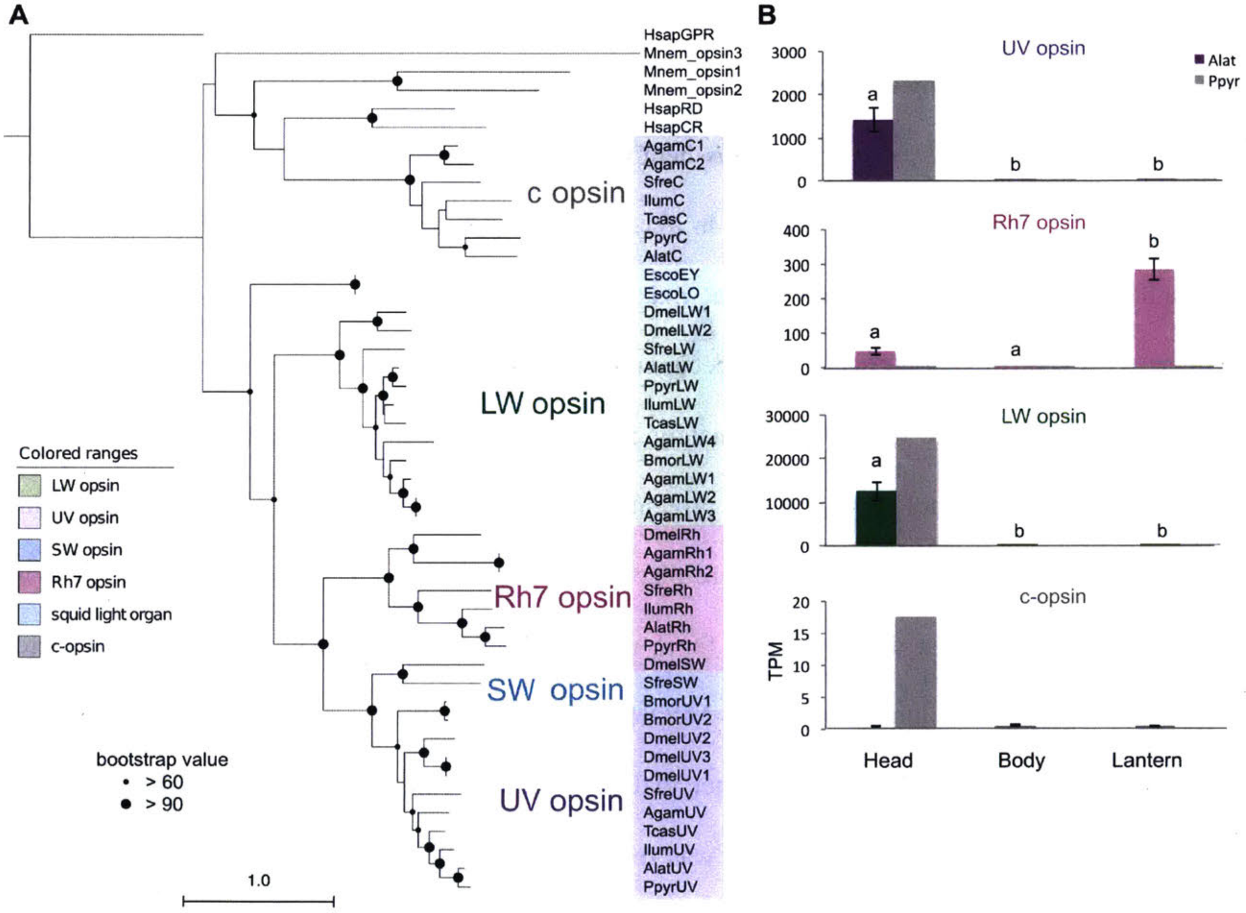
Supporting Information 4—figure 8. Maximum likelihood gene tree of the combined adenylyl-sulfate kinase and sulfate adenylyltransferase (ASKSA) orthogroup.

Peptide sequences from *P. pyralis*, *A. lateralis*, *I. luminosus*, *T. castaneum*, and *D. melanogaster* were clustered (orthogroup # 698), multiple sequence aligned, and refactored into a species rooted maximum likelihood tree, via the OrthoFinder pipeline (Supporting Information 4.2.1). As this is a genome-wide analysis where bootstrap replicates would be computationally prohibitive, no bootstrap replicates were performed to evaluate the support of the tree topology. PTS1 sequences were predicted from the peptide sequence using the PTS1 predictor server (Georg Neuberger, Sebastian Maurer-Stroh, Birgit Eisenhaber, Andreas Hartig and Frank Eisenhaber, n.d.). Figure produced with iTOL (Letunic and Bork, 2016).

4.5 Opsin analysis

Opsins are G-protein-coupled receptors that, together with a bound chromophore, form visual pigments that detect light, reviewed here (Briscoe and Chittka, 2001). While opsin genes are known for their expression in photoreceptors and function in vision, they have also been found to be expressed in other tissues, suggesting non-visual functions in some cases. Insects generally use rhabdomeric opsins (r-opsins) for vision, while mammals generally use ciliary opsins (c-opsins) for vision, products of an ancient gene duplication (Briscoe and Chittka, 2001; Porter Megan L. et al., 2012). Both insects and mammals may retain the alternate opsin type, generally in a non-visual capacity. The ancestral insect is hypothesized to have three visual opsins - one sensitive to long-wavelengths of light (LW), one to blue-wavelengths (B), and one to ultraviolet light (UV). Previously, two opsins, one with sequence similarity to other insect LW opsins and one with similarity to other insect UV opsins, were identified as highly expressed in firefly heads (Martin et al., 2015; Sander and Hall, 2015). A likely non-visual c-opsin was also detected, although not highly expressed (Martin et al., 2015; Sander and Hall, 2015).

To confirm the previously documented opsin presence and expression patterns, we collected candidate opsin genes via BLASTP searches (e-value threshold: 1 × 10−20) of the PPYR_OGS1.0, AQULA_OGS1.0 and ILUMI_OGS1.0 reference genesets against UV opsin of *P. pyralis* (Genbank Accession: ALB48839.1), as well as collected non-firefly opsin sequences via literature searches, followed by maximum likelihood phylogenetic reconstruction (Supporting Information 4—figure 9A), and expression analyses of the opsins (Supporting Information 4—figure 9B). The amino acid sequences of opsin were multiple aligned using MAFFT and trimmed using trimAL (parameters: -gt 0.5). The amino acid substitution model for ML analysis was estimated using Aminosan (v1.0.2016.11.07) (Tanabe, 2011). In *P. pyralis*, *A. lateralis*, and *I. luminosus*, we detected three r-opsins, including LW, UV, and an r-opsin homologous to Drosophila *Rh7* opsin, and one c-opsin. While LW and UV opsins were highly and differentially expressed in heads of both fireflies, c-opsin was lowly expressed, in *P. pyralis* head tissue only (Supporting Information—figure 9B). In contrast, *Rh7* was not expressed in the *P. pyralis* light organ, but was differentially expressed in the light organ of *A. lateralis* (Supporting Information—figure 9B). The detection of *Rh7* in our genomes is unusual in beetles (Feuda et al., 2016), although emerging genomic resources across the order have detected it in two taxa: *Anoplophora glabripennis* (McKenna et al., 2016) and *Leptinotarsa decemlineata* (Schoville et al., 2017). *Rh7* has an enigmatic function - a recent study in *Drosophila melanogaster* showed that *Rh7* is expressed in the brain, functions in circadian photoentrainment, and has broad UV-to-visible spectrum sensitivity (Ni et al., 2017; Sakai et al., 2017). Extraocular opsin expression has been detected in other eukaryotes: a photosensory organ is located in the genitalia at the posterior abdominal segments in butterfly (Lepidoptera) (Arikawa and Aoki, 1982). In the bioluminescent Ctenophore *Mnemiopsis leidyi,* three c-opsins are co-expressed with the luminous photoprotein in the photophores (Schnitzler et al., 2012). In the bobtail squid, *Euprymna scolopes,* one of the c-opsin isoforms is expressed in the bacterial symbiotic light organ (Pankey et al., 2014; Tong et al., 2009). Thus, it is possible that *Rh7* has a photo sensory function in the lantern of *A. lateralis*, although this putative function is seemingly not conserved in *P. pyralis*. Future study will confirm and further explore the biological, physiological, and evolutionary significance of *Rh7* expression in the light organ across firefly taxa.



Supporting Information 4—figure 9. ML tree and gene expression levels of opsin genes.

4.6 LC-HRAM-MS of lucibufagin content in *P. pyralis, A. lateralis, and I. luminosus*

We assayed the hemolymph of adult *P. pyralis* and *A. lateralis*, as well as body extracts from *P. pyralis* and *A. lateralis* larvae, and *I. luminosus* adult male thorax, for lucibufagin content using liquid-chromatography high-resolution accurate-mass mass-spectrometry (LC-HRAM-MS) and MS2 spectral similarity networking approaches. We chose to analyze extracted hemolymph from both *P. pyralis*, and *A. lateralis* for lucibufagin content, as lucibufagins are known to accumulate in the adult hemolymph and hemolymph samples give less complex extracts than tissue extracts. For *P. pyralis* and *A. lateralis* larvae, and *I. luminosus* thorax, tissue extracts were sampled as we do not have a reliable hemolymph extraction protocol for these life stages and species. Specific tissues were chosen for extracts to enable a smaller quantity of tissue to go into the metabolite extraction, and to explore possible difference in compound abundance across tissues, but we expected that defense compounds like lucibufagins would be roughly equally abundant present in all tissues.

Adult male *P. pyralis* and *A. lateralis* hemolymph was extracted by the following methods: A single live adult *P. pyralis* male was placed in a 1.5 mL microcentrifuge tube with a 5-mm-glass bead underneath the specimen, and centrifuged at maximum speed (~20,000 xg) for 30 s in a benchtop centrifuge. This centrifugation crushed the specimen on top of the bead, and allowed the hemolymph to collect at the bottom of the tube. Approximately 5 μL was obtained. The extracted hemolymph was diluted with 50 μL methanol to precipitate proteins and other macromolecules. For *A. lateralis* adult hemolymph, three adult male individuals were placed in individual 1.5 mL microcentrifuge tubes with 5-mm-glass beads, and spun at 5000 RPM for 1 min in a benchtop centrifuge. The pooled extracted hemolymph (~5 μL), was diluted with 50 μL MeOH, and air dried. The *P. pyralis* extracted hemolymph was filtered through a 0.2 μm PTFE filter (Filter Vial, P/No. 15530–100, Thomson Instrument Company), whereas the *A. lateralis* hemolymph residue was redissolved in 100 μL 50% MeOH, and then filtered through the filter vial.

For extraction of *P. pyralis* larval partial body, the posterior two abdominal segments were first cut off from a single laboratory reared larvae (Supporting Information 1.3.2), and the remaining partial body was placed in 180 μL 50% acetonitrile, and macerated with a pipette tip. The extract was sonicated in a water bath sonicator for ~10 min, not letting the temperature of the bath go above 50°C. The extract was then centrifuged (20,000 x g for 10 min), and filtered through a 0.2 μm PTFE filter (Filter Vial, P/No. 15530–100, Thomson Instrument Company).

For extraction of *A. lateralis* larval whole body, laboratory reared *A. lateralis larvae* were flash frozen in liquid N2, lyophilized, and the whole body (dry weight: 29.1 mg) was placed in 200 μL 50% methanol, and macerated with a pipette tip. The extract was sonicated in a water bath sonicator for 30 min, centrifuged (20,000xg for 10 min), and filtered through a 0.2 μm PTFE filter (Filter Vial, P/No. 15530–100, Thomson Instrument Company).

For extraction of *I. luminosus* adult thorax, the mesothorax through the two most anterior abdominal segments (ventral lantern containing segment +1 segment) of a lyophilized *I. luminosus* adult male (Supporting Information 3.3), was separated from the prothorax plus head and posterior three abdominal segments. This mesothorax + abdomen fragment was then placed in 0.5 mL 50% methanol, and macerated with a pipette tip. The extract was then sonicated in a water bath sonicator for ~10 min, not letting the temperature of the bath go above 50°C, centrifuged (20,000xg for 10 min), and filtered through a 0.2 μm PTFE filter (Filter Vial, P/No. 15530–100, Thomson Instrument Company).

Injections of these filtered extracts (*P. pyralis* adult male hemolymph 10 μL; *A. lateralis* adult male hemolymph 5 μL; *P. pyralis* partial larval body extract 5 μL; *A. lateralis* whole larval body 5 μL; *I. luminosus* thorax extract 20 μL) were separated and analyzed using an UltiMate 3000 liquid chromatography system (Thermo Scientific) equipped with a 150 mm C18 Column (Kinetex 2.6 μm silica core shell C18 100 Å pore, P/No. 00F-4462-Y0, Phenomenex, USA) coupled to a Q-Exactive mass spectrometer (Thermo Scientific, USA). Two different instrument methods were used, a slow ~44 min method, and an optimized ~28 min method. Chromatographically both methods are identical up to 20 min.

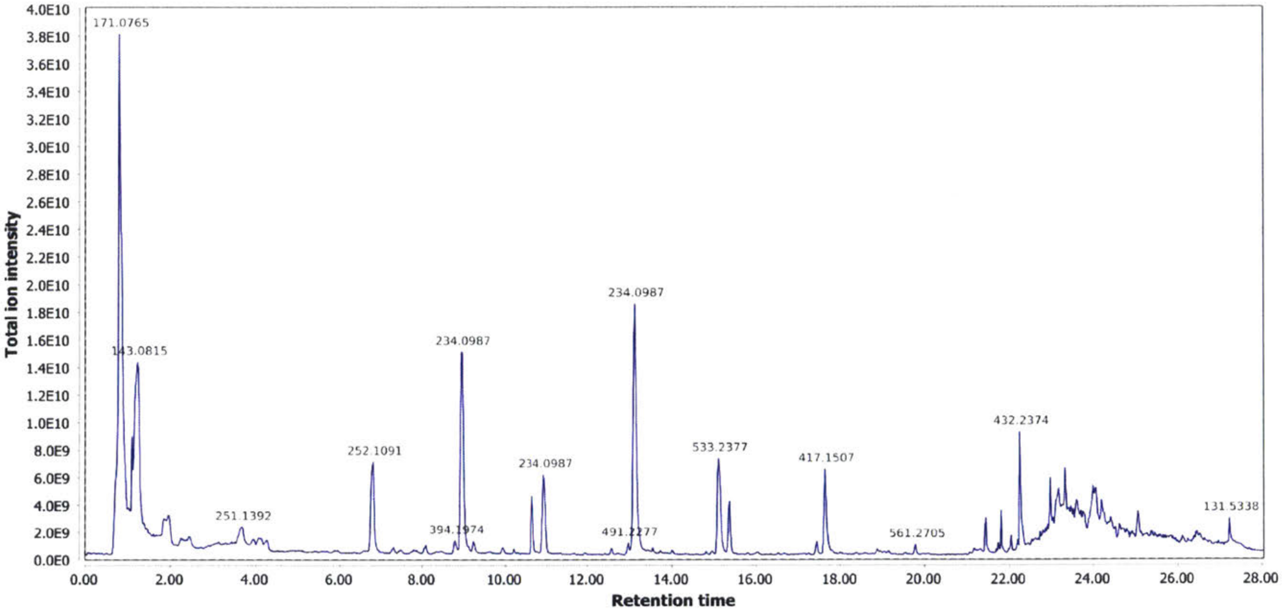
P. pyralis hemolymph compounds were separated by the optimized method (28 min), with separation via reversed-phase chromatography on a C18 column using a gradient of Solvent A (0.1%

formic acid in H2O) and Solvent B (0.1% formic acid in acetonitrile); 5% B for 2 min, 5–40% B until 20 min, 40–95% B until 22 min, 95% B for 4 min, and 5% B for 5 min; flow rate 0.8 mL/min. All other sample extracts were separated by the slow (44 min) reversed-phase chromatography method, using a C18 column with a gradient of Solvent A (0.1% formic acid in H2O) and Solvent B (0.1% formic acid in acetonitrile); 5% B for 2 min, 5–80% B until 40 min, 95% B for 4 min, and 5% B for 5 min; flow rate 0.8 mL/min.

The mass spectrometer was configured to perform one MS^1 scan from *m/z* 120–1250 followed by 1–3 data-dependent MS^2 scans using HCD fragmentation with a stepped collision energy of 10, 15, 25 normalized collision energy (NCE). Positive mode and negative mode MS^1 and MS^2 data were obtained in a single run via polarity switching for the optimized method, or in separate runs for the slow method. Data was collected as profile data. The instrument was always used within 7 days of the last mass accuracy calibration. The ion source parameters were as follows: spray voltage (+) at 3000 V, spray voltage (-) at 2000 V, capillary temperature at 275°C, sheath gas at 40 arb units, aux gas at 15 arb units, spare gas at one arb unit, max spray current at 100 (μA), probe heater temp at 350°C, ion source: HESI-II. The raw data in Thermo format was converted to mzML format using ProteoWizard MSConvert (Chambers et al., 2012). Data analysis was performed with Xcalibur (Thermo Scientific) and MZmine 2 (v2.30) (Pluskal et al., 2010)). Raw LC-MS data is available on MetaboLights (Accession: MTBLS698).

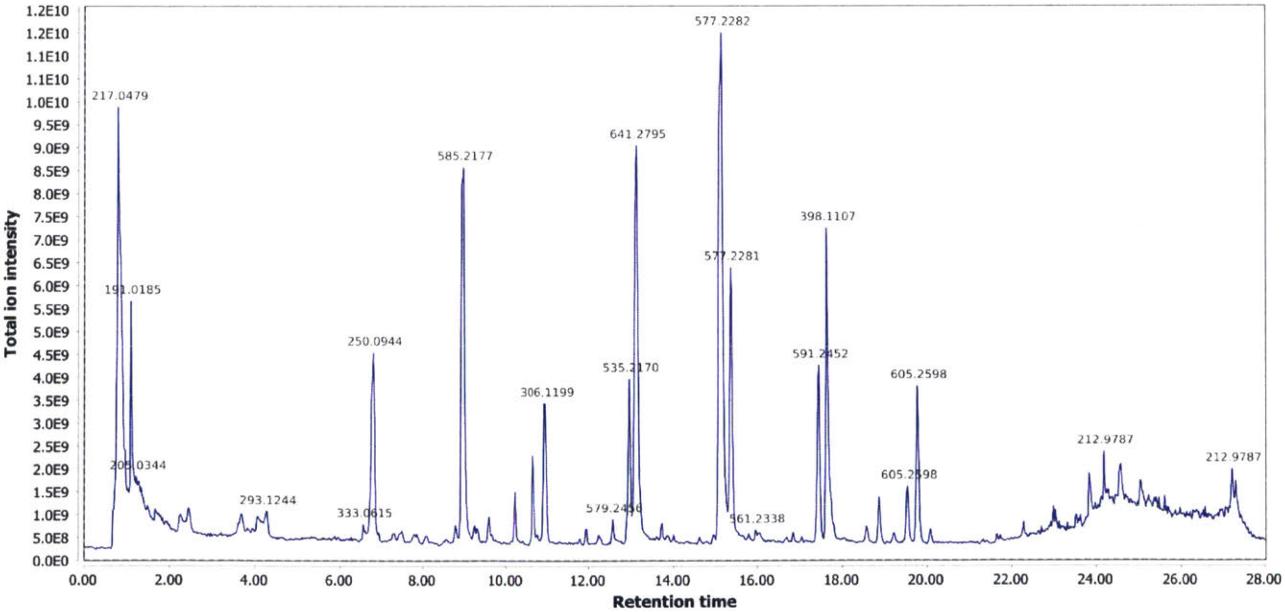
Within MZmine 2, data were from all five samples on positive mode, and were first cropped to 20 min in order to compare data which was obtained with the same LC gradient parameters. Profile MS^1 data was then converted to centroid mode with the Mass detection module(Parameters: Mass Detector = Exact mass, Noise level = 1.0E4), whereas MS^2 data was converted to centroid mode with (Noise level = 1.0E1). Ions were built into chromatograms using the Chromatogram Builder module with parameters (min_time_span = 0.10, min_height = 1.0E4, *m/z* tolerance = 0.001 *m/z* or five ppm. Chromatograms were then deconvolved using the Chromatogram deconvolution module with parameters (Algorithm =

Local Minimum Search, Chromatographic threshold = 5.0%, Search Minimum in RT range = 0.10 min, Minimum relative height = 1%, Minimum absolute height = 1.0E0, Min ratio of peak top/edge = 2, Peak duration range = 0.00–10.00). Isotopic peaks were annotated to their parent features with the Isotopic peaks grouper module with parameters (*m/z* tolerance = 0.001 or five ppm, Retention time tolerance = 0.2 min, Monotonic shape = yes, Maximum charge = 2, Representative isotope = Most intense). The five peaklists (*P. pyralis* hemolymph, *P. pyralis* larval partial body, *A. lateralis* adult hemolymph, *A. lateralis* larval whole body, *I. luminosus* thorax) were then joined and retention time aligned using the RANSAC algorithm with parameters (*m/z* tolerance = 0.001 or 10 ppm, RT tolerance = 1.0 min, RT tolerance after correction = 0.1 min, RANSAC iterations = 100, Minimum number of points = 5%, Threshold value = 0.5). These aligned peaklists were then gap-filled. Systematic mass accuracy error was determined with the endogenous tryptophan [M + H]+ ion (m/z = 205.09, RT = 3.5–4.5 mins), and was measured to be +0.6 ppm,+9.9 ppm,+1.6 ppm,+1.1 ppm, and +0.6 ppm, for *P. pyralis* adult hemolymph, *P. pyralis* partial larval body extract, *A. lateralis* adult hemolymph, *A. lateralis* larval body extract, and *I. luminosus* thorax extract, respectively.



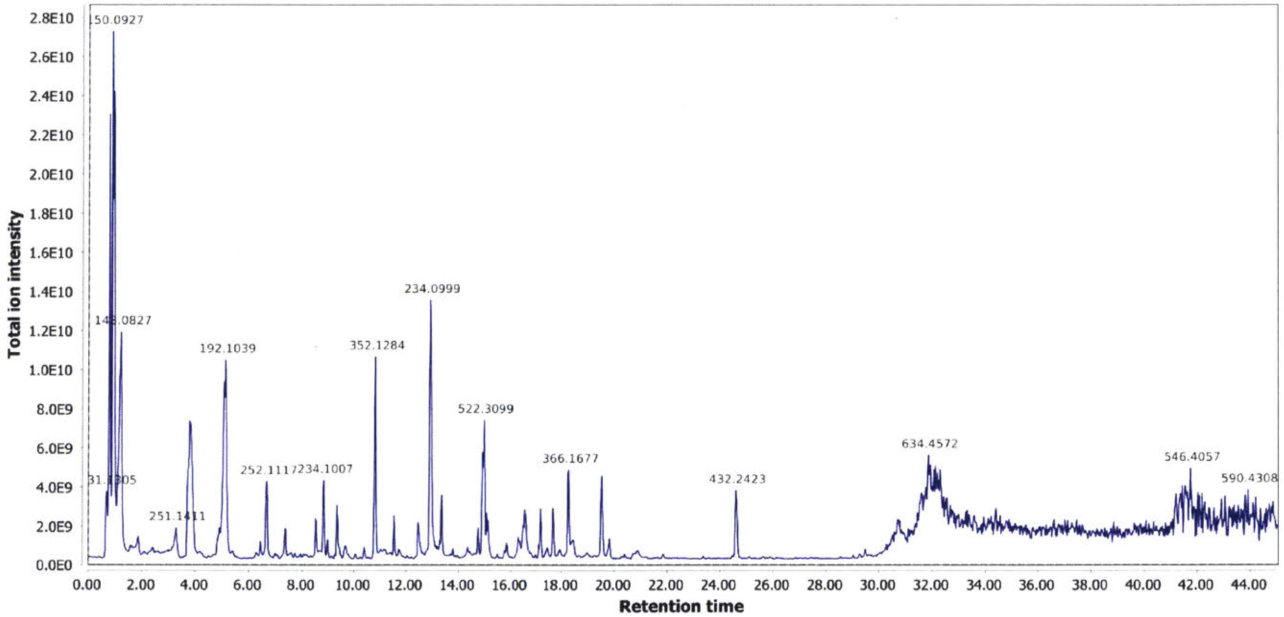
Supporting Information 4—figure 10. Positive mode MS^1 total-ion-chromatogram (TIC) of *P. pyralis* adult hemolymph LC-HRAM-MS data.

Figure produced using MZmine 2 (Pluskal et al., 2010).



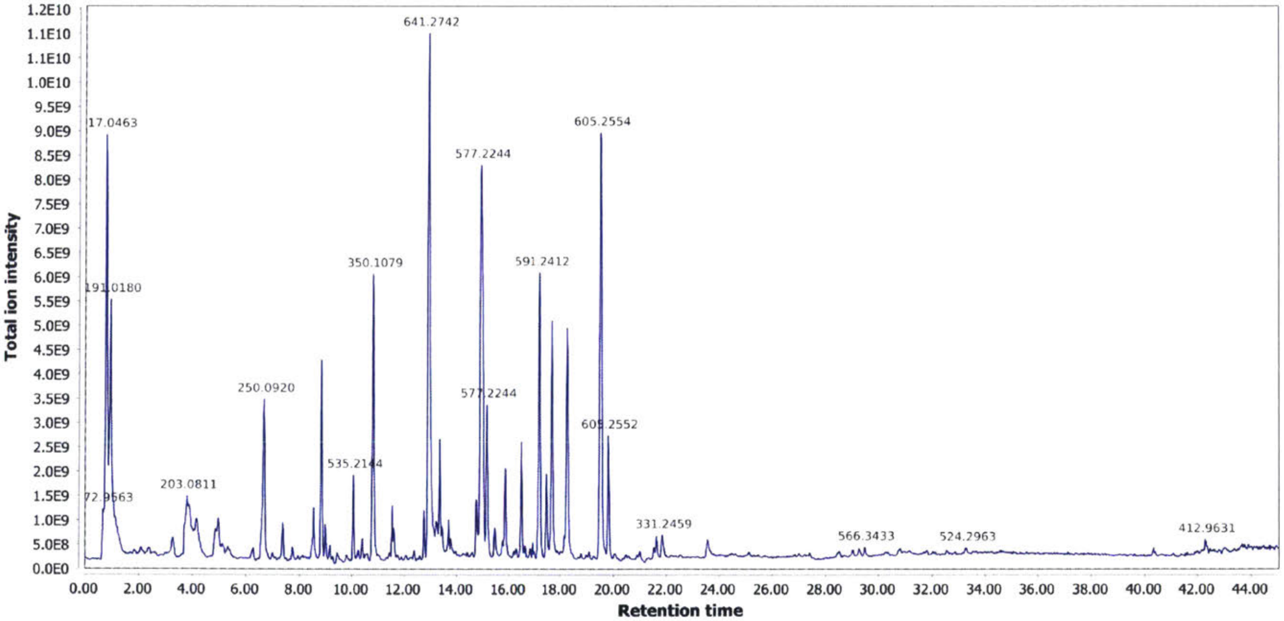
Supporting Information 4—figure 11. Negative mode MS^1 total-ion-chromatogram (TIC) of *P. pyralis* adult hemolymph LC-HRAM-MS data.

Figure produced using MZmine 2 (Pluskal et al., 2010).



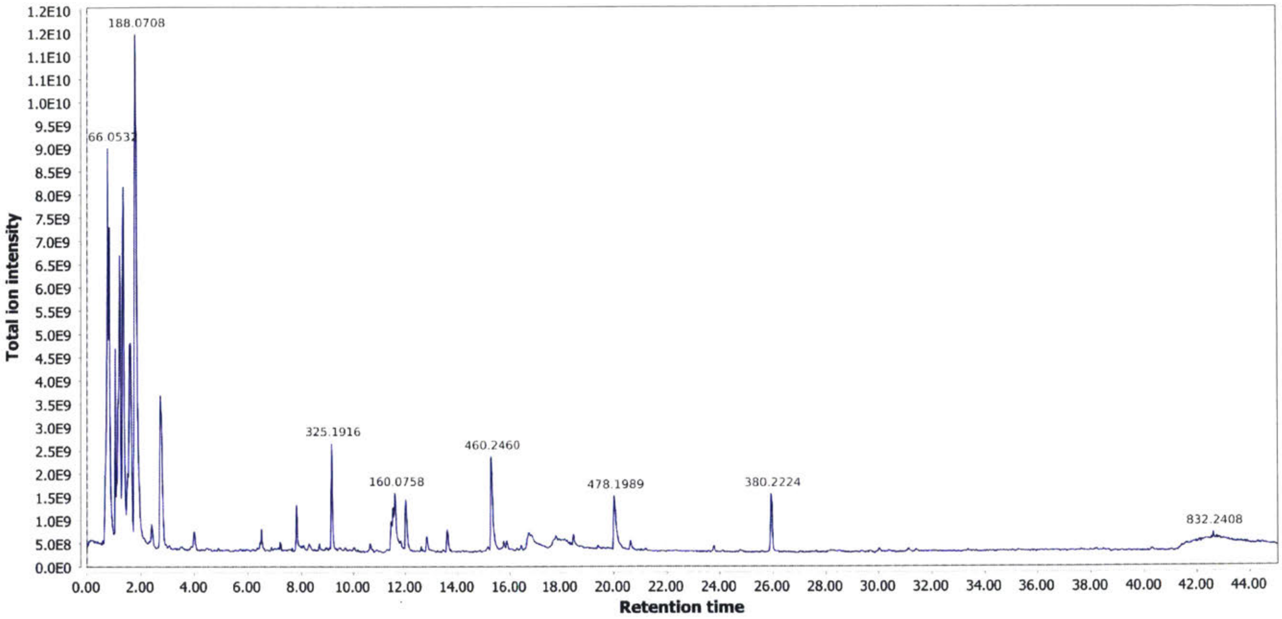
Supporting Information 4—figure 12. Positive mode MS^1 total-ion-chromatogram (TIC) of *P. pyralis* larval whole body minus two posterior segments LC-HRAM-MS data.

Figure produced using MZmine 2 (Pluskal et al., 2010).



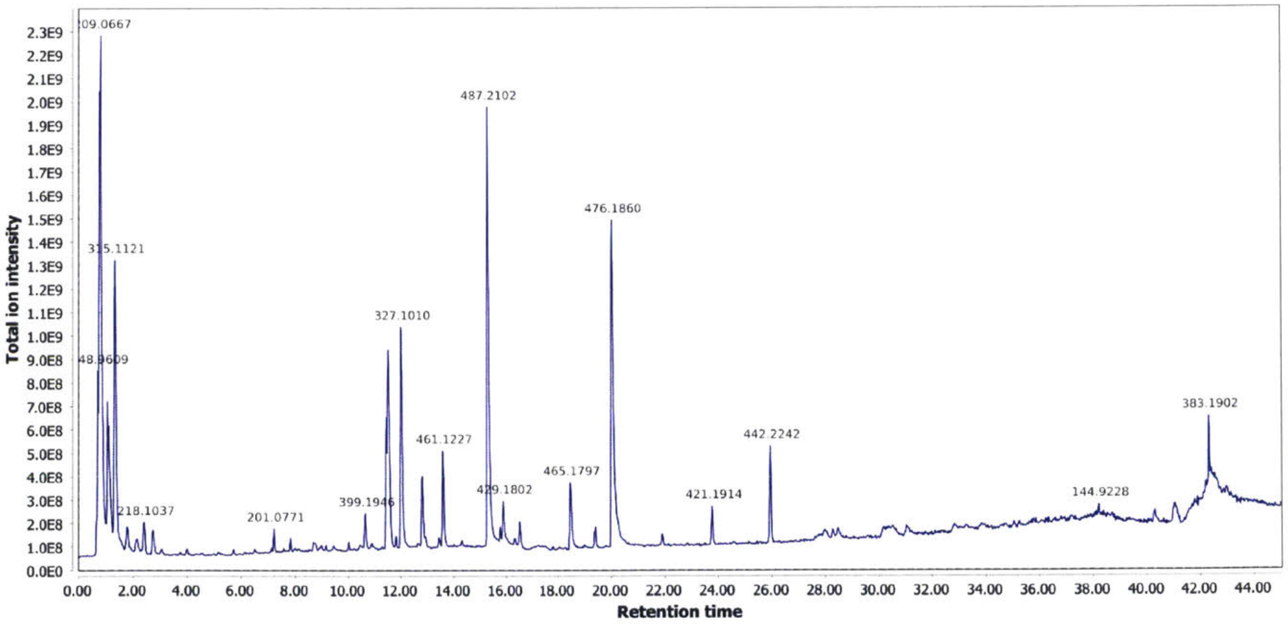
Supporting Information 4—figure 13. Negative mode MS^1 total-ion-chromatogram (TIC) of *P. pyralis* larval whole body minus two posterior segments LC-HRAM-MS data.

Figure produced using MZmine 2 (Pluskal et al., 2010).



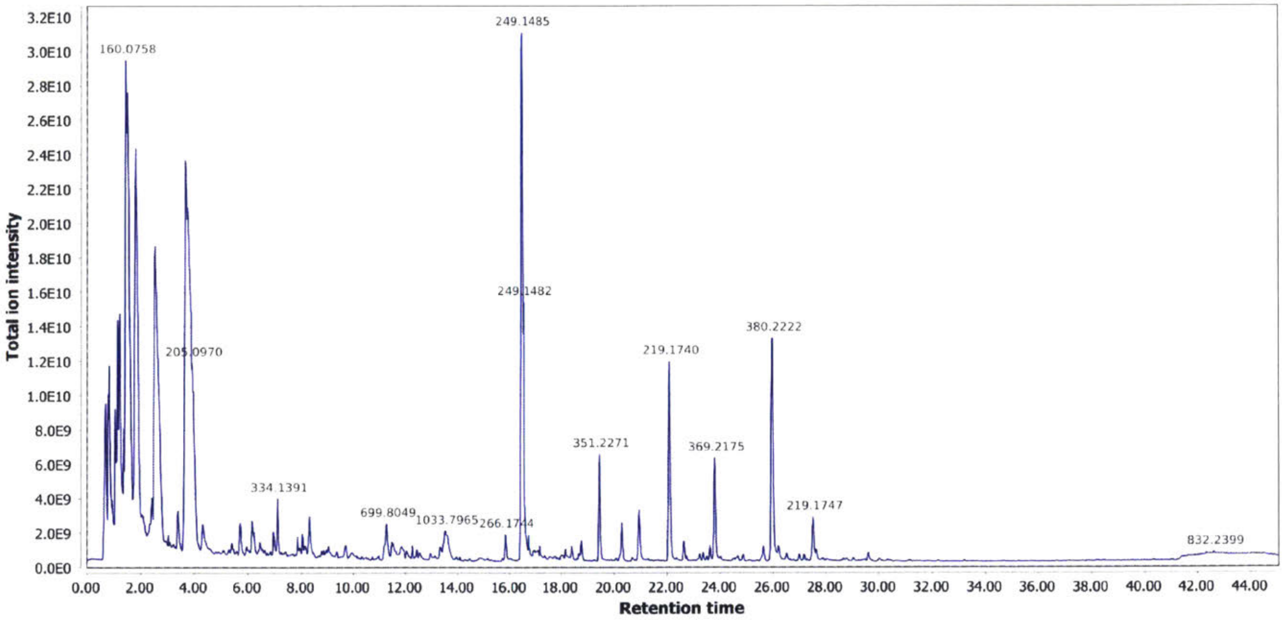
Supporting Information 4—figure 14. Positive mode MS^1 total-ion-chromatogram (TIC) of *A. lateralis* adult hemolymph LC-HRAM-MS data.

Figure produced using MZmine 2 (Pluskal et al., 2010).



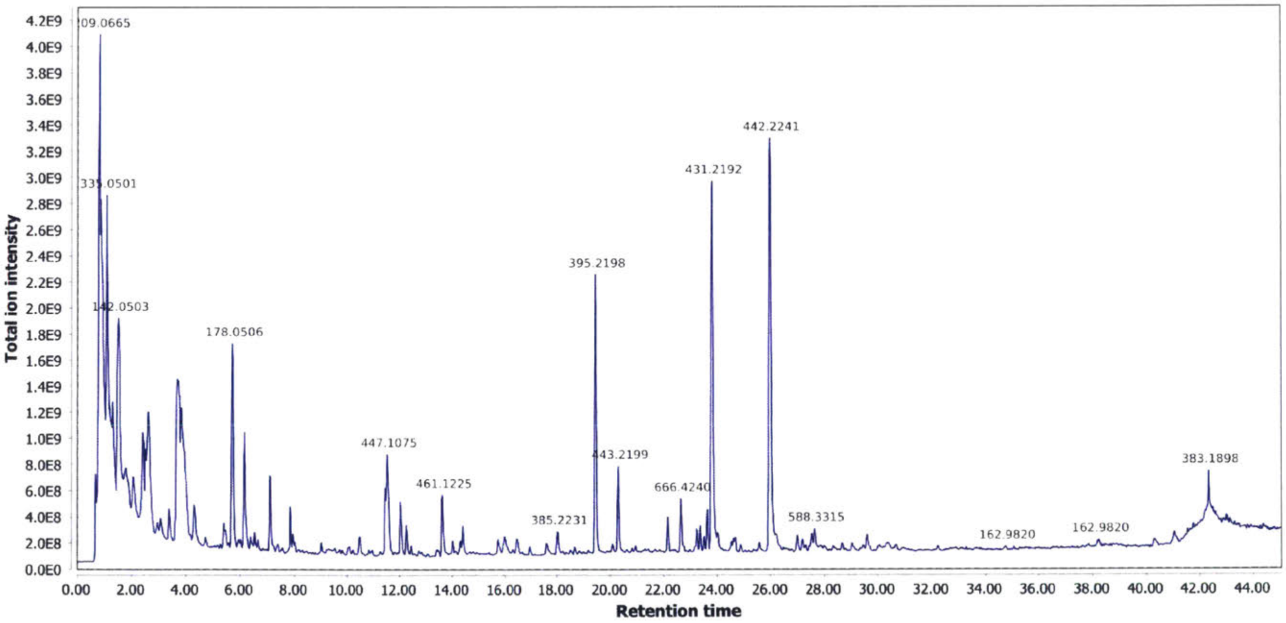
Supporting Information 4—figure 15. Negative mode MS^1 total-ion-chromatogram (TIC) of *A. lateralis* adult hemolymph LC-HRAM-MS data.

Figure produced using MZmine 2 (Pluskal et al., 2010).



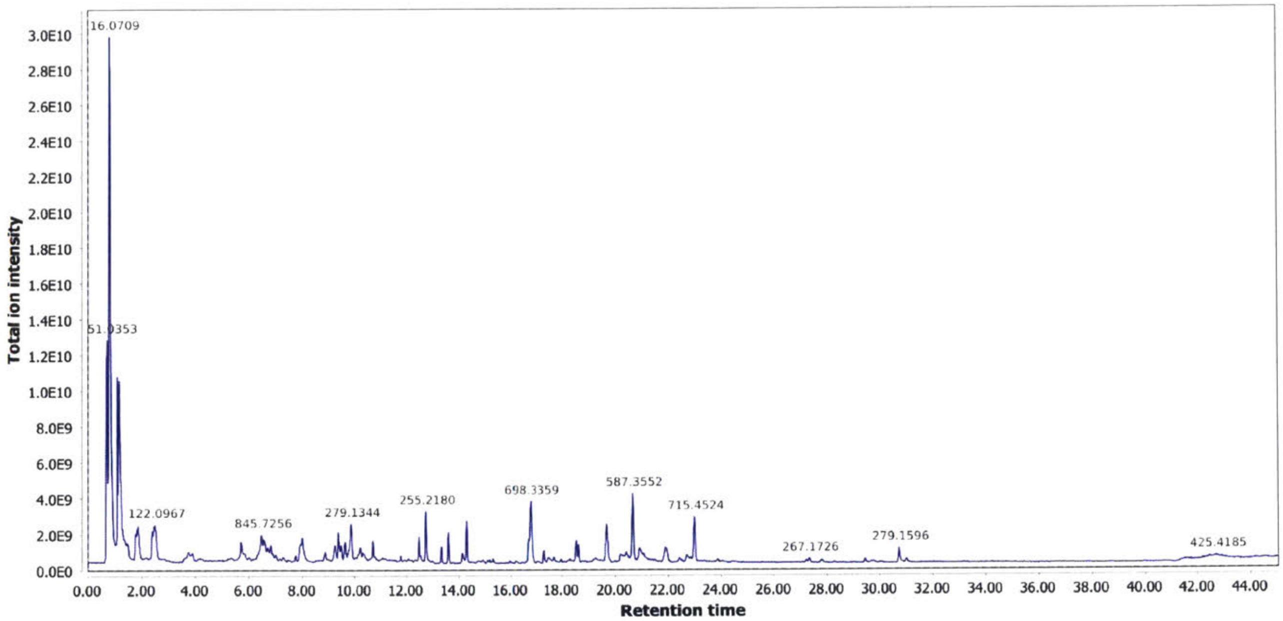
Supporting Information 4—figure 16. Positive mode MS^1 total-ion-chromatogram (TIC) of *A. lateralis* larval whole body LC-HRAM-MS data.

Figure produced using MZmine 2 (Pluskal et al., 2010).



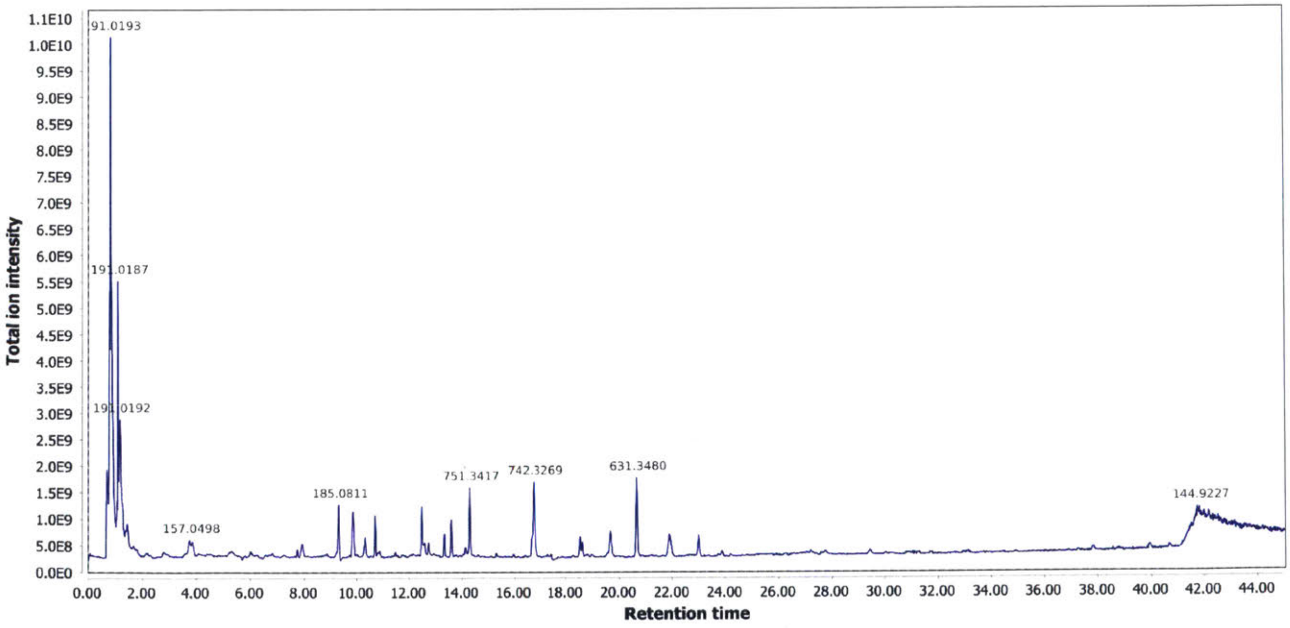
Supporting Information 4—figure 17. Negative mode MS^1 total-ion-chromatogram (TIC) of *A. lateralis* larval whole body extract LC-HRAM-MS data.

Figure produced using MZmine 2 (Pluskal et al., 2010).



Supporting Information 4—figure 18. Positive mode MS^1 total-ion-chromatogram (TIC) of *I. luminosus* mesothorax +abdomen extract LC-HRAM-MS data.

Figure produced using MZmine 2 (Pluskal et al., 2010).



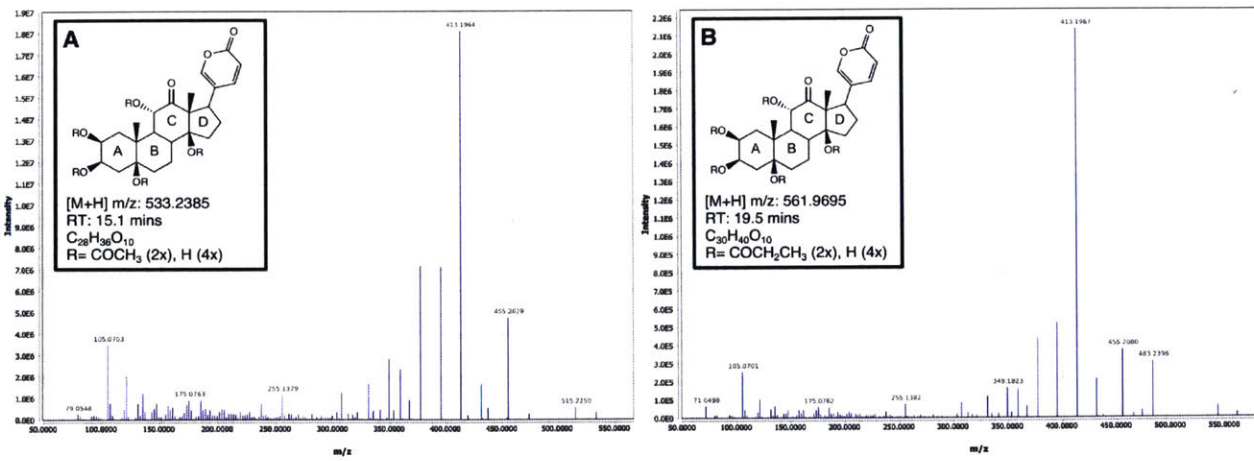
Supporting Information 4—figure 19. Negative mode MS^1 total-ion-chromatogram (TIC) of *I. luminosus* mesothorax + abdomen extract LC-HRAM-MS data.

Figure produced using MZmine 2 (Pluskal et al., 2010).

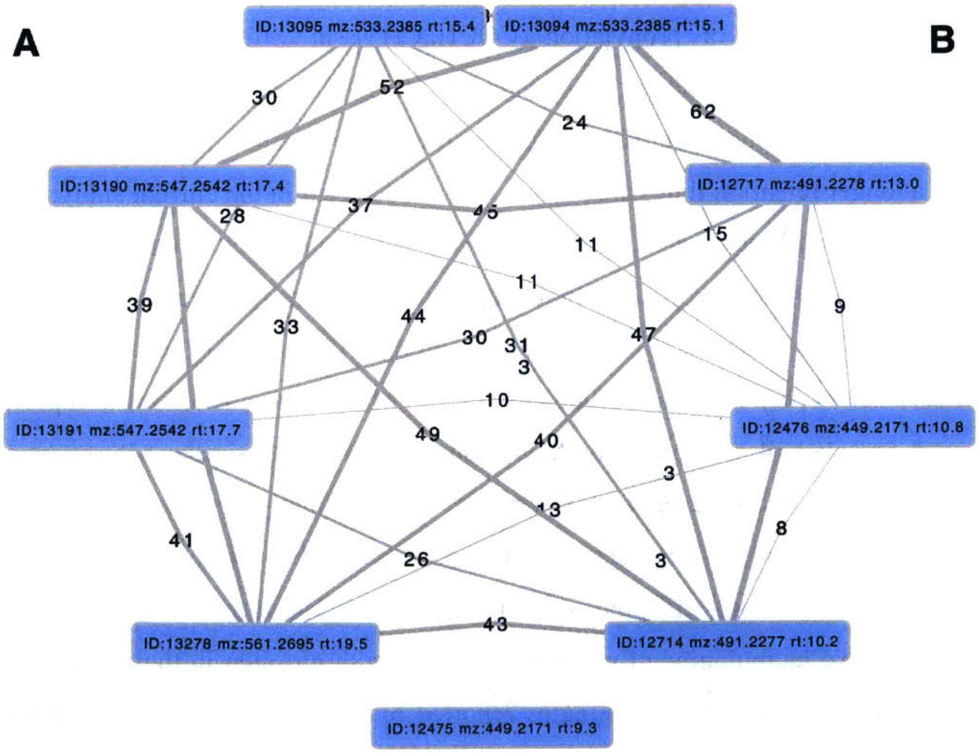
4.6.5 MS^2 similarity search for P. pyralis lucibufagins

We first performed a MS^2 similarity search within *P. pyralis* adult hemolymph for ions that showed a similar MS^2 spectra to the MS^2 spectra arising from the diacetylated lucibufagin $[M + H]^+$ ion from the same run ($[M + H]^+$ *m/z* 533.2385, RT = 15.10 mins) (Supporting Information 4—figure 20). This search was performed through the MS^2 similarity search module of MZmine 2 (v2.30) with parameters (*m/z* tolerance: 0.0004 *m/z* or 1 PPM; minimum # of ions to report: 3). This MS^2 similarity search revealed nine putative lucibufagin isomers with highly similar MS^2 spectra (Supporting Information 4—figure 21), which expanded to 17 putative lucibufagin isomers when considering features without MS^2 spectra, but with identical exact masses and close retention times (ΔRT <2 min) to the previously identified 9 (Supporting Information 4—table 7). Chemical formula prediction was assigned to each precursor ion using the Chemical formula search module of MZmine 2, whereas chemical formula predictions for product ions was performed within MZmine 2 using SIRIUS (v3.5.1) (Böcker et al., 2009). The structural identity of the nine putative lucibufagins detected via the MS^2 spectra similarity search was easily interpreted in light that the different chemical formula represented the core lucibufagins that had undergone acetylation (COCH3) or propylation (COCH2CH3), in different combinations. Notably, the most substituted isomers, dipropylated lucibufagin ($[M + H]^+$ *m/z* 561.2695, RT = 19.54 mins) were close to the edge of the cropped data (20 min), thus it may be possible that more highly substituted lucibufagins with a longer retention times are present, but not detected in the current analysis.

We then performed a MS^2 similarity search within *P. pyralis* partial body extract for ions that showed a MS^2 spectra similar to that of the dipropylated lucibufagin $[M + H]^+$ ion from the same run ([M + H]+ *m/z* 561.2738, RT = 19.53). This search was performed through the MS^2 similarity search module of MZmine 2 (v2.30) with parameters (*m/z* tolerance: 0.0004 *m/z* or 1 PPM; minimum # of ions to report: 5). This MS^2 similarity search revealed 14 putative lucibufagin isomers with highly similar MS^2 spectra (Supporting Information 4—table 7). Complexes and fragments were manually removed from the analysis. Comparison of the theoretical and observed exact mass indicated that this experimental run had an unusual degree of systematic *m/z* error, of ~+10 ppm. After manual correction m/z, chemical formula prediction revealed a several putative lucibufagins of unknown structure with nitrogen in their chemical formula, suggesting that the nitrogen containing lucibufagins reported by by Gronquist and colleagues from *Lucidota atra* (Gronquist et al., 2005) may be present in *P. pyralis* larvae.



Supporting Information 4—figure 20.
Positive mode MS^2 spectra of (A) diacetylated lucibufagin $[M + H]^+$ and (B) dipropylated lucibufagin $[M + H]^+$.


A
ID:13095 mz:533.2385 rt:15.4
ID:13094 mz:533.2385 rt:15.1
ID:13190 mz:547.2542 rt:17.4
ID:12717 mz:491.2278 rt:13.0
ID:13191 mz:547.2542 rt:17.7
ID:12476 mz:449.2171 rt:10.8
ID:13278 mz:561.2695 rt:19.5
ID:12714 mz:491.2277 rt:10.2
ID:12475 mz:449.2171 rt:9.3

B

Ion (m/z)	Chemical formula
67.0548	C_5H_6
93.0702	C_7H_9
95.0858	C_7H_{11}
121.0649	C_8H_9O
145.0649	$C_{10}H_9O$
147.0803	$C_{10}H_{11}O$
171.0805	$C_{12}H_{11}O$
173.0595	$C_{11}H_9O_2$
175.0753	$C_{11}H_{11}O_2$
225.1263	$C_{16}H_{17}O$
269.1167	$C_{17}H_{17}O_3$
341.1533	$C_{24}H_{21}O_2$
359.1641	$C_{24}H_{23}O_3$
367.1905	$C_{23}H_{27}O_4$
413.1956	$C_{24}H_{29}O_6$

Supporting Information 4—figure 21. MS^2 spectral similarity network for P. pyralis adult hemolymph lucibufagins.
(**A**) MS2 similarity network produced with the MZmine2 MS^2 similarity search module. Nodes represent MS2 spectra from the initial dataset, whereas edges represent an MS^2 similarity match between two MS^2 spectra. Thickness/label of the edge represents the number of ions matched between the two MS^2 spectra. (**B**) Table of matched ions between diacetylated lucibufagin (*m/z*: 533.2385 RT:15.1), and core (unacetylated) lucibufagin (*m/z*: 449.2171 RT:10.8 min). MS^1 adducts and complexes of the presented ions were manually removed.

Supporting Information 4—table 7. Putative lucibufagin compounds from LC-HRAM-MS of P. pyralis adult hemolymph.

Retention time and *m/z* values are not calibrated to the other samples.

Assigned ion identity	Ion type	Chemical formula	Expected *m/z*	Measured *m/z*	*m/z* error* (ppm)	Retention time (mins)	Feature area (arb)
Core lucibufagin isomer 1	$[M + H]^+$	$C_{24}H_{33}O_8$	449.2175	449.2171	−0.89	7.9	6.7E + 05
Core lucibufagin isomer 2	""	“”	“”	""	“”	9.3	1.1E + 07
Monoacetylated lucibufagin isomer 1	""	$C_{26}H_{35}O_9$	491.2281	491.2277	−0.81	10.2	4.2E + 07
Core lucibufagin isomer 3	""	$C_{24}H_{33}O_8$	449.2175	449.2171	−0.89	10.8	1.7E + 07
Monoacetylated lucibufagin isomer 2	""	$C_{26}H_{35}O_9$	491.2281	491.2277	−0.81	11.4	1.1E + 06
Monoacetylated lucibufagin isomer 3	""	“”	“”	""	“”	11.9	1.8E + 07
Monoacetylated lucibufagin isomer 4	""	“”	“”	""	“”	13.0	2.7E + 08
Monoacetylated lucibufagin isomer 5	""	“”	“”	""	“”	13.2	6.0E + 07
Monoacetylated lucibufagin isomer 6	""	“”	“”	""	“”	14.5	6.2E + 06
Diacetylated lucibufagin isomer 1	""	$C_{28}H_{37}O_{10}$	533.2387	533.2385	−0.37	15.1	4.0E + 09
Diacetylated lucibufagin isomer 2	""	“”	“”	""	“”	15.4	1.9E + 09
Monoacetylated, mono propylated lucibufagin isomer 1	""	$C_{29}H_{39}O_{10}$	547.2543	547.2542	−0.18	17.0	1.5E + 07
Monoacetylated, mono propylated lucibufagin isomer 2	""	“”	“”	""	“”	17.4	2.8E + 08
Monoacetylated, mono propylated lucibufagin isomer 3	""	“”	“”	""	“”	17.7	1.2E + 08
Dipropylated lucibufagin isomer 1	""	$C_{30}H_{41}O_{10}$	561.2700	561.2695	−0.89	18.9	1.4E + 08
Dipropylated lucibufagin isomer 2	""	“”	“”	""	“”	19.5	3.9E + 07
Dipropylated lucibufagin isomer 3	""	“”	“”	""	“”	19.8	1.8E + 08

Supporting Information 4—table 8. Putative lucibufagin compounds from LC-HRAM-MS of P. pyralis larval partial body extracts.

Retention time and *m/z* values are not calibrated to the other samples. *=*m/z* error and expected *m/z* extrapolated from ions with similar *m/z*, and chemical formula predicted from resulting extrapolated *m/z*. **=Likely chemical formula cannot be determined due to many possible chemical formula from the expected *m/z*.

Assigned ion identity	Ion type	Chemical formula	Expected *m/z*	Measured *m/z*	*m/z*error (ppm)	Retention time (mins)	Feature area (arb)
Core lucibufagin isomer 2	$[M + H]^+$	$C_{24}H_{33}O_8$	449.2175	449.2215	+8.9	9.15	8.5E + 06
Monoacetylated lucibufagin isomer 1	“”	$C_{26}H_{35}O_9$	491.2277	491.2326	+9.9	10.04	1.2E + 07
Unknown	unknown	$C_{28}H_{39}O_{10}$*	535.2543*	535.2592	+9.1*	12.40	1.6E + 07
Unknown	unknown	$C_{24}H_{38}NO_6$*	436.2695*	436.2735	+9.1*	13.30	2.2E + 07
Unknown	unknown	$C_{27}H_{45}N_2O_8$*	525.3173*	525.3221	+9.1*	13.35	1.3E + 08
Unknown	unknown	$C_{24}H_{40}NO_7$*	454.2799*	454.2840	+9.1*	13.73	1.3E + 07
Diacetylated lucibufagin isomer 1	[M + H]+	$C_{28}H_{37}O_{10}$	533.2387	533.2426	+7.3	14.93	1.7E + 09
Diacetylated lucibufagin isomer 2	[M + H]+	“”	“”	533.2426	+7.3	15.16	3.5E + 08
Unknown	Unknown	$C_{29}H_{46}NO_8$*	536.3216*	536.3256	+7.3*	16.57	4.1E + 07
Unknown	Unknown	Unknown**	563.2854*	563.2896	+7.3*	16.80	1.3E + 07
Unknown	Unknown	$C_{26}H_{31}O_7$	455.2056	455.2097	+9.1*	17.22	5.8E + 07
Dipropylated lucibufagin isomer 3	Unknown	$C_{30}H_{41}O_{10}$	561.2700	561.2738	+6.7	19.53	2.0E + 09
Dipropylated lucibufagin isomer 4	Unknown	$C_{30}H_{41}O_{10}$	561.2700	561.2738	+6.7	19.82	2.2E + 08

Supporting Information 4—table 9. Putative lucibufagin [M + H]$^+$ exact masses adjusted for instrument run specific systematic *m/z* error (Figure 6B).
Used for multi-ion-chromatogram (MIC) traces in Figure 6B.

Chemical formula	Predicted exact mass	Exact mass adjusted to *P. pyralis* hemolymph data (+0.6 ppm)	Exact mass adjusted to *P. pyralis* partial larval body data (+9.9 ppm)	Exact mass adjusted to *A. lateralis* hemolymph data (+1.6 ppm)	Exact mass adjusted to *A. lateralis* larval body data (+1.1 ppm)	Exact mass adjusted to *I. luminosus* thorax data (+0.6 ppm)
$C_{24}H_{33}O_8$	449.2175	449.2178	449.2219	449.2182	449.2180	449.2178
$C_{24}H_{38}NO_6$*	436.2699	436.2702	436.2742	436.2706	436.2704	436.2702
$C_{24}H_{40}NO_7$*	454.2804	454.2807	454.2849	454.2811	454.2809	454.2807
$C_{26}H_{31}O_7$	455.2069	455.2072	455.2114	455.2076	455.2074	455.2072
$C_{26}H_{35}O_9$	491.2281	491.2284	491.2330	491.2289	491.2286	491.2284
$C_{27}H_{45}N_2O_8$*	525.3175	525.3178	525.3227	525.3183	525.3181	525.3178
$C_{28}H_{37}O_{10}$	533.2386	533.2389	533.2439	533.2395	533.2392	533.2389
$C_{28}H_{39}O_{10}$*	535.2543	535.2546	535.2596	535.2552	535.2549	535.2546
$C_{29}H_{39}O_{10}$	547.2543	547.2546	547.2597	547.2552	547.2549	547.2546
$C_{29}H_{46}NO_8$*	536.3223	536.3226	536.3276	536.3232	536.3229	536.3226
$C_{30}H_{41}O_{10}$	561.2699	561.2702	561.2755	561.2708	561.2705	561.2702

*=Chemical formula assigned for structurally unclear putative lucibufagins

4.6.7 MS2 similarity search for A. lateralis lucibufagins

Although our earlier LC-HRAM-MS analysis (Figure 6B; Supporting Information 4.6) indicated *A. lateralis* adult male hemolymph does not contain detectable quantities of the *P. pyralis* lucibufagins, this does not exclude that structurally unknown lucibufagins with chemical formula not present in *P. pyralis*, are present in *A. lateralis*. To address this, we performed a MS^2 similarity search against the *A. lateralis* adult male hemolymph MS^2 spectra, with the MS^2 spectra of lucibufagin C (*m/z* 533.2385, RT = 15.1) as bait, using the MZmine 2 similarity search module with parameters (*m/z* tolerance = 0.001 or 10 ppm, Minimum # of matched ions = 10). After filtering to those precursors that were mostly likely to be the [M + H]$^+$ of a lucibufagin-like molecule (*m/z* 350–800, RT = 8–20 mins), 9 MS^2 spectra were matched

(Supporting Information 4—table 10). None of these features were detected in *P. pyralis* (Supporting Information 4—table 10). Chemical formula prediction was difficult due to the high *m/z* of the ions, but in those cases where it was successful, the additions of nitrogens and/or phosphorus to the chemical formula was confident. Notably, the most confident chemical formula predictions reported ≤23 carbons, and as the core lucibufagin of *P. pyralis* contains 24 carbons, it is unlikely that these ions derive from lucibufagins. The notable degree of MS^2 similarity may be due to the *A. lateralis compounds* also being steroid derived compounds. That being said, the identity and role of the compound giving rise to ion 460.2462 is intriguing, as it is highly abundant in the *A. lateralis* adult hemolymph, is absent from the *P. pyralis* adult hemolymph, and is possibly a steroidal compound.

Supporting Information 4—table 10
Relative quantification of *A. lateralis* features identified by lucibufagin MS^2 similarity search.

Assigned identity	***m/z***	**Chemical formula**	**RT (mins)**	**Similarity score**	**# of ions matched**	***A. lateralis* feature area (arb)**	***P. pyralis* feature area (arb)**
Unknown	460.2462	$C_{22}H_{38}NO_7P$*; $C_{25}H_{29}N_7O_2$*	15.27	4.10E + 11	34	7.04E + 08	0.00E + 00
""	657.2229	N.D.	12.01	9.50E + 11	29	6.13E + 07	""
""	414.2043	N.D.	18.07	1.20E + 11	25	5.61E + 06	""
""	381.2176	$C_{23}H_{28}N_2O_3$*	15.77	3.80E + 11	18	1.22E + 08	""
""	476.1839	N.D.	15.93	3.80E + 11	16	9.87E + 06	""
""	456.2148	N.D.	19	2.30E + 11	14	5.03E + 06	""
""	351.228	N.D.	19.42	2.60E + 11	13	1.56E + 07	""
""	479.1948	N.D.	19.83	2.20E + 11	12	1.11E + 07	""

*Determined with Sirius (MS^2 analysis), and MZmine 2 (isotope pattern analysis).
N.D., Not determined

Supporting Information 5

Microbiome analyses

5.1 Assembly and annotation of the complete *Entomoplasma luminosum* subsp. pyralis genome

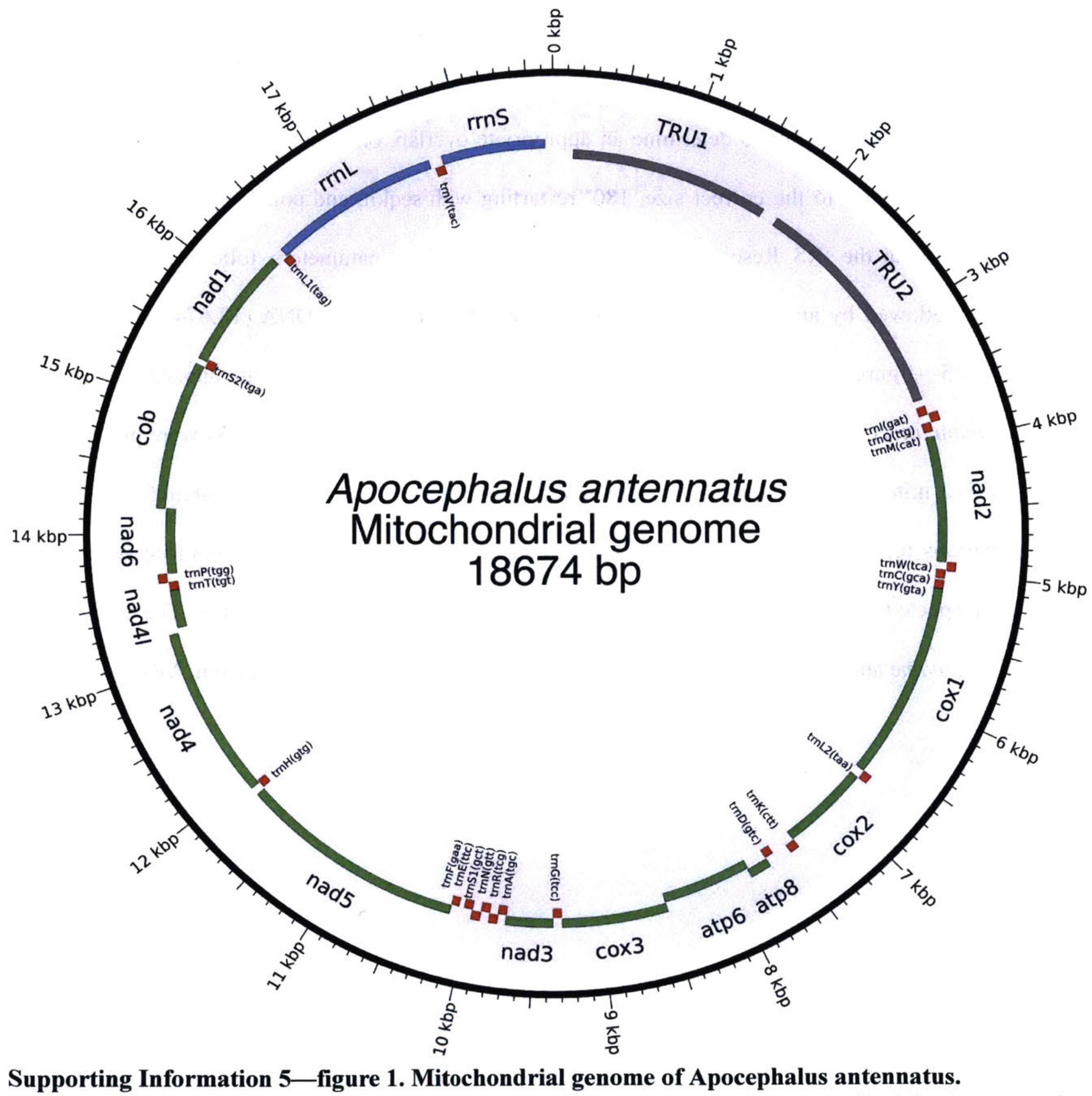
The complete genome of the molicute (Phylum: Tenericutes) *Entomoplasma luminosum* subsp. pyralis was constructed by a long-read metagenomic sequencing and assembly approach from the *P. pyralis* PacBio data. First, BUSCO v.3 with the bacterial BUSCO set was used to identify those contigs from the PacBio only Canu assembly (Ppyr0.1-PB) which contained conserved bacterial genes. A single 1.04 Mbp contig with 73 bacterial BUSCO genes was the only contig identified with more than 1 BUSCO hit. Inspection of the Canu produced assembly graph with Bandage v0.8.1 (Wick et al., 2015), revealed that the contig had a circular assembly path. BLASTN alignment of the contig to the NCBI nt database indicated that this contig had a high degree of similarity to annotated Mycoplasmal genomes. Together this data suggested that this contig represented a complete Mycoplasmal genome. Polishing of the contig was performed by mapping and PacBio consensus-calling using SMRTPortal v2.3.0.140893 with the 'RS_Resequencing.1' protocol with default parameters. The median coverage was ~50x. The resulting consensus sequence was restarted with seqkit (Shen et al., 2016) to place the FASTA record junction 180° across the circular chromosome, and reentered into the polishing process to enable efficient mapping across the circular junction. This mapping, consensus calling, and rotation process was repeated three times total, after which no additional nucleotide changes occurred. The genome was 'restarted' with seqkit such that the FASTA start position began between the ribosomal RNAs, and annotation was conducted through NCBI using their prokaryotic gene annotation pipeline (PGAP). Analysis with BUSCO v.3 of the peptides produced from the aforementioned genome annotation indicated that 89.8% of expected Tenericutes single-copy conserved orthologs were captured in the annotation (C:89.8%[S:89.8%,D:0.0%], F:2.4%, M:7.8%, n:166). Comparison of the predicted 16S rRNA gene sequence to the NCBI 16S rRNA gene database indicated that this gene had 99% identity to the *E. luminosum* 16S sequence (ATCC 49195 - formerly *Mycoplasma luminosum;* NCBI Assembly ID

ASM52685v1)(Kyrpides et al., 2014; Williamson et al., 1990), leading to our description of this genome as the genome of *Entomoplasma luminosum* subspecies (subsp.) pyralis. Protein overlap comparisons using the OrthoFinder pipeline (v1.1.10) (Emms and Kelly, 2015) between our predicted protein geneset for *E. luminosum* var. pyralis and the protein geneset of *Entomoplasma luminosum* (ATCC 49195 - formerly *M. luminosum;* NCBI Assembly ID ASM52685v1), indicated that 94% (670/709) of the previously annotated *E. luminosum* proteins are present in our genome of *E. luminosum* subsp. pyralis.

5.2 Assembly and annotation of Phorid mitochondrial genome

The complete mitochondrial genome of the dipteran parasitoid *Apocephalus antennatus,* first detected via BLASTN of mtDNAs as a concatemerized sequence in the Canu PacBio only assembly (Ppyr0.1-PB) was constructed in full by a long-read metagenomic sequencing and assembly approach. First, PacBio reads were mapped to the NCBI set of mitochondrial genomes concatenated with the *P. pyralis mitochondrial* genome assembly reported in this manuscript (NCBI accession KY778696.1), using GraphMap v0.5.2 with parameters 'align -C -t 4 -P'. Of the mitochondrially mapped reads (45949 reads), 98% (45267 reads) were partitioned to the *P. pyralis* mtDNA. The next most abundant category at 1.1% (531 reads), was partitioned to the mtDNA of the Phorid fly *Megaselia scalaris* (NCBI accession: KF974742.1). The next most abundant category at 0.11% (53 reads) was partitioned to the mitochondrion of the Red algae *Galdieria sulphuraria* (NCBI accession: NC_024666.1). The reads were then split into three partitions: *P. pyralis* mapping, *M. scalaris* mapping, and other, and input into Canu (v1.6+44) (Koren et al., 2017) for assembly. Each partitioned assembly by Canu produced a single circular contig, notably the 'other' and Megaselia partitions produced highly similar sequences, whereas the *P. pyralis* partition produced a circular sequence that was highly similar to the *P. pyralis* mtDNA. We inspected the *M. scalaris* partition further as it was produced with more reads. Notably, although an inspection of the contig was circular, and showed a high degree of similarity upon BLASTN to the *M. scalaris* mtDNA, the contig was ~2x larger than expected (29,821 bp). An analysis of contig's self-complementarity with

Gepard (v1.40) (Krumsiek et al., 2007), indicated that this contig had 2x tandem repetitive regions, and was duplicated overall twice. Similarly, the. GFA output of Canu noted an overlap of 29,821, indicating that the assembler was unable to determine an appropriate overlap, other than the entire contig. Manual trimming of the contig to the correct size, 180° restarting with seqkit, and polishing using SMRTPortal v2.3.0.140893 with the 'RS_Resequencing.1' protocol with default parameters, followed by 180° seqkit 'restarting', followed by another round of polishing, produced the final mtDNA (18,674 bp; Supporting Information 5—figure 1). This mtDNA was taxonomically identified in a separate analysis to originate from *A. antennatus* (Supporting Information 5.3). Coding regions, tRNAs, and rRNAs were predicted via the MITOSv2 mitochondrial genome annotation web server (http://mitos2.bioinf.uni-leipzig.de/). Small mis-annotations (e.g. low scoring additional predictions of already annotated mitochondrial genes) were manually inspected and removed. Tandem repetitive regions were manually annotated. The complete *A. antennatus genome* annotation plus assembly is available on NCBI Genbank (Accession: MG546669).



Supporting Information 5—figure 1. Mitochondrial genome of Apocephalus antennatus.
The mitochondrial genome of *A. antennatus* was assembled and annotated as described in the Supporting Information 5.2, and taxonomically identified as described in Supporting Information 5.3. Figure produced with Circos (Krzywinski et al., 2009).

5.3 Taxonomic identification of Phorid mitochondrial genome origin

After the successful metagenomic assembly of the mitochondrial genome of an unknown Phorid fly species from the *P. pyralis* PacBio library (Supporting Information 5.2), we sought to characterize the species of origin for this mitochondrial genome. We planned to achieve this by collecting the Phorid flies which emerged from adult *P. pyralis*, taxonomically identifying them, and performing targeted mitochondrial PCR and sequencing experiments to correlate their mitochondrial genome sequence to our mtDNA assembly. We successfully obtained phorid fly larvae emerging from *P. pyralis* adult males collected from MMNJ (identical field site to PacBio collection), and Rochester, NY (RCNY), in the summer of 2017. The MMNJ phorid larvae did not successfully pupate, however we obtained five adult specimens from successful pupations of the RCNY larvae. Two adults from this batch were identified as *A. antennatus* (Malloch), by Brian V. Brown, Entomology Curator of the Natural History Museum of Los Angeles County. DNA was extracted from one of the remaining three specimens and a COI fragment was PCR-amplified and Sanger sequenced. The forward primer was 5'-TTTGATTCTTCGGCCACCCA-3', the reverse primer 5'-AGCATCGGGGTAGTCTGAGT-3'. This COI fragment from had 99% identity (558/563 nt) to the COI gene of our mitochondrial assembly. This sequenced COI fragment has been submitted to GenBank (GenBank Accession: MG517481). We conclude that this is sufficient evidence to denote that our assembled Phorid mitochondrial genome is the mitochondrial genome of *A. antennatus*. Notably, *A. antennatus* was previously reported by (Lloyd, 1973) to be a parasite of several firefly species in genera *Photuris*, *Photinus*, and *Pyractomena*, from collection sites ranging from Florida to New York. To our knowledge, this is the first report of a mitochondrial genome which was first assembled via an untargeted metagenomic approach and then later correlated to its species of origin.

5.4 *Photinus pyralis* orthomyxo-like viruses

We identified the first two viruses associated to *P. pyralis* and the Lampyridae family. The proposed *Photinus pyralis* orthomyxo-like virus 1 and 2 (PpyrOMLV1 and 2) present a multipartite genome conformed by five RNA segments encoding a putative nucleoprotein (NP), hemagglutinin-like glycoprotein (HA) and a heterotrimeric viral RNA polymerase (PB1, PB2 and PA). The viral genomes for Photinus pyralis orthomyxo-like virus 1 and 2 are available on NCBI Genbank with accessions MG972985-MG972994. Expression analyses on 24 RNA libraries of diverse individuals/developmental stages/tissues and geographic origins of *P. pyralis* indicate a dynamic presence, widespread prevalence, a pervasive tissue tropism, a low isolate variability, and a persistent life cycle through transovarial transmission of PpyrOMLV1 and 2. Genomic and phylogenetic studies suggest that the detected viruses correspond to a new lineage within the *Orthomyxoviridae* family (ssRNA(-)) (Supporting Information 5—figure 2A-I). The concomitant occurrence in the *P. pyralis* genome of species-specific signatures of Endogenous viral-like elements (EVEs) associated to retrotransposons linked to the identified Orthomyxoviruses, suggest a past evolutionary history of host-virus interaction (Supporting Information 5.5, Supporting Information 5—figure 2J). This tentative interface is correlated to low viral RNA levels, persistence and no apparent phenotypes associated with infection. We suggest that the identified viruses are potential endophytes of high prevalence as a result of potential evolutionary modulation of viral levels associated to EVEs. Photinus pyralis orthomyxo-like virus 1 and 2 (PpyrOMLV1 and PpyrOMLV2) share their genomic architecture and evolutionary clustering (Supporting Information 5—figure 2A-H, Supporting Information 5—figure 3). They are multipartite linear ssRNA negative strand viruses, conformed by five genome segments generating a ca. 10.8 Kbp total RNA genome. Genome segments one through three (ca. 2.3–2.5 Kbp long) encode a heterotrimeric viral polymerase constituted by subunit Polymerase Basic protein 1 - PB1 (PpyrOMLV1: 801 aa, 91 kDA; PpyrOMLV2: 802 aa, 91.2 kDA), Polymerase Basic protein 2 - PB2 (PpyrOMLV1: 804 aa, 92.6 kDA; PpyrOMLV2: 801 aa, 92.4 kDA) and Polymerase Acid protein - PA (PpyrOMLV1: 754 aa, 86.6 kDA; PpyrOMLV2: 762 aa, 87.9 kDA). PpyrOMLV1 and PpyrOMLV2 PB1 present a Flu_PB1 functional domain (Pfam: pfam00602; PpyrOMLV1: interval = 49–741, e-value = 2.93e-69; PpyrOMLV2: interval = 49–763, e-value = 1.42e-62) which is the RNA-directed RNA polymerase catalytic subunit, responsible for replication and transcription of virus RNA segments, with two nucleotide-binding GTP domains. PpyrOMLV1 and PpyrOMLV2 PB2 present a typical Flu_PB2 functional domain (Pfam: pfam00604; PpyrOMLV1: interval = 26–421, e-value = 5.10e-13; PpyrOMLV2: interval = 1–692, e-value = 1.57e-11) which is involved in 5' end cap RNA structure recognition and binding to further initiate virus transcription. PpyrOMLV1 and PpyrOMLV2 PA subunits share a characteristic Flu_PA domain (Pfam: pfam00603; PpyrOMLV1: interval = 122–727, e-value = 3.73e-07; PpyrOMLV2: interval = 117–732, e-value = 5.63e-10) involved in viral endonuclease activity, necessary for the cap-snatching process (Guilligay et al., 2014). Genome segment four (1.6 Kbp size) encodes a Hemagglutinin protein – HA (PpyrOMLV1: 526 aa, 59.7 kDA; PpyrOMLV2: 525 aa, 58.6 kDA) presenting a Baculo_gp64 domain (Pfam: pfam03273; PpyrOMLV1: interval = 108–462, e-value = 2.16e-15; PpyrOMLV2: interval = 42–460, e-value = 1.66e-23), associated with the gp64 glycoprotein from baculovirus as well as other viruses, such as Thogotovirus (*Orthomyxoviridae* - OMV) which was postulated to be related to the arthropod-borne nature of these specific Orthomyxoviruses. In addition, HA as expected, presents an N-terminal signal domain, a C terminal transmembrane domain, and a putative glycosylation site. Lastly, genome segment five (ca. 1.8 Kbp size) encodes a putative nucleocapsid protein – NP (PpyrOMLV1: 562 aa, 62.3 kDA; PpyrOMLV2: 528 aa, 58.5 kDA) with a Flu_NP structural domain (Pfam: pfam00506; PpyrOMLV1: interval = 145–322, e-value = 1.32e-01; PpyrOMLV2: interval = 94–459, e-value = 1.47e-04) this single-strand RNA-binding protein is associated to encapsidation of the virus genome for the purposes of RNA transcription, replication and packaging (Supporting Information 5—figure 2E). Despite sharing genome architecture and structural and functional domains of their predicted proteins, PpyrOMLV1 and PpyrOMLV2 pairwise identity of ortholog gene products range between 21.4% (HA) to 49.8% (PB1), suggesting although a common evolutionary history, a strong divergence indicating separated species, borderline to be considered even members of different virus genera (Supporting Information 5—figure 3). The conserved 3' sequence termini of the viral genomic RNAs are (vgRNA ssRNA(-) 3'-end) 5'-GUUCUUACU-3' for PpyrOMLV1, and and 5'-(G/A)U(U/G)(G/U/C)(A/C/U)UACU-3'. for PpyrOMLV2. The 5' termini of the vgRNAs are partially complementary to the 3' termini, supporting a panhandle structure and a hook like structure of the 5' end by a terminal short stem loop. PpyrOMLV1 and PpyrOMLV2 genome segments present an overall high identity in their respective RNA segments ends (Supporting Information 5—figure 2F). These primary and secondary sequence cues are associated to polymerase binding and promotion of both replication and transcription. In influenza viruses, and probably every OMV, the first 10 nucleotides of the 3′ end form a stem-loop or 'hook' with four base-pairs (two canonical base-pairs flanked by an A-A base-pair). This compact RNA structure conforms the promoter, which activates polymerase initiation of RNA synthesis (Reich et al., 2017). The presence of eventual orthologs of *OMV* additional genome segments and proteins, such as Neuraminidase (NA), Matrix (M) and Non-structural proteins (NS1, NS2) was assessed retrieving no results by TBLASTN relaxed searches, nor with *in silico approaches* involving co-expression, expression levels, or conserved terminis. Given that the presence of those additional segments varies among diverse OMV genera, and that 35 related tentative new virus species identified in TSA did not present any additional segments, we believe that these lineages of viruses are conformed by five genome segments. Further experiments based on specific virus particle purification and target sequencing could corroborate our results. Based on sequence homology to best BLASTP hits, amino acid sequence alignments, predicted proteins and domains, and phylogenetic comparisons to reported species we assigned PpyrOMLV1 and PpyrOMLV2 to the OMV virus family. These are the first viruses that have been associated with the *Lampyridae* beetle family, which includes over 2000 species. The OMV virus members share diverse structural, functional and biological characters that define and restrict the family. OMV virions are 80–120 nm in diameter, of spherical or pleomorphic morphology. The virion envelope is derived from the host cell membrane, incorporating virus glycoproteins and eventually non-glycosylated proteins (one or two in number). Typical virion surface glycoprotein projections are 10–14 nm in length and 4–6 nm in diameter. The virus genome is multisegmented, has a helical-like symmetry, consisting of different size ribonucleoproteins (RNP), 50–150 nm in length. Influenza RNPs can perform either replication or transcription of the same template. Virions of each genus contain different numbers of linear ssRNA (-) genome segments (King et al., 2011). Influenza A virus (FLUAV), influenza B virus (FLUBV) and infectious salmon anemia virus (ISAV) are conformed of eight segments. Influenza C virus (FLUCV), Influenza D virus (FLUDV) and Dhori virus (DHOV) have seven segments. Thogoto virus (THOV) and Quaranfil virus (QUAV) have six segments. Johnston Atoll virus (JAV) genome is still incomplete, and only two segments have been described. Segment lengths range from 736 to 2396 nt. Genome size ranges from 10.0 to 14.6 Kbp (King et al., 2011). As described previously, every OMV RNA segment possess conserved and partially complementary 5′- and 3′-end sequences with promoter activity (Hsu et al., 1987). OMV structural proteins are tentatively common to all genera involving the three polypeptides subunits that form the viral RdRP (PA, PB1, PB2) (Pflug et al., n.d.)); a nucleoprotein (NP), which binds with each genome ssRNA segment to form RNPs; and the hemagglutinin protein (HA, HE or GP), which is a type I membrane integral glycoprotein involved in virus attachment, envelope fusion and neutralization. In addition, a non-glycosylated matrix protein (M) is present in most species. There are some species-specific divergence in some structural OMVs proteins. For instance, HA of FLUAV is acylated at the membrane-spanning region and has widespread N-linked glycans (Eisfeld et al., 2015). The HA protein of FLUCV, besides its hemagglutinating and envelope fusion function, has an esterase activity that induces host receptor enzymatic destruction (King et al., 2011). In contrast, the HA of THOV is divergent to influenza virus HA proteins, and presents high sequence similarity to a baculovirus surface glycoprotein (Leahy et al., 1997). The HA protein has been described to have an important role in determining OMV host specificity. For instance, human infecting Influenza viruses selectively bind to glycolipids that contain terminal sialyl-galactosyl residues with a 2–6 linkage, in contrast, avian influenza viruses bind to sialyl-galactosyl residues with a 2–3 linkage (King et al., 2011). Furthermore, FLUAV and FLUBV share a neuraminidase protein (NA), which is an integral, type II envelope glycoprotein containing sialidase activity. Some OMVs possess additional small integral membrane proteins (M2, NB, BM2, or CM2) that may be glycosylated and have diverse functions. As an illustration, M2 and BM2 function during un-coating and fusion by equilibrating the intralumenal pH of the trans-Golgi apparatus and the cytoplasm. In addition, some viruses encode two nonstructural proteins (NS1, NS2) (King et al., 2011). OMV share replication properties, which have been studied mostly in Influenza viruses. It is important to note that gene reassortment has been described to occur during mixed OMV infections, involving viruses of the same genus, but not between viruses of different genera (Kimble, 2013). This is used also as a criteria for OMV genus demarcation. Influenza virus replication and transcription occurs in the cell nucleus and comprises the production of the three types of RNA species (i) genomic RNA (vRNA) which are found in virions; (ii) cRNA molecules which are complementary RNA in sequence and identical in length to vRNA; and also (iii) virus mRNA molecules which are 5' capped by cap snatching of host RNAs and 3' polyadenylated by polymerase stuttering on U rich stretches. These remarkable dynamic multifunction characters of OMV polymerases are associated with its complex tertiary structure, of this modular heterotrimeric replicase (Te Velthuis and Fodor, 2016). We explored in detail the putative polymerase subunits of the identified firefly viruses. The PB1 subunit catalyzes RNA synthesis in its internal active site opening, which is formed by the highly conserved polymerase motifs I-III. Motifs I and III (Supporting Information 5—figure 2H) present three conserved aspartates (PpyrOMLV1: Asp 346, Asp 491 and Asp 492; PpyrOMLV2: Asp 348, Asp 495 and Asp 496) which coordinate and promote nucleophilic attack of the terminal 3' OH from the growing transcript on the alpha-phosphate of the inbound NTP (Pflug et al., n.d.). Besides presenting, with high confidence, the putative functional domains associated with their potential replicase/transcriptase function, we assessed whether the potential spatial and functional architecture was conserved at least in part in FOML viruses. In this direction we employed the SWISS-MODEL automated protein structure homology-modelling server to generate a 3D structure of PpyrOMLV1 heterotrimeric polymerase. The SWISS server selected as best-fit template the trimeric structure of Influenza A virus polymerase, generating a structure for each polymerase subunit of PpyrOMLV1. The generated structure shared structural cues related to its multiple role of RNA nucleotide binding, endonuclease, cap binding, and nucleotidyl transferase (Supporting Information 5—figure 2G-H). The engendered subunit structures suggest a probable conservation of PpyrOMLV1 POL, that could allow the predicted functional enzymatic activity of this multiple gene product. The overall polymerase rendered structure presents a typical U shape with two upper protrusions corresponding to the PA endonuclease and the PB2 cap-binding domain. The PB1 subunit appears to plug into the interior of the U and has the distinctive fold of related viral RNA polymerases with fingers, palm and thumb adjacent to a tentative central active site opening where RNA synthesis may occur (Hengrung et al., 2015; Reich et al., 2017). OMV Pol activity is central in the virus cycle of OMVs, which have been extensively studied. The life cycle of OMVs starts with virus entry involving the HA by receptor-mediated endocytosis. For Influenza, sialic acid bound to glycoproteins or glycolipids function as receptor determinants of endocytosis. Fusion between viral and cell membranes occurs in endosomes. The infectivity and fusion of influenza is associated to the post-translational cleavage of the virion HA. Cleavability depends on the number of basic amino acids at the target cleavage site (King et al., 2011). In thogotoviruses, no requirement for HA glycoprotein cleavage have been demonstrated (Leahy et al., 1997). Integral membrane proteins migrate through the Golgi apparatus to localized regions of the plasma membrane. New virions form by budding, incorporating matrix proteins and viral RNPs. Viral RNPs are transported to the cell nucleus where the virion polymerase complex synthesizes mRNA species (Hara et al., 2017). Another tentative function of the NP could be associated to the potential interference of the host immune response in the nucleus mediated by capsid proteins of some RNA virus, which could inhibit host transcription and thus liberate and direct it to viral RNA synthesis (Wulan et al., 2015). mRNA synthesis is primed by capped RNA fragments 10–13 nt in length that are generated by cap snatching from host nuclear RNAs which are sequestered after cap recognition by PB2 and incorporated to vRNA by PB1 and PA proteins which present viral endonuclease activity (Sikora et al., 2017). In contrast, thogotoviruses have capped viral mRNA without host-derived sequences at the 5′ end. Virus mRNAs are polyadenylated at the 3′ termini through iterative copying by the viral polymerase stuttering on a poly U track in the vRNA template. Some OMV mRNAs are spliced generating alternative gene products with defined functions. Protein synthesis of influenza viruses occurs in the cytoplasm. Partially complementary vRNA molecules act as templates for new viral RNA synthesis and are neither capped nor polyadenylated. These RNAs exist as RNPs in infected cells. Given the diverse hosts of OMV, biological properties of virus infection diverge between species. Influenzavirus A infect humans and cause respiratory disease, and they have been found to infect a variety of bird species and some mammalian species. Interspecies transmission, although rare, is well documented. Influenzavirus B infect humans and cause epidemics, and have been rarely found in seals. Influenzavirus C causes limited outbreaks in humans and have been occasionally found on dogs. Influenza spreads globally in a yearly outbreak, resulting in about three to five million cases of severe illness and about 250,000 to 500,000 human deaths (Thompson et al., 2009). Influenzavirus D has been recently reported and accepted and infects cows and swine (Hause et al., 2013). Natural transmission of influenzaviruses is by aerosol (human and non-aquatic hosts) or is water-borne (avians). In contrast, Thogoto and Dhori viruses which also infect humans, are transmitted by, and able to replicate in ticks. Thogoto virus was identified in *Rhipicephalus sp.* ticks collected from cattle in the Thogoto forest in Kenya, and Dhori virus was first isolated in India from *Hyalomma dromedarii*, a species of camel ticks (Anderson and Casals, 1973; Haig et al., 1965). Dhori virus infection in humans causes a febrile illness and encephalitis. Serological evidence suggests that cattle, camel, goats, and ducks might be also susceptible to this virus. Experimental hamster infection with THOV may be lethal. Unlike influenzaviruses, these viruses do not cause respiratory disease. The transmission of fish infecting isaviruses (ISAV) is via water, and virus infection induces the agglutination of erythrocytes of many fish species, but not avian or mammalian erythrocytes (Mjaaland et al., 1997). Quaranfil and Johnston Atoll are transmitted by ticks and infect avian species (Presti et al., 2009).

We have limited biological data of the firefly detected viruses. Nevertheless, a significant consistency in the genomic landscape and predicted gene products of the detected viruses in comparison with accepted OMV species sufficed to suggest for PpyrOMLV1 and PpyrOMLV2 a tentative taxonomic assignment within the OMV family. Besides relying on the OMV structural and functional signatures determined by virus genome annotation, we explored the evolutionary clustering of the detected viruses by phylogenetic insights. We generated MAFFT alignments and phylogenetic trees of the predicted viral polymerase of firefly viruses and the corresponding replicases of all 493 proposed and accepted species of ssRNA(-) virus. The generated trees consistently clustered the diverse sequences to their corresponding taxonomical niche, at the level of genera. Interestingly, PpyrOMLV1 and PpyrOMLV2 replicases were placed unequivocally within the OMV family (Supporting Information 5—figure 2B). When the genetic distances of firefly viruses proteins and ICTV accepted OMV species were computed, a strong similarity was evident (Supporting Information 5—figure 2B-D). Overall similarity levels of PpyrOMLV polymerase subunits ranged between 11.03% to as high as 37.30% among recognized species, while for the more divergent accepted OMV (ISAV - *Isavirus* genus) these levels ranged only from 8.54% to 20.74%, illustrating that PpyrOMLV are within the OMV by genetic standards. Phylogenetic trees based on aa alignments of structural gene products of recognized species and PpyrOMLV supported this assignment, placing ISAV and issavirus as the most distant species and genus within the family, and clustering PpyrOMLV1 and PpyrOMLV2 in a distinctive lineage within OMV, more closely related to the *Quaranjavirus* and *Thogotovirus* genera than the *Influenza* A-D or *Isavirus* genera (Supporting Information 5—figure 3). Furthermore, it appears that virus genomic sequence data, while it has been paramount to separate species, in the case of genera, there are some contrasting data that should be taken into consideration. For instance, DHOV and THOV are both members of the *Thogotovirus* genus, sharing a 61.9% and a 34.9% identity at PB1 and PB2, respectively. However, FLUCV and FLUDV are assigned members of two different genus, *Influenzavirus C* and *Influenzavirus D*, while sharing a higher 72.2% and a 52.2% pairwise identity at PB1 and PB2, respectively (Supporting Information 5—figure 3). In addition, FLUAV and FLUBV, assigned members of two different genus, *Influenzavirus A* and *Influenzavirus D* present a comparable identity to that of DHOV and THOV thogotoviruses, sharing a 61% and a 37.9% identity at PB1 and PB2, respectively. It is worth noting that similarity thresholds and phylogenetic clustering based in genomic data have been used differently to demarcate OMV genera, hence there is a need to eventually re-evaluate a series of consensus values, which in addition to biological data, would be useful to redefine the OMV family. Perhaps, these criteria discrepancies are more related to a historical evolution of the OMV taxonomy than to pure biological or genetic standards. In contrast to FLUDV, JOV and QUAV, the other virus members of OMV have been described, proposed and assigned at least 34 years ago.

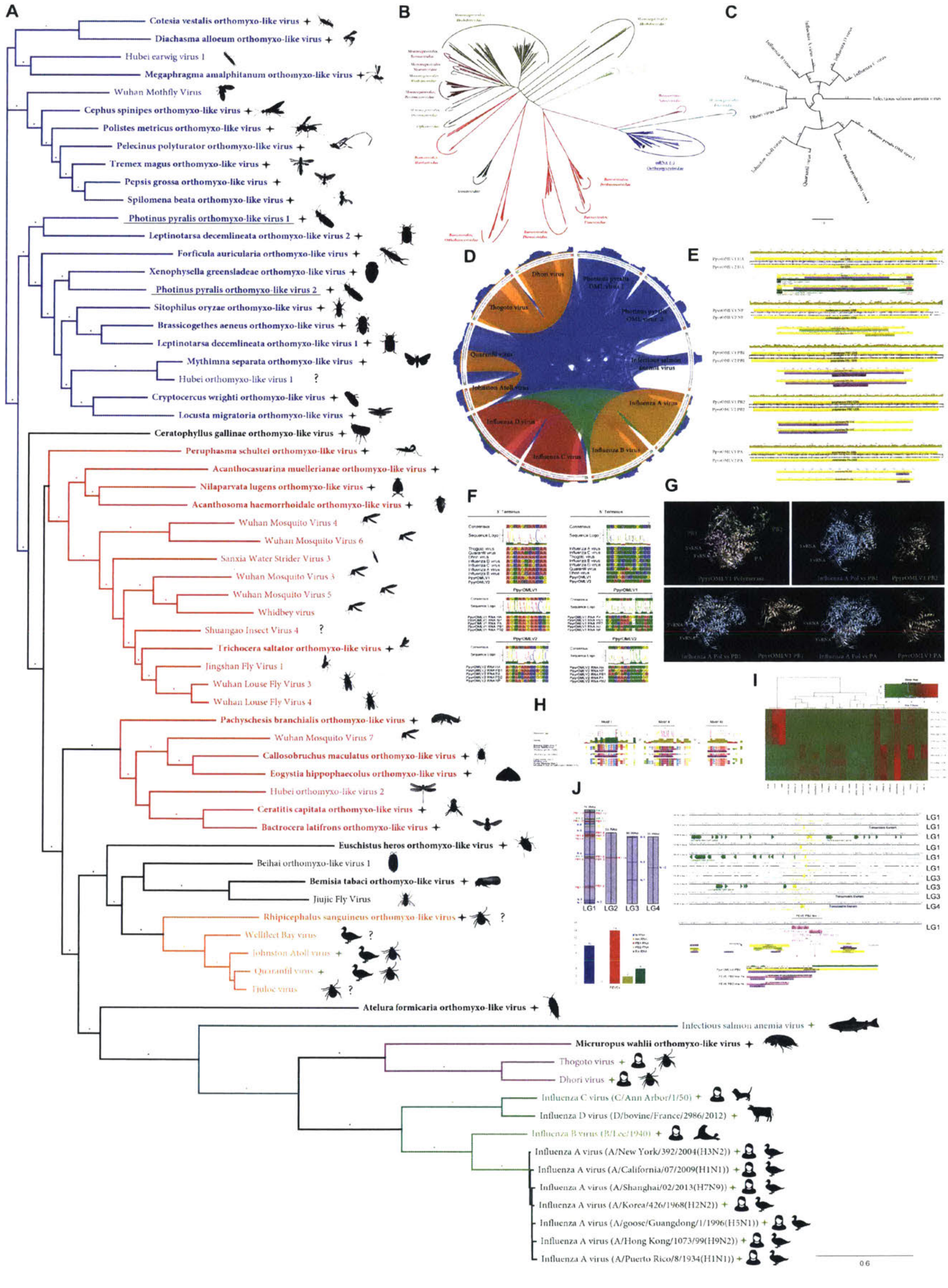
The potential prevalence, tissue/organ tropism, geographic dispersion and lifestyle of PpyrOMLV1 and 2 were assessed by the generation and analyses of 29 specific RNA-Seq libraries of *P. pyralis* (Supporting Information 1—table 1). As RNA was isolated from independent *P. pyralis* individuals of diverse origin, wild caught or lab reared, the fact that we found at least one of the PpyrOMLV present in 82% of the libraries reflects a widespread presence and potentially a high prevalence of these viruses in *P. pyralis* (Supporting Information 5—figure 2J, Supporting Information 5—table 3, Supporting Information 5.4.6). Wild caught individuals were collected in period spanning six years, and locations separated as much as 900 miles (New Jersey – Georgia, USA). Interestingly PpyrOMLV1 and 2 were found in individuals of both location, and the corresponding assembled isolate virus sequences presented negligible differences, with an inter-individual variability equivalent to that of isolates (0.012%). A similar result was observed for virus sequences identified in RNA libraries generated from samples collected in different years. We were not able to identified fixed mutations associated to geographical or chronological cues. Further experiments should explore the mutational landscape of PpyrOMLV1 and 2, which appears to be significantly lower than of Influenzaviruses, specifically

Influenza A virus, which are characterized by high mutational rate (ca. one mutation per genome replication) associated to the absence of RNA proofreading enzymes (Pauly et al., 2017). In addition we evaluated the presence of PpyrOMLV1 and 2 on diverse tissues and organs of *P. pyralis*. Overall virus RNA levels were generally low, with an average of 9.47 FPKM on positive samples. However, PpyrOMLV1 levels appear to be consistently higher than PpyrOMLV2, with an average of 20.50 FPKM for PpyrOMLV1 versus 4.22 FPKM for PpyrOMLV2 on positive samples. When the expression levels are scrutinized by genome segment, HA and NP encoding segments appear to be, for both viruses, at higher levels, which would be in agreement with other OMV such as Influenzaviruses, in which HA and NP proteins are the most expressed proteins, and thus viral mRNAs are consistently more expressed (King et al., 2011). Nevertheless, these preliminary findings related to expression levels should be taken cautiously, given the small sample size. Perhaps, the more remarkable allusion derived from the analyses of virus presence is related to tissue and organ deduced virus tropism. Strikingly, we found virus transcripts in samples exclusively obtained from light organs, complete heads, male or female thorax, female spermatheca, female spermatophore digesting glands and bursa, abdominal fat bodies, male reproductive spiral gland, and other male reproductive accessory glands (Supporting Information 5—table 3, Supporting Information 5.4.6), indicating a widespread tissue/organ tropism of PpyrOMLV1 and 2. This tentatively pervasive tropism of PpyrOMLV1 and 2 emerges as a differentiation character of these viruses and accepted OMV. For instance, influenza viruses present a epithelial cell-specific tropism, restricted typically to the nose, throat, and lungs of mammals, and intestines of birds. Tropism has consequences on host restriction. Human influenza viruses mainly infect ciliated cells, because attachment of all *influenza A virus* strains to cells requires sialic acids. Differential expression of sialic acid residues in diverse tissues may prevent cross-species or zoonotic transmission events of avian influenza strains to man (Zeng et al., 2013). Tropism has also influence in disease associated effects of OMV. Some *influenza A virus* strains are more present in tracheal and bronchial tissue which is associated with the primary lesion of tracheobronchitis observed in typical epidemic influenza. Other *influenza A virus* strains are more prevalent in type II pneumocytes and alveolar macrophages in the lower respiratory tract, which is correlated to diffuse alveolar damage with avian influenza (Mansfield, 2007). The presence of PpyrOMLV1 and 2 virus RNA in reproductive glands raises some potential of the involvement of sex in terms of prospective horizontal transmission. Given that most libraries corresponded to 3–6 pooled individuals samples of specific organs/tissue, direct comparisons of virus RNA levels were not always possible. However, this valuable data gives important insights into the widespread potential presence of the viruses in every analyzed organ/tissue. Importantly, RNA levels of the putative virus segments shared co-expression levels and a systematic pattern of presence/absence, supporting the suggested multipartite nature of the viruses. We observed the presence of virus RNA of both PpyrOMLV1 and 2 in eight of the RNA-Seq libraries, thus mixed infections appear to be common. Interestingly, we did not observe in any of the 24 virus positive samples evidence of reassortment. Reassortment is a common event in OMV, a process by which influenza viruses swap gene segments. Genetic exchange is possible due to the segmented nature of the OMV viral genome and may occur during mixed infections. Reassortment generates viral diversity and has been associated to host gain of Influenzavirus (Steel and Lowen, 2014). Reassorted Influenzavirus have been reported to occasionally cross the species barrier, into birds and some mammalian species like swine and eventually humans. These infections are usually dead ends, but sporadically, a stable lineage becomes established and may spread in an animal population (Kimble, 2013). Besides its evolutionary role, reassortment has been used as a criterion for species/genus demarcation, thus the lack of observed gene swap in our data supports the phylogenetic and sequence similarity insights that indicates species separation of PpyrOMLV1 and 2.

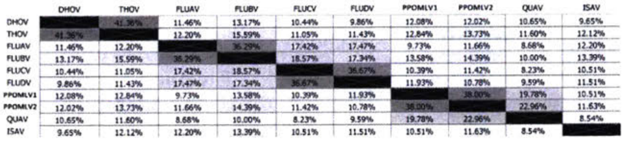
In light of the presence of virus RNA in reproductive glands, we further explored the potential life style of PpyrOMLV1 and 2 related to eventual vertical transmission. Vertical transmission is extremely exceptional for OMV, and has only been conclusively described for the *Infectious salmon anemia virus*

(*Isavirus)(Marshall et al., 2014)*. In this direction, we were able to generate a strand-specific RNA-Seq library of one *P. pyralis* adult female PpyrOMLV1 virus positive (parent), another library from seven eggs of this female at ~13 days post fertilization, and lastly an RNA-Seq library of four 1 st instar larvae (offspring). When we analyzed the resulting RNA reads, we found as expected virus RNA transcripts of every genome segment of PpyrOMLV1 in the adult female library. Remarkably, we also found PpyrOMLV1 sequence reads of every genome segment of PpyrOMLV1 in both the eggs and larvae samples. Moreover, virus RNA levels fluctuated among the different developmental stages of the samples. The average RNA levels of the adult female were 41.10 FPKM, in contrast, the fertilized eggs sample had higher levels of virus related RNA, averaging at 61.61 FPKM and peaking at the genome segment encoding NP (104.49 FPKM). Interestingly, virus RNA levels appear to drop in first instar larvae, in the sequenced library average virus RNA levels were of 10.42 FPKM. Future experiments should focus on PpyrOMLV1 and 2 virus titers at extended developmental stages to complement these preliminary results. However, it is interesting to note that the tissue specific library corresponding to female spermatheca, where male sperm are stored prior to fertilization, presented relatively high levels of both PpyrOMLV1 and 2 virus RNAs, suggesting that perhaps during early reproductive process and during egg development virus RNAs tend to raise. This tentatively differential and variable virus RNA titers observed during development could be associated to an unknown mechanism of modulation of latent antiviral response that could be repressed in specific life cycle stages. Further studies may validate these results and unravel a mechanistic explanation of this phenomenon. Nevertheless, besides the preliminary developmental data, the consistent presence of PpyrOMLV1 in lab-reared, isolated offspring of an infected *P. pyralis* female is robust evidence demonstrating mother-to-offspring vertical transmission for this newly identified OMV.

One of many questions that remains elusive here is whether PpyrOMLV1 and 2 are associated with any potential alteration of phenotype of the infected host. We failed to unveil any specific effect of the presence of PpyrOMLV1 and 2 on fireflies. It is worth noting that subtle alterations or symptoms would be difficult to pinpoint in these insects. Future studies should enquire whether PpyrOMLV1 and 2 may have any influence in biological attributes of fireflies such as fecundity, life span or life cycle. Nevertheless, we observed in our data some hints that could be indicative of a chronic state status, cryptic or latent infection of firefly individuals: (i) virus positive individuals presented in general relatively low virus RNA levels. (ii) virus RNA was found in every assessed tissue/organ. (iii) vertical transmission of the identified viruses. The first hint is hardly conclusive, it is difficult to define what a relatively low RNA level is, and high virus RNA loads are not directly associated with disease on reported OMV. The correlation of high prevalence, prolonged host infection, and vertical transmission observed in several new mosquito viruses has resulted in their classification as 'commensal' microbes. A shared evolutionary history of viruses and host, based in strategies of immune evasion of the viruses and counter antiviral strategies of the host could occasionally result in a modulation of viral loads and a chronic but latent state of virus infection (Hall et al., 2016).

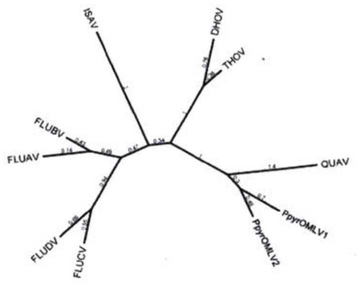

A
Cotesia vestalis orthomyxo-like virus +
Diachasma alloeum orthomyxo-like virus +
Hubei earwig virus 1
Megaphragma amalphitanum orthomyxo-like virus +
Wuhan Mothfly Virus
Cephus spinipes orthomyxo-like virus +
Polistes metricus orthomyxo-like virus +
Pelecinus polyturator orthomyxo-like virus +
Tremex magus orthomyxo-like virus +
Pepsis grossa orthomyxo-like virus +
Spilomena beata orthomyxo-like virus +
Photinus pyralis orthomyxo-like virus 1 +
Leptinotarsa decemlineata orthomyxo-like virus 2 +
Forficula auricularia orthomyxo-like virus +
Xenophysella greensladeae orthomyxo-like virus +
Photinus pyralis orthomyxo-like virus 2 +
Sitophilus oryzae orthomyxo-like virus +
Brassicogethes aeneus orthomyxo-like virus +
Leptinotarsa decemlineata orthomyxo-like virus 1 +
Mythimna separata orthomyxo-like virus +
Hubei orthomyxo-like virus 1 ?
Cryptocercus wrighti orthomyxo-like virus +
Locusta migratoria orthomyxo-like virus +
Ceratophyllus gallinae orthomyxo-like virus +
Peruphasma schultei orthomyxo-like virus +
Acanthocasuarina muellerianae orthomyxo-like virus +
Nilaparvata lugens orthomyxo-like virus +
Acanthosoma haemorrhoidale orthomyxo-like virus +
Wuhan Mosquito Virus 4
Wuhan Mosquito Virus 6
Sanxia Water Strider Virus 3
Wuhan Mosquito Virus 3
Wuhan Mosquito Virus 5
Whidbey virus
Shuangao Insect Virus 4 ?
Trichocera saltator orthomyxo-like virus +
Jingshan Fly Virus 1
Wuhan Louse Fly Virus 3
Wuhan Louse Fly Virus 4
Pachyschesis branchialis orthomyxo-like virus +
Wuhan Mosquito Virus 7
Callosobruchus maculatus orthomyxo-like virus +
Eogystia hippophaecolus orthomyxo-like virus +
Hubei orthomyxo-like virus 2
Ceratitis capitata orthomyxo-like virus +
Bactrocera latifrons orthomyxo-like virus +
Euschistus heros orthomyxo-like virus +
Beihai orthomyxo-like virus 1
Bemisia tabaci orthomyxo-like virus +
Jiujie Fly Virus
Rhipicephalus sanguineus orthomyxo-like virus + ?
Wellfleet Bay virus ?
Johnston Atoll virus +
Quaranfil virus +
Tjuloc virus ?
Atelura formicaria orthomyxo-like virus +
Infectious salmon anemia virus +
Micruropus wahlii orthomyxo-like virus +
Thogoto virus +
Dhori virus +
Influenza C virus (C/Ann Arbor/1/50) +
Influenza D virus (D/bovine/France/2986/2012) +
Influenza B virus (B/Lee/1940) +
Influenza A virus (A/New York/392/2004(H3N2)) +
Influenza A virus (A/California/07/2009(H1N1)) +
Influenza A virus (A/Shanghai/02/2013(H7N9)) +
Influenza A virus (A/Korea/426/1968(H2N2)) +
Influenza A virus (A/goose/Guangdong/1/1996(H5N1)) +
Influenza A virus (A/Hong Kong/1073/99(H9N2)) +
Influenza A virus (A/Puerto Rico/8/1934(H1N1)) +
0.6
B
ssRNA (-)
Orthomyxoviridae
C
Influenza A virus
Influenza D virus
Influenza B virus
Influenza C virus
Thogoto virus
Infectious salmon anemia virus
Dhori virus
Photinus pyralis OML virus 2
Johnston Atoll virus
Quaranfil virus
Photinus pyralis OML virus 1
D
Dhori virus
Photinus pyralis OML virus 1
Thogoto virus
Photinus pyralis OML virus 2
Quaranfil virus
Infectious salmon anemia virus
Johnston Atoll virus
Influenza A virus
Influenza D virus
Influenza B virus
Influenza C virus
E
PpyrOMLV1 NP
PpyrOMLV2 NP
PpyrOMLV1 PB1
PpyrOMLV2 PB1
PpyrOMLV1 PB2
PpyrOMLV2 PB2
PpyrOMLV1 PA
PpyrOMLV2 PA
F
3' Terminus
5' Terminus
Consensus
Sequence Logo
Thogoto virus
Quaranfil virus
Dhori virus
Influenza D virus
Influenza A virus
Influenza C virus
Influenza B virus
PpyrOMLV1
PpyrOMLV2
G
PB1
PB2
PA
5'vRNA
3'vRNA
PpyrOMLV1 Polymerase
Influenza A Pol vs PB2
PpyrOMLV1 PB2
Influenza A Pol vs PB1
PpyrOMLV1 PB1
Influenza A Pol vs PA
PpyrOMLV1 PA
H
Motif I
Motif II
Motif III
I
J
LG1
LG2
LG3
LG4
Transposable Element
LG1
LG1
LG1
LG1
LG1
LG1
LG3
LG3
LG3
LG4
LG1

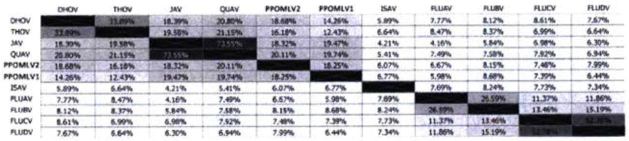
Supporting Information 5—figure 2. Photinus pyralis viruses and endogenous viral-like elements. (**A**) Phylogenetic tree based in MAFFT alignments of predicted replicases of *Orthomyxoviridae* (OMV) ICTV accepted viruses (green stars), new *Photinus pyralis* viruses (underlined) and tentative OMV-like virus species (black stars). ICTV recognized OMV genera: *Quaranjavirus* (orange), *Thogotovirus* (purple), *Issavirus* (turquoise), *Influenzavirus A-D* (green). Silhouettes correspond to host species. Asterisk denote FastTree consensus support >0.5. Question marks depict viruses with unidentified or unconfirmed host. (**B**) Phylogenetic tree of OMV proposed and recognized species in the context of all ssRNA (-) virus species, based on MAFFT alignments of refseq replicases. *Photinus* pyralis viruses are portrayed by black stars. (**C**) Phylogenetic tree of ICTV recognized OMV species and PpyrOMLV1 and 2. Numbers indicate FastTree consensus support. (**D**) Genetic distances of concatenated gene products of OMV depicted as circoletto diagrams. Proteins are oriented clockwise in N-HA-PB1-PB2-PA order when available. Sequence similarity is expressed as ribbons ranging from blue (low) to red (high). (**E**) Genomic architecture, predicted gene products and structural and functional domains of PpyrOLMV1 and 2. (**F**) Virus genomic noncoding termini analyses of PpyrOLMV1 and 2 in the context of ICTV OMV. The 3' and 5' end, A and U rich respectively, partially complementary sequences are associated to tentative panhandle polymerase binding and replication activity, typical of OMV. (**G**) 3D renders of the heterotrimeric polymerase of PpyrOMLV1 based on Swiss-Expasy generated models using as template the Influenza A virus polymerase structure. Structure comparisons were made with the MatchAlign tool of the Chimera suite, and solved in PyMOL. (**H**) Conserved functional motifs of PpyrOLMV1 and 2 PB1 and related viruses. Motif I-III are essential for replicase activity of viral polymerase. (**I**) Dynamic and prevalent virus derived RNA levels of the corresponding PpyrOMLV1 and 2 genome segments, determined in 24 RNA libraries of diverse individuals/developmental stages/tissues and geographic origins. RNA levels are expressed as normalized TPM, heatmaps were generated by Shinyheatmap. Values range from low (green) to high (red). (**J**) Firefly EVEs (FEVEs) identified in the *P. pyralis* genome assembly mapped to the corresponding pseudo-molecules. A 15 Kbp region flanking nucleoprotein like FEVES are depicted, enriched in transposable elements. Representative products of a putative PB2 FEVE are aligned to the corresponding protein of PpyrOMLV 2.



	DHOV	THOV	FLUAV	FLUBV	FLUCV	FLUDV	PPOMLV1	PPOMLV2	QUAV	ISAV
DHOV		41.36%	11.46%	13.17%	10.44%	9.86%	12.08%	12.02%	10.65%	9.65%
THOV	41.36%		12.20%	15.59%	11.05%	11.43%	12.84%	13.73%	11.60%	12.12%
FLUAV	11.46%	12.20%		36.29%	17.42%	17.47%	9.73%	11.66%	8.68%	12.20%
FLUBV	13.17%	15.59%	36.29%		18.57%	17.34%	13.58%	14.39%	10.00%	13.39%
FLUCV	10.44%	11.05%	17.42%	18.57%		36.67%	10.39%	11.42%	8.23%	10.51%
FLUDV	9.86%	11.43%	17.47%	17.34%	36.67%		11.93%	10.78%	9.59%	11.51%
PPOMLV1	12.08%	12.84%	9.73%	13.58%	10.39%	11.93%		38.00%	19.78%	10.51%
PPOMLV2	12.02%	13.73%	11.66%	14.39%	11.42%	10.78%	38.00%		22.96%	11.63%
QUAV	10.65%	11.60%	8.68%	10.00%	8.23%	9.59%	19.78%	22.96%		8.54%
ISAV	9.65%	12.12%	12.20%	13.39%	10.51%	11.51%	10.51%	11.63%	8.54%	

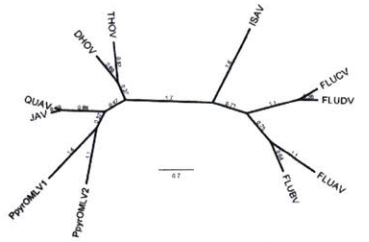
Nucleoprotein

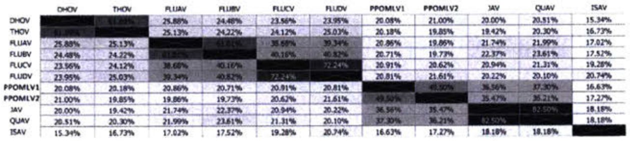




	DHOV	THOV	JAV	QUAV	PPOMLV2	PPOMLV1	ISAV	FLUAV	FLUBV	FLUCV	FLUDV
DHOV		33.89%	18.39%	20.80%	18.68%	14.26%	5.89%	7.77%	8.12%	8.61%	7.67%
THOV	33.89%		19.58%	21.15%	16.18%	12.43%	6.64%	8.47%	8.37%	6.99%	6.64%
JAV	18.39%	19.58%		73.55%	18.32%	19.47%	4.21%	4.16%	5.84%	6.98%	6.30%
QUAV	20.80%	21.15%	73.55%		20.11%	19.74%	5.41%	7.49%	7.58%	7.92%	6.94%
PPOMLV2	18.68%	16.18%	18.32%	20.11%		18.25%	6.07%	6.67%	8.15%	7.48%	7.99%
PPOMLV1	14.26%	12.43%	19.47%	19.74%	18.25%		6.77%	5.98%	8.68%	7.39%	6.44%
ISAV	5.89%	6.64%	4.21%	5.41%	6.07%	6.77%		7.69%	8.24%	7.73%	7.34%
FLUAV	7.77%	8.47%	4.16%	7.49%	6.67%	5.98%	7.69%		26.59%	11.37%	11.86%
FLUBV	8.12%	8.37%	5.84%	7.58%	8.15%	8.68%	8.24%	26.59%		13.46%	15.19%
FLUCV	8.61%	6.99%	6.98%	7.92%	7.48%	7.39%	7.73%	11.37%	13.46%		52.78%
FLUDV	7.67%	6.64%	6.30%	6.94%	7.99%	6.44%	7.34%	11.86%	15.19%	52.78%	

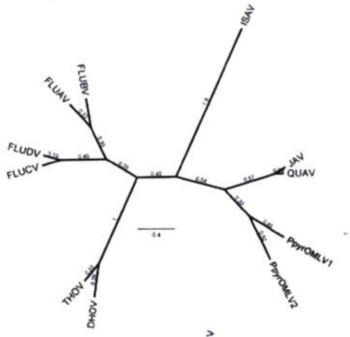
Hemaglutinin

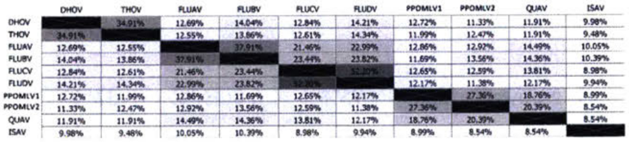




	DHOV	THOV	FLUAV	FLUBV	FLUCV	FLUDV	PPOMLV1	PPOMLV2	JAV	QUAV	ISAV
DHOV		[illegible]	25.88%	24.48%	23.56%	23.95%	20.08%	21.00%	20.00%	20.51%	15.34%
THOV	[illegible]		25.13%	24.22%	24.12%	25.03%	20.18%	19.85%	19.42%	20.30%	16.73%
FLUAV	25.88%	25.13%		[illegible]	38.68%	39.34%	20.86%	19.86%	21.74%	21.99%	17.02%
FLUBV	24.48%	24.22%	[illegible]		40.16%	40.82%	20.71%	19.73%	22.37%	23.61%	17.52%
FLUCV	23.56%	24.12%	38.68%	40.16%		72.24%	20.91%	20.62%	20.94%	21.31%	19.28%
FLUDV	23.95%	25.03%	39.34%	40.82%	72.24%		20.81%	21.61%	20.22%	20.10%	20.74%
PPOMLV1	20.08%	20.18%	20.86%	20.71%	20.91%	20.81%		49.50%	36.56%	37.30%	16.63%
PPOMLV2	21.00%	19.85%	19.86%	19.73%	20.62%	21.61%	49.50%		35.47%	36.21%	17.27%
JAV	20.00%	19.42%	21.74%	22.37%	20.94%	20.22%	36.56%	35.47%		82.50%	18.18%
QUAV	20.51%	20.30%	21.99%	23.61%	21.31%	20.10%	37.30%	36.21%	82.50%		18.18%
ISAV	15.34%	16.73%	17.02%	17.52%	19.28%	20.74%	16.63%	17.27%	18.18%	18.18%	

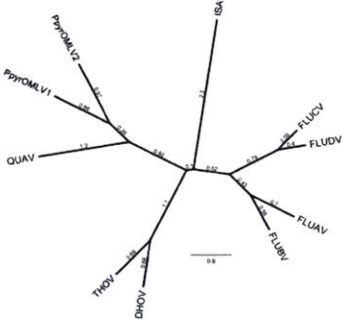
PB1 Polymerase

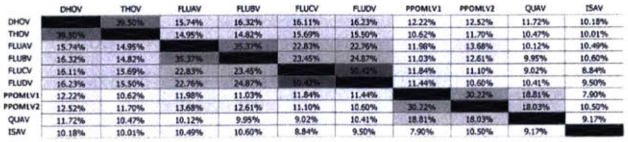




	DHOV	THOV	FLUAV	FLUBV	FLUCV	FLUDV	PPOMLV1	PPOMLV2	QUAV	ISAV
DHOV		34.91%	12.69%	14.04%	12.84%	14.21%	12.72%	11.33%	11.91%	9.98%
THOV	34.91%		12.55%	13.86%	12.61%	14.34%	11.99%	12.47%	11.91%	9.48%
FLUAV	12.69%	12.55%		37.91%	21.46%	22.99%	12.86%	12.92%	14.49%	10.05%
FLUBV	14.04%	13.86%	37.91%		23.44%	23.82%	11.69%	13.56%	14.36%	10.39%
FLUCV	12.84%	12.61%	21.46%	23.44%		52.20%	12.65%	12.59%	13.81%	8.98%
FLUDV	14.21%	14.34%	22.99%	23.82%	52.20%		12.17%	11.38%	12.17%	9.94%
PPOMLV1	12.72%	11.99%	12.86%	11.69%	12.65%	12.17%		27.36%	18.76%	8.99%
PPOMLV2	11.33%	12.47%	12.92%	13.56%	12.59%	11.38%	27.36%		20.39%	8.54%
QUAV	11.91%	11.91%	14.49%	14.36%	13.81%	12.17%	18.76%	20.39%		8.54%
ISAV	9.98%	9.48%	10.05%	10.39%	8.98%	9.94%	8.99%	8.54%	8.54%	

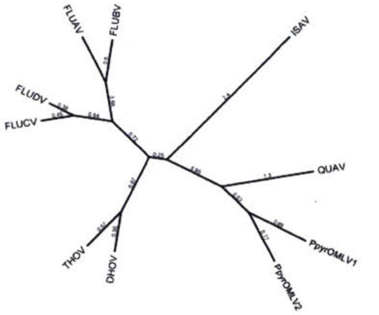
PB2 Polymerase





	DHOV	THOV	FLUAV	FLUBV	FLUCV	FLUDV	PPOMLV1	PPOMLV2	QUAV	ISAV
DHOV		39.50%	15.74%	16.32%	16.11%	16.23%	12.22%	12.52%	11.72%	10.18%
THOV	39.50%		14.95%	14.82%	15.69%	15.50%	10.62%	11.70%	10.47%	10.01%
FLUAV	15.74%	14.95%		35.37%	22.83%	22.76%	11.98%	13.68%	10.12%	10.49%
FLUBV	16.32%	14.82%	35.37%		23.45%	24.87%	11.03%	12.61%	9.95%	10.60%
FLUCV	16.11%	15.69%	22.83%	23.45%		50.42%	11.84%	11.10%	9.02%	8.84%
FLUDV	16.23%	15.50%	22.76%	24.87%	50.42%		11.44%	10.60%	10.41%	9.50%
PPOMLV1	12.22%	10.62%	11.98%	11.03%	11.84%	11.44%		30.22%	18.81%	7.90%
PPOMLV2	12.52%	11.70%	13.68%	12.61%	11.10%	10.60%	30.22%		18.03%	10.50%
QUAV	11.72%	10.47%	10.12%	9.95%	9.02%	10.41%	18.81%	18.03%		9.17%
ISAV	10.18%	10.01%	10.49%	10.60%	8.84%	9.50%	7.90%	10.50%	9.17%	

PA Polymerase



Supporting Information 5—figure 3. Pairwise identity of OMLV viral proteins amongst identified OMLV viruses.

Supporting Information 5—table 1. Best hits from BLASTP of PpyrOMLV proteins against the NCBI database

Genome segment	Size (nt)	Gene product (aa)	Best hit	Best hit taxonomy	Query cover	E value	Identity
PpyrOMLV1-PB1	2510	801 PB1	Wuhan Mothfly Virus	Orthomyxoviridae	83%	0.0	**51%**
PpyrOMLV1-PA	2346	754 PA	Hubei earwig virus 1	Orthomyxoviridae	98%	4.00E-137	**35%**
PpyrOMLV1-HA	1667	526 HA	Tjuloc virus	Orthomyxoviridae	91%	9.00E-25	**25%**
PpyrOMLV1-PB2	2517	804 PB2	Hubei earwig virus 1	Orthomyxoviridae	91%	3.00E-118	**31%**
PpyrOMLV1-N	1835	562 N	Hubei earwig virus 1	Orthomyxoviridae	93%	8.00E-74	**30%**
PpyrOMLV2-PB1	2495	802 PB1	Hubei orthomyxo-like virus 1	Orthomyxoviridae	93%	0.0	**48%**
PpyrOMLV2-PA	2349	762 PA	Hubei earwig virus 1	Orthomyxoviridae	98%	1.00E-107	**31%**
PpyrOMLV2-HA	1668	525 HA	Wellfleet Bay virus	Orthomyxoviridae	82%	3.00E-40	**26%**
PpyrOMLV2-PB2	2506	801 PB2	Hubei earwig virus 1	Orthomyxoviridae	96%	3.00E-86	**27%**
PpyrOMLV2-N	1738	528 N	Hubei earwig virus 1	Orthomyxoviridae	95%	6.00E-82	**32%**

Supporting Information 5—table 2. InterProScan domain annotation of PpyrOMLV proteins.

Genome product	Annotation	Start	End	Length	Database	Id	InterPro ID	InterPro name
PpyrOMLV1-PB1	Flu_PB1	48	752	705	PFAM	PF00602	IPR001407	RNA_pol_PB1_influenza
	RDRP_SSRNA	330	529	200	PROSITE_PROFILES	PS50525	IPR007099	RNA-dir_pol_NSvirus
PpyrOMLV2-PB1	Flu_PB1	54	766	713	PFAM	PF00602	IPR001407	RNA_pol_PB1_influenza
	RDRP_SSRNA	337	539	203	PROSITE_PROFILES	PS50525	IPR007099	RNA-dir_pol_NSvirus

PpyrOMLV1-PB2	Flu_PB2	13	421	409	PFAM	PF00604	IPR001591	RNA_pol_PB2_orthomyxovir
PpyrOMLV2-PB2	Flu_PB2	13	415	403	PFAM	PF00604	IPR001591	RNA_pol_PB2_orthomyxovir
PpyrOMLV1-HA	SignalP-noTM	1	19	19	SIGNALP_EUK	SignalP-noTM		Unintegrated
	Baculo_gp64	108	432	325	PFAM	PF03273	IPR004955	Baculovirus_Gp64
PpyrOMLV2-HA	SignalP-noTM	1	21	21	SIGNALP_EUK	SignalP-noTM		Unintegrated
	Baculo_gp64	66	426	361	PFAM	PF03273	IPR004955	Baculovirus_Gp64
PpyrOMLV1-PA	Flu_PA	663	736	74	PFAM	PF00603	IPR001009	RNA-dir_pol_influenzavirus
PpyrOMLV2-PA	Flu_PA	667	740	74	PFAM	PF00603	IPR001009	RNA-dir_pol_influenzavirus
PpyrOMLV1-PB1	flu NP-like	94	459	366	SUPERFAMILY	SSF161003		Unintegrated
PpyrOMLV2-PB1	flu NP-like	363	483	121	SUPERFAMILY	SSF161003		Unintegrated

Supporting Information 5—table 3. Total reads mapped to PpyrOMLV genome segments from *P. pyralis* RNA-Seq datasets.

	SRR 3883773	**SRR 3883772**	**SRR 3883758**	**SRR 3883771**	**SRR 3883770**	**SRR 3883769**	**SRR 3883768**	**SRR 3883767**	**SRR 3883765**	**SRR 3883764**	**SRR 3883763**	**SRR 3883762**
Ppyr OMLV1 HA	11	541	2	160	0	4	881	2	0	2	199	2848
Ppyr OMLV1 NP	0	321	0	141	0	0	523	0	0	0	120	1460
Ppyr OMLV1 PA	3	256	0	95	0	0	306	1	0	5	100	660

Ppyr OMLV1 PB1	2	364	2	208	0	4	820	0	0	0	669	1464
Ppyr OMLV1 PB2	5	194	0	152	2	0	319	2	0	0	106	696
Ppyr OMLV2 HA	12	444	266	124	54	247	549	38	22	10	232	710
Ppyr OMLV2 NP	29	526	275	144	66	299	653	24	205	57	274	1067
Ppyr OMLV2 PA	12	88	216	72	40	204	97	18	15	8	50	838
Ppyr OMLV2 PB1	9	115	75	72	26	78	76	8	74	57	146	493
Ppyr OMLV2 PB2	5	50	57	67	47	131	110	22	85	72	173	728

	SRR3883761	**SRR3883760**	**SRR3883759**	**SRR3883757**	**SRR3883756**	**SRR3883766**	**SRR2103867**	**SRR2103849**	**SRR2103848**	**Ppyr_larvae**	**Ppyr_Female**	**Ppyr_eggs**
Ppyr OMLV1 HA	0	578	2	6	867	0	0	0	0	1664	7826	15586
Ppyr OMLV1 NP	0	289	0	3	647	0	2	0	0	644	5216	6562
Ppyr OMLV1 PA	0	124	0	2	626	0	0	0	0	1264	3692	9564
Ppyr OMLV1 PB1	2	460	0	3	1607	2	0	0	0	2824	7144	15952
Ppyr	0	188	0	2	848	0	0	0	0	648	2562	10568

OMLV1 PB2												
Ppyr OMLV2 HA	13	236	23	546	337	286	43	190	415	0	0	0
Ppyr OMLV2 NP	32	248	22	501	482	196	51	127	432	0	0	0
Ppyr OMLV2 PA	14	93	6	234	222	131	75	54	97	0	0	0
Ppyr OMLV2 PB1	29	90	4	168	180	63	22	96	190	0	0	0
Ppyr OMLV2 PB2	49	90	6	256	230	94	22	57	96	0	0	0

Supporting Information 5—table 4. FPKM of reads mapped to PpyrOMLV genome segments from *P. pyralis* RNA-Seq datasets.

	SRR	**SRR**	**SRR**	**SRR**	**SRR**	**SRR**	**SRR**	**SRR**	**SRR**	**SRR**		
	38837	**3883**	**38837**	**3883**	**3883**	**38837**	**38837**	**3883**	**38837**	**388376**	**SRR**	**SRR**
	73	**772**	**58**	**771**	**770**	**69**	**68**	**767**	**65**	**4**	**3883763**	**3883762**
Ppyr OMLV 1 HA	19.10	0.32	0.05	6.46	0.00	0.11	30.69	0.05	0.00	0.08	4.07	69.54
Ppyr OMLV 1 NP	10.37	0.00	0.00	5.21	0.00	0.00	16.66	0.00	0.00	0.00	2.24	32.61
Ppyr OMLV 1 PA	6.46	0.06	0.00	2.74	0.00	0.00	7.62	0.02	0.00	0.13	1.46	11.52
Ppyr OMLV 1 PB1	8.53	0.04	0.04	5.57	0.00	0.07	18.95	0.00	0.00	0.00	9.07	23.72

Ppyr OMLV 1 PB2	4.50	0.10	0.00	4.03	0.05	0.00	7.29	0.03	0.00	0.00	1.42	11.16
Ppyr OMLV 2 HA	16.13	0.36	7.41	5.15	2.31	6.80	19.68	0.90	1.05	0.39	4.88	17.84
Ppyr OMLV 2 NP	17.36	0.79	6.96	5.44	2.57	7.48	21.27	0.52	8.87	2.01	5.24	24.36
Ppyr OMLV 2 PA	2.21	0.25	4.17	2.07	1.19	3.89	2.41	0.30	0.49	0.21	0.73	14.58
Ppyr OMLV 2 PB1	2.73	0.18	1.37	1.95	0.73	1.40	1.78	0.12	2.30	1.44	2.01	8.10
Ppyr OMLV 2 PB2	1.18	0.10	1.03	1.81	1.31	2.34	2.56	0.34	2.63	1.81	2.36	11.88

	SRR 3883761	**SRR 3883760**	**SRR 3883759**	**SRR 3883757**	**SRR 3883756**	**SRR 3883766**	**SRR 2103867**	**SRR 2103849**	**SRR 2103848**	**Ppyr_ larvae**	**Ppyr_ Female**	**Ppyr_ eggs**
Ppyr OMLV 1 HA	0.00	18.29	0.08	0.21	23.44	0.00	0.00	0.00	0.00	15.89	74.25	104.49
Ppyr OMLV 1 NP	0.00	8.37	0.00	0.09	16.00	0.00	0.04	0.00	0.00	5.62	45.27	40.24
Ppyr OMLV 1 PA	0.00	2.81	0.00	0.05	12.10	0.00	0.00	0.00	0.00	8.63	25.05	45.85
Ppyr OMLV 1 PB1	0.04	9.66	0.00	0.07	28.83	0.04	0.00	0.00	0.00	17.89	44.97	70.96
Ppyr	0.00	3.91	0.00	0.05	15.05	0.00	0.00	0.00	0.00	4.06	15.96	46.51

OMLV 1 PB2												
Ppyr OMLV 2 HA	0.43	7.68	0.95	19.30	9.38	9.74	1.02	4.94	8.95	0.00	0.00	0.00
Ppyr OMLV 2 NP	0.97	7.34	0.82	16.09	12.19	6.07	1.10	3.00	8.47	0.00	0.00	0.00
Ppyr OMLV 2 PA	0.32	2.10	0.17	5.73	4.28	3.09	1.23	0.97	1.45	0.00	0.00	0.00
Ppyr OMLV 2 PB1	0.63	1.92	0.11	3.88	3.27	1.40	0.34	1.63	2.68	0.00	0.00	0.00
Ppyr OMLV 2 PB2	1.06	1.90	0.16	5.88	4.16	2.08	0.34	0.96	1.35	0.00	0.00	0.00

5.5 *P. pyralis* Endogenous virus-like Elements (EVEs)

To gain insights on the potential shared evolutionary history of *P. pyralis* and the IOMV PpyrOMLV1 and 2, we examined our assembly of *P. pyralis* for putative signatures or paleovirological traces (Ballinger et al., 2014; Feschotte and Gilbert, 2012; Metegnier et al., 2015) that would indicate ancestral integration of virus related sequences into the firefly host. Remarkably, we found Endogenous virus-like Elements (EVEs) (Katzourakis and Gifford, 2010), sharing significant sequence identity with most PpyrOMLV1 and 2 genome segments, spread along four *P. pyralis* linkage-groups. Virus integration into host genomes is a frequent event derived from reverse transcribing RNA viruses (*Retroviridae*). Retroviruses are the only animal viruses that depend on integration into the genome of the host cell as an obligate step in their replication strategy (Temin, 1985). Viral infection of germ line cells may lead to viral gene fragments or genomes becoming integrated into host chromosomes and subsequently inherited as host genes.

Animal genomes are paved by retrovirus insertions (Bushman et al., 2005). These insertions, which are eventually eliminated from the host gene pool within a few generations, and may, in some cases, increase in frequency, and ultimately reach fixation. This fixation in the host species can be mediated by drift or positive selection, depending on their selective value. On the other hand, genomic integration of non-retroviral viruses, such as PpyrOMLV1 and 2, is less common. Viruses with a life cycle characterized by no DNA stage, such as OMV, do not encode a reverse transcriptase or integrase, thus are not retro transcribed nor integrated into the host genome. However, exceptionally and recently, several non-retroviral sequences have been identified on animal genomes; these insertions have been usually associated with the transposable elements machinery of the host, which provided a means to genome integration (Gilbert and Cordaux, 2017; Palatini et al., 2017). Interestingly, when we screened our *P. pyralis* genome assembly Ppyr1.2 by BLASTX searches (E-value <1e10−6) of PpyrOMLV1 and 2 genome segments, we identified several genome regions that could be defined as Firefly EVEs, which we termed FEVEs (Supporting Information 5—figure 2J; Supporting Information 5—table 5-8). We found 30 OMV related FEVEs, which were mostly found in linkage group one (LG1, 83% of pinpointed FEVEs). The majority of the detected FEVEs shared sequence identity to the PB1 encoding region of genome segment one of PpyrOMLV1 and 2 (ca. 46% of FEVEs; Supporting Information 5—table 5), followed by NP encoding genome segment five (ca. 33% of detected FEVEs; Supporting Information 5—table 8). In addition we identified four FEVEs related to genome segment three (PA region; Supporting Information 5—table 7) and two FEVEs associated to genome segment two (PB2 encoding region; Supporting Information 5—table 6). We found no evidence of FEVEs related to the hemagglutinin coding genome segment four (HA) via BLASTX. The detected *P. pyralis* FEVEs represented truncated fragments of virus like sequences, generally presenting frameshift mutations, early termination codons, lacking start codons, and sharing diverse mutations that altered the potential translation of eventual gene products. FEVEs shared sequence similarity to the coding sequence of specific genome segments of the cognate FOLMV.

We generated best/longest translation products of the corresponding FEVEs, which presented an average length of ca. 21.86% of the corresponding PpyrOMLV genome segment encoding gene region (Supporting Information 5—table 5-5.5.5), and an average pairwise identity to the FOLMV virus protein of 55.08%. Nevertheless, we were able to identify FEVEs that covered as high as ca. 60% of the corresponding gene product, and in addition, although at specific short protein regions of the putative related FOLMV, similarity values were as high as 89% pairwise identity. In addition, most of the detected FEVEs were flanked by Transposable Elements (TE) (Supporting Information 5—figure 2J) suggesting that integration followed ectopic recombination between viral RNA and transposons. We found several conserved domains associated to reverse transcriptases and integrases adjacent to the corresponding FEVEs, which supports the hypothesis that these virus-like elements could be reminiscent of an OMV-like ancestral virus that could have been integrated into the genome by occasional sequestering of viral RNAs by the TE machinery. The finding of EVEs in the *P. pyralis* genome is not trivial, OMV EVEs are extremely rare. There has been only one report of OMV like sequences integrated into animal host genomes, which is the case of *Ixodes scapularis*, the putative vector of *Quaranfil virus* and *Johnston Atoll virus* corresponding to genus *Quaranjavirus* (Katzourakis and Gifford, 2010). The fact that besides FEVEs, the only other OMV EVE corresponded to an Arthropod genome, given the ample studies of bird and mammal genomes, is suggestive that perhaps OMV EVEs are restricted to Arthropod hosts. Sequence similarity of FEVEs and firefly viruses suggest that these viral 'molecular fossils' could have been tightly associated to PpyrOLMV1 and 2 ancestors. Moreover, we found potential NP and PB1 EVEs in our genome of light emitting click beetle *Ignelater luminosus* (Elateridae), an evolutionary distant coleoptera. Sequence similarity levels of the corresponding EVEs averaging 52%, could not be related with evolutionary distances of the hosts. We were not able to generate conclusive phylogenetic insights of the detected EVEs, given their partial, truncated and altered nature of the virus like sequences. In specific cases such as PB1-like EVEs there appears to be a trend suggesting an indirect relation between sequence identity and evolutionary status of the firefly host, but this preceding findings should be taken cautiously until more gathered data is available. The widespread presence of DNA sequences significantly similar to OMV in the explored firefly and related genomes are an interesting and intriguing result. At this stage is prudently not to venture to suggest more likely one of the two plausible explanations of the presence of these sequences in related beetles genomes: (i) Ancestral OMV like virus sequences were retrotranscribed and incorporated to an ancient beetle, followed by speciation and eventual stabilization or lost of EVEs in diverse species. (ii) Recent and recursive integration of OMV like virus sequences in fireflies and horizontal transmission between hosts. These propositions are not mutually exclusive, and may be indistinctly applied to specific cases. Future studies should enquire in this genome dark matter to better understand this interesting phenomenon. When more data is available EVE sequences may be combined with phylogenetic data of host species to expose eventual patterns of inter-class virus transmission. Either way, more studies are needed to explore these proposals, Katzourakis and Gifford (Katzourakis and Gifford, 2010) suggested that EVEs could reveal novel virus diversity and indicate the likely host range of virus clades.

After identification and confirmation that firefly related EVEs are present in the host DNA genome, an obvious question follows: Are these EVEs just signatures of an evolutionary vestige of stochastic past infections; or could they be associated with an intrinsic function? It has been suggested that intensity and prevalence of infection may be a determinant of EVEs integration, and that exposure to environmental viruses may not (Olson and Bonizzoni, 2017). Previous reports have suggested that EVEs may firstly function as restriction factors in their hosts by conferring resistance to infection by exogenous viruses, and the eventual counter-adaptation of virus populations of EVE positive hosts, could reduce the EVE restriction mechanism to a non-functional status (Aiewsakun and Katzourakis, 2015). Recently, in mosquitoes, a new mechanism of antiviral immunity against RNA viruses has been proposed, relying in the production and expression of EVEs DNA (Goic et al., 2016). Alternatively, eventual EVE expression could lend to the production viral like truncated proteins that may compete in trans with virus proteins from infecting viruses and limit viral replication, transcription or virion assembly (Aaskov et al., 2006). In addition, integration and eventual modulation in the host genome may be associated with an interaction between viral RNA and the mosquito RNAi machinery (Goic et al., 2013). The piRNA pathway mediates through small RNAs and Piwi-Argonaute proteins the repression of TE-derived nucleic acids based on sequence complementarity, and has also been associated to regulation of arbovirus viral-related RNA, suggesting a functional connection among resistance mechanisms against RNA viruses and TEs(Miesen et al., 2016; Palatini et al., 2017). Furthermore, arbovirus EVEs have been linked to the production of viral-derived piRNAs and virus-specific siRNA, inducing host cell immunity without limiting viral replication, supporting persistent and chronic infection (Goic et al., 2016). Perhaps, an EVE-dependent mechanism of modulation of virus infection could have some level of reminiscence to the paradigmatic CRISPR/Cas system which mediates bacteriophage resistance in prokaryotic hosts.

In sum, genomic studies are a great resource for the understanding of virus and host evolution. Here, we glimpsed an unexpected hidden evolutionary tale of firefly viruses and related FEVEs. Animal genomes appear to reflect as a book, with many dispersed sentences, an antique history of ancestral interaction with microbes, and EVEs functioning as virus related bookmarks. The exponential growth of genomic data would help to further understand this complex and intriguing interface, in order to advance not only in the apprehension of the phylogenomic insights of the host, but also explore a multifaceted and dynamic virome that has accompanied and even might have shifted the evolution of the host.

Supporting Information 5—table 5. FEVE hits from BLASTX of PpyrOMLV PB1.

Scaffold	Start	End	Strand	Id with PpOMLV	E value	Coverage	FEVE
Ppyr1.2_LG1	12787323	12786796	(-)	56.30%	8.22E-50	39.10%	EVE PB1 like-1
Ppyr1.2_LG1	13016647	13016120	(-)	56.30%	8.22E-50	39.10%	EVE PB1 like-2
Ppyr1.2_LG1	34701480	34701560	(+)	37.00%	2.88E-26	26.70%	EVE PB1 like-3
Ppyr1.2_LG1	34701562	34701774	(+)	37.60%	2.88E-26	30.20%	EVE PB1 like-3

Ppyr1.2_LG1	34701801	34702214	(+)	45.30%	2.88E-26	34.00%	EVE PB1 like-3
Ppyr1.2_LG1	35094645	35095094	(+)	28.10%	2.15E-10	9.50%	EVE PB1 like-4
Ppyr1.2_LG1	35110084	35109956	(-)	53.50%	2.37E-14	4.40%	EVE PB1 like-5
Ppyr1.2_LG1	35110214	35110107	(-)	75.00%	2.37E-14	14.70%	EVE PB1 like-5
Ppyr1.2_LG1	35110347	35110213	(-)	42.60%	2.37E-14	2.90%	EVE PB1 like-5
Ppyr1.2_LG1	50031464	50031330	(-)	64.40%	1.18E-09	10.00%	EVE PB1 like-6
Ppyr1.2_LG1	50031498	50031457	(-)	71.40%	1.18E-09	11.60%	EVE PB1 like-6
Ppyr1.2_LG1	50613130	50612921	(+)	49.40%	3.71E-11	4.90%	EVE PB1 like-7
Ppyr1.2_LG1	50673211	50673621	(+)	38.50%	1.03E-12	9.70%	EVE PB1 like-8
Ppyr1.2_LG1	51208464	51207634	(-)	77.20%	0	56.40%	EVE PB1 like-9
Ppyr1.2_LG1	51209399	51208467	(-)	68.50%	0	53.60%	EVE PB1 like-9
Ppyr1.2_LG1	51209556	51209398	(-)	71.70%	0	39.20%	EVE PB1 like-9
Ppyr1.2_LG1	61871682	61872158	(+)	31.10%	2.84E-23	36.00%	EVE PB1 like-10
Ppyr1.2_LG1	61872158	61872319	(+)	46.30%	2.84E-23	28.30%	EVE PB1 like-10
Ppyr1.2_LG1	61872355	61872456	(+)	41.20%	2.84E-23	27.00%	EVE PB1 like-10
Ppyr1.2_LG1	61930528	61930205	(-)	38.00%	3.58E-27	30.90%	EVE PB1 like-11
Ppyr1.2_LG1	61930686	61930504	(-)	63.60%	3.58E-27	35.90%	EVE PB1 like-11
Ppyr1.2_LG1	68038999	68039073	(+)	60.00%	7.73E-12	6.60%	EVE PB1 like-12
Ppyr1.2_LG1	68039072	68039314	(+)	40.70%	7.73E-12	5.00%	EVE PB1 like-12
Ppyr1.2_LG1	68039289	68039330	(+)	64.30%	7.73E-12	8.00%	EVE PB1 like-12
Ppyr1.2_LG1	68128820	68129008	(+)	51.50%	1.89E-06	4.90%	EVE PB1 like-13
Ppyr1.2_LG2	34545814	34545680	(-)	58.70%	3.84E-06	7.20%	EVE PB1 like-14
Ppyr1.2_LG2	34546169	34545801	(-)	52.80%	1.16E-31	34.10%	EVE PB1 like-14

Supporting Information 5—table 6. FEVE hits from BLASTX of PpyrOMLV PB2.

Scaffold	Start	End	Strand	Id with PpOMLV	E value	Coverage	FEVE
Ppyr1.2_LG1	50313869	50314219	(+)	82.10%	6.91E-54	48.30%	EVE PB2 like-1
Ppyr1.2_LG1	50314216	50315016	(+)	82.40%	1.92E-142	57.90%	EVE PB2 like-1
Ppyr1.2_LG1	50315772	50315002	(-)	89.10%	9.97E-145	60.60%	EVE PB2 like-1
Ppyr1.2_LG1	58707403	58706942	(-)	52.60%	6.19E-42	35.80%	EVE PB2 like-2

Supporting Information 5—table 7. FEVE hits from BLASTX of PpyrOMLV PA.

Scaffold	Start	End	Strand	Id with PpOMLV	E value	Coverage	FEVE
Ppyr1.2_LG1	34977392	34977231	(-)	48.10%	7.73E-07	3.50%	EVE PA like-1
Ppyr1.2_LG1	62052289	62052023	(-)	28.70%	8.92E-11	7.10%	EVE PA like-2
Ppyr1.2_LG1	62117077	62116811	(-)	28.70%	1.22E-10	7.10%	EVE PA like-3
Ppyr1.2_LG1	62117493	62117101	(-)	26.30%	1.22E-10	8.60%	EVE PA like-3
Ppyr1.2_LG1	68122348	68122440	(+)	77.40%	3.40E-06	15.70%	EVE PA like-4

Supporting Information 5—table 8. FEVE hits from BLASTX of PpyrOMLV NP

Scaffold	Start	End	Strand	Id with PpOMLV	E value	Coverage	FEVE
Ppyr1.2_LG1	181303	181404	(+)	79.40%	7.01E-09	17.90%	EVE NP like-1
Ppyr1.2_LG1	1029425	1029568	(+)	93.80%	9.59E-21	27.40%	EVE NP like-2
Ppyr1.2_LG1	2027860	2027438	(-)	35.50%	3.00E-21	30.80%	EVE NP like-3
Ppyr1.2_LG1	36568324	36568551	(+)	42.10%	8.99E-11	7.20%	EVE NP like-4
Ppyr1.2_LG1	52877256	52877086	(-)	68.40%	3.87E-15	14.60%	EVE NP like-5
Ppyr1.2_LG1	59927414	59927271	(+)	93.80%	5.60E-20	26.40%	EVE NP like-6
Ppyr1.2_LG3	17204346	17204122	(-)	46.70%	7.60E-13	7.10%	EVE NP like-7
Ppyr1.2_LG3	31635344	31635030	(-)	35.80%	3.30E-08	10.00%	EVE NP like-8
Ppyr1.2_LG3	50175821	50175922	(+)	79.40%	7.01E-09	17.90%	EVE NP like-9
Ppyr1.2_LG4	27811681	27811758	(+)	38.50%	3.22E-13	2.50%	EVE NP like-10
Ppyr1.2_LG4	27811853	27812179	(+)	39.00%	3.22E-13	10.90%	EVE NP like-10

Supporting Information 6

Data availability

6.1 Files on FigShare

1. *Photinus pyralis* sighting records (Excel spreadsheet) - (10.6084/m9.figshare.5688826)
2. Ilumi1.0 Blobtools results - (10.6084/m9.figshare.5688952)
3. Alat1.2 Blobtools results - (10.6084/m9.figshare.5688928)
4. Ppyr1.2 Blobtools results - (10.6084/m9.figshare.5688982)
5. Protein multiple sequence alignment for P450 tree - Supporting Information 1—figure 13 - (10.6084/m9.figshare.5697643)
6. Photinus pyralis orthomyxo-like virus 1 sequence and annotation - (10.6084/m9.figshare.5714806)
7. Photinus pyralis orthomyxo-like virus 2 sequence and annotation - (10.6084/m9.figshare.5714812)
8. OrthoFinder protein clustering analysis (Orthogroups) - (10.6084/m9.figshare.5715136)
9. PPYR_OGS1.1 kallisto RNA-Seq expression quantification (TPM) - (10.6084/m9.figshare.5715139)
10. AQULA_OGS1.0 kallisto RNA-Seq expression quantification (TPM) - (10.6084/m9.figshare.5715142)
11. Figure 5. PPYR_OGS1.1+AQULA_OGS1.0 Sleuth/differential expression Venn diagram analysis (BSN-TPM) - (10.6084/m9.figshare.5715151)
12. Ilumi_OGS1.2 kallisto RNA-Seq expression quantification (TPM) - (10.6084/m9.figshare.5715157)
13. Supporting Information 4—figure 2: DNA and tRNA methyltransferase gene phylogeny - (10.6084/m9.figshare.6531311)
14. Supporting Information 4—figure 6 Preliminary maximum likelihood phylogeny of luciferase homologs - (10.6084/m9.figshare.6687086)
15. Supporting Information 4—figure 9A Opsin gene tree - (10.6084/m9.figshare.5723005)
16. Testing for ancestral selection of elaterid ancestral luciferase (Figure 4B): MEME selected site analysis - (10.6084/m9.figshare.6626651)
17. Testing for ancestral selection of elaterid ancestral luciferase (Figure 4B): PAML-BEB selected site analysis - (10.6084/m9.figshare.6725081)

6.2 Files on www.fireflybase.org/www.github.org

6.2.1 Photinus pyralis genome and associated files

- Ppyr1.3 genome assembly - (http://www.fireflybase.org/firefly_data/Ppyr1.3.fasta.zip)
- *P. pyralis* Official Geneset (OGS) GFF3 files - (https://github.com/photocyte/ PPYR_OGS; copy archived at https://github.com/elifesciences-publications/ PPYR_OGS)
    - Official geneset gene-span nucleotide FASTA files
    - Official geneset mRNA nucleotide FASTA files
    - Official geneset CDS nucleotide FASTA files
    - Official geneset peptide FASTA files
- Supporting Non-OGS files - (https://github.com/photocyte/PPYR_OGS/tree/master/Supporting_non-OGS_data)
    - Trinity/PASA direct coding gene models (DCGM) GFF3 file
        - DCGM CDS FASTA file

    - DCGM peptide FASTA file
  - Stringtie stranded direct coding gene model (DCGM) GFF3 file
    - DCGM CDS FASTA file
    - DCGM peptide FASTA file
  - Stringtie unstranded direct coding gene model (DCGM) GFF3 file
    - DCGM CDS FASTA file
    - DCGM peptide FASTA file
  - Expression quantification (TPM)
  - InterProScan OGS functional annotation
  - PTS1 OGS annotation
  - Gaps GFF3 file
  - Repeat library FASTA and aligned GFF3 file.
  - Ab-initio gene models

6.2.2 Aquatica lateralis genome and associated files

- Alat1.3 genome assembly - (http://www.fireflybase.org/firefly_data/Alat1.3.fasta.zip)
- *A. lateralis* Official Geneset (OGS) GFF3 files - (https://github.com/photocyte/ AQULA_OGS; copy archived at https://github.com/elifesciences-publications/ AQULA_OGS)
  - Official geneset gene-span nucleotide FASTA files
  - Official geneset mRNA nucleotide FASTA files
  - Official geneset CDS nucleotide FASTA files
  - Official geneset peptide FASTA files
- Supporting Non-OGS files - (https://github.com/photocyte/AQULA_OGS/tree/master/Supporting_non-OGS_data)
  - Trinity/PASA direct coding gene models (DCGM) GFF3 file
    - DCGM CDS FASTA file
    - DCGM peptide FASTA file
  - Stringtie unstranded direct coding gene model (DCGM) GFF3 file
    - DCGM CDS FASTA file
    - DCGM peptide FASTA file
  - Expression quantification (TPM)
  - InterProScan OGS functional annotation
  - PTS1 OGS annotation
  - Gaps GFF3 file
  - Repeat library FASTA and aligned GFF3 file.

6.2.3 Ignelater luminosus genome and associated files

- Ilumi1.2 genome assembly - (http://www.fireflybase.org/firefly_data/Ilumi1.2.fasta.zip)
- *I. luminosus* Official Geneset (OGS) GFF3 files - (https://github.com/photocyte/ ILUMI_OGS; copy archived at https://github.com/elifesciences-publications/ ILUMI_OGS)
  - Official geneset gene-span nucleotide FASTA files
  - Official geneset mRNA nucleotide FASTA files
  - Official geneset CDS nucleotide FASTA files
  - Official geneset peptide FASTA files
- Supporting Non-OGS files - (https://github.com/photocyte/ILUMI_OGS/tree/master/Supporting_non-OGS_data)
  - Trinity/PASA direct coding gene models (DCGM) GFF3 file

    - DCGM CDS FASTA file
    - DCGM peptide FASTA file
- Stringtie unstranded direct coding gene model (DCGM) GFF3 file
    - DCGM CDS FASTA file
    - DCGM peptide FASTA file
- Expression quantification (TPM)
- o InterProScan OGS functional annotation
- o PTS1 OGS annotation
- Gaps GFF3 file
- Repeat library FASTA and aligned GFF3 file.
- Ab-initio gene models

6.3 Tracks on www.fireflybase.org JBrowse (Skinner et al., 2009) **genome browser**

For each genome:

1. Gaps
2. Repeats
3. Direct gene-models (Stringtie)
4. Direct gene-models (Trinity)
5. Official geneset gene-models

REFERENCES

Aaskov J, Buzacott K, Thu HM, Lowry K, Holmes EC. 2006. Long-term transmission of defective RNA viruses in humans and Aedes mosquitoes. *Science* **311**:236–238.

Aiewsakun P, Katzourakis A. 2015. Endogenous viruses: Connecting recent and ancient viral evolution. *Virology* **479-480**:26–37.

Al-Wathiqui N, Fallon TR, South A, Weng J-K, Lewis SM. 2016. Molecular characterization of firefly nuptial gifts: a multi-omics approach sheds light on postcopulatory sexual selection. *Sci Rep* **6**:38556.

Amaral DT, Silva JR, Viviani VR. 2017. Transcriptional comparison of the photogenic and non-photogenic tissues of Phrixothrix hirtus (Coleoptera: Phengodidae) and non-luminescent Chauliognathus flavipes (Coleoptera: Cantharidae) give insights on the origin of lanterns in railroad worms. *Gene Reports* **7**:78–86.

Anderson CR, Casals J. 1973. Dhori virus, a new agent isolated from Hyalomma dromedarii in India. *Indian J Med Res* **61**:1416–1420.

Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. 2016. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics* **32**:3380–3387.

Arias-Bohart ET. 2015. Malalcahuelloocaresi gen. & sp. n. (Elateridae, Campyloxeninae). *Zookeys* 1–13.

Arikawa K, Aoki K. 1982. Response characteristics and occurrence of extraocular photoreceptors on lepidopteran genitalia. *J Comp Physiol* **148**:483–489.

Arnoldi FGC, da Silva Neto AJ, Viviani VR. 2010. Molecular insights on the evolution of the lateral and head lantern luciferases and bioluminescence colors in Mastinocerini railroad-worms (Coleoptera: Phengodidae). *Photochem Photobiol Sci* **9**:87–92.

Arnoldi F, Ogoh K, Ohmiya Y, Viviani VR. 2007. Mitochondrial genome sequence of the Brazilian luminescent click beetle Pyrophorus divergens (Coleoptera: Elateridae): mitochondrial genes utility to investigate the evolutionary history of Coleoptera and its bioluminescence. *Gene* **405**:1–9.

Bae JS, Kim I, Sohn HD, Jin BR. 2004. The mitochondrial genome of the firefly, Pyrocoelia rufa:

complete DNA sequence, genome organization, and phylogenetic analysis with other insects. *Mol Phylogenet Evol* **32**:978–985.

Ballinger MJ, Bruenn JA, Hay J, Czechowski D, Taylor DJ. 2014. Discovery and evolution of bunyavirids in arctic phantom midges and ancient bunyavirid-like sequences in insect genomes. *J Virol* **88**:8783–8794.

Bao Z, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12**:1269–1276.

Bechara EJH, Stevani CV. 2018. Brazilian Bioluminescent Beetles: Reflections on Catching Glimpses of Light in the Atlantic Forest and Cerrado. *An Acad Bras Cienc* **90**:663–679.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**:573–580.

Bessho-Uehara M, Konishi K, Oba Y. 2017. Biochemical characteristics and gene expression profiles of two paralogous luciferases from the Japanese firefly Pyrocoelia atripennis (Coleoptera, Lampyridae, Lampyrinae): insight into the evolution of firefly luciferase genes. *Photochem Photobiol Sci* **16**:1301–1310.

Bessho-Uehara M, Oba Y. 2017. Identification and characterization of the Luc2-type luciferase in the Japanese firefly, Luciola parvula, involved in a dim luminescence in immobile stages. *Luminescence* **32**:924–931.

Bewick AJ, Vogel KJ, Moore AJ, Schmitz RJ. 2017. Evolution of DNA Methylation across Insects. *Mol Biol Evol* **34**:654–665.

Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Gallo Cassarino T, Bertoni M, Bordoli L, Schwede T. 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res* **42**:W252–8.

Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, Burton JN, Huson HJ, Nystrom JC, Kelley CM, Hutchison JL, Zhou Y, Sun J, Crisà A, Ponce de León FA, Schwartz JC, Hammond JA, Waldbieser GC, Schroeder SG, Liu GE, Dunham MJ, Shendure J, Sonstegard TS, Phillippy AM, Van Tassell CP, Smith TPL. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet* **49**:643–650.

Bitler B, McElroy WD. 1957. The preparation and properties of crystalline firefly luciferin. *Arch Biochem Biophys* **72**:358–368.

Blum MS, Sannasi A. 1974. Reflex bleeding in the lampyrid Photinus pyralis: Defensive function. *J Insect Physiol* **20**:451–460.

Bocak L, Motyka M, Bocek M, Bocakova M. 2018. Incomplete sclerotization and phylogeny: The phylogenetic classification of Plastocerus (Coleoptera: Elateroidea). *PLoS One* **13**:e0194026.

Bocakova M, Bocak L, Hunt T, Teraväinen M, Vogler AP. 2007. Molecular phylogenetics of Elateriformia (Coleoptera): evolution of bioluminescence and neoteny. *Cladistics* **23**:477–496.

Böcker S, Letzel MC, Lipták Z, Pervukhin A. 2009. SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics* **25**:218–224.

Branchini BR, Southworth TL, Salituro LJ, Fontaine DM, Oba Y. 2017. Cloning of the Blue Ghost (Phausis reticulata) Luciferase Reveals a Glowing Source of Green Light. *Photochem Photobiol* **93**:473–478.

Branham MA, Wenzel JW. 2003. The origin of photic behavior and the evolution of sexual communication in fireflies (Coleoptera: Lampyridae). *Cladistics* **19**:1–22.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**:525–527.

Briscoe AD, Chittka L. 2001. The evolution of color vision in insects. *Annu Rev Entomol* **46**:471–510.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat*

Methods **12**:59–60.

Buck J, Case J. 2002. Physiological Links in Firefly Flash Code Evolution. *J Insect Behav* **15**:51–68.

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**:1119–1125.

Buschman LL. 1988. Larval Development and Its Photoperiodic Control in the Firefly Pyractomena lucifera (Coleoptera: Lampyridae). *Ann Entomol Soc Am* **81**:82–90.

Bushman F, Lewinski M, Ciuffi A, Barr S, Leipzig J, Hannenhalli S, Hoffmann C. 2005. Genome-wide analysis of retroviral DNA integration. *Nat Rev Microbiol* **3**:848–858.

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. 2008. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res* **18**:810–820.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**:1972–1973.

Case JF. 2004. Flight studies on photic communication by the firefly Photinus pyralis. *Integr Comp Biol* **44**:250–258.

Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak M-Y, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P. 2012. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **30**:918–920.

Chang H, Kirejtshuk AG, Ren D, Shih C. 2009. First Fossil Click Beetles from the Middle Jurassic of Inner Mongolia, China (Coleoptera: Elateridae). *Annal Zool* **59**:7–14.

Charles Darwin. 1872. The Origin of Species, 6th ed. PF Collier & Son, New York.

Costa C. 1984. Note on the bioluminescence of Balgus schnusei (Heller, 1974)(Trixagidae, Coleoptera). *Rev Bras Entomol.*

Costa C. 1975. Systematics and evolution of the tribes Pyrophorini and Heligmini, with description of Campyloxeninae, new subfamily (Coleoptera, Elateridae). *Arq Zool* **26**:49.

Costa, C., Lawrence, J. F. & Rosa, S. P. 2010. Elateridae Leach, 1815 In: Leschen, R. A. B., Beutel, R. G. & Lawrence, J. F., editor. Handbook of Zoology, Vol. IV, Arthropoda: Insecta, Teilband 39, Coleoptera, Beetles. Vol. 2: Morphology and Systematics. Walter de Gruyter, Berlin. pp. 75–103.

Costa C, Vanin SA. 2010. Coleoptera Larval Fauna Associated with Termite Nests (Isoptera) with Emphasis on the "Bioluminescent Termite Nests" from Central Brazil. *Psyche* **2010**. doi:10.1155/2010/723947

Cratsley CK, Rooney JA, Lewis SM. 2003. Limits to Nuptial Gift Production by Male Fireflies, Photinus ignitus. *J Insect Behav* **16**:361–370.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14**:1188–1190.

Cunningham CB, Ji L, Wiberg RAW, Shelton J, McKinney EC, Parker DJ, Meagher RB, Benowitz KM, Roy-Zokan EM, Ritchie MG, Brown SJ, Schmitz RJ, Moore AJ. 2015. The Genome and Methylome of a Beetle with Complex Social Behavior, Nicrophorus vespilloides (Coleoptera: Silphidae). *Genome Biol Evol* **7**:3383–3396.

Darzentas N. 2010. Circoletto: visualizing sequence similarity with Circos. *Bioinformatics* **26**:2620–2621.

Day JC. 2013. The role of gene duplication in the evolution of beetle bioluminescence. *Trends Entomol* **9**:55–63.

De Cock R, Matthysen E. 1999. Aposematism and Bioluminescence: Experimental evidence from

Glow-worm Larvae(Coleoptera: Lampyridae). *Evol Ecol* **13**:619–639.

Dettner K. 1987. Chemosystematics and Evolution of Beetle Chemical Defenses. *Annu Rev Entomol* **32**:17–48.

De Wet JR, Wood KV, DeLuca M, Helinski DR, Subramani S. 1987. Firefly luciferase gene: structure and expression in mammalian cells. *Mol Cell Biol* **7**:725–737.

de Wet JR, Wood KV, Helinski DR, DeLuca M. 1985. Cloning of firefly luciferase cDNA and the expression of active luciferase in Escherichia coli. *Proc Natl Acad Sci U S A* **82**:7870–7873.

Dias CM, Schneider MC, Rosa SP, Costa C, Cella DM. 2007. The first cytogenetic report of fireflies (Coleoptera, Lampyridae) from Brazilian fauna: The first cytogenetic report of Brazilian fireflies. *Acta Zool* **88**:309–316.

Dinkel H, Michael S, Weatheritt RJ, Davey NE, Van Roey K, Altenberg B, Toedt G, Uyar B, Seiler M, Budd A, Jödicke L, Dammert MA, Schroeter C, Hammer M, Schmidt T, Jehl P, McGuigan C, Dymecka M, Chica C, Luck K, Via A, Chatr-aryamontri A, Haslam N, Grebnev G, Edwards RJ, Steinmetz MO, Meiselbach H, Diella F, Gibson TJ. 2012. ELM—the database of eukaryotic linear motifs. *Nucleic Acids Res* **40**:D242–D251.

Dolin VG. 1978. Phylogeny of the click beetles (Coleoptera, Elateridae). *Vestn Zool* **May/June 1978, 3**.

Douglas H. 2011. Phylogenetic relationships of Elateridae inferred from adult morphology, with special reference to the position of Cardiophorinae. *Zootaxa* **2900**:1–45.

Dubois R. 1886. Les Élatérides lumineux: contribution a l'étude de la production de la lumière par les êtres vivants. la Société zoologique de France.

Dubois R. 1885. Fonction photogénique des Pyrophores. *CR Seances Soc Biol Fil* **37**:559–562.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. 2016a. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**:99–101.

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016b. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**:95–98.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**:1792–1797.

Eisfeld AJ, Neumann G, Kawaoka Y. 2015. At the centre: influenza A virus ribonucleoproteins. *Nat Rev Microbiol* **13**:28–41.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**:157.

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**:e47768.

Fallon TR, Li F-S, Vicent MA, Weng J-K. 2016. Sulfoluciferin is Biosynthesized by a Specialized Luciferin Sulfotransferase in Fireflies. *Biochemistry* **55**:3341–3344.

Faust L, De Cock R, Lewis S. 2012. Thieves in the Night: Kleptoparasitism by Fireflies in the Genus Photuris Dejean (Coleoptera: Lampyridae). *Coleopt Bull* **66**:1–6.

Faust LF. 2017. Fireflies, Glow-worms, and Lightning Bugs: Identification and Natural History of the Fireflies of the Eastern and Central United States and Canada. University of Georgia Press.

Faust L, Faust H. 2014. The Occurrence and Behaviors of North American Fireflies (Coleoptera: Lampyridae) on Milkweed, Asclepias syriaca L. *Coleopt Bull* **68**:283–291.

Feder JL, Velez S. 2009. Intergenic exchange, geographic isolation, and the evolution of bioluminescent color for Pyrophorus click beetles. *Evolution* **63**:1203–1216.

Feschotte C, Gilbert C. 2012. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet* **13**:283–296.

Feschotte C, Pritham EJ. 2007. DNA Transposons and the Evolution of Eukaryotic Genomes. *Annu Rev Genet* **41**:331–368.

Feuda R, Marlétaz F, Bentley MA, Holland PWH. 2016. Conservation, Duplication, and Divergence of Five Opsin Genes in Insect Evolution. *Genome Biol Evol* **8**:579–587.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**:D279–85.

Fraga H. 2008. Firefly luminescence: a historical perspective and recent developments. *Photochem Photobiol Sci* **7**:146–158.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**:3150–3152.

Fu XH, Ballantyne LA, Lambkin CL. 2010. Aquatica gen. nov. from mainland China with a description of Aquatica wuhana sp. nov.(Coleoptera: Lampyridae: Luciolinae). *Zootaxa* **2530**:1–18.

Fu X, Li J, Tian Y, Quan W, Zhang S, Liu Q, Liang F, Zhu X, Zhang L, Wang D, Hu J. 2017. Long-read sequence assembly of the firefly Pyrocoelia pectoralis genome. *Gigascience*. doi:10.1093/gigascience/gix112

Fu X, Vencl FV, Nobuyoshi O, Benno Meyer-Rochow V, Lei C, Zhang Z. 2007. Structure and function of the eversible glands of the aquatic firefly Luciola leii (Coleoptera: Lampyridae). *Chemoecology* **17**:117–124.

Georg Neuberger, Sebastian Maurer-Stroh, Birgit Eisenhaber, Andreas Hartig and Frank Eisenhaber. n.d. The PTS1 Predictor. http://mendel.imp.ac.at/pts1/

Ghiradella H, Schmidt JT. 2004. Fireflies at one hundred plus: a new look at flash control. *Integr Comp Biol* **44**:203–212.

Gilbert C, Cordaux R. 2017. Viruses as vectors of horizontal transfer of genetic material in eukaryotes. *Curr Opin Virol* **25**:16–22.

Glastad KM, Arsenault SV, Vertacnik KL, Geib SM, Kay S, Danforth BN, Rehan SM, Linnen CR, Kocher SD, Hunt BG. 2017. Variation in DNA Methylation Is Not Consistently Reflected by Sociality in Hymenoptera. *Genome Biol Evol* **9**:1687–1698.

Glastad KM, Gokhale K, Liebig J, Goodisman MAD. 2016. The caste- and sex-specific DNA methylome of the termite Zootermopsis nevadensis. *Sci Rep* **6**:37110.

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* **108**:1513–1518.

Goetz MA, Meinwald J, Eisner T. 1981. Lucibufagins, IV. New defensive steroids and a pterin from the firefly,Photinus pyralis (coleoptera: Lampyridae). *Experientia* **37**:679–680.

Goh K-S, Li C-W. 2011. A photocytes-associated fatty acid-binding protein from the light organ of adult Taiwanese firefly, Luciola cerata. *PLoS One* **6**:e29576.

Goic B, Stapleford KA, Frangeul L, Doucet AJ, Gausson V, Blanc H, Schemmel-Jofre N, Cristofari G, Lambrechts L, Vignuzzi M, Saleh M-C. 2016. Virus-derived DNA drives mosquito vector tolerance to arboviral infection. *Nat Commun* **7**:12410.

Goic B, Vodovar N, Mondotte JA, Monot C, Frangeul L, Blanc H, Gausson V, Vera-Otarola J, Cristofari G, Saleh M-C. 2013. RNA-mediated interference and reverse transcription control the persistence of RNA viruses in the insect model Drosophila. *Nat Immunol* **14**:396–403.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**:644–652.

Green JW. 1956. Revision of the nearctic species of Photinus (Coleoptera: Lampyridae). *Proc Calif Acad Sci* **28**:561–613.

Gremme G, Steinbiss S, Kurtz S. 2013. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans Comput Biol Bioinform* **10**:645–656.

Grimaldi D, Engel MS. 2005. Evolution of the Insects. Cambridge University Press.

Gronquist M, Meinwald J, Eisner T, Schroeder FC. 2005. Exploring uncharted terrain in nature's structure space using capillary NMR spectroscopy: 13 steroids from 50 fireflies. *J Am Chem Soc* **127**:10810–10811.

Guilligay D, Kadlec J, Crépin T, Lunardi T, Bouvier D, Kochs G, Ruigrok RWH, Cusack S. 2014. Comparative structural and functional analysis of orthomyxovirus polymerase cap-snatching domains. *PLoS One* **9**:e84973.

Guittard E, Blais C, Maria A, Parvy J-P, Pasricha S, Lumb C, Lafont R, Daborn PJ, Dauphin-Villemant C. 2011. CYP18A1, a key enzyme of Drosophila steroid hormone inactivation, is essential for metamorphosis. *Dev Biol* **349**:35–45.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**:1072–1075.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**:R7.

Hackett KJ, Whitcomb RF, Tully JG, Lloyd JE, Anderson JJ, Clark TB, Henegar RB, Roset DL, Clark EA, Vaughn JL. 1992. Lampyridae (Coleoptera): A plethora of mollicute associations. *Microb Ecol* **23**:181–193.

Haig DA, Woodall JP, Danskin D. 1965. THOGOTO VIRUS: A HITHERTO UNDERSCRIBED AGENT ISOLATED FROM TICKS IN KENYA. *J Gen Microbiol* **38**:389–394.

Hall RA, Bielefeldt-Ohmann H, McLean BJ, O'Brien CA, Colmant AMG, Piyasena TBH, Harrison JJ, Newton ND, Barnard RT, Prow NA, Deerain JM, Mah MGKY, Hobson-Peters J. 2016. Commensal Viruses of Mosquitoes: Host Restriction, Transmission, and Interaction with Arboviral Pathogens. *Evol Bioinform Online* **12**:35–44.

Hamberger Björn, Bak Søren. 2013. Plant P450s as versatile drivers for evolution of species-specific chemical diversity. *Philos Trans R Soc Lond B Biol Sci* **368**:20120426.

Han Y, Wessler SR. 2010. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res* **38**:e199.

Hara K, Kashiwagi T, Hamada N, Watanabe H. 2017. Basic amino acids in the N-terminal half of the PB2 subunit of influenza virus RNA polymerase are involved in both transcription and replication. *J Gen Virol* **98**:900–905.

Harvey EN. 1952. Bioluminescence. Academic Press.

Harvey EN, Stevens KP. 1928. THE BRIGHTNESS OF THE LIGHT OF THE WEST INDIAN ELATERID BEETLE, PYROPHORUS. *J Gen Physiol* **12**:269–272.

Hause BM, Ducatez M, Collin EA, Ran Z, Liu R, Sheng Z, Armien A, Kaplan B, Chakravarty S, Hoppe AD, Webby RJ, Simonson RR, Li F. 2013. Isolation of a novel swine influenza virus from Oklahoma in 2011 which is distantly related to human influenza C viruses. *PLoS Pathog* **9**:e1003176.

Heberle H, Meirelles GV, da Silva FR, Telles GP, Minghim R. 2015. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics* **16**:169.

Helvig C, Koener JF, Unnithan GC, Feyereisen R. 2004. CYP15A1, the cytochrome P450 that catalyzes epoxidation of methyl farnesoate to juvenile hormone III in cockroach corpora allata. *Proc Natl Acad Sci U S A* **101**:4024–4029.

Hengrung N, El Omari K, Serna Martin I, Vreede FT, Cusack S, Rambo RP, Vonrhein C, Bricogne G, Stuart DI, Grimes JM, Fodor E. 2015. Crystal structure of the RNA-dependent RNA polymerase from influenza C virus. *Nature* **527**:114–117.

Herb BR, Wolschin F, Hansen KD, Aryee MJ, Langmead B, Irizarry R, Amdam GV, Feinberg AP. 2012.

Reversible switching between epigenetic states in honeybee behavioral subcastes. *Nat Neurosci* **15**:1371–1373.

Hess WN. 1920. Notes on the biology of some common Lampyridae. *Biol Bull* **38**:39–76.

Higuchi H, editor. 1996. Conservation of EcosystemConservation Biology. Univ. Tokyo Press, Tokyo. pp. 71–102.

Ho J-Z, Chiang P-H, Wu C-H, Yang P-S. 2010. Life cycle of the aquatic firefly Luciola ficta (Coleoptera: Lampyridae). *J Asia Pac Entomol* **13**:189–196.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**:491.

Hsu MT, Parvin JD, Gupta S, Krystal M, Palese P. 1987. Genomic RNAs of influenza viruses are held in a circular conformation in virions and in infected cells by a terminal panhandle. *Proc Natl Acad Sci U S A* **84**:8140–8144.

Huson DH, Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst Biol* **61**:1061–1067.

Hyslop JA. 1917. The phylogeny of the Elateridae based on larval characters. *Ann Entomol Soc Am* **10**:241–263.

Ikeya H. 2016. Melanic strain of Luciola lateralis. *Bull Firefly Mus Toyota Town* **8**:175–177.

Imperial Palace Outer Garden Management Office. 2017. Efforts to protect heike fireflies living in the moat of the Imperial Palace Gardens. http://www.env.go.jp/garden/kokyogaien/topics/post_134.html

Inoue M, Yamamoto H. 1987. Cytological studies of family Lampyridae I. Karyotypes of Luciola lateralis and L. cruciata. *La Kromosomo* **II-45**:1440–1443.

Johnson PJ. 2002. 58. Elateridae Leach 1815. *American beetles* **2**:160–173.

Käll L, Krogh A, Sonnhammer ELL. 2004. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**:1027–1036.

Kanda S. 1935. Firefly.

Kanie S, Nakai R, Ojika M, Oba Y. 2018. 2-S-cysteinylhydroquinone is an intermediate for the firefly luciferin biosynthesis that occurs in the pupal stage of the Japanese firefly, Luciola lateralis. *Bioorg Chem*. doi:10.1016/j.bioorg.2018.06.028

Kanie S, Nishikawa T, Ojika M, Oba Y. 2016. One-pot non-enzymatic formation of firefly luciferin in a neutral buffer from p-benzoquinone and cysteine. *Sci Rep* **6**:24794.

Kapitonov VV, Jurka J. 2001. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci U S A* **98**:8714–8719.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**:772–780.

Katzourakis A, Gifford RJ. 2010. Endogenous viral elements in animal genomes. *PLoS Genet* **6**:e1001191.

Kawashima I, Suzuki H, Sato M. 2003. A check-list of Japanese fireflies (Coleoptera, Lampyridae and Rhagophthalmidae). *Jpn J syst Ent, Matsuyama* **9**:241–261.

Kazantsev SV. 2015.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**:656–664.

Khomtchouk BB, Hennessy JR, Wahlestedt C. 2017. shinyheatmap: Ultra fast low memory heatmap web interface for big data genomics. *PLoS One* **12**:e0176334.

Kimble JB. 2013. Zoonotic Transmission of Influenza H9 subtype through Reassortment. Ann Arbor, United States: University of Maryland, College Park.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**:357–360.

King AMQ, Lefkowitz E, Adams MJ, Carstens EB. 2011. Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier.

Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Vesztrocy AW, Naldi A, Mungall CJ, Yunes JM, Botvinnik O, Weigel M, Dampier W, Dessimoz C, Flick P, Tang H. 2018. GOATOOLS: A Python library for Gene Ontology analyses. *Scientific Reports*. doi:10.1038/s41598-018-28948-z

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**:722–736.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**:59.

Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, Maroon H, Thomas F, Aboobaker AA, Blaxter M. 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. *Proc Natl Acad Sci U S A* **113**:5053–5058.

Kretsch E. 2000. Courtship Behavior of Ignelater luminosus.

Krumsiek J, Arnold R, Rattei T. 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**:1026–1028.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**:1639–1645.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**:1870–1874.

Kundrata R, Bocak L. 2011. The phylogeny and limits of Elateridae (Insecta, Coleoptera): is there a common tendency of click beetles to soft-bodiedness and neoteny? *Zool Scr* **40**:364–378.

Kyrpides NC, Woyke T, Eisen JA, Garrity G, Lilburn TG, Beck BJ, Whitman WB, Hugenholtz P, Klenk H-P. 2014. Genomic Encyclopedia of Type Strains, Phase I: The one thousand microbial genomes (KMG-I) project. *Stand Genomic Sci* **9**:1278–1284.

Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies. *F1000Res* **6**. doi:10.12688/f1000research.12232.1

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**:357–359.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**:R25.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**:2947–2948.

Larracuente AM, Ferree PM. 2015. Simple method for fluorescence DNA in situ hybridization to squashed chromosomes. *J Vis Exp* 52288.

Leahy MB, Dessens JT, Weber F, Kochs G, Nuttall PA. 1997. The fourth genus in the Orthomyxoviridae: sequence analyses of two Thogoto virus polymerase proteins and comparison with influenza viruses. *Virus Res* **50**:215–224.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol* **25**:1307–1320.

Letunic I, Bork P. 2018. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res* **46**:D493–D496.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**:W242–5.

Levy HC. 1998. Greatest bioluminescence In: Walker TJ, editor. Book of Insect Records. Univ. Florida, Florida. pp. 72–73.

Lewis SM, Cratsley CK. 2008. Flash signal evolution, mate choice, and predation in fireflies. *Annu Rev Entomol* **53**:293–321.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**:289–293.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**:3094–3100.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**:1658–1659.

Lloyd J. 1996. Where can I find information on raising fireflies? *Fireflyer Companion* **1**:20.

Lloyd JE. 2008. Fireflies (Coleoptera: Lampyridae) In: Capinera JL, editor. Encyclopedia of Entomology. Dordrecht: Springer Netherlands. pp. 1429–1452.

Lloyd JE. 1973. Firefly Parasites and Predators. *Coleopt Bull* **27**:91–106.

Lloyd JE. 1966. Studies on the flash communication system in Photinus fireflies.

Lower SS, Johnston JS, Stanger-Hall KF, Hjelmen CE, Hanrahan SJ, Korunes K, Hall D. 2017. Genome Size in North American Fireflies: Substantial Variation Likely Driven by Neutral Processes. *Genome Biol Evol* **9**:1499–1512.

Luk SPL, Marshall SA, Branham MA. 2011. The fireflies of Ontario (Coleoptera: Lampyridae). *Can J Arthropod Identif* **16**:1–105.

Maddison WP, Maddison DR. 2017. Mesquite: a modular system for evolutionary analysis.

Maeda J, Kato D-I, Arima K, Ito Y, Toyoda A, Noguchi H. 2017. The complete mitochondrial genome sequence and phylogenetic analysis of Luciola lateralis, one of the most famous firefly in Japan (Coleoptera: Lampyridae). *Mitochondrial DNA Part B* **2**:546–547.

Mansfield KG. 2007. Viral tropism and the pathogenesis of influenza in the Mammalian host. *Am J Pathol* **171**:1089.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**:764–770.

Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Geer LY, Bryant SH. 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res* **45**:D200–D203.

Marshall SH, Ramírez R, Labra A, Carmona M, Muñoz C. 2014. Bona fide evidence for natural vertical transmission of infectious salmon anemia virus in freshwater brood stocks of farmed Atlantic salmon (Salmo salar) in Southern Chile. *J Virol* **88**:6012–6018.

Martin GJ, Branham MA, Whiting MF, Bybee SM. 2017. Total evidence phylogeny and the evolution of adult bioluminescence in fireflies (Coleoptera: Lampyridae). *Mol Phylogenet Evol* **107**:564–575.

Martin GJ, Lord NP, Branham MA, Bybee SM. 2015. Review of the firefly visual system (Coleoptera: Lampyridae) and evolution of the opsin genes underlying color vision. *Org Divers Evol* **15**:513–526.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**:10–12.

Masuda T, Tatsumi H, Nakano E. 1989. Cloning and sequence analysis of cDNA for luciferase of a Japanese firefly, Luciola cruciata. *Gene* **77**:265–270.

McKenna DD, Farrell BD. 2009. Beetles (Coleoptera). *The timetree of life* **278**:289.

McKenna DD, Scully ED, Pauchet Y, Hoover K, Kirsch R, Geib SM, Mitchell RF, Waterhouse RM, Ahn S-J, Arsala D, Benoit JB, Blackmon H, Bledsoe T, Bowsher JH, Busch A, Calla B, Chao H, Childers AK, Childers C, Clarke DJ, Cohen L, Demuth JP, Dinh H, Doddapaneni H, Dolan A, Duan JJ, Dugan S, Friedrich M, Glastad KM, Goodisman MAD, Haddad S, Han Y, Hughes DST, Ioannidis P, Johnston JS, Jones JW, Kuhn LA, Lance DR, Lee C-Y, Lee SL, Lin H, Lynch JA, Moczek AP, Murali SC, Muzny DM, Nelson DR, Palli SR, Panfilio KA, Pers D, Poelchau MF, Quan H, Qu J, Ray AM, Rinehart JP, Robertson HM, Roehrdanz R, Rosendale AJ, Shin S, Silva C, Torson AS, Jentzsch IMV, Werren JH, Worley KC, Yocum G, Zdobnov EM, Gibbs RA, Richards S. 2016. Genome of the Asian longhorned beetle (Anoplophora glabripennis), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface. *Genome Biol* **17**:227.

McKenna DD, Wild AL, Kanda K, Bellamy CL, Beutel RG, Caterino MS, Farnum CW, Hawks DC, Ivie MA, Jameson ML, Leschen RAB, Marvaldi AE, Mchugh JV, Newton AF, Robertson JA, Thayer MK, Whiting MF, Lawrence JF, Ślipiński A, Maddison DR, Farrell BD. 2015. The beetle tree of life reveals that Coleoptera survived end-Permian mass extinction to diversify during the Cretaceous terrestrial revolution. *Syst Entomol* **40**:835–880.

McLean M, Buck J, Hanson FE. 1972. Culture and Larval Behavior of Photurid Fireflies. *Am Midl Nat* **87**:133–145.

Meinwald J, Wiemer DF, Eisner T. 1979. Lucibufagins. 2. Esters of 12-oxo-2.beta.,5.beta.,11.alpha.-trihydroxybufalin, the major defensive steroids of the firefly Photinus pyralis (Coleoptera: Lampyridae). *J Am Chem Soc* **101**:3055–3060.

Metegnier G, Becking T, Chebbi MA, Giraud I, Moumen B, Schaack S, Cordaux R, Gilbert C. 2015. Comparative paleovirological analysis of crustaceans identifies multiple widespread viral groups. *Mob DNA* **6**:16.

Miesen P, Joosten J, van Rij RP. 2016. PIWIs Go Viral: Arbovirus-Derived piRNAs in Vector Mosquitoes. *PLoS Pathog* **12**:e1006017.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, Niehuis O, Petersen M, Izquierdo-Carrasco F, Wappler T, Rust J, Aberer AJ, Aspöck U, Aspöck H, Bartel D, Blanke A, Berger S, Böhm A, Buckley TR, Calcott B, Chen J, Friedrich F, Fukui M, Fujita M, Greve C, Grobe P, Gu S, Huang Y, Jermiin LS, Kawahara AY, Krogmann L, Kubiak M, Lanfear R, Letsch H, Li Y, Li Z, Li J, Lu H, Machida R, Mashimo Y, Kapli P, McKenna DD, Meng G, Nakagaki Y, Navarrete-Heredia JL, Ott M, Ou Y, Pass G, Podsiadlowski L, Pohl H, von Reumont BM, Schütte K, Sekiya K, Shimizu S, Slipinski A, Stamatakis A, Song W, Su X, Szucsich NU, Tan M, Tan X, Tang M, Tang J, Timelthaler G, Tomizuka S, Trautwein M, Tong X, Uchifune T, Walzl MG, Wiegmann BM, Wilbrandt J, Wipfler B, Wong TKF, Wu Q, Wu G, Xie Y, Yang S, Yang Q, Yeates DK, Yoshizawa K, Zhang Q, Zhang R, Zhang W, Zhang Y, Zhao J, Zhou C, Zhou L, Ziesmann T, Zou S, Li Y, Xu X, Zhang Y, Yang H, Wang J, Wang J, Kjer KM, Zhou X. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**:763–767.

Mjaaland S, Rimstad E, Falk K, Dannevig BH. 1997. Genomic characterization of the virus causing infectious salmon anemia in Atlantic salmon (Salmo salar L.): an orthomyxo-like virus in a teleost. *J Virol* **71**:7681–7686.

Mofford DM, Liebmann KL, Sankaran GS, Reddy GSKK, Reddy GR, Miller SC. 2017. Luciferase Activity of Insect Fatty Acyl-CoA Synthetases with Synthetic Luciferins. *ACS Chem Biol* **12**:2946–2951.

Nathanson JA, Kantham L, Hunnicutt EJ. 1989. Isolation and N-terminal amino acid sequence of an octopamine ligand binding protein. *FEBS Lett* **259**:117–120.

Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. Oxford University Press.

Newman DJ, Cragg GM. 2015. Endophytic and epiphytic microbes as "sources" of bioactive agents. *Front Chem* **3**:34.

Ni JD, Baik LS, Holmes TC, Montell C. 2017. A rhodopsin in the brain functions in circadian photoentrainment in Drosophila. *Nature* **545**:340–344.

Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, Prjibelsky A, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, McLean J, Lasken R, Clingenpeel SR, Woyke T, Tesler G, Alekseyev MA, Pevzner PA. 2013. Assembling Genomes and Mini-metagenomes from Highly Chimeric ReadsResearch in Computational Molecular Biology. Springer Berlin Heidelberg. pp. 158–170.

Oba Y. 2009. On the origin of beetle luminescence. *Bioluminescence in focus Res Signpost, India*

277–290.

Oba Y, Furuhashi M, Bessho M, Sagawa S, Ikeya H, Inouye S. 2013a. Bioluminescence of a firefly pupa: involvement of a luciferase isotype in the dim glow of pupae and eggs in the Japanese firefly, Luciola lateralis. *Photochem Photobiol Sci* **12**:854–863.

Oba Y, Hoffmann KH. 2014. Insect Bioluminescence in the Post-Molecular Biology Era. *Insect Molecular Biology and Ecology* 94–120.

Oba Y, Kumazaki M, Inouye S. 2010a. Characterization of luciferases and its paralogue in the Panamanian luminous click beetle Pyrophorus angustus: a click beetle luciferase lacks the fatty acyl-CoA synthetic activity. *Gene* **452**:1–6.

Oba Y, Mori N, Yoshida M, Inouye S. 2010b. Identification and characterization of a luciferase isotype in the Japanese firefly, Luciola cruciata, involving in the dim glow of firefly eggs. *Biochemistry* **49**:10788–10795.

Oba Y, Ojika M, Inouye S. 2003. Firefly luciferase is a bifunctional enzyme: ATP-dependent monooxygenase and a long chain fatty acyl-CoA synthetase. *FEBS Lett* **540**:251–254.

Oba Y, Sato M, Ohta Y, Inouye S. 2006. Identification of paralogous genes of firefly luciferase in the Japanese firefly, Luciola cruciata. *Gene* **368**:53–60.

Oba Y, Yoshida M, Shintani T, Furuhashi M, Inouye S. 2012. Firefly luciferase genes from the subfamilies Psilocladinae and Ototretinae (Lampyridae, Coleoptera). *Comp Biochem Physiol B Biochem Mol Biol* **161**:110–116.

Oba Y, Yoshida N, Kanie S, Ojika M, Inouye S. 2013b. Biosynthesis of firefly luciferin in adult lantern: decarboxylation of L-cysteine is a key step for benzothiazole ring formation in firefly luciferin synthesis. *PLoS One* **8**:e84023.

O'Connell J, Schulz-Trieglaff O, Carlson E, Hims MM, Gormley NA, Cox AJ. 2015. NxTrim: optimized trimming of Illumina mate pair reads. *Bioinformatics* **31**:2035–2037.

Ohba N. 2004. Mystery of Fireflies. Yokosuka City Mus. Yokosuka, Japan (In Japanese).

Ohba N. 1983. Studies on the communication system of Japanese fireflies. *Sci Rept Yokosuka City Mus* **30**:1–62.

Ohba N, Hidaka T. 2002. Reflex bleeding of fireflies and prey-predator relationship. *Science Report of the Yokosuka City Museum* **49**:1–12.

Ôhira H. 2013. Illustrated key to click beetles of Japan. *Jpn Soc Environ Entomol Zool, editor An Illustrated Guide to Identify Insects Osaka, Japan: Bunkyo Shuppan* 227–251.

Ôhira H. 1962. Morphological and taxonomic study on the larvae of Elateridae in Japan (Coleoptera). *H Ohira, Okazaki City, Japan* **61**.

Ohmiya Y, Sumiya M, Viviani VR, Ohba N. 2000. Comparative aspects of a luciferase molecule from the Japanese luminous beetle, Rhagophthalmus ohbai. *Sci Rep Yokosuka City Mus* **47**:31–38.

Okada K, Iio H, Kubota I, Goto T. 1974. Firefly bioluminescence III. Conversion of oxyluciferin to luciferin in firefly. *Tetrahedron Lett* **15**:2771–2774.

Okonechnikov K, Conesa A, García-Alcalde F. 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**:292–294.

Olson DM, Dinerstein E, Wikramanayake ED, Burgess ND, Powell GVN, Underwood EC, D'amico JA, Itoua I, Strand HE, Morrison JC, Others. 2001. Terrestrial Ecoregions of the World: A New Map of Life on Earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *Bioscience* **51**:933–938.

Olson KE, Bonizzoni M. 2017. Nonretroviral integrated RNA viruses in arthropod vectors: an occasional event or something more? *Curr Opin Insect Sci* **22**:45–53.

Ow DW, DE Wet JR, Helinski DR, Howell SH, Wood KV, Deluca M. 1986. Transient and stable expression of the firefly luciferase gene in plant cells and transgenic plants. *Science* **234**:856–859.

Palatini U, Miesen P, Carballar-Lejarazu R, Ometto L, Rizzo E, Tu Z, van Rij RP, Bonizzoni M. 2017.

Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors Aedes aegypti and Aedes albopictus. *BMC Genomics* **18**:512.

Pankey MS, Minin VN, Imholte GC, Suchard MA, Oakley TH. 2014. Predictable transcriptome evolution in the convergent and complex bioluminescent organs of squid. *Proc Natl Acad Sci U S A* **111**:E4736–42.

Pauly MD, Procario M, Lauring AS. 2017. The mutation rates and mutational bias of influenza A virus. *bioRxiv*. doi:10.1101/110197

Perez-Gelabert DE. 2008. Arthropods of Hispaniola (Dominican Republic and Haiti): A checklist and bibliography. Magnolia Press.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**:290–295.

Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8**:785–786.

Pettersen EF, Goddard TD, Huang CC. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of*.

Pflug A, Lukarska M, Resa-Infante P, Reich S, Cusack S. n.d. Structural insights into RNA synthesis by the influenza virus transcription-replication machine. 2017. *Virus Res* 30782–30781.

Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. 2017. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods* **14**:687–690.

Pluskal T, Castillo S, Villar-Briones A, Oresic M. 2010. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**:395.

Poelchau M, Childers C, Moore G, Tsavatapalli V, Evans J, Lee C-Y, Lin H, Lin J-W, Hackett K. 2015. The i5k Workspace@NAL--enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic Acids Res* **43**:D714–9.

Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**:676–679.

Porter Megan L., Blasic Joseph R., Bok Michael J., Cameron Evan G., Pringle Thomas, Cronin Thomas W., Robinson Phyllis R. 2012. Shedding new light on opsin evolution. *Proceedings of the Royal Society B: Biological Sciences* **279**:3–14.

Presti RM, Zhao G, Beatty WL, Mihindukulasuriya KA, da Rosa APAT, Popov VL, Tesh RB, Virgin HW, Wang D. 2009. Quaranfil, Johnston Atoll, and Lake Chad viruses are novel members of the family Orthomyxoviridae. *J Virol* **83**:11599–11606.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**:i351–8.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**:e9490.

Priyam A, Woodcroft BJ, Rai V, Munagala A, Moghul I, Ter F, Gibbins MA, Moon H, Leonard G, Rumpf W, Wurm Y. 2015. Sequenceserver: a modern graphical user interface for custom BLAST databases. *bioRxiv*. doi:10.1101/033142

Pryszcz LP, Gabaldón T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res* **44**:e113.

Reich S, Guilligay D, Cusack S. 2017. An in vitro fluorescence based study of initiation of RNA synthesis by influenza B polymerase. *Nucleic Acids Res* **45**:3353–3368.

Reijden ED van der, Monchamp JD, Lewis SM. 1997. The formation, transfer, and fate of spermatophores in Photinus fireflies (Coleoptera: Lampyridae). *Can J Zool* **75**:1202–1207.

Rewitz KF, O'Connor MB, Gilbert LI. 2007. Molecular evolution of the insect Halloween family of cytochrome P450s: phylogeny, gene organization and functional conservation. *Insect Biochem Mol*

Biol **37**:741–753.

Reyes N, Lee V. 2010. Behavioral and morphological observations of Ignelater luminosus in Dominica.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**:276–277.

Rooney J, Lewis SM. 2002. Fitness advantage from nuptial gifts in female fireflies. *Ecol Entomol* **27**:373–377.

Rosa SP. 2010. New species of Ignelater Costa (Coleoptera, Elateridae, Pyrophorini). *Pap Avulsos Zool* **50**:445–449.

Rosa SP. 2007. Análise filogenética e revisão taxonômica da tribo Pyrophorini Candeze, 1863 (Coleoptera, Elateridae, Agrypninae). Universidade de São Paulo.

Sagegami-Oba R, Oba Y, Ohira H. 2007. Phylogenetic relationships of click beetles (Coleoptera: Elateridae) inferred from 28S ribosomal DNA: insights into the evolution of bioluminescence in Elateridae. *Mol Phylogenet Evol* **42**:410–421.

Sakai K, Tsutsui K, Yamashita T, Iwabe N, Takahashi K, Wada A, Shichida Y. 2017. Drosophila melanogaster rhodopsin Rh7 is a UV-to-visible light sensor with an extraordinarily broad absorption spectrum. *Sci Rep* **7**:7349.

Sander SE, Hall DW. 2015. Variation in opsin genes correlates with signalling ecology in North American fireflies. *Mol Ecol* **24**:4679–4696.

Schnitzler CE, Pang K, Powers ML, Reitzel AM, Ryan JF, Simmons D, Tada T, Park M, Gupta J, Brooks SY, Blakesley RW, Yokoyama S, Haddock SH, Martindale MQ, Baxevanis AD. 2012. Genomic organization, evolution, and expression of photoprotein and opsin genes in Mnemiopsis leidyi: a new view of ctenophore photocytes. *BMC Biol* **10**:107.

Schoville SD, Chen YH, Andersson MN, Benoit JB, Bhandari A, Bowsher JH, Brevik K, Cappelle K, Chen M-JM, Childers AK, Childers C, Christiaens O, Clements J, Didion EM, Elpidina EN, Engsontia P, Friedrich M, Garcia-Robles I, Gibbs RA, Goswami C, Grapputo A, Gruden K, Grynberg M, Henrissat B, Jennings EC, Jones JW, Kalsi M, Khan SA, Kumar A, Li F, Lombard V, Ma X, Martynov A, Miller NJ, Mitchell RF, Munoz-Torres M, Muszewska A, Oppert B, Palli SR, Panfilio KA, Pauchet Y, Perkin LC, Petek M, Poelchau MF, Record E, Rinehart JP, Robertson HM, Rosendale AJ, Ruiz-Arroyo VM, Smagghe G, Szendrei Z, Thomas GWC, Torson AS, Vargas Jentzsch IM, Weirauch MT, Yates AD, Yocum GD, Yoon J-S, Richards S. 2017. A model species for agricultural pest genomics: the genome of the Colorado potato beetle, Leptinotarsa decemlineata (Coleoptera: Chrysomelidae). *bioRxiv*. doi:10.1101/192641

Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, Lin S, Lin Y, Jung I, Schmitt AD, Selvaraj S, Ren B, Sejnowski TJ, Wang W, Ecker JR. 2015. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**:212–216.

Schultz MD, Schmitz RJ, Ecker JR. 2012. "Leveling" the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet* **28**:583–585.

Sezutsu Hideki, Le Goff Gaëlle, Feyereisen René. 2013. Origins of P450 diversity. *Philos Trans R Soc Lond B Biol Sci* **368**:20120428.

Shear WA. 2015. The chemical defenses of millipedes (diplopoda): Biochemistry, physiology and ecology. *Biochem Syst Ecol* **61**:78–117.

Sheiman IM, Shkutin MF, Terenina NB, Gustafsson MKS. 2006. A behavioral study of the beetle Tenebrio molitor infected with cysticercoids of the rat tapeworm Hymenolepis diminuta. *Naturwissenschaften* **93**:305–308.

Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11**:e0163962.

Shi G, Grimaldi DA, Harlow GE, Wang J, Wang J, Yang M, Lei W, Li Q, Li X. 2012. Age constraint on

Burmese amber based on U–Pb dating of zircons. *Cretaceous Res* **37**:155–163.

Shimomura O. 2012. Bioluminescence: Chemical Principles and Methods. World Scientific.

Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* **3**:265–274.

Sikora D, Rocheleau L, Brown EG, Pelchat M. 2017. Influenza A virus cap-snatches host RNAs based on their abundance early after infection. *Virology* **509**:167–177.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**:3210–3212.

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. 2009. JBrowse: a next-generation genome browser. *Genome Res* **19**:1630–1638.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**:31.

Slipinski, S. A., Leschen, R. A. B. & Lawrence, J. F. 2011. Order Coleoptera Linnaeus, 1758 In: Zhang Z –Q, editor. Animal Biodiversity: An Outline of Higher-Level Classification and Survey of Taxonomic Richness. Magnolia Press, Auckland. pp. 203–208.

Smedley SR, Risteen RG, Tonyai KK, Pitino JC, Hu Y, Ahmed ZB, Christofel BT, Gaber M, Howells NR, Mosey CF, Rahim FU, Deyrup ST. 2017. Bufadienolides (lucibufagins) from an ecologically aberrant firefly (Ellychnia corrusca). *Chemoecology* **27**:141–153.

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol* **32**:1342–1353.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**:2688–2690.

Stanger-Hall KF, Lloyd JE. 2015. Flash signal evolution in Photinus fireflies: character displacement and signal exploitation in a visual communication system. *Evolution* **69**:666–682.

Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**:62.

Steel J, Lowen AC. 2014. Influenza A virus reassortment. *Curr Top Microbiol Immunol* **385**:377–401.

Stibick JNL. 1979. Classification of the Elateridae (Coleoptera). *Relationships and classification of the subfamilies and tribes Pacific Insects* **20**:145–186.

Stolz U, Velez S, Wood KV, Wood M, Feder JL. 2003. Darwinian natural selection for orange bioluminescent color in a Jamaican click beetle. *Proc Natl Acad Sci U S A* **100**:14955–14959.

Suzuki MM, Kerr ARW, De Sousa D, Bird A. 2007. CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Res* **17**:625–631.

Tanabe AS. 2011. Kakusan4 and Aminosan: two programs for comparing nonpartitioned, proportional and separate models for combined molecular phylogenetic analyses of multilocus sequence data. *Mol Ecol Resour* **11**:914–921.

Tatsumi H, Kajiyama N, Nakano E. 1992. Molecular cloning and expression in Escherichia coli of a cDNA clone encoding luciferase of a firefly, Luciola lateralis. *Biochim Biophys Acta* **1131**:161–165.

Team RC, Others. 2013. R: A language and environment for statistical computing.

Temin HM. 1985. Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. *Mol Biol Evol* **2**:455–468.

Te Velthuis AJW, Fodor E. 2016. Influenza virus RNA polymerase: insights into the mechanisms of viral RNA synthesis. *Nat Rev Microbiol* **14**:479–493.

Thompson WW, Weintraub E, Dhankhar P, Cheng P-Y, Brammer L, Meltzer MI, Bresee JS, Shay DK. 2009. Estimates of US influenza-associated deaths made using four different methods. *Influenza Other Respi Viruses* **3**:37–49.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV):

high-performance genomics data visualization and exploration. *Brief Bioinform* **14**:178–192.

Timmermans MJTN, Dodsworth S, Culverwell CL, Bocak L, Ahrens D, Littlewood DTJ, Pons J, Vogler AP. 2010. Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Res* **38**:e197.

Timmermans MJTN, Vogler AP. 2012. Phylogenetically informative rearrangements in mitochondrial genomes of Coleoptera, and monophyly of aquatic elateriform beetles (Dryopoidea). *Mol Phylogenet Evol* **63**:299–304.

Tong D, Rozas NS, Oakley TH, Mitchell J, Colley NJ, McFall-Ngai MJ. 2009. Evidence for light perception in a bioluminescent organ. *Proceedings of the National Academy of Sciences* **106**:9836–9841.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**:511–515.

Tribolium Genome Sequencing Consortium, Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Beeman RW, Brown SJ, Bucher G, Friedrich M, Grimmelikhuijzen CJP, Klingler M, Lorenzen M, Richards S, Roth S, Schröder R, Tautz D, Zdobnov EM, Muzny D, Gibbs RA, Weinstock GM, Attaway T, Bell S, Buhay CJ, Chandrabose MN, Chavez D, Clerk-Blankenburg KP, Cree A, Dao M, Davis C, Chacko J, Dinh H, Dugan-Rocha S, Fowler G, Garner TT, Garnes J, Gnirke A, Hawes A, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Jackson L, Kovar C, Kowis A, Lee S, Lewis LR, Margolis J, Morgan M, Nazareth LV, Nguyen N, Okwuonu G, Parker D, Richards S, Ruiz S-J, Santibanez J, Savard J, Scherer SE, Schneider B, Sodergren E, Tautz D, Vattahil S, Villasana D, White CS, Wright R, Park Y, Beeman RW, Lord J, Oppert B, Lorenzen M, Brown S, Wang L, Savard J, Tautz D, Richards S, Weinstock G, Gibbs RA, Liu Y, Worley K, Weinstock G, Elsik CG, Reese JT, Elhaik E, Landan G, Graur D, Arensburger P, Atkinson P, Beeman RW, Beidler J, Brown SJ, Demuth JP, Drury DW, Du Y-Z, Fujiwara H, Lorenzen M, Maselli V, Osanai M, Park Y, Robertson HM, Tu Z, Wang J-J, Wang S, Richards S, Song H, Zhang L, Sodergren E, Werner D, Stanke M, Morgenstern B, Solovyev V, Kosarev P, Brown G, Chen H-C, Ermolaeva O, Hlavina W, Kapustin Y, Kiryutin B, Kitts P, Maglott D, Pruitt K, Sapojnikov V, Souvorov A, Mackey AJ, Waterhouse RM, Wyder S, Zdobnov EM, Zdobnov EM, Wyder S, Kriventseva EV, Kadowaki T, Bork P, Aranda M, Bao R, Beermann A, Berns N, Bolognesi R, Bonneton F, Bopp D, Brown SJ, Bucher G, Butts T, Chaumot A, Denell RE, Ferrier DEK, Friedrich M, Gordon CM, Jindra M, Klingler M, Lan Q, Lattorff HMG, Laudet V, von Levetsow C, Liu Z, Lutz R, Lynch JA, da Fonseca RN, Posnien N, Reuter R, Roth S, Savard J, Schinko JB, Schmitt C, Schoppmeier M, Schröder R, Shippy TD, Simonnet F, Marques-Souza H, Tautz D, Tomoyasu Y, Trauner J, Van der Zee M, Vervoort M, Wittkopp N, Wimmer EA, Yang X, Jones AK, Sattelle DB, Ebert PR, Nelson D, Scott JG, Beeman RW, Muthukrishnan S, Kramer KJ, Arakane Y, Beeman RW, Zhu Q, Hogenkamp D, Dixit R, Oppert B, Jiang H, Zou Z, Marshall J, Elpidina E, Vinokurov K, Oppert C, Zou Z, Evans J, Lu Z, Zhao P, Sumathipala N, Altincicek B, Vilcinskas A, Williams M, Hultmark D, Hetru C, Jiang H, Grimmelikhuijzen CJP, Hauser F, Cazzamali G, Williamson M, Park Y, Li B, Tanaka Y, Predel R, Neupert S, Schachtner J, Verleyen P, Raible F, Bork P, Friedrich M, Walden KKO, Robertson HM, Angeli S, Forêt S, Bucher G, Schuetz S, Maleszka R, Wimmer EA, Beeman RW, Lorenzen M, Tomoyasu Y, Miller SC, Grossmann D, Bucher G. 2008. The genome of the model beetle and pest Tribolium castaneum. *Nature* **452**:949–955.

Tyler J, Mckinnon W, Lord GA, Hilton PJ. 2008. A defensive steroidal pyrone in the Glow-worm Lampyris noctiluca L. (Coleoptera: Lampyridae). *Physiol Entomol* **33**:167–170.

Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR. 2015. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat Protoc* **10**:475–483.

Vélez S. 2006. Biogeographic and Genetic Approaches to the Natural History of the Bioluminescent Jamaican Click Beetle, Pyrophorus plagiophthalamus (Coleoptera: Elateridae) (Master of Science). University of Notre Dame.

Velez S, Feder JL. 2006. Integrating biogeographic and genetic approaches to investigate the history of bioluminescent colour alleles in the Jamaican click beetle, Pyrophorus plagiophthalamus. *Mol Ecol* **15**:1393–1404.

Virkki N, Flores M, Escudero J. 1984. Structure, orientation, and segregation of the sex trivalent in Pyrophorus luminosus III. (Coleoptera, Elateridae). *Can J Genet Cytol* **26**:326–330.

Viviani VR, Amaral D, Prado R, Arnoldi FGC. 2011. A new blue-shifted luciferase from the Brazilian Amydetes fanestratus (Coleoptera: Lampyridae) firefly: molecular evolution and structural/functional properties. *Photochem Photobiol Sci* **10**:1879–1886.

Viviani VR, Bechara EJ, Ohmiya Y. 1999a. Cloning, sequence analysis, and expression of active Phrixothrix railroad-worms luciferases: relationship between bioluminescence spectra and primary structures. *Biochemistry* **38**:8271–8279.

Viviani VR, Silva AC, Perez GL, Santelli RV, Bechara EJ, Reinach FC. 1999b. Cloning and molecular characterization of the cDNA for the Brazilian larval click-beetle Pyrearinus termitilluminans luciferase. *Photochem Photobiol* **70**:254–260.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**:2202–2204.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**:e112963.

Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJM, Birol I. 2015. LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *Gigascience* **4**:35.

Wasserman M, Ehrman L. 1986. Firefly Chromosomes, II. (Lampyridae: Coleoptera). *Fla Entomol* **69**:755–757.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res* **27**:757–767.

Weng J-K. 2014. The evolutionary paths towards complexity: a metabolic perspective. *New Phytol* **201**:1141–1149.

Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L. 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Res* **31**:28–33.

Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**:3350–3352.

Williams FX. 1917. Notes on the life-history of some North American Lampyridae. *J N Y Entomol Soc* **25**:11–33.

Williamson DL, Tully JG, Rose DL, Hackett KJ, Henegar R, Carle P, Bové JM, Colflesh DE, Whitcomb RF. 1990. Mycoplasma somnilux sp. nov., Mycoplasma luminosum sp. nov., and Mycoplasma lucivorax sp. nov., new sterol-requiring mollicutes from firefly beetles (Coleoptera: Lampyridae). *Int J Syst Bacteriol* **40**:160–164.

Willingham AT, Keil T. 2004. A tissue specific cytochrome P450 required for the structure and function of Drosophila sensory organs. *Mech Dev* **121**:1289–1297.

Wolcott GN. 1950. The Rise and Fall of the White Grub in Puerto Rico. *Am Nat* **84**:183–193.

Wolcott GN. 1948. The Insects of Puerto Rico: Coleoptera. University of Puerto Rico, Agricultural Experiment Station.

Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15**:R46.

Wood KV, de Wet JR, Dewji N, DeLuca M. 1984. Synthesis of active firefly luciferase by in vitro translation of RNA obtained from adult lanterns. *Biochem Biophys Res Commun* **124**:592–596.

Wood KV, Lam YA, Seliger HH, McElroy WD. 1989. Complementary DNA coding click beetle luciferases can elicit bioluminescence of different colors. *Science* **244**:700–702.

World Wildlife Fund. 2017. Terrestrial ecoregions of the world. https://www.worldwildlife.org/publications/terrestrial-ecoregions-of-the-world

Wulan WN, Heydet D, Walker EJ, Gahan ME, Ghildyal R. 2015. Nucleocytoplasmic transport of nucleocapsid proteins of enveloped RNA viruses. *Front Microbiol* **6**:553.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**:1859–1875.

Xiang H, Zhu J, Chen Q, Dai F, Li X, Li M, Zhang H, Zhang G, Li D, Dong Y, Zhao L, Lin Y, Cheng D, Yu J, Sun J, Zhou X, Ma K, He Y, Zhao Y, Guo S, Ye M, Guo G, Li Y, Li R, Zhang X, Ma L, Kristiansen K, Guo Q, Jiang J, Beck S, Xia Q, Wang W, Wang J. 2010. Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. *Nat Biotechnol* **28**:516–520.

Ye L, Buck LM, Schaeffer HJ, Leach FR. 1997. Cloning and sequencing of a cDNA for firefly luciferase from Photuris pennsylvanica. *Biochim Biophys Acta* **1339**:39–52.

Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva EV. 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res* **45**:D744–D749.

Zeng H, Goldsmith CS, Maines TR, Belser JA, Gustin KM, Pekosz A, Zaki SR, Katz JM, Tumpey TM. 2013. Tropism and infectivity of influenza virus, including highly pathogenic avian H5N1 virus, in ferret tracheal differentiated primary epithelial cell cultures. *J Virol* **87**:2597–2607.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**:2669–2677.

Zimin AV, Puiu D, Luo M-C, Zhu T, Koren S, Marçais G, Yorke JA, Dvořák J, Salzberg SL. 2017. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res* **27**:787–792.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**:3406–3415.

CHAPTER 4.

Stable isotope tracing reveals a lack of active *de novo* luciferin biosynthesis in firefly adult or larval light organs

Authors

Timothy R. Fallon[1,2],Fu-Shuang Li[2], Jing-Ke Weng[1,2]

Author Affiliations

[1]Whitehead Institute for Biomedical Research, 455 Main Street, Cambridge, MA 02142
[2]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

Author contributions:

I performed experiments and analyses of all figures, and wrote the chapter with input and minor edits from Jing-Ke Weng. Chemical syntheses were performed in collaboration with Fu-Shuang Li.

This chapter is unpublished as of May 2019

ABSTRACT

Firefly luciferin is a specialized metabolite naturally found only in the fireflies (Lampyridae), American railroad worms (Phengodidae), Asian starworms (Rhagophthalmidae), and bioluminescent click beetles (Elateridae). Luciferin is widely used in biomedical research and biotechnology as a substrate for the specialized light producing luciferase enzymes of these 4 families, either *in vitro* using recombinant luciferases, or *in vivo* where these luciferases are expressed transgenically in heterologous hosts. Of note, the *de novo* biosynthetic enzymes of firefly luciferin remain unknown, impeding total reconstitution of the firefly bioluminescent system in heterologous hosts and necessitating continuous supplementation of chemically synthesized luciferin in experiments. Here, we describe stable isotope tracing experiments on live North American adult and larval fireflies, and live larval Asian fireflies, aimed at elucidating the biosynthetic pathway of firefly luciferin. Our results suggest that, contrary to previous hypotheses, the specialized light organs of adult and larval of fireflies do not appear to significantly *de novo* biosynthesize luciferin. Luciferin may instead be produced from a recycling program with re-incorporates cysteine into the thiazoline of luciferin postoxidation to oxyluciferin.

INTRODUCTION

Bioluminescence, the production of light by a chemical reaction in a biological context, is found in diverse lineages across the tree of life in both terrestrial and marine environments. In the well-described cases, bioluminescence consists of the oxidation of a reduced small molecule, known as luciferin, by an enzyme, known as luciferase, with molecular O_2. This luciferin oxidation typically produces a high-energy peroxy-dioxetane intermediate that decays with production of an electronically excited molecule, dubbed oxyluciferin. Excited oxyluciferin then returns to its ground state emitting a photon in a process analogous to fluorescent emission. Despite the shared nomenclature of luciferin and luciferase, known bioluminescence consists of at least 7 independently evolved systems with structurally unique luciferins and non-homologous luciferases (Shimomura 2012).

Bioluminescence is useful for biotechnology. Of known bioluminescence, that of the bioluminescent beetles (Fireflies, Lampyridae; Click beetles, Elateridae; Railroad Worms, Phengodidae; and Starworms, Rhagophthalmidae) has been the most widely applied. But beetle luminescence applications still require external supplementation of luciferin, and therefore description of the *de novo* biosynthetic pathway of firefly luciferin has been a long sought goal, both for the basic biological interest of understanding how the unique benzothiazole-thiazole structure of luciferin is synthesized (Figure 2), and for the biotechnological applications of understanding luciferin biosynthesis, so that it can be reconstituted in heterologous hosts and enable large-scale or long-term luminescence which may not currently be possible.

The similarity of the luciferin thiazoline to D-cysteine led to the early suggestion for cysteine as a biosynthetic precursor (McCapra and Razavi, 1975), but the benzothiazole of luciferin did not have a clear biosynthetic analog. The presence of hydroquinone and its oxidative product benzoquinone in various lineages of beetles, for example, the bombardier beetle (Dettner, 1987), led to the hypothesis that benzoquinone and cysteine were the biosynthetic precursors of firefly luciferin.

Results from radioactive isotope (Okada et al., 1976) and stable isotope tracing experiments (Oba et al., 2013) of hydroquinone or benzoquinone in live fireflies have been consistent with the hypothesis that benzoquinone and cysteine are the biosynthetic precursors of firefly luciferin (Figure 1).

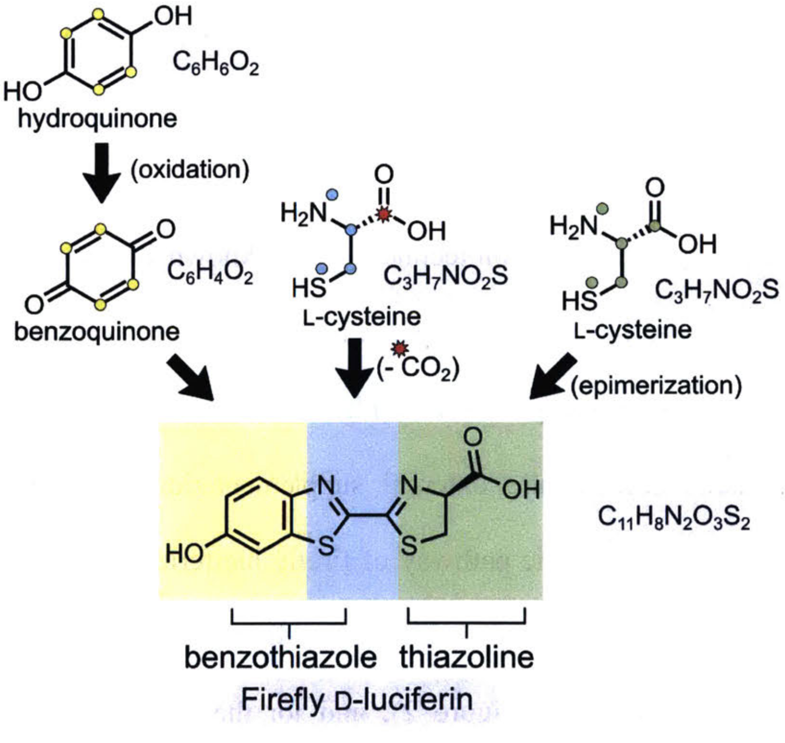


Figure 1: Hypothesized biosynthetic precursors of firefly luciferin

Firefly luciferin has two structural components, the benzothiazole (shown on the left in yellow/blue), and the thiazoline (shown on the right, in green). The thiazoline is almost certainly derived from a single molecule L-cysteine, however the stereochemistry of the stereocenter is opposite that of the natural L-cysteine. The N-S heterocycle portion of the luciferin benzothiazole (blue) is likely derived from an L-cysteine molecule, but a carbon is lost in the process, whereas the rest of the benzothiazole (yellow) is presumably derived from benzoquinone.

But observations that the presumed biosynthetic intermediates of firefly luciferin are made non-enzymatically by the "adventitious" melanin-polymerization-like redox chemistry of cysteine and benzoquinone (Crescenzi et al., 1988), combined with the observation that luciferin itself can be produced non-enzymatically in reactions of benzoquinone and cysteine (Kanie et al., 2018), has cast doubt upon the interpretation of these *in vivo* benzoquinone tracing results. Here we present stable isotope tracing results in adults and larvae of North American Lampyrinae subfamily fireflies, and larvae of the Luciolinae subfamily, aimed at testing the evidence that hydroquinone/benzoquinone is the true biosynthetic precursor of the benzothiazole ring of firefly D-luciferin, and the hypothesis that fireflies actively *de novo* biosynthesize luciferin in their light organs.

RESULTS

Establishing a method for stable isotope tracing in live fireflies

In order to better understand and ultimately identify the enzymes of the *de novo* biosynthetic pathway for firefly luciferin, we first sought to first identify the biosynthetic intermediates of luciferin through an untargeted liquid-chromatography high-resolution accurate-mass mass-spectrometry (LC-HRAM-MS) based metabolomics experiment. We surmised that if *de novo* luciferin biosynthesis occured in the adult male firefly lantern, that injection of supraphysiologic levels of the presumed biosynthetic precursors of firefly luciferin into live fireflies, namely injection of cysteine and hydroquinone, would produce detectable levels of biosynthetic intermediates (Figure 2A). We hypothesized that these biosynthetic intermediates would become more abundant when compared to the non-injection condition, allowing for unbiased detection through untargeted metabolomics (Figure 2C).

Furthermore, as untargeted metabolomics experiments typically give 100s to 1000s of detected differentially abundant compounds, within which it would be difficult to identify the subset of differentially abundant features that might correspond to luciferin biosynthetic intermediates, we sought to exploit stable isotope tracing to limit the candidate features to the features most relevant to the luciferin biosynthetic pathway. The availability of stable isotope labeled forms of both the biosynthetic precursors, namely $^{15}N^{13}C_3$ cysteine & $^2H_6/D_6$ hydroquinone (Figure 2B), allows for the filtering of the differentially abundant ions to those that were synthetically derived from both hydroquinone and cysteine, as we expected for firefly luciferin and structurally related metabolites, rather than ions which were derived from just one tracer or the other (Figure 2B, 2C). It is important to note that ^{15}N and ^2D have distinct mass defects compared to ^{13}C (+0.9970 Da and +1.0062 Da, versus +1.0033 Da), allowing for distinct quantitation of these resulting isotopologues via fine isotopic analysis on a high resolution mass spectrometer, such as the Q-Exactive mass spectrometer used in our study. Filtering to those features derived from both cysteine and hydroquinone therefore reduces the hundreds to thousands of differentially abundant compounds detected in a typical untargeted metabolomics experiment, to a short list of the most interesting features that may be involved in luciferin biosynthesis.

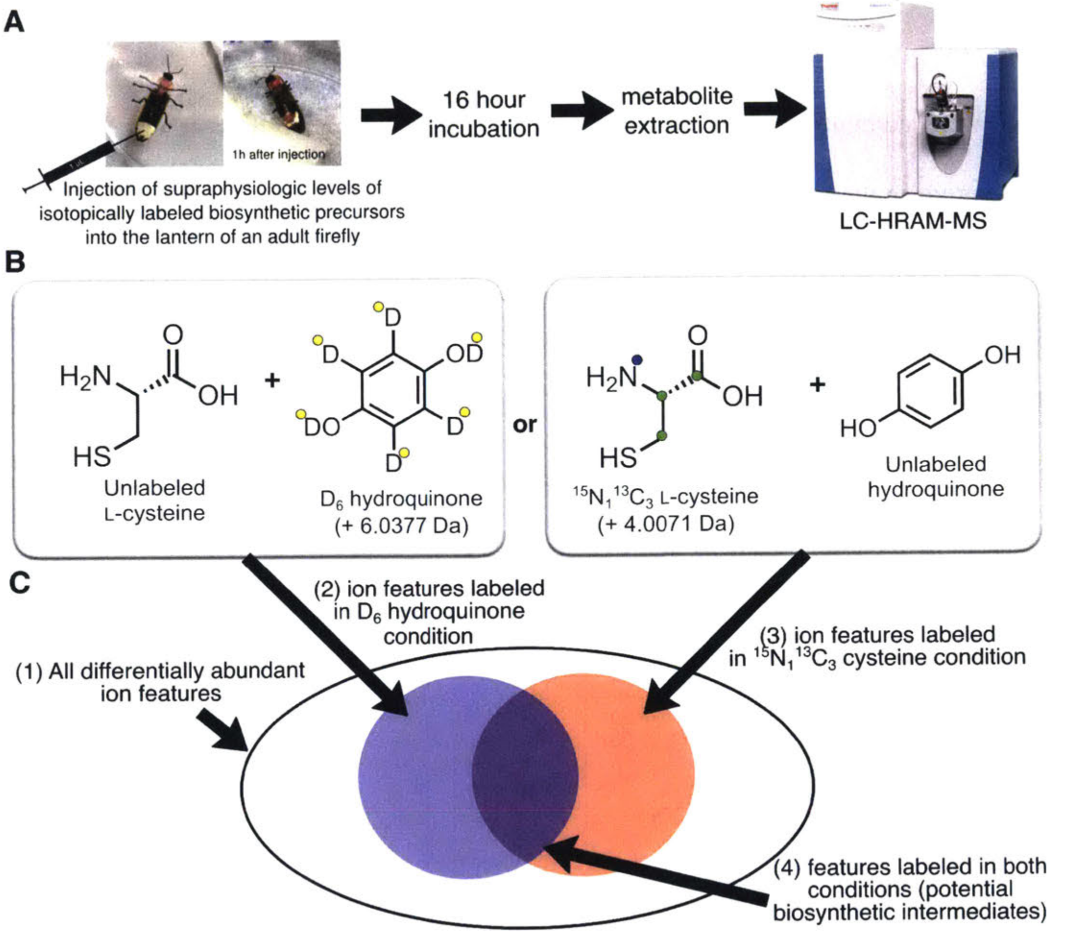

A
1h after injection
Injection of supraphysiologic levels of isotopically labeled biosynthetic precursors into the lantern of an adult firefly
16 hour incubation
metabolite extraction
LC-HRAM-MS
B
H_2N
O
OH
HS
Unlabeled L-cysteine
+
D
OD
DO
D_6 hydroquinone (+ 6.0377 Da)
or
$^{15}N_1{}^{13}C_3$ L-cysteine (+ 4.0071 Da)
OH
HO
Unlabeled hydroquinone
C
(1) All differentially abundant ion features
(2) ion features labeled in D_6 hydroquinone condition
(3) ion features labeled in $^{15}N_1{}^{13}C_3$ cysteine condition
(4) features labeled in both conditions (potential biosynthetic intermediates)

Figure 2: A stable isotope tracing untargeted metabolomics experiment the adult firefly lantern.

(A) Experimental scheme for stable isotope tracing of cysteine and hydroquinone in adult fireflies. A 1 μL solution of 550 mM cysteine and 550 mM hydroquinone was injected into the lantern of an adult male firefly (either genus *Photinus*, or *Pyractomena*). After a 16 hour incubation, metabolites were extracted and analyzed by LC-HRAM-MS **(B)** Tracing conditions used in the experiment, both cysteine and hydroquinone were injected simultaneously, but only one compound in the pair was stable isotope labeled at a time **(C)** Venn diagram representation of ions resulting from the injection experiment including (1) all differentially abundant ions, (2) ions labeled in the heavy hydroquinone tracing condition, (3) ions labeled in the heavy cysteine tracing condition, and (4) ions labeled in both tracing conditions, and therefore potential compounds that are structurally related to firefly luciferin.

The results of an initial tracing experiment confirmed that after injection of labeled biosynthetic precursors into live fireflies, stable isotope labels were incorporated into firefly D-luciferin (Figure 3). These stable isotope incorporation patterns were consistent with those previously reported (Oba et al., 2013), including a +3 Da ($+^{15}N_1{}^{13}C_2$ or $+D_3$) indicating cysteine or hydroquinone incorporation into the luciferin benzothiazole (Figure 3C,3D), the +4 Da ($+^{15}N_1$ $^{13}C_3$) incorporation of cysteine into the thiazoline of luciferin (Figure 3C), and the +7 Da (+$^{15}N_2$$^{13}C_5$) incorporation of cysteine into the benzothiazole and thiazoline of luciferin (Figure 3C). These result suggested that cysteine and hydroquinone were authentic synthetic precursors, and that the *de novo* biosynthetic pathway of firefly luciferin was active in the adult light organ. However, in these experiments we noted that hydroquinone and cysteine injection had dramatic effects on firefly coloration, including a red (Figure 2A), and then later black color development in the typically yellow firefly lantern emanating from the site of injection, which had not been reported in previously published firefly tracing experiments. This color development was reminiscent of the reported pheomelanin-like polymerization chemistry that been reported for the direct coupling of benzoquinone and cysteine (Crescenzi et al., 1988).

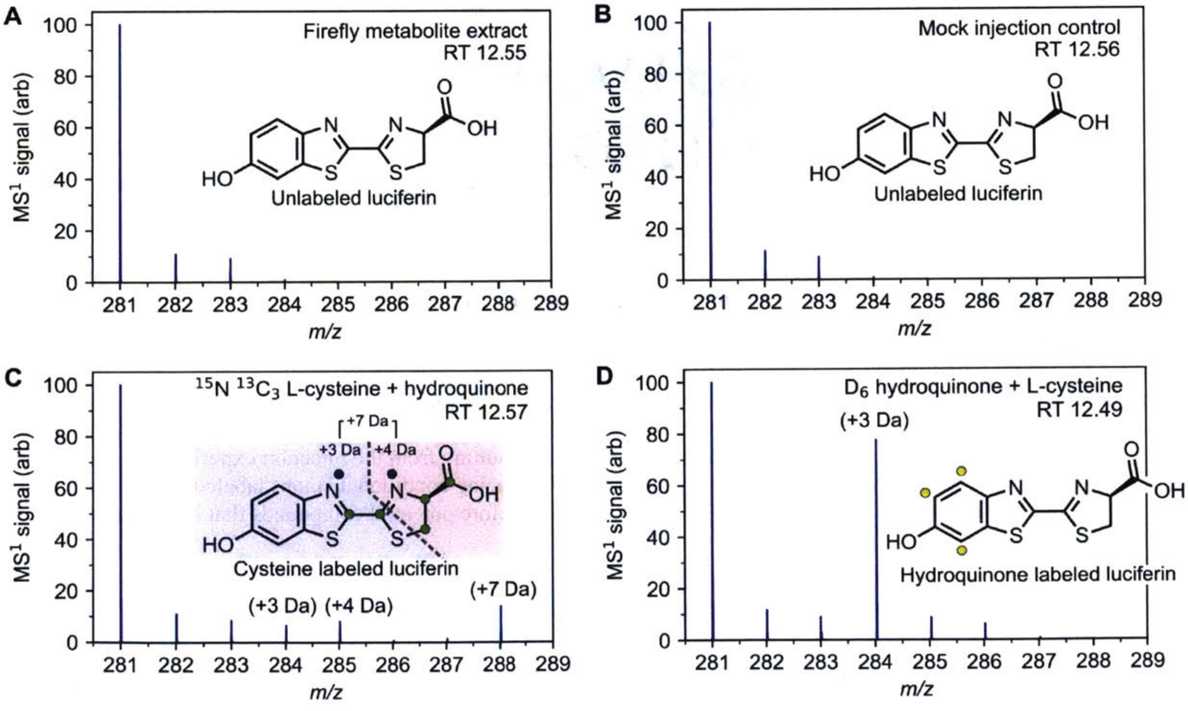


Figure 3: Representative luciferin MS^1 spectra from preliminary stable isotope tracing lantern injection experiments.

(A) No injection condition **(B)** Mock injection condition **(C)** Injection of ^{15}N ^{13}C$_3$ L-cysteine + unlabeled hydroquinone. Structural interpretation of the tracing pattern is shown in the inset panel. Blue circles represent ^{15}N isotopic incorporation, green circles represent ^{13}C isotopic incorporation **(D)** Injection of D_6 hydroquinone + unlabeled L-cysteine. Yellow circles represent incorporation of deuterium bound to the indicated carbon.

Identification of putative biosynthetic intermediates via intersectional-tracing stable-isotope-assisted-metabolomics

With confirmation that stable isotope tracing of cysteine and hydroquinone into luciferin was reproducible and achievable using our experimental approach, we next performed a large-scale replicated LC-HRAM-MS firefly-injection & stable isotope tracing experiment, using several tracing conditions. These tracers included heavy cysteine ($^{15}N^{13}C_3$) with unlabeled hydroquinone, heavy hydroquinone ($^2H_6$) with unlabeled cysteine, heavy cysteine alone, and heavy hydroquinone alone. We subjected the resulting LC-HRAM-MS data to a computational analysis designed to isolate those differentially abundant features which were labeled in both the heavy hydroquinone with unlabeled cysteine and the heavy cysteine with unlabeled hydroquinone conditions (Figure 2C), through an approach which we dub intersectional-tracing stable-isotope-assisted-metabolomics (IT-SIAM). This IT-SIAM analysis reduced the 4956 features detected as differentially abundant in the tracing conditions, down to 51 features that were labeled in both tracing conditions (Figure 4). Via comparison to standard compounds and comparison to MS/MS spectra of known compounds, we were able to confirm that our analysis detected several known luciferin-related metabolites, including luciferin itself, and sulfoluciferin (Fallon et al., 2016), leading us to believe that our approach was a reliable detector of luciferin related metabolites (Figure 4D). Many of the 51 isolated features were alternate ionization adducts (e.g. $[M+Na]^+$ vs $[M+H]^+$), or simple isotopologues from natural isotopes (e.g. $+^{13}C_1$, $+^{13}C_2$), making the true number of compounds detected likely less than 51. To elucidate the features within the set of 51 features with potential relevance to luciferin biosynthesis, we first inspected the most abundant features in the set which had unknown structures: *m/z* 230.04812 at RT 1.44 mins, *m/z* 196.0063 at RT 5.8, *m/z* 152.0165 at RT 7.7, and *m/z* 210.0219 at RT 12.2. Calculation of possible chemical formula within 5 ppm mass accuracy and comparison to in-silico calculated MS^1 isotopic patterns (data not shown), confirmed these features as nitrogen-sulfur containing, with unionized chemical formula of $C_9H_{11}NO_4S$, $C_8H_5NO_3S$, C_7H_5NOS, and $C_9H_7NO_3S$ respectively. Manual inspection and interpretation of the MS^2 fragmentation spectra of these features indicated they were benzothiazole containing compounds (data not shown), which led to the structural hypotheses that the four features arose from the compounds 2-amino-3-[(2,5-dihydroxyphenyl)sulfanyl]propanoic acid, 6-hydroxy-1,3-benzothiazole-2-carboxylic acid, 1,3-benzothiazol-6-ol, and methyl 6-hydroxy-1,3-benzothiazole-2-carboxylate, which we respectively dubbed cysteine-quinol, 6-hydroxybenzothiazole-carboxylic acid (6-HBZ-CA), 6-hydroxybenzothiazole (6-HBZ), and 6-hydroxybenzothiazole-methyl ester (6-HBZ-CA-ME)(Figure 4D).

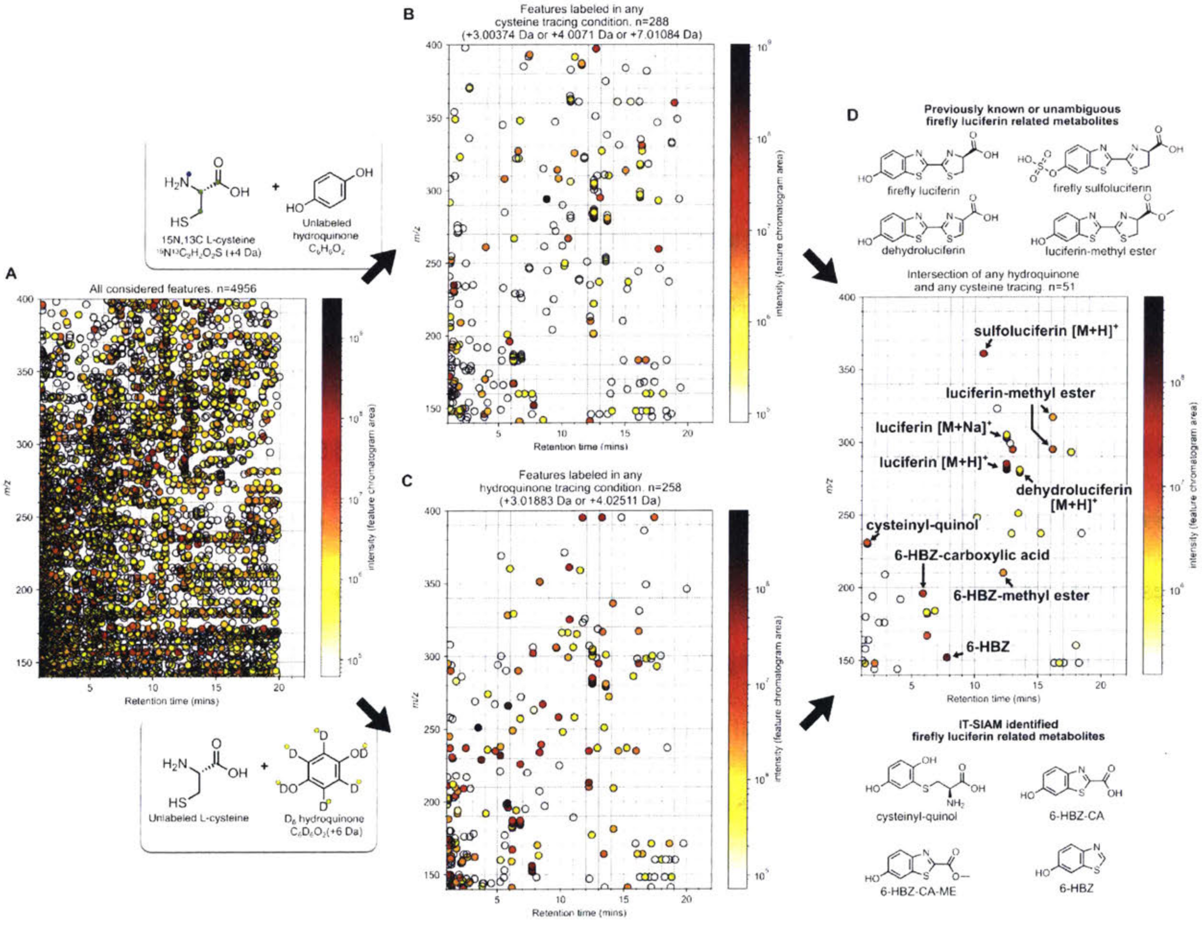


Figure 4: Elucidation of features structurally related to luciferin by intersectional tracing stable isotope assisted metabolomics (IT-SIAM).

(A) All features considered in the IT-SIAM analysis, prior to reduction **(B)** All features (n=288) which show a $^{15}N^{13}C_2$ (+3.003474 Da), $^{15}N^{13}C_3$ (+4.0071 Da), or $^{15}N_2{}^{13}C_5$ (+7.01084 Da) tracing signal. **(C)** All features (n=258) which show a $^2H_3$ (+3.01883 Da) or $^2H_4$ (+4.02511 Da) tracing signal. **(D)** Resulting features (n=51) which show tracing signals in both the heavy cysteine with hydroquinone and cysteine with heavy hydroquinone tracing conditions. Known or unambiguously related firefly luciferin metabolites are shown above the panel, whereas the structural interpretation of newly identified firefly luciferin related metabolites are shown below the panel.

Tracing experiments with putative biosynthetic intermediates identified by IT-SIAM

With the identification of 6-HBZ-CA and its analogs in our set of IT-SIAM highlighted features, we hypothesized a working model for the biosynthesis of luciferin (Figure 5). The presence of 6-HBZ-CA in our injections was particularly promising, as it a simple substitution of cysteine onto the carboxylic acid would allow for synthesis of luciferin (Figure 5).

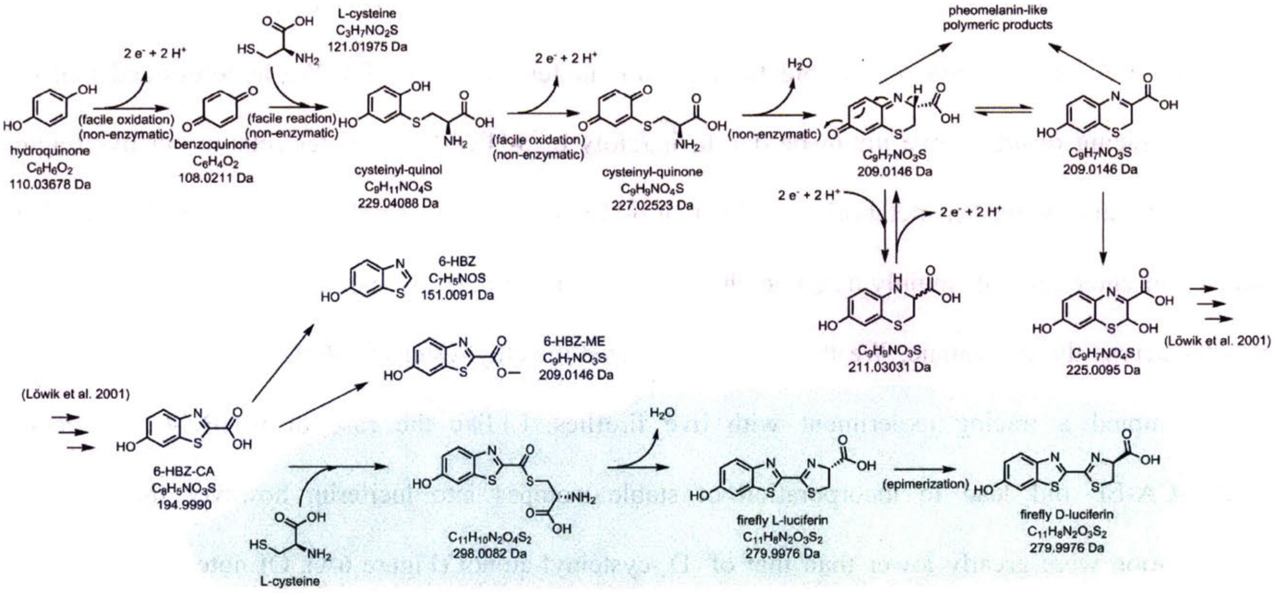


Figure 5: Working model of luciferin biosynthesis.

Hydroquinone and cysteine, through non-enzymatic spontaneous coupling and oxidation (red background), produce bicyclic benzothiazine compounds. These benzothiazines can spontaneously and non-enzymatically produce 6-HBZ-CA (Löwik et al., 2001), and other simple derivatives including 6-HBZ-ME & 6-HBZ observed in our IT-SIAM analysis (gray background). These benzothiazines can also produce off-pathway polymeric products akin to pheomelanin (Crescenzi et al., 1988). 6-HBZ-CA is then enzymatically activated, allowing the nucleophilic substitution of L-cysteine, after which an intramolecular ring closure forms L-luciferin, which is then epimerized to form the final substrate D-luciferin (green background).

Notably, 6-HBZ-CA was previously synthesized by Löwik and colleagues (Löwik et al., 2001), who demonstrated that the contraction of the 6-membered benzothiazine, resulting from the direct conjugation of benzoquinone and cysteine, to the 6-HBZ-CA-like 5-membered benzothiazole could proceeded non-enzymatically and spontaneously in near stoichiometric yield with the simple addition of NaOH. Furthermore, there have been unpublished tests if 6-HBZ-CA could act as a precursor to firefly luciferin in firefly lantern extracts (Day et al., 2004), with negative results and undescribed experimental parameters. We therefore sought to test if D_3-6-HBZ-CA and D_3-cysteinyl-quinol could trace into luciferin in our experimental system, and thereby provide clear evidence of their role as biosynthetic intermediates. We synthesized the ethyl ester of D_3-cysteinyl-quinol (D_3-cysteinyl-quinol-EE), D_3-HBZ-CA, and the ethyl ester of D_3-6-HBZ-CA (D_3-6-HBZ-CA-EE). The tracing experiment with D_3-cysteinyl-quinol-EE showed robust incorporation into luciferin as expected (Figure 6A). To our surprise however, D_3-6-HBZ-CA showed zero detectable tracing into luciferin (Figure 6B). We hypothesized that our negative tracing results above might be due to inability of 6-HBZ-CA to enter cells in our live tracing experiments and be further metabolized. Although 6-HBZ-CA is quite small (<200 Da), making it likely that it can enter cells, its mainly negative charge at neutral pH due to it carboxylic acid may prevent its transfer across the membrane. We therefore synthesized the ethyl-ester of 6-HBZ-CA (6-HBZ-CA-EE), and attempted a tracing experiment with live fireflies. Unlike the case of 6-HBZ-CA injection, 6-HBZ-CA-EE did lead to incorporation of stable isotopes into luciferin, however the levels of incorporation were greatly lower than that of D_3-cysteinyl-quinol (Figure 6A). Of note, we co-injected cysteine alongside D_3-cysteinyl-quinol-EE in the injection, so this low level tracing may be due to non-enzymatic transesterification with cysteine.

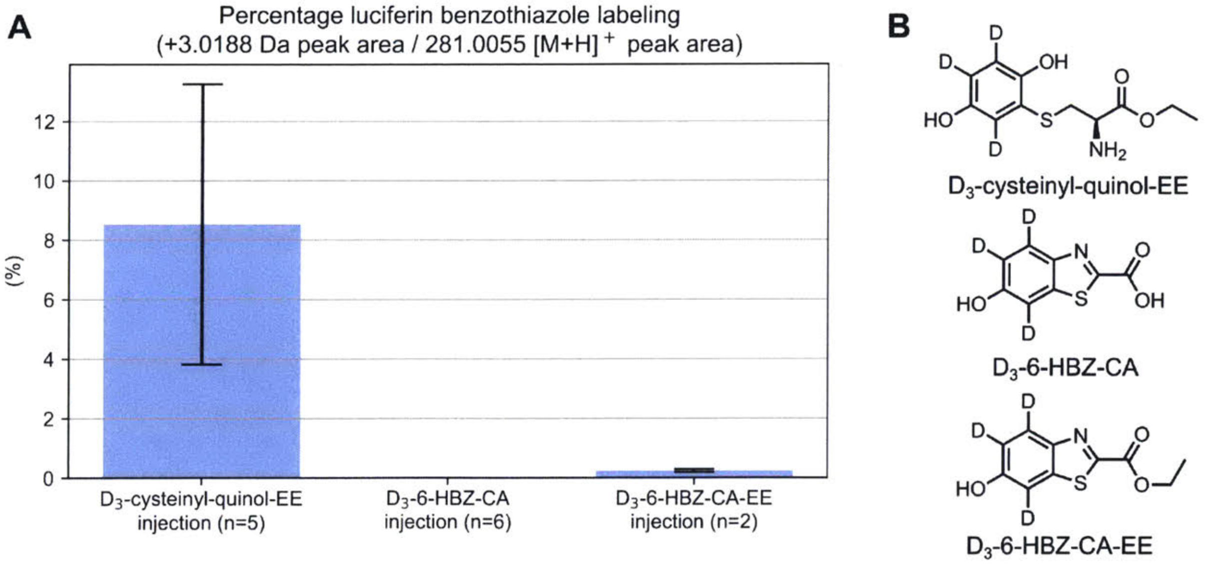


Figure 6: Tracing with putative luciferin biosynthetic precursors.

(A) Incorporation of +3 Da $^2H_3$ tracing signal into firefly luciferin after injection of D_3-cysteinyl-quinol-EE, or D_3-6-HBZ-CA, or D_3-6-HBZ-CA-EE into adult male fireflies **(B)** Chemical structures of D_3-cysteinyl-quinol-EE, D_3-6-HBZ-CA, D_3-6-HBZ-CA-EE.

Tracing results in the absence of hydroquinone

Perplexed by our incongruous tracing result between D_3-6-HBZ-CA, D_3-6-HBZ-CA-EE, and D_3-cysteinyl-quinol, we more closely analyzed our previous heavy cysteine tracing alone, and heavy hydroquinone alone tracing data to better understand the presumed biosynthetic process. For heavy hydroquinone alone injections, we noted a robust +3.0188 Da incorporation into the benzothiazole of firefly luciferin, as expected (Figure 7A). For heavy cysteine alone injections however, we noted a substantial difference between the incorporation results compared to when hydroquinone was co-injected (Figure 7A), and when heavy cysteine was alone (Figure 7B). In the cysteine alone injection condition, we only observed reasonable +4 Da ($^{15}N^{13}C_3$) tracing (~5%) indicative of tracing of cysteine into the thiazoline of luciferin, and observed negligible tracing (~0.05%) into the benzothiazole of luciferin. A calculation of the natural isotopic abundances of the firefly luciferin $[M+H]^+$ ion in MZmine 2 (Pluskal et al., 2010) indicated the natural isotopologues indistinguishable from the *m/z* value of the $^{15}N_1{}^{13}C_2$

isotopologue (trace natural abundance) had a combined abundance of ~0.03%, indicating that the observed ~0.05% $^{15}N_1{}^{13}C_2$ signal in the cysteine alone injection condition is likely background signal. These results are in contrast to our expectation, that if luciferin was being actively *de novo* biosynthesized during our injection experiments, and if cysteine was the source of the thiazole portion of luciferin's benzothiazole, then the benzothiazole should be expected to become robustly labeled with a +3 Da $^{15}N^{13}C_2$ label.

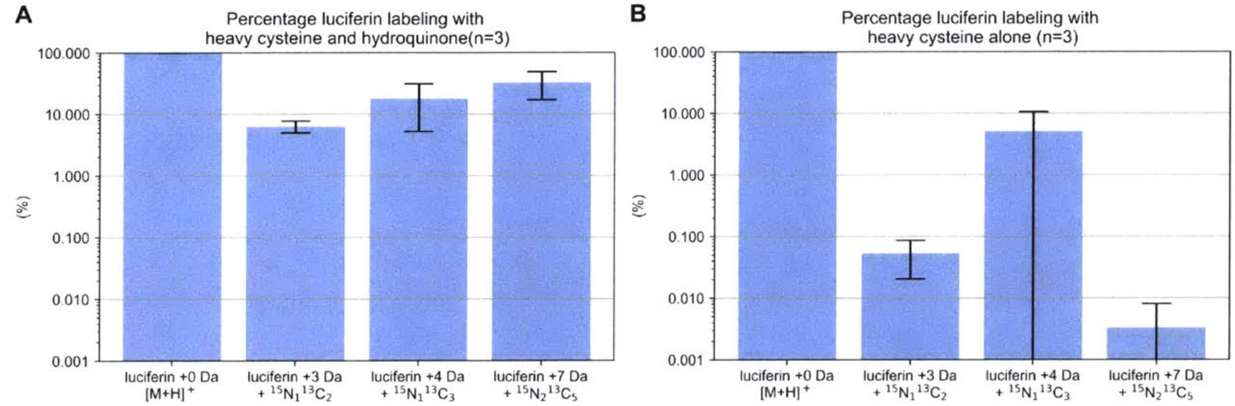


Figure 7: Tracing with cysteine alone in adult fireflies

(A) Incorporation of tracing signals into firefly luciferin, after injection of heavy cysteine with hydroquinone. Note the robust tracing, especially of the +3 Da $^{15}N_1{}^{13}C_2$ indicative of cysteine incorporation into the benzothiazole **(B)** Incorporation of tracing signals into firefly luciferin, after injection of heavy cysteine alone. Note the greatly reduced cysteine incorporation into the +3 Da $^{15}N_1{}^{13}C_2$ indicative of cysteine incorporation into the benzothiazole, but a still quite robust incorporation into the luciferin thiazoline, as indicated by the +4 Da $^{15}N_1{}^{13}C_3$ signal.

Stable isotope tracing experiments in adult and larval light organs stimulated to continuously luminescence

Given that hydroquinone is easily oxidized to benzoquinone, and that benzoquinone and cysteine have been demonstrated to non-enzymatically and spontaneously react to produce low yields of luciferin *in vitro* at neutral pH (Kanie et al., 2016), we were then concerned that our previously robust tracing results when cysteine is injected alongside hydroquinone, may not be due to luciferin synthesis by a biosynthetic route, but rather could be due to non-enzymatic synthesis from a "messy chemistry" or "adventitious" synthesis of luciferin through the intrinsic reactivity of cysteine and benzoquinone. A simple calculation indicated that indeed, even the low-yield quantities of luciferin produced by what we dub the "messy chemistry" synthesis route described by Kanie and colleagues (~0.1%), combined with the comparatively high quantities of cysteine and benzoquinone injected in our and others (Oba et al., 2013) tracing experiments, would produce roughly the same quantity of luciferin as is present in a single firefly specimen, potentially leading to a false positive tracing signal.

To explain the near lack of tracing into the luciferin benzothiazole when heavy cysteine was injected alone, we hypothesized that this result was due to the near absence of *de novo* luciferin biosynthesis in the adult male firefly lantern. This was in contrast to seemingly present oxyluciferin recycling or thiazoline cysteine exchange pathway, which was leading to robust incorporation of the +4 Da $^{15}N_1{}^{13}C_3$ signal (Figure 7B). We further hypothesized that if *de novo* luciferin biosynthesis was nearly absent in the adult, that the biosynthetic pathway may only become active in response to a depletion of luciferin. Therefore, to test this hypothesis, we conducted heavy cysteine only tracing in adult male *Photinus pyralis* firefly light organs induced to continuously luminesce. Although injection of the insect neurotransmitter synephrine, a mono-methylated analog of octopamine, the known effector neurotransmitter of firefly bioluminescence (Carlson, 1972; Ghiradella and Schmidt, 2004), does induce light production in adult male fireflies, this light production is short lived, only lasting around 15 minutes per injection. We therefore sought an alternative method to induce luminescence. As the production of light by the adult firefly light organ is thought to is gated by oxygen availability, we dissected the adult male light organ of *Photinus pyralis* fireflies, and placed them in an explant tissue culture. Indeed, separated from the control of oxygen exposure mediated by the physiology of the intact firefly, these explanted light organs glowed continuously, with an intensity that was easily observable by the naked eye. These light organs could be maintained in culture for about 48 hours, throughout which they were continuously luminous, although decreasing in intensity. After about 48 hours of culture, they invariably succumbed to a microbial growth with a striking pink coloration. This microbial growth occurred despite inclusion of the "Primocin" antibiotic mixture in the culture media, and an involved dissection procedure aimed at removing microbial contaminants. Heavy cysteine tracing experiments in these explanted light organs showed a very robust labeling of the +4 Da thiazoline peak, with again little detectable tracing into the benzothiazole of luciferin as indicated by the +3 Da $^{15}N_1{}^{13}C_2$ signal (Figure 8C). Notably, there is a background signal (~0.05%, relative tracing) of the +3 Da $^{15}N_1{}^{13}C_2$ signal in a no tracer added control (Figure 8B), likely due to the $^{15}N_1{}^{13}C_2$ isotopologue arising from natural isotopic abundances as previously mentioned, however in the cysteine alone tracing condition, the +3 Da $^{15}N_1{}^{13}C_2$ signal (~0.4%) is clearly significantly greater than that found in the no tracer added control (Figure 8B), or in the condition where cysteine alone was injected into fireflies without stimulation (Figure 7B). These results may support that authentic *de novo* biosynthesis of the luciferin benzothiazole is occuring, albeit at a rate that is roughly ~1000x less than the incorporation of cysteine into the luciferin thiazoline. Overall these results support the presence of a luciferin recycling, or thiazoline cysteine exchanging pathway, and falsified the hypothesis that *de novo* luciferin biosynthesis could be induced to occur at a high level in the firefly adult male light organ.

Given the relatively weak tracing +3 Da $^{15}N_1{}^{13}C_2$ signal observed, we hypothesized that cysteine may still not be the true biosynthetic source of the thiazole in the luciferin benzothiazole. We therefore attempted a tracing experiment with D_4-L-tyrosine, L-tyrosine being the most likely biosynthetic source of the phenolic ring of the benzothiazole. In this experiment we did not observe detectable incorporation of the +3 Da signal into firefly luciferin (Figure 8D). Taken together, these results suggest that adult male Lampyrinae fireflies do not appreciably *de novo* biosynthesize firefly luciferin, and that cysteine may be the biosynthetic precursor, but that tyrosine, at least in our experimental conditions, gives no detectable indication of incorporation into luciferin. Rather that highlighting the *de novo* biosynthetic pathway for firefly luciferin, our evidence suggests that luciferin recycling, namely the replacement of the thiazole of oxyluciferin with cysteine to reproduce the thiazoline of luciferin, is the main route by which luciferin is maintained over time in the adult male firefly light organ.

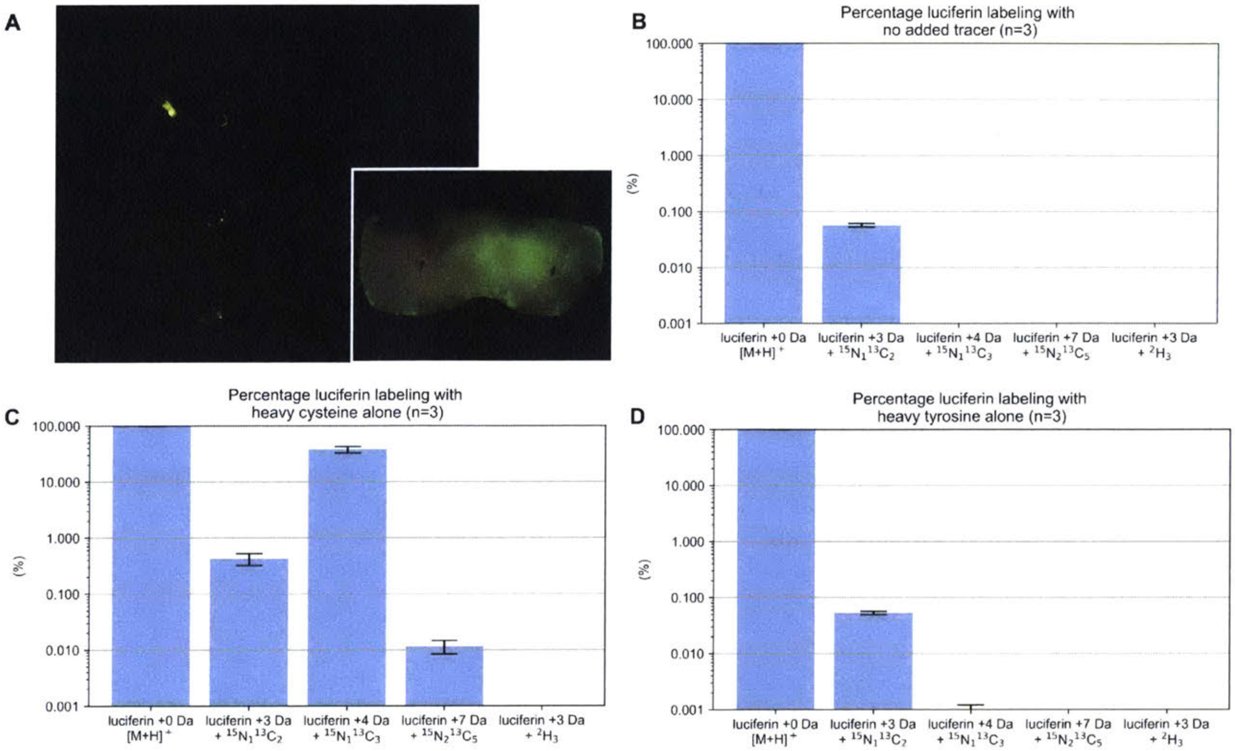


Figure 8. Cysteine and tyrosine stable isotope tracing in explanted adult male *P. pyralis* light organs.

(A) Light organs glowing in culture. Inset: Representative dissected anterior adult male *P. pyralis* light organ. Imaged under its own light (green), and brightfield. **(B)** Tracing results after 16 hr, with no added tracer to the media. **(C)** Tracing results in explanted adult male *P. pyralis* light organs stimulated to luminesce for 16 hr, after addition of heavy L-cysteine ($^{15}N_1{}^{13}C_3$) to the culture media. **(D)** Tracing results in explanted adult male *P. pyralis* light organs stimulated to luminesce for 16 hr after addition of heavy L-tyrosine ($^2H_4$) to the culture media.

We next hypothesized that if robust luciferin *de novo* biosynthesis was not occuring in the adult life stage, that it might instead be occurring in the long-lived larval life stage. We therefore attempted the heavy cysteine alone tracing experiment in *P. pyralis* larvae that had been induced to luminescence continuously. Like adult fireflies, injections of synephrine into larvae induced luminescence, but unlike adult fireflies, firefly larvae are extremely ill-suited to repeated injection, with hydrostatic pressure of the larval body leading to a large extrusion and loss of hemolymph with each injection. We therefore sought to find an experimental protocol where firefly larvae could be induced to luminesce without injection. As nitric oxide (NO) is reported to be the ultimate effector molecule in the control of firefly luminescence, we first attempted chemical production of nitric oxide to induce luminescence. Exposure of the spontaneously NO evolving compound DEA NONOate to larvae had mixed results, with 50% death of the larvae, with no observed induction of light emission, and therefore was not further investigated. Physical stimulation of the larvae had mixed results, as even larvae that had initially luminescenced on physical stimulation, would eventually stop responding to repeated stimulation. We ultimately found the simplest and most effective method: Covering larvae in solid DL-synephrine that had been suspended in water. This treatment induced long lasting luminescence (>8 hours), and surprisingly, appeared to have little to no toxicity (Figure 9A, 9B, 9C).

With a well-performing larval luminescence induction method established, we repeated our heavy cysteine only tracing experiment. Again, this tracing showed robust tracing into the luciferin thiazoline, with little detectable tracing into the luciferin benzothiazole, albeit tracing that is above the background level expected for the $^{15}N_1{}^{13}C_2$ signal (Figure 9).

We next hypothesized that the lack of tracing in larvae, may be different in other subfamilies of fireflies. We therefore sought to also perform the heavy cysteine tracing experiment in firefly larvae of the distantly related subfamily Luciolinae, most commonly found in Asia and Africa, and Australasia. We obtained *Aquatica ficta* larval fireflies, and repeated the injection-tracing experiment. Like our results with North American fireflies, this experiment showed robust tracing of heavy cysteine into the luciferin thiazoline, with little detectable tracing into the luciferin benzothiazole. D_4-Tyrosine tracing was not undertaken in either *Aquatica ficta* or *Photinus pyralis* larvae, due to a lack of specimens.

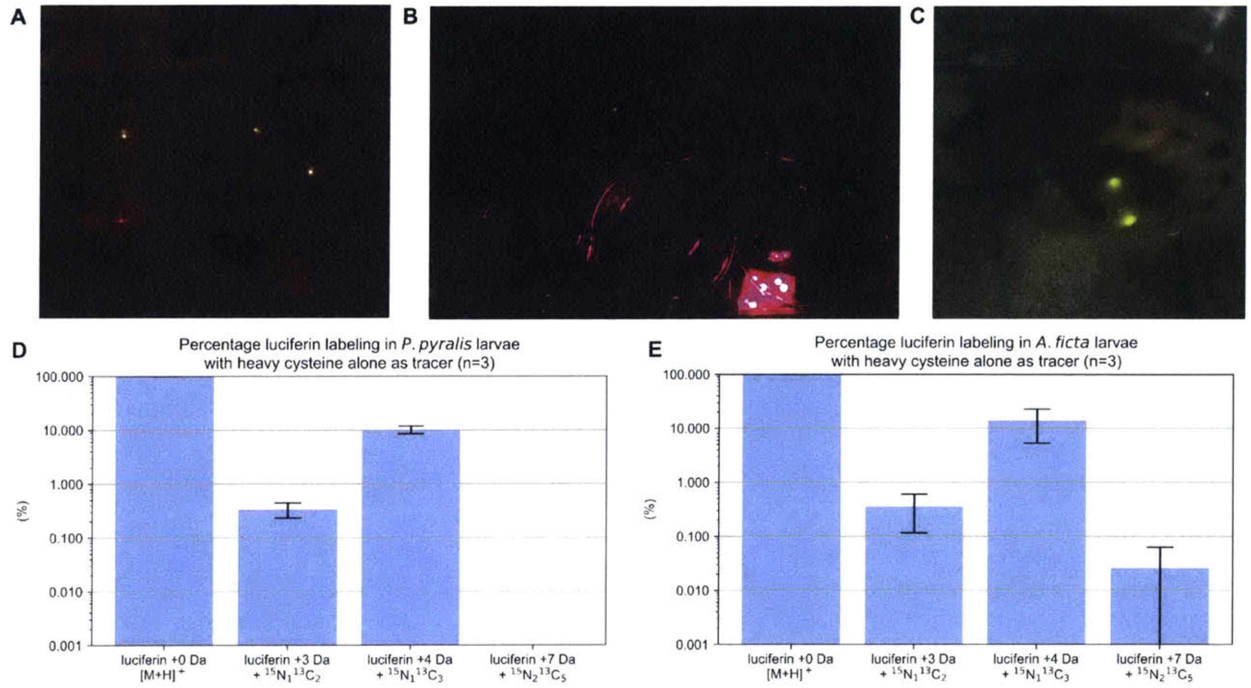


Figure 9: Tracing with cysteine alone in artificially stimulated larvae.

(A) *P. pyralis* larvae induced to continuously luminesce using topical DL-synephrine treatment **(B)** *P. pyralis* (top row), and *A. ficta* (bottom row) larvae stimulated to luminesce continually, which were used for experimental injection. Note that the light organs of *A. ficta* appear to be smaller and less bright than those of *P. pyralis* **(C)** Representative artificially stimulated *P. pyralis* larva glowing from larval light organs. Image is a composite of two photos, and the luminescence is brighter than it appeared through the microscope. Blue grid behind larvae is 1 cm **(D)** Tracing results in *P. pyralis* larvae stimulated to luminesce for 16 hr, after injection of heavy L-cysteine ($^{15}N_1{}^{13}C_3$) **(E)** Tracing results in *A. ficta* larvae stimulated to luminesce for 16 hr, after injection of heavy L-cysteine ($^{15}N_1{}^{13}C_3$)

DISCUSSION

The biosynthesis of firefly luciferin has been discussed for decades, but there are relatively few experimental results regarding luciferin biosynthesis. To the authors' knowledge, the total set of published experiments in firefly luciferin biosynthesis are discussed below. The first discussion of luciferin biosynthesis was an *in vitro* study of the synthesis of benzothiazoles from benzoquinone and cysteine by McCapra and colleagues (McCapra and Razavi, 1975). McCapra were successfully able to synthesize 6-HBZ-CA-EE from benzoquinone and cysteine, with relatively mild synthetic steps, which established the plausibility of luciferin biosynthesis from cysteine and benzoquinone. In the first biological experiment, Okada and colleagues injected ^{14}C radiolabeled benzoquinone, hydroquinone, tyrosine, and acetate into *Luciola cruciata* fireflies, and measured the specific incorporation (specific activity of the desired product divided by the specific activity of the starting material)(David Morgan, 2007) of the radiolabel into luciferin (Okada et al., 1976). Okada reported a specific incorporation, of ~0.3-0.4% for hydroquinone and benzoquinone, to 0.005-0.02% for tyrosine, and 0.0023-0.0036% for acetate. These radiolabeled tracing experiments can be interpreted in absolute terms regarding the efficiency of label incorporation ('specific incorporation'), whereas the stable isotope tracing experiments presented here and by other authors (Oba et al., 2013) are interpreted in relative terms where the abundance of the labeled peak is compared to that of unlabeled peak. In the Okada 1975 experiments, the 0.3-0.4% label incorporation of hydroquinone/benzoquinone into luciferin, is quite close to the 0.1%-0.45% yield of the non-enzymatic "messy-chemistry" luciferin synthesis reported by Kanie and colleagues (Kanie et al., 2016). Therefore, it reinterpreting these results in light of knowledge of the "messy-chemistry", it seems unjustified to conclude that the robust tracing by hydroquinone and benzoquinone observed by Okada et al., is mainly due to authentic biosynthesis, rather than non-enzymatic luciferin synthesis.

In the second biological experiment, McCapra and colleagues fed cystine which was radiolabeled on the carboxylic carbons (1-^{14}C), to adult specimens of the bioluminescent click beetle *Pyrophorus pellucens* (now named *Pyrophorus noctilucus*). The specific incorporation could not be determined in this experiment, but incorporation of the radiolabel into luciferin was confirmed. Importantly, via degradation experiments, the radiolabel was only found on the thiazoline carboxylic carbon of luciferin. This result has several interpretations, e.g., *de novo* biosynthesis was occurring but cysteine carboxylic carbon was lost specifically upon incorporation into the benzothiazole, or, there was a complete lack of *de novo* biosynthesis, but a presence of a thiazoline exchanging recycling pathway, akin to the results we observe in fireflies.

As a third biological experiment Colepicolo and colleagues injected universally labeled radiolabeled cystine (U-^{14}C) into larval specimens of the Brazilian bioluminescent click beetle *Pyrearinus* *termitilluminans* (Colepicolo et al., 1988). In this case, the specific incorporation was reported to be 3% after 24 hours. Given that universally labeled cysteine was used, incorporation into specific forms of luciferin (e.g. benzothiazole vs thiazoline) could not be determined.

In 2013, Oba and colleagues performed the second luciferin firefly biosynthesis experiment in fireflies, when they injected stable labeled cysteine, hydroquinone, and benzoquinone into the Japanese firefly *Aquatica lateralis* (Oba et al., 2013). They observed up to ~40% luciferin labeling ranging (relative to unlabeled luciferin), when cysteine and hydroquinone were co-injected. However, like the results we present here, when cysteine was injected without hydroquinone or benzoquinone, they only observe the +3 Da ($+^{13}C_3$) signal indicative of cysteine incorporation into the luciferin thiazoline, and they did not observe the +2 Da peak ($^{13}C_2$) indicative of cysteine incorporation into the luciferin thiazoline.

The literature on pheomelanin polymerization chemistry, which used benzoquinone and cysteine as model compounds, demonstrated that the "messy" redox chemistry of benzoquinone and cysteine could produce a huge variety of compounds (Crescenzi et al., 1988). This suggested that luciferin could be produced *in vitro* purely by the non-enzymatic reaction of benzoquinone and cysteine, but this hypothesized "messy" synthesis of luciferin was not rigorously demonstrated until the work of Kanie and colleagues, which carefully quantified this phenomena (Kanie et al., 2016). Their results showed that rather than being a simply trace phenomena, under neutral conditions cysteine and benzoquinone could make experimentally misleading quantities of luciferin (~0.1-0.45% yield)

Finally, Kanie and colleagues recently found that D_3-2-S-cysteinylhydroquinone (in our terminology, D_3-cysteinyl-quinol) injected into the *A. lateralis* could robustly trace into firefly luciferin, akin to our reported result (Figure 6A).

How does one interpret this disparate data? Clearly, the fireflies and click-beetles experiments should be regarded separately, as there is currently no evidence that luciferin biosynthesis in click-beetles and fireflies follows a homologous mechanism. Although the involvement of cysteine in the synthesis of at least the luciferin thiazoline seems clear (in both fireflies and click beetles), it is unfortunate to say that due to the contribution of the "messy" non-enzymatic synthesis of luciferin when hydroquinone or benzoquinone is involved, that the evidence for hydroquinone being a biosynthetic precursor of luciferin is not well substantiated. We believe a skeptical summary of the existing data, including our own, would be: cysteine can incorporate into the thiazoline of luciferin, and when hydroquinone or cysteinyl-quinol, the immediate downstream product of cysteine-hydroquinone coupling is included, non-enzymatic "messy-chemistry" leads to low level luciferin synthesis. This low-level luciferin synthesis, when compared terms of the signal of the stable isotope tracer relative to the quantity of unlabeled luciferin, or in absolute 'specific-incorporation' terms relative to the quantity of injected radioactive tracer, has led to an erroneous interpretation over multiple decades that the authentic *de novo* luciferin biosynthetic pathway is active, when the vast majority of the tracing signal is likely due to this "messy chemistry". One intriguing datapoint that argues against this pessimistic interpretation is the report by Oba and colleagues that ~140 pmol of arbutin, the glycosylated analog of hydroquinone, is found per *Aquatica lateralis* firefly (Oba et al., 2013). Given that about ~500 pmol of luciferin is found in adult *A. lateralis* (Oba et al., 2008), this quantity of arbutin is compatible with a potential role of hydroquinone in *de novo* luciferin synthesis.

In our experiments, although we chose to interpret our hydroquinone tracing results as essentially indicative of only "messy chemistry" luciferin synthesis, and uninterpretable from the perspective of the presence of authentic luciferin *de novo* biosynthesis, in our results we do see unambiguous signals indicating incorporation of cysteine into the benzothiazole of luciferin under near-natural tracing conditions (e.g., without hydroquinone co-injection). This signal is however, extremely limited (~0.35% relative tracing after background subtraction, Figure 8B vs 8C), and detectable only due to the very high sensitivity and mass resolution of the Q-Exactive mass spectrometer used in this study. Whether this signal of cysteine incorporation into the benzothiazole cysteine incorporation signal represents a low-level, authentic biosynthesis (e.g., arbutin being hydrolyzed to hydroquinone, and undergoing an enzyme catalyzed biosynthesis with cysteine producing luciferin), or a low-level unintended reaction, is unclear. Such unintended reactions could include cysteine being incorporated into some non-natural benzothiazole precursor, by a non-enzymatic chemical reaction, or, our $^{15}N_1{}^{13}C_3$ cysteine tracer losing a labeled carbon or nitrogen (e.g. via transamination), becoming incorporated into the luciferin thiazoline, & increasing the abundance of an isotopologue with a indistinguishable *m/z* to that of our expected $^{15}N_1{}^{13}C_2$ signal. If it is the latter case, it may even be that cysteine is not the "true" precursor of the thiazole portion of the luciferin benzothiazole, but that another sulfur containing compound, or multiple step-process where the sulfur is separated from the $^{15}N_1{}^{13}C_3$ label on cysteine, is responsible for the luciferin benzothiazole. In metazoan metabolism, cysteine is metabolically linked to all the compounds which seem most likely to be involved if cysteine itself was not directly involved (e.g. glutathione, coenzyme A), but perhaps not all labeled atoms would be conserved, e.g. ^{15}N could be lost via transamination of cysteine. Alternatively, if the sulfur chemistry of microbes is involved, for example through a stable firefly symbiont such as the tenericutes symbionts reported in *Photinus* fireflies (Fallon et al., 2018), then presumably a broader array of biosynthetic routes should be considered.

We believe that our results demonstrate that the adult and larval light organ is not the site of *de novo* luciferin biosynthesis, either under homeostatic conditions, or under conditions where the light organs have been induced to luminescence continuously for some time. It then seems clear that fundamental questions must be answered before substantial progress can be made in firefly luciferin biosynthesis. These questions come down to the when, where, and what of luciferin biosynthesis, which is to say: (1) When does *de novo* luciferin biosynthesis mainly take place? Is a particular life stage the best candidate for studying this phenomena? (2) Where does luciferin *de novo* biosynthesis take place? Is there a particular tissue or cell type that performs this task? (3) What are the precursors of firefly luciferin, including the source of the carbon as well as nitrogen-sulfur atoms of both the benzothiazole and thiazoline?

In the simpler, "localized synthesis" scenario, *de novo* luciferin biosynthesis would mainly take place in the developing adult lantern, during the pupal metamorphosis. In the absence of other evidence, we would hypothesize that the luciferin is derived from hydroquinone (stored as arbutin), that is presumably derived from tyrosine. In the Japanese firefly *A. lateralis,* luciferin levels reportedly increase ~5.7x (Niwa et al., 2006), to ~12x (Kanie et al., 2018), in the transition from larvae to adult, supporting that this could be the stage where luciferin is biosynthesized. That being said, these papers do not measure the levels of sulfoluciferin over the development of the firefly. Stable isotope tracing in pupae, either using cysteine alone, or using likely primary metabolic precursors of the luciferin benzothiazole, such as tyrosine, or glucose, could provide evidence for which compounds serve as the biosynthetic precursors of luciferin. These tracing experiments have not yet been tested in firefly pupae however, as in our case North American firefly pupae are extremely difficult to obtain. We are working with Dr. Yuichi Oba (Chubu Univ.) to establish our own colony of the lab rearable Japanese firefly, *Aquatica lateralis* strain Ikeya-Y90, but to date we have not obtained enough specimens for a reasonable stable isotope tracing experiment in pupae.

There is of course no guarantee that fireflies utilize the simplest pathway to biosynthesize luciferin. An alternative pathway is a situation we dub the "diffuse synthesis" scenario of firefly luciferin biosynthesis. In this scenario, the *de novo* biosynthesis of luciferin may take place at a low level, continuously or sporadically throughout the life of the firefly, in tissues which are not the specialized light organs. Luciferin from this low level luciferin synthesis could then be concentrated into the larval and adult light organs by transport mechanisms, either through specific expression of vectorial luciferin transporters, or through mechanisms like trapping luciferin in a state with altered transport properties, such as sulfoluciferin. This hypothesized scenario could be supported by the fact that fireflies have a dim, unlocalized luminescence, distinct from that of the specialized light organs, throughout their entire life cycle, but at its highest intensity during the pupal metamorphosis (Strause et al., 1979). We hypothesize that specialized enzymes would perform the biosynthetic reactions, but if the concept of "messy-chemistry" is extrapolated to analogous *in vivo* processes, it is possible that quinone-oxidation processes with promiscuous substrate specificities such as the melaninization-like sclerotization (tanning) of the cuticle during molts and metamorphosis (Asano et al., 2019), or in the melanin deposition upon insect wound healing (Lu et al., 2014), could provide a oxidative source producing benzoquinone for low level non-enzymatic or unspecialized luciferin synthesis.

MATERIALS AND METHODS

Firefly collection for preliminary stable isotope tracing experiments (Figure 3)

Pyractomena sp. specimens were collected as larvae and reared to adults from a collection in October 2014, from the Rock Meadow Conservation Area in Belmont, MA (42° 24' 6.65" N, 71° 11' 50.40" W). Dried *Photinus pyralis* specimens (Figure 3A) were obtained from commercial sources (P/N: FFW-5G, Sigma-Aldrich). Firefly collections from Rock Meadow were approved by the Belmont Conservation Commission. Firefly larvae were collected from ~6-inch tall grass in the Rock Meadow at night by hand on the basis of sporadic glowing behavior. Identifications of firefly genera, both as larvae and adult, were assisted by Dr. Sara Lewis (Tufts University), and through comparisons to firefly photographs on BugGuide.net. Firefly larvae were kept in continual darkness in plastic containers with airholes & moistened kimwipes. Larvae were fed on a weekly diet of moistened cat food (Friskies), as well as occasional live Bladder snails (*Physella* sp.). Food was provided to the larvae overnight, and was removed the next day. Under these conditions firefly larvae survived for multiple months, although a minority of larvae did die during rearing. Larvae were resistant to starvation for at least 1 month. No specific manipulation was made to induce pupation of the larvae. Pupation occurred stochastically after about two months in captivity and not at all in some specimens. Live adult firefly specimens were maintained in the laboratory for less than 2 weeks in petri-dishes with regularly moistened kimwipes (Kimtech) and slices of apple (replaced when browned).

Preliminary stable isotope tracing experiments (Figure 3)

Adult *Pyractomena* sp. male specimens were injected using a 701RN 10 μL syringe (Hamilton Company) with 2 μL of 550 mM free-acid $^{15}N_1{}^{13}C_3$-L-Cysteine (P/N: CNLM-3871-H, Cambridge Isotope laboratories) with 550 mM unlabeled hydroquinone (P/N: H9003, Sigma-Aldrich), or 2 μL of 550 mM unlabeled free-acid L-Cysteine (P/N: W326305, Sigma-Aldrich) with 550 mM D_6-hydroquinone (Cambridge Isotope laboratories). No pH adjustment of the injection mixtures was performed. A dried *Photinus pyralis* firefly was included as a negative control (Figure 3A), and was processed identically to the injected fireflies. After injection, fireflies were incubated overnight (~16 hours) at room temperature. The following day, the firefly was frozen in liquid N_2, the abdominal segments containing the lantern were broken off with a razor blade, and the abdomen plus lantern extracted with 150 μL 50% methanol. The extracted lantern was vortexed, placed in a water bath sonicator for 20 minutes, and centrifuged at 16,000 x g for 10 minutes. 20 μL of the centrifuged extract was separated on an UltiMate 3000 liquid chromatography system (Dionex) equipped with a 150 mm C18 Column (Kinetex 2.6 μm silica core shell C18 100Å pore, P/No. 00F-4462-Y0, Phenomenex) coupled to a Q-Exactive mass spectrometer (Thermo Scientific). Compounds were separated by reversed-phase chromatography on the C18 column by a gradient of Solvent A (0.1% formic acid in H2O) and Solvent B (0.1% formic acid in acetonitrile); 5% B for 2 min, 5-80% B over 40 min, 95% B for 4 min, and 5% B for 5 min; flow rate 0.8 mL/min.

Positive and negative ionization runs were performed in separate injections. The mass spectrometer was configured to perform 1 MS^1 scan from *m/z* 120-1250 followed by 1-3 data-dependent MS^2 scans using HCD fragmentation with a stepped collision energy of 10, 15, 25 normalized collision energy (NCE). Data was collected as profile data. The instrument was always used within 7 days of the last mass accuracy calibration. The ion source parameters were as follows: spray voltage (+) at 3000 V, spray voltage (-) at 2000 V, capillary temperature at 275 °C, sheath gas at 40 arb units, aux gas at 15 arb units, spare gas at 1 arb unit, max spray current at 100 (μA), probe heater temp at 350 °C, ion source: HESI-II. The raw data in Thermo format was converted to mzML format using ProteoWizard MSConvert (Chambers et al., 2012). Data analysis was performed with Xcalibur (Thermo Scientific), MZmine 2 (v2.38) (Pluskal et al., 2010), and custom Python 3 analyses in the Jupyter programming environment (Kluyver et al., 2016) using the pyMZML library (Kösters et al., 2018), and Matplotlib (Hunter, 2007).

Firefly collection for IT-SIAM, D_3-6-HBZ-CA, D_3-6-HBZ-CA-EE and cysteine alone tracing experiments (Figure 4, Figure 6, Figure 7)

Adult male *Pyractomena* sp., and *Photinus* sp. were collected on May 25th and June 25th 2015, from the Rock Meadow Conservation Area in Belmont, MA (42° 24' 6.65" N, 71° 11' 50.40" W). Firefly collections from Rock Meadow were approved by the Belmont Conservation Commission. Fireflies were captured in flight on the basis of flashing behavior.

Synthesis of D_3-cysteinyl-quinol-ethyl-ester, D_3-6-HBZ-CA-EE, and D_3-6-HBZ-CA

First, p-[2,3,5,6-D]-Benzoquinone (D_4-benzoquinone) was prepared from D_6-hydroquinone (Cambridge Isotope Laboratories) as previously reported (Derikvand et al., 2010). In brief, Ag_2O (270 mg) plus hydroquinone (1.5g) was dissolved in methanol (30 mL) with stirring for 5 minutes. 3.3 mL of 30% aq. H_2O_2 in methanol (45 mL) was then added dropwise to the reaction mixture with stirring of the reaction mixture. The reaction mixture was allowed to stir for 40 minutes at room temperature. The reaction mixture was diluted to 135 mL with ddH_2O, and a liquid-liquid separation was performed with diethyl ether (135 mL). The diethyl ether fraction was then decanted into a beaker and allowed to evaporate at room temp. The resulting large yellow crystals were filtered. GC-MS analysis confirmed the identity, chemical purity (>99%), and isotopic purity (>99%) of the resulting D_4-benzoquinone (data not shown). Labeled D_3-6-HBZ-CA was synthesized as previously reported (Löwik et al., 2001), starting with D_4-benzoquinone (prepared above; 0.301 g) and L-cysteine ethyl ester hydrochloride (Sigma-Aldrich; 0.51 g). Methanol in the initial reaction mixture was removed by evaporation under reduced pressure, and the reaction mixture was lyophilized, after which 1.743 g of cysteinyl-quinol-ethyl-ester was obtained as a reddish-brown solid. 763 mg of the cysteinyl-quinol-ethyl-ester was dissolved in 8.76 mL of methanol. 1M $K_3Fe(CN)_6$ (15.7 mL) was added to the reaction mixture with continuous stirring. A 1.09 mL of 4M NaOH was added to 10.9 mL of methanol, and added dropwise into the reaction mixture, and then the reaction was incubated at room temperature for 2 hours. The reaction mixture was then diluted to 100 mL, and extracted 3 times with 1 volume of ethyl acetate. The combined ethyl acetate fractions were washed with concentrated brine, and dried with $MgSO_4$. Evaporation under reduced pressure gave a solid 1.247 g of a dark red-brown solid. This solid was dissolved in 1:1 MeOH:$CHCl_3$ and fractionated isocratically with 1:1 MeOH:$CHCl_3$ on a 40 cm x 5 cm diameter LH20 column. Resulting LH20 fractions were checked via TLC using F_{254} plates. The fractions with high purity 6-HBZ-CA-EE were combined and evaporated under reduced pressure, followed by drying with nitrogen gas. 34.9 mg of 6-HBZ-CA-EE were obtained. 6-HBZ-CA was hydrolyzed from 6-HBZ-CA-EE following previously reported procedures (Löwik et al., 2001). 30 mg of 6-HBZ-CA was obtained.

Replicated intersectional-tracing stable-isotope-assisted-metabolomics (IT-SIAM) tracing experiments (Figure 4, Figure 6, Figure 7)

Adult male *Pyractomena* sp. and *Photinus* sp. fireflies were injected using a 701RN 10 μL syringe (Hamilton Company) with 5 μL of tracing solutions. The tracing solutions included (1) 50 mM D_3-6-HBZ-CA in 1x phosphate buffered saline (PBS), (2) 50 mM D_3-6-HBZ-CA with 50 mM unlabeled L-cysteine in 1x PBS, (3) 100 mM $^{15}N^{13}C_3$ L-cysteine in 1x PBS, (4) 100 mM $^{15}N^{13}C_3$ L-cysteine with 100 mM unlabeled hydroquinone in 1X PBS, (5) 100 mM D_6-hydroquinone in 1x PBS, and (6) 100 mM D_6-hydroquinone with 100 mM unlabeled L-cysteine in 1x PBS. The number of biological replicates (independently injected firefly specimens) for each condition was 3, 6, 3, 3, 3, and 2, respectively. In a separate experiment 1 µL of 550 mM cysteinyl-quinol-ethyl-ester (cysteinyl-quinol-EE) was injected in two separate biological replicates. Fireflies were incubated overnight (~16 hours) at room temperature. The following day, the fireflies were frozen in liquid N_2, the abdominal segments containing the lantern were broken off with a razor blade, and the abdomen plus lantern extracted with 150 µL 50% methanol. The extracted lantern was vortexed, placed in a water bath sonicator for 20 minutes, and centrifuged at 16,000 x g for 10 minutes and filtered through a 0.2 µm PTFE filter (Filter Vial, P/No. 15530-100, Thomson Instrument Company). 20 µL of this filtered extract was separated on an UltiMate 3000 liquid chromatography system (Dionex) equipped with a 150 mm C18 Column (Kinetex 2.6 µm silica core shell C18 100Å pore, P/No. 00F-4462-Y0, Phenomenex) coupled to a Q-Exactive mass spectrometer (Thermo Scientific). Compounds were separated by reversed-phase chromatography on the C18 column by a gradient of Solvent A (0.1% formic acid in H2O) and Solvent B (0.1% formic acid in acetonitrile); 5% B for 2 min, 5-80% B over 40 min, 95% B for 4 min, and 5% B for 5 min; flow rate 0.8 mL/min.

Positive and negative ionization runs were performed in separate injections. The mass spectrometer was configured to perform 1 MS^1 scan from *m/z* 120-1250 followed by 1-3 data-dependent MS^2 scans using HCD fragmentation with a stepped collision energy of 10, 15, 25 normalized collision energy (NCE). Data was collected as profile data. The instrument was always used within 7 days of the last mass accuracy calibration. The ion source parameters were as follows: spray voltage (+) at 3000 V, spray voltage (-) at 2000 V, capillary temperature at 275 °C, sheath gas at 40 arb units, aux gas at 15 arb units, spare gas at 1 arb unit, max spray current at 100 (µA), probe heater temp at 350 °C, ion source: HESI-II. The raw data in Thermo format was converted to mzML format using ProteoWizard MSConvert (Chambers et al., 2012). Data analysis was performed with Xcalibur (Thermo Scientific), MZmine 2 (v2.38) (Pluskal et al., 2010). A custom MZmine 2 batch-mode based mzML to joined mzTab feature calling pipeline with chromatogram deconvolution was parallelized on the high-performance computing cluster of Whitehead Institute using NextFlow (Di Tommaso et al., 2017). The intersectional tracing stable isotope assisted metabolomics analysis of the resulting mzTab file was implemented via a custom program in a Python3 Jupyter programming environment (Kluyver et al., 2016) using the pandas library (McKinney and Others, 2010), and Matplotlib (Hunter, 2007).

Firefly collection for stable isotope tracing in explanted *Photinus pyralis* adult male firefly light organs (Figure 8)

Photinus pyralis adult males were collected as adults on August 2nd, 2017, from the Fred Wolfe Park Soccer Fields in New Haven, CT (41°16'14.7"N 73°01'57.2"W). Adult males were captured in flight, on the bases of the characteristic rising “J” flash of *P. pyralis*.

Isolation of *Photinus pyralis* adult male photophores (Figure 8)

To remove potentially adherent microorganisms and/or the waxy cuticle layer, adult *P. pyralis* fireflies were first placed in a 1.5 mL eppendorf tube with a 1 mL cleaning mixture of artificial freshwater (distilled water) with 1% no-tears shampoo (Johnson and Johnson), and vortexed for 30 seconds. The cleaning mixture was then decanted, and 1 mL of ddH_2O was added to wash the firefly. The above cleaning procedure was repeated a total of 2 times. After the initial cleaning procedure, 1.5 mL of ddH_2O was placed in the eppendorf tube, and a moderate pressure was applied to the tube cap with the investigators finger, for approximately 2 minutes, until the “drowned” firefly stopped moving, which we interpret as water pressure from the investigator’s finger compressing and/or displacing the air in the spiracles of the firefly respiratory system, leading to a lack of available O_2 and eventual cessation of movement. We then decanted the water, and added 1 mL of 70% ethanol. The firefly was vortexed for 30 seconds, and then the ethanol was decanted. Anecdotally, the previously described "drowning" procedure improved survival of the fireflies after exposure to ethanol, perhaps due to a protective effect of ddH_2O that had been forced into the spiracles. The firefly was then washed with ddH_2O twice, and placed on a dry kimwipe in a closed plastic box. Typically the firefly would recover and begin moving again within 10 minutes. The firefly was then transferred to a 2.5 cm sterile tissue culture dish, and the anterior abdominal segments containing the light organs were torn from the rest of the specimen using Dumont #5 biology thickness dissecting forceps (P/N: 11252-40, Fine Scientific Tools). The abdominal fragment was then placed in a 1.5 mL eppendorf with 500 µL EX-CELL 420 media with L-glutamine (P/N: 14420C, Sigma-Aldrich) with 100 µg / mL Primocin antibiotic mixture (InvivoGen), and shaken vigorously to remove the weakly bound fatbody. The tissue containing the light organs was then moved back to the 2.5 cm petri dish, with 500 µL of media + antibiotics, however the media was solely for humidity in the dish, as it was easier to dissect the tissue when it was not directly contained in the media. The tissue was then dissected using the dissecting forceps, taking special care to avoid the light organ and to not grasp the tissue close to the light organ. After removal of the extraneous tissue, the edge of the cuticle would often become separated from the light organ surface, and could be carefully peeled off. We dub this isolated light organ after cuticle removal, the photophore. Once the photophore was isolated from the majority of the extraneous tissue, and the cuticle had been removed, it was transferred to a new 500 µL aliquot of media + antibiotic, and shaken vigorously to attempt to remove the remaining pieces of adherent tissue. The media with suspended tissue fragments (not including the photophore), was aspirated with a P200 pipette tip attached to a vacuum source. The new-media transfer, shaking, and aspiration procedure described above was repeated again, for a total of 2 times, then the cleaned photophores (1 anterior and 1 posterior, per specimen) were transferred to a new 500 µL media + antibiotic aliquot.

Stable isotope tracing in in *P. pyralis* adult male photophores (Figure 8).

Cleaned anterior and posterior photophores were obtained via dissection, and placed in a single well of a polystyrene 24-well tissue culture dish with 300 µL of EX-CELL 420 media with L-glutamine (P/N: 14420C, Sigma-Aldrich) and 100 µg/mL Primocin antibiotic mixture (InvivoGen). Three tracing conditions were used: media with antibiotics and heavy $^{15}N_1{}^{13}C_3$ L-cysteine at a 10 mM concentration, media with antibiotics and heavy D_4-tyrosine at a 10 mM concentration, and media with antibiotics and no added tracer. There were 3 biological replicates per condition, with 2 photophores (the anterior and posterior photophores from the same individual) per biological replicate. The explanted photophores in media with tracers were incubated in a dark room in a sealed humid box at room temperature (~22°C) for 10 hours. After 10 hours, it was observed that the investigator's shaking of the tissue culture dish induced an elevated level of luminescence, likely due to an increased level of oxygen in the media, so for the remaining 6 hours of the experiment, the tissue was shaken gently at 60 RPM in a 25°C incubator, under ambient light conditions. Observation at 14 hours indicated that the explanted photophores of each replicate of the 3 biological conditions were still glowing. At 16 hours after the start of the experiment, the photophores from each biological replicate were moved to independent 1.5 mL eppendorf tubes, and 60 µL 50% methanol was added. The tissues in methanol were water bath sonicated for 10 minutes, vortexed at max speed for 30 seconds, and water bath sonicated again for 10 minutes, after which all the tissue appeared to be dispersed. The tissue extract was centrifuged for 10 minutes at 16,000 x g, and filtered through a 0.2 µm PTFE filter (Filter Vial, P/No. 15530-100, Thomson Instrument Company). 5 µL injections of these filtered extracts were separated and analyzed using an UltiMate 3000 liquid chromatography system (Thermo Scientific) equipped with a 150 mm C18 Column (Kinetex 2.6 µm silica core shell C18 100 Å pore, P/No. 00F-4462-Y0, Phenomenex, USA) coupled to a Q-Exactive mass spectrometer (Thermo Scientific). Compounds were separated via reversed-phase chromatography on the C18 column using a gradient of Solvent A (0.1% formic acid in H_2O) and Solvent B (0.1% formic acid in acetonitrile); 5% B for 2 min, 5–40% B until 20 min, 40–95% B until 22 min, 95% B for 4 min, and 5% B for 5 min; flow rate 0.8 mL/min. The mass spectrometer was configured to perform one MS^1 scan from *m/z* 120–1250 followed by 1 data-dependent MS^2 scans using HCD fragmentation with a stepped collision energy of 10, 15, 25 normalized collision energy (NCE). Positive mode and negative mode MS^1 and MS^2 data were obtained in a single run via polarity switching. Data was collected as profile data. The instrument was always used within 7 days of the last mass accuracy calibration. The ion source parameters were as follows: spray voltage (+) at 3000 V, spray voltage (-) at 2000 V, capillary temperature at 275˚C, sheath gas at 40 arb units, aux gas at 15 arb units, spare gas at one arb unit, max spray current at 100 (μA), probe heater temp at 350˚C, ion source: HESI-II. The raw data in Thermo format was converted to mzML format using ProteoWizard MSConvert (Chambers et al., 2012). Data analysis was performed with Xcalibur (Thermo Scientific) and MZmine 2 (v2.38) (Pluskal et al., 2010).

Firefly collection for stable isotope tracing in firefly larvae (Figure 9)

P. pyralis larvae were raised from eggs laid by 2 adult female *P. pyralis* mated with 2 adult males, which were collected on July 27th, 2017 from Waveny Park in Fairfield CT (41°07'09.9"N 73°29'32.9"W). Males were captured during flight on the basis of their rising "J" flash, whereas females were collected from the grass below by mimicking the *Photinus pyralis* adult male "J" advertising flash with a penlight, thereby stimulating the stereotyped female reponse flash and allowing collection. The *P. pyralis* larvae were raised to their ~4th instar following a previously published draft rearing protocol (Fallon et al., 2018). Approximately 4th instar *Aquatica ficta* larvae were obtained from the entomology department of the Taipei Zoo, Taipei, Taiwan, and kept in the laboratory in 15 cm petri dishes with distilled water with 1:1000 diluted artificial seawater. *A. ficta* were periodically fed frozen bloodworms (Hikari).

Stable isotope tracing in Lampyrinae and Luciolinae larvae stimulated to luminesce continuously (Figure 9)

Three *Photinus pyralis* and three *Aquatica ficta* ~4th instar larvae were injected using a 701RN 10 µL syringe (Hamilton Company) with 2 µL of 750 mM heavy ($^{15}N_1{}^{13}C_3$) L-cysteine in 0.5x PBS. Larvae were temporarily restrained using double-stick tape on glass microscope slides, and injections were made on the ventral-lateral surface, posterior to the most posterior legs. Post injection, larva were allowed to recover for 5-10 minutes to allow the injection wound to clot, then placed in a ~20 mg suspension of DL-synephrine (P/N: S0752, Sigma-Aldrich) in 1 mL distilled water. The larvae was shaken until they were fully coated in the DL-synephrine suspension. Larvae were left in the synephrine suspension for 10-30 minutes, after which they were individually transferred to a single well of a polystyrene 6-well tissue culture dish, which was placed in a humid box within a dark room for overnight incubation. As the larvae were moved without washing, a "thin-film" of synephrine was left on their body. 11 hours after the first injection, the *Photinus pyralis* larvae were still observed to be glowing, however the *Aquatica ficta* larvae were no longer glowing. At 13 hour 35 minutes after the first injection, only 2 of the *P. pyralis* larvae were still observed to be glowing, and at this point all larvae were placed back in the synephrine suspension. After 5 minutes of exposure to synephrine, the *P. pyralis* larvae again began to glow, but the *A. ficta* larvae did not start glowing again. At the 16 hour mark, the posterior 2 abdominal fragments containing the larval light organ (dubbed the "tail") were removed with a razor blade, and placed in 60 µL 50% acetonitrile (ACN). The tail extract was sonicated in a water bath sonicator for 10 minutes, vortexed, sonicated again, centrifuged at 16,000 x g, and filtered through a 0.2 µm PTFE filter (Filter Vial, P/No. 15530-100, Thomson Instrument Company). 5 µL injections of these filtered extracts were separated and analyzed using an UltiMate 3000 liquid chromatography system (Thermo Scientific) equipped with a 150 mm C18 Column (Kinetex 2.6 µm silica core shell C18 100 Å pore, P/No. 00F-4462-Y0, Phenomenex, USA) coupled to a Q-Exactive mass spectrometer (Thermo Scientific, USA). Compounds were separated via reversed-phase chromatography on the C18 column using a gradient of Solvent A (0.1% formic acid in H_2O) and Solvent B (0.1% formic acid in acetonitrile); 5% B for 2 min, 5–40% B until 20 min, 40–95% B until 22 min, 95% B for 4 min, and 5% B for 5 min; flow rate 0.8 mL/min. The mass spectrometer was configured to perform one MS^1 scan from *m/z* 120–1250 followed by 1 data-dependent MS^2 scans using HCD fragmentation with a stepped collision energy of 10, 15, 25 normalized collision energy (NCE). Positive mode and negative mode MS^1 and MS^2 data were obtained in a single run via polarity switching. Data was collected as profile data. The instrument was always used within 7 days of the last mass accuracy calibration. The ion source parameters were as follows: spray voltage (+) at 3000 V, spray voltage (-) at 2000 V, capillary temperature at 275°C, sheath gas at 40 arb units, aux gas at 15 arb units, spare gas at one arb unit, max spray current at 100 (μA), probe heater temp at 350°C, ion source: HESI-II. The raw data in Thermo format was converted to mzML format using ProteoWizard MSConvert (Chambers et al., 2012). Data analysis was performed with Xcalibur (Thermo Scientific) and MZmine 2 (v2.38) (Pluskal et al., 2010).

REFERENCES

Asano T, Seto Y, Hashimoto K, Kurushima H. 2019. Mini-review an insect-specific system for terrestrialization: Laccase-mediated cuticle formation. *Insect Biochem Mol Biol* **108**:61–70.

Carlson AD. 1972. A comparison of transmitter and synephrine on luminescence induction in the firefly larva. *J Exp Biol* **57**:737–743.

Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak M-Y, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P. 2012. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **30**:918–920.

Colepicolo P, Pagni D, Bechara EJH. 1988. Luciferin biosynthesis in larval Pyrearinus termitilluminans (Coleoptera: elateridae). *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry* **91**:143–147.

Crescenzi O, Prota G, Schultz T, Wolfram LJ. 1988. The reaction of cysteine with 1, 4-benzoquinone: a revision. *Tetrahedron* **44**:6447–6450.

David Morgan E. 2007. Biosynthesis in Insects. Royal Society of Chemistry.

Day JC, Tisi LC, Bailey MJ. 2004. Evolution of beetle bioluminescence: the origin of beetle luciferin.

Luminescence **19**:8–20.

Derikvand F, Bigi F, Maggi R, Piscopo CG, Sartori G. 2010. Oxidation of hydroquinones to benzoquinones with hydrogen peroxide using catalytic amount of silver oxide under batch and continuous-flow conditions. *J Catal* **271**:99–103.

Dettner K. 1987. Chemosystematics and Evolution of Beetle Chemical Defenses. *Annu Rev Entomol* **32**:17–48.

Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017. Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35**:316–319.

Fallon TR, Li F-S, Vicent MA, Weng J-K. 2016. Sulfoluciferin is Biosynthesized by a Specialized Luciferin Sulfotransferase in Fireflies. *Biochemistry* **55**:3341–3344.

Fallon TR, Lower SE, Chang C-H, Bessho-Uehara M, Martin GJ, Bewick AJ, Behringer M, Debat HJ, Wong I, Day JC, Suvorov A, Silva CJ, Stanger-Hall KF, Hall DW, Schmitz RJ, Nelson DR, Lewis SM, Shigenobu S, Bybee SM, Larracuente AM, Oba Y, Weng J-K. 2018. Firefly genomes illuminate parallel origins of bioluminescence in beetles. *Elife* **7**. doi:10.7554/eLife.36495

Ghiradella H, Schmidt JT. 2004. Fireflies at one hundred plus: a new look at flash control. *Integr Comp Biol* **44**:203–212.

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* **9**:90–95.

Kanie S, Nakai R, Ojika M, Oba Y. 2018. 2-S-cysteinylhydroquinone is an intermediate for the firefly luciferin biosynthesis that occurs in the pupal stage of the Japanese firefly, Luciola lateralis. *Bioorg Chem*. doi:10.1016/j.bioorg.2018.06.028

Kanie S, Nishikawa T, Ojika M, Oba Y. 2016. One-pot non-enzymatic formation of firefly luciferin in a neutral buffer from p-benzoquinone and cysteine. *Sci Rep* **6**:24794.

Kluyver T, Ragan-Kelley B, Pérez F, Granger BE, Bussonnier M, Frederic J, Kelley K, Hamrick JB, Grout J, Corlay S, Others. 2016. Jupyter Notebooks-a publishing format for reproducible computational workflowsELPUB. pp. 87–90.

Kösters M, Leufken J, Schulze S, Sugimoto K, Klein J, Zahedi RP, Hippler M, Leidel SA, Fufezan C. 2018. pymzML v2.0: introducing a highly compressed and seekable gzip format. *Bioinformatics* **34**:2513–2514.

Löwik DW, Tisi LC, Murray JAH, Lowe CR. 2001. Synthesis of 6-hydroxybenzothiazole-2-carboxylic acid. *Synthesis* **2001**:1780–1783.

Lu A, Zhang Q, Zhang J, Yang B, Wu K, Xie W, Luan Y-X, Ling E. 2014. Insect prophenoloxidase: the view beyond immunity. *Front Physiol* **5**:252.

McCapra F, Razavi Z. 1975. A model for firefly luciferin biosynthesis. *J Chem Soc Chem Commun* 42b–43.

McKinney W, Others. 2010. Data structures for statistical computing in pythonProceedings of the 9th Python in Science Conference. Austin, TX. pp. 51–56.

Niwa K, Nakamura M, Ohmiya Y. 2006. Stereoisomeric bio-inversion key to biosynthesis of firefly d-luciferin. *FEBS Lett* **580**:5283–5287.

Oba Y, Shintani T, Nakamura T, Ojika M, Inouye S. 2008. Determination of the luciferin contents in luminous and non-luminous beetles. *Biosci Biotechnol Biochem* **72**:1384–1387.

Oba Y, Yoshida N, Kanie S, Ojika M, Inouye S. 2013. Biosynthesis of firefly luciferin in adult lantern: decarboxylation of L-cysteine is a key step for benzothiazole ring formation in firefly luciferin synthesis. *PLoS One* **8**:e84023.

Okada K, Iio H, Goto T. 1976. Biosynthesis of firefly luciferin. Probable formation of benzothiazole from p-benzoquinone and cysteine. *J Chem Soc Chem Commun* 32–32.

Pluskal T, Castillo S, Villar-Briones A, Oresic M. 2010. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**:395.

Strause LG, DeLuca M, Case JF. 1979. Biochemical and morphological changes accompanying light organ development in the firefly, Photuris pennsylvanica. *J Insect Physiol* **25**:339–347.

CHAPTER 5.

Discussion and future directions

In this thesis I have investigated four unanswered questions in firefly bioluminescence: (1) Do fireflies possess a storage form of their luciferin? (2) What is the evolutionary relationship of bioluminescence amongst the bioluminescent beetles, and has this trait independently evolved multiple times? (3) How is firefly luciferin biosynthesized? And finally (4) are there accessory genes from the bioluminescent beetles which act in bioluminescent metabolism, and if so, might they be useful for biotechnological applications?

For the first question, my discovery of sulfoluciferin and luciferin-sulfotransferase as described in Chapter 2, suggests that fireflies, like certain other bioluminescent organisms, do store their luciferin as a sulfonated form. But unlike the sulfonated version of the marine luciferin coelenterazine, coelenterazine enol-sulfate, which is made considerably more stable to air oxidation by sulfonation (Hori et al., 1972), I found that sulfoluciferin is not noticeably more stable than luciferin to air oxidation, and although careful quantitation was not performed, *in vitro* it appears to be less stable to air oxidation than luciferin. This begs the question if the description of sulfoluciferin as a storage form is appropriate. I believe it is, as we found that sulfoluciferin is more abundant than luciferin in fireflies in absolute molar terms. Furthermore, air-stability alone may not be a relevant characteristic for a compound to serve as a storage form. Sulfonation could have multiple benefits, such as changing the membrane transport or diffusion characteristics of the compound, thereby helping to trap the compound in a particular cell, or a particular membrane compartment.

An alternative hypothesis to sulfoluciferin simply being a storage form, is that sulfonated luciferin metabolites act as intermediates, e.g., in the unknown oxyluciferin recycling pathway of firefly luciferin, perhaps sulfo-oxyluciferin is the substrate in a recycling pathway, which later leads to sulfoluciferin, and then luciferin. Or, if dehydroluciferin is able to be reduced into luciferin, perhaps sulfo-dehydroluciferin is the appropriate substrate. Based on structural considerations, especially given that LST operates on both L-luciferin and D-luciferin, it seems reasonable to assume that LST could also operate on compounds which are structurally related to luciferin, such as dehydroluciferin, and oxyluciferin, and thereby catalyze the interconversion between their respective sulfonated forms. Although I have not detected these hypothesized alternate forms of sulfonated luciferin, if they are truly intermediates in physiologically important pathways, their homeostatic concentration may be extremely low. Furthermore, sulfonated compounds behave poorly in the reversed-phase chromatography conditions that I have employed throughout this work, so it may be difficult to detect such compounds as a discrete chromatographic peak. As described in Chapter 3, the high expression of the LST in the *Photinus pyralis* and *Aquatica lateralis* adult male light organ, combined with high expression of the LST-cofactor PAP synthesizing enzyme adenylyl-sulfate kinase and sulfate adenylyltransferase enzyme (ASKSA), suggests that flux through LST is important in the adult male light organ, but Chapter 4 demonstrated that the adult male light organ does not appreciably *de novo* biosynthesize firefly luciferin. Therefore LST / sulfonated forms of luciferin are participating in at least a non-*de novo* biosynthesis pathway. Ultimately, the cleanest test of LST function would come from a RNAi knock-down or CRISPR/Cas9 knockout experiment. If LST were in fact essential to an easily recognized process in the firefly lantern, such as catabolism or recycling of a competitive inhibitor of luciferase (e.g. dehydroluciferin or oxyluciferin), such genetic experiments should produce a clean phenotype between conditions.

For the second question, in Chapter 3 I considered the shared or parallel origins of the firefly and click beetle bioluminescent systems, and ultimately concluded the luciferases of fireflies and click-beetles neofunctionalized independently. Although Darwin was remarkably prescient in "The Origin of Species" when he declared that the luminous organs amongst the various families of beetles were not inherited from a common progenitor (Charles Darwin, 1872), the question of whether the four families of bioluminescent beetles (Lampyridae, Phengodidae, Rhagophthalmidae, Elateridae), were independent or homologous continued to be a somewhat unclear in the modern literature, even amongst experts in the field. This lack of clarity is somewhat unsurprising, as firefly and click beetle luciferins are structurally identical, and their luciferases extremely similar. Indeed, our genomic analyses demonstrated that when the firefly and click beetle luciferases were put in the context of the full phylogenetic tree of the peroxisomal acyl-CoA synthetase (PACS) enzymes detected in all 3 genomes, although the luciferases were clearly independently neofunctionalized, they were descended from relatively closely related branch of the greater PACS family. Our results demonstrate that genomic comparisons and selection analyses are effective tools to elucidate difficult cases of molecular evolution where parallel evolution of a given activity arose from relatively closely related ancestral genes.

For the third question, my stable isotope tracing in live fireflies as described in Chapter 3 argued that there is, so far, no evidence that the adult or larval light organs of fireflies *de novo* biosynthesize large quantities of luciferin. In contrast, we found a robust activity which exchanged cysteine into the thiazoline of luciferin. This activity was also moderately enhanced upon stimulation of light emission. The simplest interpretation of this thiazoline incorporation explained by the recycling of oxyluciferin back into luciferin, as has been suggested by Okada and colleagues (Okada et al., 1974). Such an oxyluciferin recycling activity would be biotechnologically valuable, as it could lead to increased efficiencies of luminescence in transgenic applications of luciferase. That being said, the existence of a recycling pathway at all has been somewhat doubted in the past. In the recorded notes of the question and answer session after the first presentation of oxyluciferin recycling by Okada and colleagues (Cormier et al., 1973), which presumably later led to their 1974 paper (Okada et al., 1974), Seliger countered that adult fireflies emerge with about 10^{15} quanta of luciferin (~1 nmol) and ~10^{15} quanta of luciferase (~1 nmol), and according to his extrapolation, will emit about 10^{15} photons of light over the course of their adult lifespan. Therefore a recycling pathway for oxyluciferin, or presumably a *de novo* synthesis pathway for luciferin, would not be necessary in adult fireflies, given the large store of luciferin relative to the quantity of photons that need to be produced. More recent measurements of the absolute quantity of luciferin and luciferase within adult fireflies (Strause et al., 1979)(Oba et al., 2008), combined with absolute quantitation of light emitted during a flash (Case, 2004), are roughly consistent with Seliger's claim. Furthermore my discovery of sulfoluciferin in Chapter 2, a major unmeasured component of the total stored luciferin which naturally would have been overlooked in all past work, presumably strengthens Seliger's argument. The experimental support for oxyluciferin recycling is limited. Okada's tracing experiments injected a rather large quantity of oxyluciferin (~200 nmol), and reported a relatively limited specific incorporation into luciferin (~0.4% *in vivo* after 6 hours). Follow up cell-free extracts experiments demonstrated that the oxyluciferin to luciferin recycling rate did not change quantity of extract in the experiment was changed (Okada et al., 1974), arguing that the observed recycling may be a non-enzymatic phenomena. Okada and colleagues also established that at neutral pH oxyluciferin would non-enzymatically degrade to the nitrile compound 2-cyano-6-hydroxybenzothiazole (CHBT) with a half-life of approximately 2.5 hours. Given that CHBT and cysteine can rapidly couple in high-yield and at near neutral pH to form luciferin (White et al., 1963), a fully non-enzymatic pathway which recycles from luciferin from oxyluciferin is possible, further complicating interpretation of their *in vivo* tracing experiments. Falsification or support of the hypothesized oxyluciferin recycling activity may be most easily obtained by experiments to determine what happens to oxyluciferin post-oxidation *in vivo*. Something should happen to oxyluciferin, at is a competitive inhibitor of luciferase, and it is made in stoichiometric quantities during the luminescent reaction. The possibilities for oxyluciferin's fate include recycling to luciferin, catabolism (breakdown to small molecule products), and storage or export (presumably as large molecule conjugates like glucosides or sulfonated forms). Experiments in dissected firefly light organs which have been stimulated to luminescence, such as I demonstrated in Chapter 4, combined with time courses of absolute quantitation of luciferin, sulfoluciferin, and approaches such as untargeted LC-HRAM-MS, NMR, provide an experiment platform which should allow for careful quantitation of the fate of oxyluciferin *in vivo*, thereby falsifying or supporting the hypotheses of the biotechnological valuable oxyluciferin recycling pathway.

For the fourth question, I stipulated the question of whether accessory metabolic genes of the firefly light organ might be useful for biotechnology. Amongst the 4 questions posed, this is the question which had the least direct experimentation directed towards it, but the one which I believe is the most promising direction for future research. In some sense, the discovery of LST in Chapter 2, represents just such a hypothesized accessory metabolic gene, however, to date there has been no demonstration of LST's direct use in biotechnology. Given that the firefly bioluminescent system has been put to some rather creative uses, such as the use of luciferase for the detection of pyrophosphatase release in 454 "pyrosequencing" (Ronaghi, 2001), or the directed evolution of luciferase to use various luciferin analogs and thereby produce altered emission wavelengths (Kaskova et al., 2016), I believe sulfoluciferin and LST can be put to use to enhance applications of firefly bioluminescence. I propose one such application here: it may be possible to evolve a luciferase to use sulfoluciferin as a substrate with an altered emission color. Although sulfoluciferin was shown to not be an efficient substrate for firefly luciferase in Chapter 2, that fact does not rule out that an evolved variant of luciferase could produce light from sulfoluciferin. Given the electron withdrawing nature of the sulfo group of sulfoluciferin (apparent in the chemical shift of the phenolic hydrogens in ^{1}H NMR spectra of sulfoluciferin), it is likely that sulfoluciferin would have an altered fundamental emission wavelength from D-luciferin. Other luciferin-analogs which have steric hindrances in the same location of the 6' phenolic hydroxyl, such as CycLuc (Reddy et al., 2010), suggest that a sulfoluciferin-utilizing luciferase variant could be evolved.

In Chapter 3, I presented a lantern expression analysis which highlighted the enzymes that were highly expressed in the light organ, differentially expressed in the light organ when compared to a non-light organ tissue, and evolutionarily conserved between both *P. pyralis* (Lampyrinae), and *A.*

lateralis (Luciolinae) fireflies. This analyses identified several enzymes, which in addition to sharing the aforementioned expression characteristics, also had evolutionary indicators of specialization including having direct orthologs only in fireflies, and having detectable positive selection when compared to their closely related homologs within fireflies (data not shown). Work is ongoing to demonstrate possible functions for these enzymes.

REFERENCES

Case JF. 2004. Flight studies on photic communication by the firefly Photinus pyralis. *Integr Comp Biol* **44**:250–258.

Charles Darwin. 1872. The Origin of Species, 6th ed. PF Collier & Son, New York.

Cormier MJ, Hercules DM, Lee J, editors. 1973. Chemiluminescence and Bioluminescence. Springer, Boston, MA.

Hori K, Nakano Y, Cormier MJ. 1972. Studies on the bioluminescence of Renilla reniformis. XI. Location of the sulfate group in luciferyl sulfate. *Biochim Biophys Acta* **256**:638–644.

Kaskova ZM, Tsarkova AS, Yampolsky IV. 2016. 1001 lights: luciferins, luciferases, their mechanisms of action and applications in chemical analysis, biology and medicine. *Chem Soc Rev* **45**:6048–6077.

Oba Y, Shintani T, Nakamura T, Ojika M, Inouye S. 2008. Determination of the luciferin contents in luminous and non-luminous beetles. *Biosci Biotechnol Biochem* **72**:1384–1387.

Okada K, Iio H, Kubota I, Goto T. 1974. Firefly bioluminescence III. Conversion of oxyluciferin to luciferin in firefly. *Tetrahedron Lett* **15**:2771–2774.

Reddy GR, Thompson WC, Miller SC. 2010. Robust light emission from cyclic alkylaminoluciferin substrates for firefly luciferase. *J Am Chem Soc* **132**:13586–13587.

Ronaghi M. 2001. Pyrosequencing sheds light on DNA sequencing. *Genome Res* **11**:3–11.

Strause LG, DeLuca M, Case JF. 1979. Biochemical and morphological changes accompanying light organ development in the firefly, Photuris pennsylvanica. *J Insect Physiol* **25**:339–347.

White EH, McCapra F, Field GF. 1963. The Structure and Synthesis of Firefly Luciferin. *J Am Chem Soc* **85**:337–343.

APPENDIX A.

Accessory metabolic genes of firefly bioluminescence

Authors

Timothy R. Fallon[1,2], Jing-Ke Weng[1,2]

Author Affiliations

[1]Whitehead Institute for Biomedical Research, 455 Main Street, Cambridge, MA 02142
[2]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

Author contributions:

I performed experiments and analyses of all figures, and wrote the section with input and minor edits from Jing-Ke Weng.

This work is unpublished as of May 2019

Accessory genes in bioluminescent metabolism.

Firefly luciferase is widely used in biotechnology, however it is unknown if there are other accessory metabolic genes from the firefly bioluminescent system, including catabolic or anabolic enzymes, transporters, or luciferin storage proteins, which could also be useful tools for biological research. Here, we describe preliminary experiments with accessory genes in firefly metabolism, including a theorized D/L-luciferin epimerization pathway, and a reduction pathway for firefly dehydroluciferin (Figure 1).

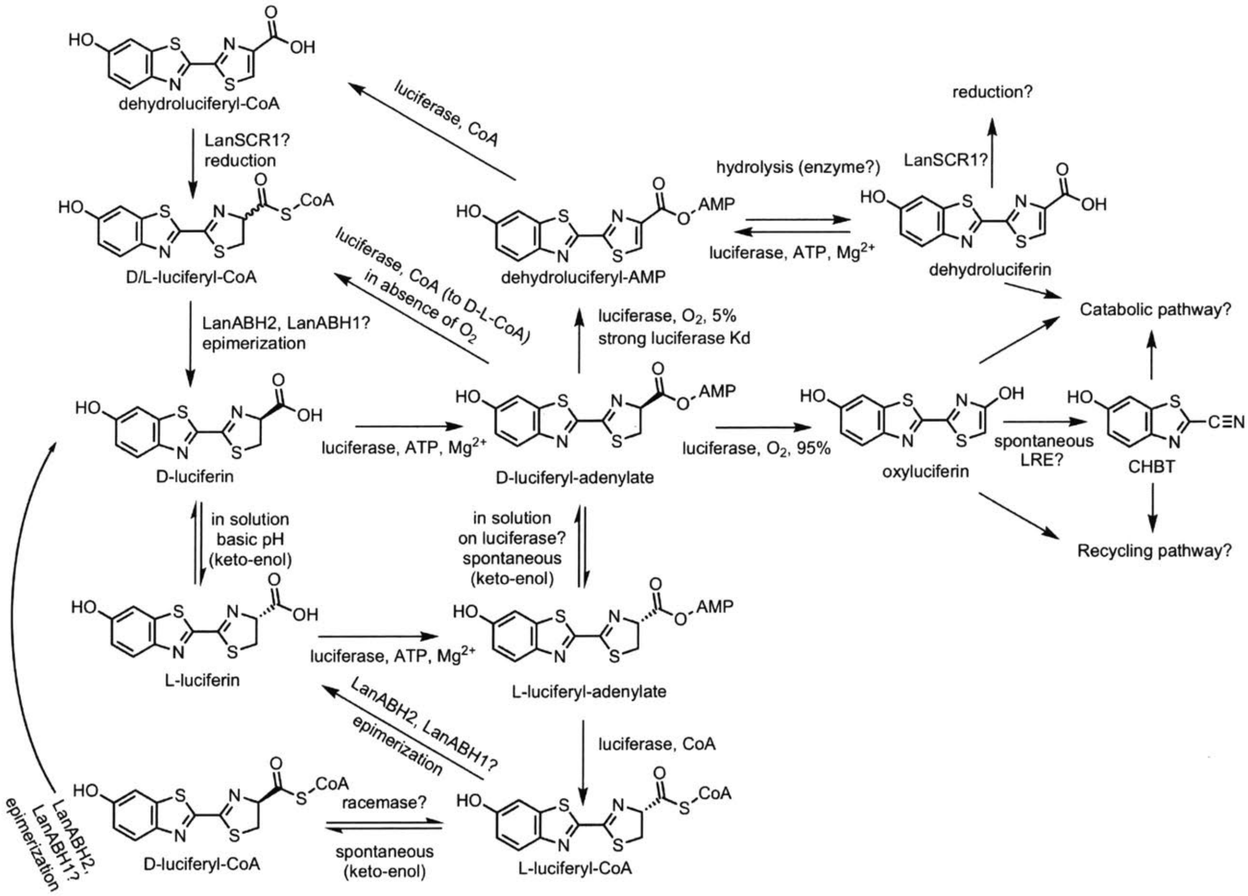


Figure 1: Known and theorized metabolic transformations of firefly luciferin. Two pathways, a L-luciferin to D-luciferin epimerization pathway, and a dehydroluciferin reductase pathway, were hypothesized.

Epimerization pathway genes in firefly luciferin metabolism

Firefly D-luciferin is active for light production, while its enantiomer L-luciferin is inactive for light production (White et al., 1963). The stereochemistry of the luciferin thiazoline ring is likely biosynthetically derived from natural L-cysteine, making it more likely that L-luciferin is the particular enantiomer produced by a *de novo* biosynthetic pathway. There is also a hypothesized luciferin recycling pathway, where the nitrile catabolic product of oxyluciferin, 2-cyano-6-hydroxybenzothiazole (CHBT), couples directly with cysteine, reforming luciferin (Okada et al., 1974). In this case the stereochemistry of the thiazoline ring is again derived from the stereochemistry of cysteine. Given that L-cysteine is the natural and greatly more common epimer of cysteine, and that fireflies do not contain large amounts of D-cysteine (Niwa et al., 2006), L-luciferin is likely made both by a recycling pathway and a *de novo* biosynthetic pathway. Therefore a L-luciferin to D-luciferin epimerization pathway should exist in fireflies to produce the D-luciferin needed for light emission. The *in vivo* presence of an ATP and CoA dependent L-luciferin to D-luciferin epimerization pathway was first demonstrated by Niwa and colleagues (Niwa et al., 2006). As luciferase is able to catalyze the formation of CoA thioesters (Oba et al., 2003), including the formation of L-luciferyl-CoA from L-luciferin (Nakamura et al., 2005), it was hypothesized that enhanced epimerization of L-luciferyl-CoA, either through non-enzymatic keto-enol tautomerization, or through catalytic racemization, could allow for facile racemization of luciferyl-CoA. This production of racemic luciferyl-CoA, when combined with a thioesterase with a preference for D-luciferyl-CoA hydrolysis, could then lead to the accumulation of D-luciferin. More recent work has further supported this model through the reconstitution of L-luciferin to D-luciferin epimerization activity using a combination of luciferase and non-firefly thioesterases and racemases (Maeda et al., 2017). However, the native enzymes of the firefly which might mediate this epimerization activity *in vivo* remain unknown. In (Fallon et al., 2018), a combined expression analysis was used to highlight genes that had similar characteristics to LST (Fallon et al., 2016) and luciferase (Wood et al., 1984), which to date are the only genes that have unambiguous roles in firefly luciferin metabolism. Within this list of enzymes was an alpha-beta hydrolase (PPYR_06194-PA) that was a potential for a candidate D-luciferyl-CoA thioesterase. We dubbed this enzyme lantern alpha-beta hydrolase 1 (LanABH1). A related paralog of LanABH1, LanABH2 (PPYR_10586-PA), was also highly expressed in the adult male *Photinus pyralis* firefly lantern, although it was not contained in the more stringent list of candidate enzymes from (Fallon et al., 2018). We therefore sought to test if LanABH1 and LanABH2 were luciferyl-CoA thioesterases. LanABH1 and LanABH2 were cloned from cDNA, and expressed recombinantly in *E. coli*, however only a small quantity of LanABH1 was obtained, whereas workable quantities of LanABH2 could not be obtained (Figure 2).

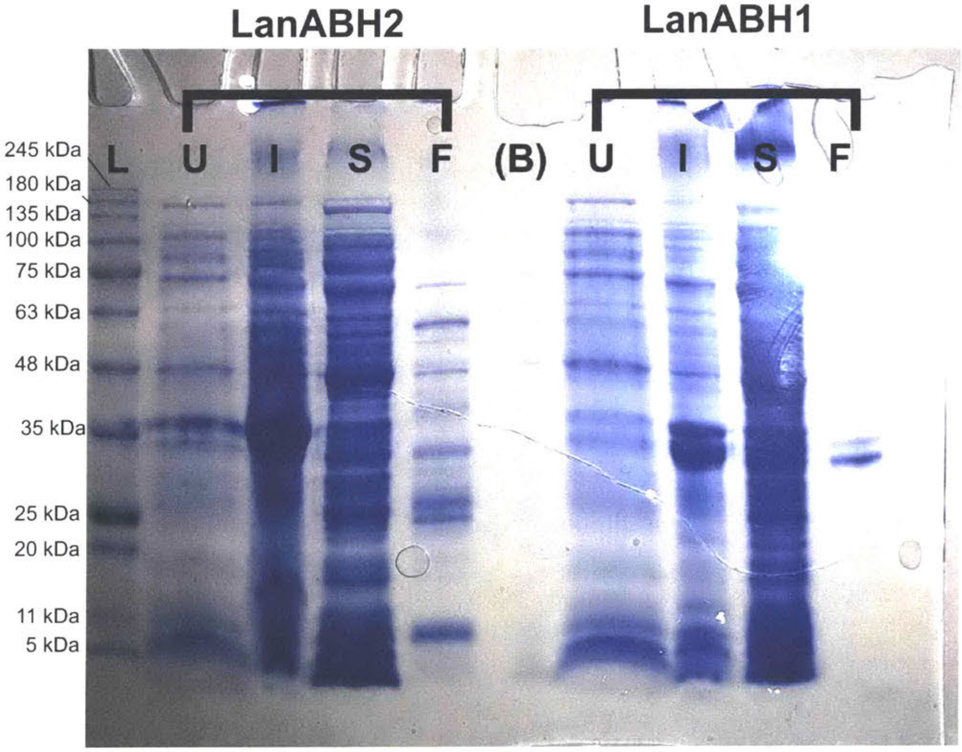


Figure 2: Recombinant expression of LanABH1 and LanABH2. L=Protein molecular weight standard ladder, U=uninduced - *E. coli* culture before addition of IPTG. I=insoluble - protein extract from centrifuged particulate post induction & E. coli lysis , S=soluble - supernatant from post induction E. coli lysis, F=final purified protein - post Ni-NTA IMAC and Sephadex size-exclusion chromatography. (B)=blank lane. Ladder is the Gold Biotechnology BlueStain Protein ladder (P/N: P007-500). Note the

large bands in the insoluble fraction compatible with LanABH1 (36.3 kDa) and LanABH2 (36.7 kDa), but the negligible soluble protein. The two bands in the LanABH1 final protein represent protein with and without the 6x histidine tag, which is intended to be cleaved off in our protein purification protocol but in this case did not go to full completion.

With workable quantities of LanABH1, we then performed directed enzymology experiments with LanABH1 and enzymatically synthesized luciferyl-CoA. These experiments demonstrated that LanABH1 could hydrolyze luciferyl-CoA, and that the chirality of the resulting luciferin was D-, suggesting that LanABH1 is a D-luciferyl-CoA specific thioesterase (Figure 3). That being said, these experiments used relatively large quantities of LanABH1 (~1 mg/mL). More careful kinetic characterization of LanABH1 is needed to determine whether the activity of LanABH1 is specific and rapid, as would be expected if LanABH1 were a positively selected specialized hydrolase involved in firefly luciferin metabolism.

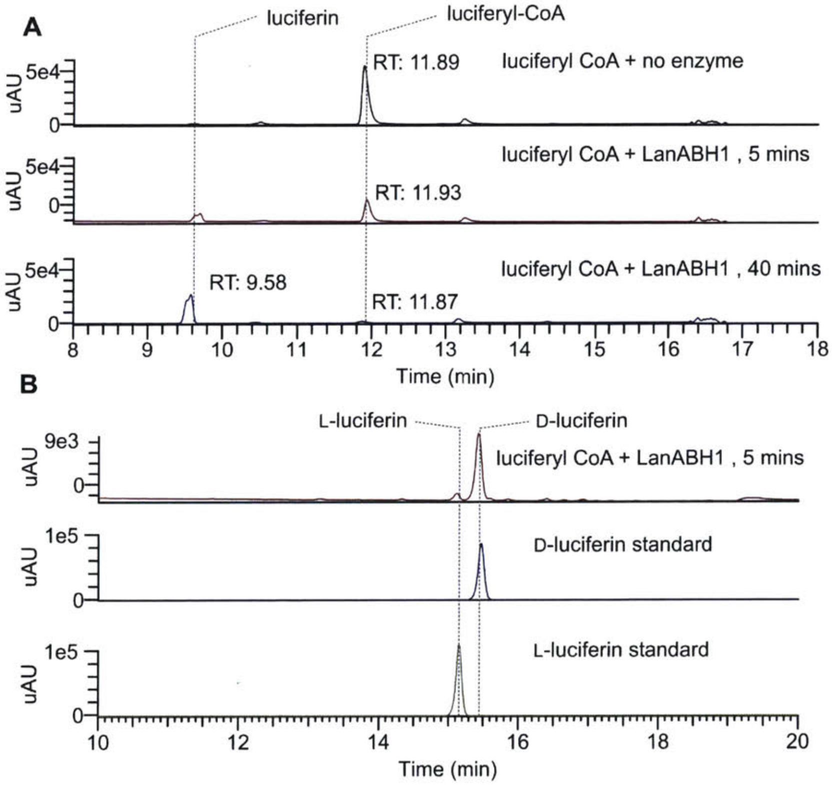


Figure 3: Test of the D-luciferyl-CoA thioesterase activity of LanABH1.
Luciferyl-CoA was first produced in an enzymatic reaction of luciferase and L-luciferin, and was provided to LanABH1. Aliquots of the reaction mixtures were taken at defined time periods, and were were assayed by (A) reverse phase chromatography, and (B) Chiral chromatography. Note that mostly D-luciferin was produced from luciferyl-CoA from LanABH1.

Beyond a D-luciferyl-CoA thioesterase, we hypothesized that if the non-enzymatic epimerization of luciferyl-CoA is not sufficient, there would also be an enzyme which catalyzes the racemization of D/L-luciferyl-CoA. The alpha-methylacyl CoA racemases, an enzyme family which operate in the catabolism of branched chain fatty acids (Lloyd et al., 2008), were strong candidates for a luciferyl-CoA racemase (Figure 4).

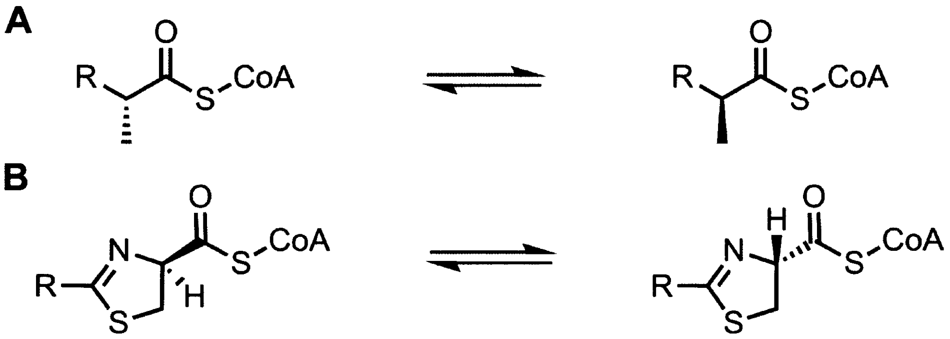


Figure 4: Chemical scheme of racemization of alpha-methylacyl CoA thioesters and luciferyl-CoA
The racemization reaction of (A) an alpha-methylacyl branched fatty-acyl CoA thioester, and (B) a luciferyl-CoA thioester, are highly similar, both consisting of a stereochemical inversion at the alpha carbon.

Within the candidate enzyme list from the *Photinus pyralis* lantern (Fallon et al., 2018), indeed there is a specialized alpha-methylacyl racemase which we dub lantern racemase (LanRac / PPYR_09240). Work is ongoing to determine if this enzyme does in fact catalyze the interconversion of D/L-luciferyl-CoA.

Dehydroluciferyl-CoA reductase genes in the firefly lantern

Dehydroluciferin (DHL) and dehydroluciferyl-AMP (DHL-AMP) are oxidized metabolites of firefly luciferin (Figure 1). DHL and DHL-AMP can both result from nonenzymatic oxidation of luciferin and luciferyl-AMP respectively, but DHL-AMP can also be produced in a non-light emitting side-reaction of luciferase (Fraga et al., 2006). Dehydroluciferyl-AMP is a strong binder and competitive inhibitor of luciferase (IC50 ~= 5 nM) (Fraga et al., 2006). The competitive inhibition of luciferase with DHL-AMP, combined its unavoidable "off-pathway" synthesis by luciferase, leads to the rapid "flash" inhibition kinetics observed in typical bioluminescence experiments (Fontes et al., 2008). Two biological metabolites, pyrophosphate and coenzyme A can remove dehydroluciferyl-AMP from luciferase and relieve this strong competitive inhibition. Pyrophosphate relieves inhibition by “pyrophosphorolysis”, where the reaction of dehydroluciferyl-AMP and pyrophosphate works in the reverse direction of the typical dehydoluciferin adenylation reaction to produce ATP and the less strongly inhibiting product dehydroluciferin (Fraga et al., 2005). In the case of coenzyme A, luciferase is able to catalyze the formation of dehydroluciferyl-CoA from dehydroluciferyl-AMP and CoA. It stands to reason that *in vivo*, fireflies must relieve the tight-binding inhibition of dehydroluciferyl-AMP, and that either pyrophosphorolysis or CoA ligation provide routes to do so. In the case of pyrophosphorolysis, available evidence such as the presence of a peroxisome targeted and highly and differentially expressed inorganic pyrophosphatase (LanPPase / PPYR_06392-PB) in the adult male *Photinus pyralis* light organ, suggests that pyrophosphate is rapidly catabolized, precluding removal of DHL-AMP via pyrophosphorolysis. In contrast, the CoA ligation to DHL-AMP seems likely to exist *in vivo*, as luciferase has a well established DHL-AMP CoA ligation activity, and free CoA would likely be present in the peroxisomes of the firefly light organ. Although CoA ligation of DHL-AMP would solve the immediate problem of luciferase inhibition, DHL-CoA is still a competitive inhibitor, albeit a weaker one (da Silva and da Silva, 2011). It seems likely that the efficient luminescence of fireflies requires the removal of DHL-CoA i*n vivo*, either by catabolism, transport & storage, or conversion of DHL-CoA into non-inhibitory products. The most appealing and efficient route for removal of DHL-CoA would be the reduction of DHL-CoA back into luciferin. Within the previously published candidate accessory metabolic geneset (Fallon et al., 2018), there was a very highly expressed short chain reductase, which we dub lantern short chain reductase 1 (LanSCR1 / PPYR_04899). A closely related paralog to LanSCR1 (PPYR_04900), which we dub LanSCR2 was found tandem to LanSCR1 locus in the *P. pyralis* genome. Interestingly, LanSCR2 was not highly expressed in the *P. pyralis* adult male light organ, but it was highly expressed in the larval light organ (Figure 5), suggesting that these two closely related genes are isofunctional in their catalytic role but have undergone subfunctionalization in terms of their expression patterns.

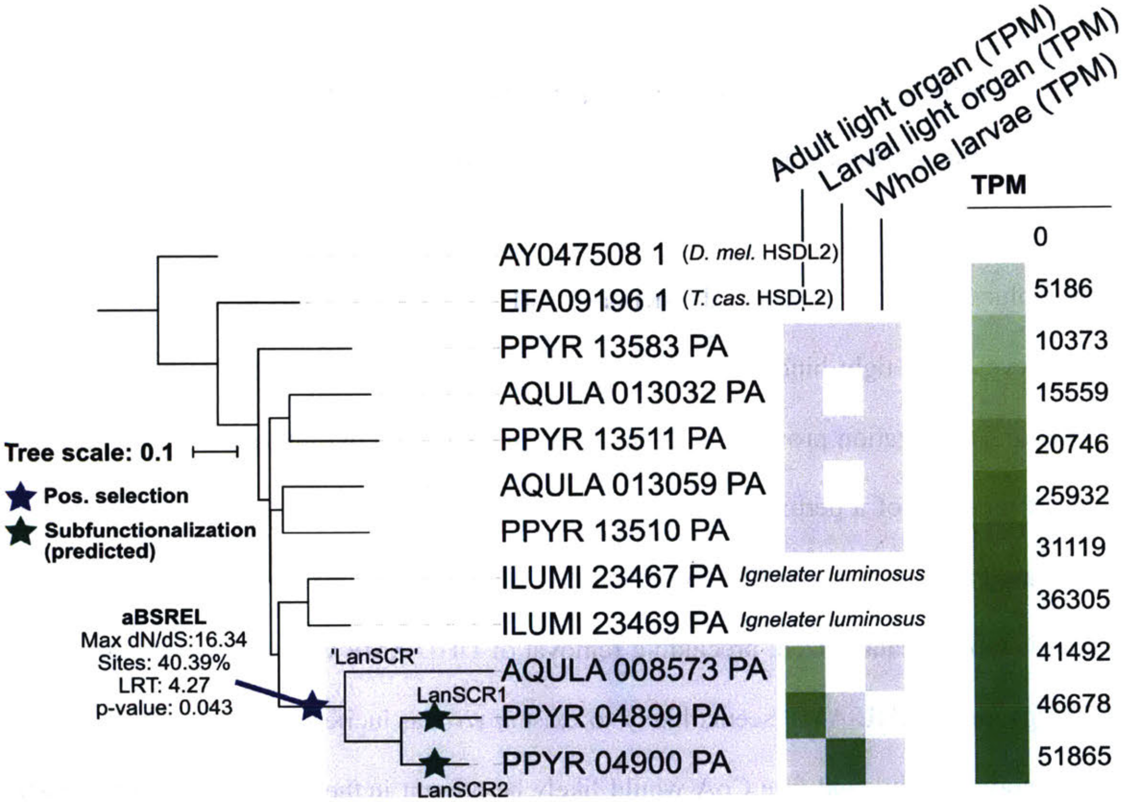


Figure 5: Maximum likelihood Gene tree of the HSDL2 orthogroup in select insects
CDS sequences of HSDL2 orthogroup genes were selected from *Drosophila melanogaster*, *Tribolium castaneum*, *Ignelater luminosus*, *Aquatica lateralis*, and *Photinus pyralis*, and aligned using MAFFT (Katoh and Standley, 2013). A maximum likelihood gene tree was then inferred from this multiple sequence alignment using the Mega7 package(Kumar et al., 2016), and the selection analysis using the aBSREL module (Smith et al., 2015) of the Hyphy molecular evolution software package (Pond et al., 2005).

To help elucidate the function of LanSCR1, we turned to the annotations of its most closely related enzyme in humans, the so-called hydroxysteroid dehydrogenase-like protein 2 (HSDL2). Despite its name, steroids have never been demonstrated as substrate of HSDL2, and the function of this enzyme remains unknown (Kowalik et al., 2009). In terms of domain structure and sequence similarity, HSDL2 is most similar to the peroxisomal enoyl-CoA reductase peroxisomal trans-2-enoyl-CoA reductase (PECR)(Gloerich et al., 2006), peroxisomal 2,4-dienoyl-CoA reductase (DECR2)(De Nys et al., 2001), and the retinal reductase (DHRS4)(Rattner et al., 2000). Each of these enzymes perform NADPH dependent reductions on linear lipids, such as branched chain fatty acids. Notably, PECR reduces a double bond in an analogous position (the 2,3 position) to that of the double bond in dehydroluciferin, supporting that these short chain reductase family enzymes could operate on dehydroluciferin or dehydroluciferyl-CoA. That being said, PECR is likely not the enzyme, as it reportedly cannot utilize substrates which are sterically hindered at the alpha carbon with methyl groups or presumably other bulky groups (Gloerich et al., 2006). In addition to an N-terminal short chain reductase domain, HSDL2 and LanSCR1 also have a C-terminal sterol carrier protein (SCP) domain. We hypothesize that HSDL2 also performs reductions on linear lipids, possibly those with very large sizes or structural rigidity which cause them to project from the enzyme active side and necessitate use of an external SCP domain. That being said, another peroxisomal enzyme, the thiolase sterol carrier protein X, also has a similar domain structure where an N-terminal enzymatic domain is contiguous with a C-terminal sterol carrier domain (Ferdinandusse et al., 2006), however it is believed that sterol carrier protein X is cleaved into its separate domains after import into the peroxisome (Ferdinandusse et al., 2000). HSDL2 may undergo a similar post-translational processing. Curiously, enzymes with confirmed roles in primary metabolism such as PECR, do not have clear direct orthologs between mammals and insects, but HSDL2 is single copy and has clear direct orthologs in most metazoans as well as some single celled eukaryotes. Furthermore, HSDL2 is widely expressed across all human tissues (Bastian et al., 2008). This argues that HSDL2 plays an important, albeit still unknown, role in primary metabolism within the peroxisome of metazoans.

In bioluminescent beetles, the HSDL2 orthogroup has undergone a significant expansion and positive selection, leading to LanSCR1 (Figure 5). Across the majority of insects, there are 1 or 2 genes within the HSDL2 orthogroup (Zdobnov et al., 2017), however in the North American firefly *P. pyralis*, there are 5 HSDL2 orthogroup genes, two of which (LanSCR1 and LanSCR2) show evidence of positive selection versus their outgroup paralogs. Similarly, the Japanese firefly *Aquatica lateralis* has 3 HSDL2 orthogroup genes, with 1 showing a orthologous relationship to LanSCR1 and LanSCR2 within the positively selected clade. This evolutionary evidence supports that LanSCR1 has undergone positive selection, perhaps to become a reductase of DHL-CoA.

We sought to directly test LanSCR1 via recombinant expression. Heterologous expression of LanSCR1 in *E. coli* gave workable quantities of protein (Figure 6).

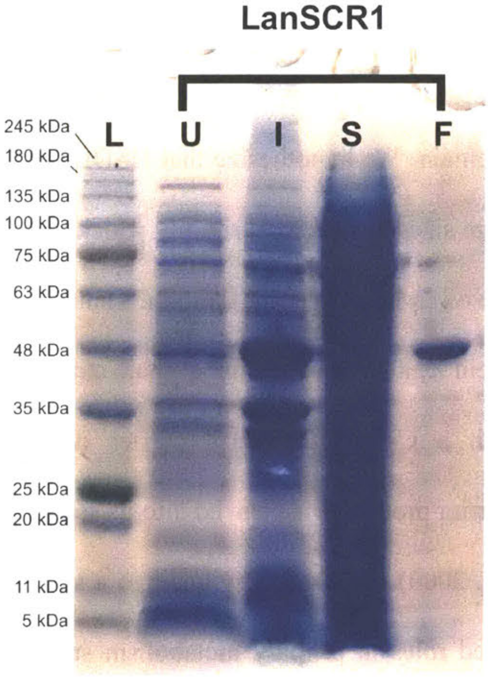


Figure 6: Recombinant expression of LanSCR1. L=Protein molecular weight standard ladder, U=uninduced - *E. coli* culture before addition of IPTG. I=insoluble - protein extract from centrifuged particulate post induction & E. coli lysis , S=soluble - supernatant from post induction E. coli lysis, F=final purified protein - post Ni-NTA IMAC and Sephadex size-exclusion chromatography. (B)=blank lane. Ladder is the Gold Biotechnology BlueStain Protein ladder (P/N: P007-500). The observed band is roughly consistent with the calculated molecular mass of LanSCR1 (45.6 kDa).

We then tested our recombinant LanSCR1 in directed enzymology experiments, testing the hypothesis that it acted as reductase of either dehydroluciferin and dehydroluciferin, using both NADH and NADPH as co-substrates. However, all tests gave negative results (data not shown). It has however been reported in the literature that the PECR homolog of LanSCR1 is very labile. For example, PECR could not be frozen without a complete loss of activity, and lost activity throughout the activity guided purification process although it maintained solubility and ability to be purified via a NADPH affinity column (Das et al., 2000). Given that we did freeze LanSCR1 before use and did not take special steps to preserve its activity during our purification protocol, it may be possible that these negative results with recombinant LanSCR1 are false positives, due to the unusual lability of this enzyme family. LanSCR1 likely accomplished a specialized role in the firefly lantern, but further experiments are needed to decipher its catalytic role.

MATERIALS AND METHODS

Recombinant protein expression

Single-strand cDNA was prepared from *P. pyralis* total RNA extracted from the lantern by poly-T primed reverse transcriptase using the SuperScript III First-Strand Synthesis System for RT-PCR (Invitrogen), following the manufacturer's instructions. LanABH1, LanABH2, and LanSCR1 were cloned from this first strand cDNA pool of the *Photinus pyralis* adult male light organ via PCR, and inserted via Gibson assembly into the T7 expression plasmid pHis8-4. Single-strand cDNA was prepared from *P. pyralis* total RNA extracted from the lantern by poly-T primed reverse transcriptase using the SuperScript III First-Strand Synthesis System for RT-PCR (Invitrogen), following the manufacturer's instructions. pHis8-4 is an *E. coli* T7 expression plasmid descended from pHIS8-3 (Weng and Noel, 2012), that harbors an N-terminal 8xHis tag followed by a TEV protease cleavage site for His-tag removal. The resulting plasmids were dubbed pJKW 0642 (LanABH1), pJKW 1199 (LanABH2), and pJKW 0631 (LanSCR1) respectively. The expression plasmids were transformed into BL21(DE3) *E. coli,* and protein was purified via Ni-NTA immobilized metal affinity chromatography (IMAC). Proteins were purified as previously described (Fallon et al., 2016).

Enzymology of recombinant LanABH1

Luciferyl-CoA was first synthesized in a 500 µL reaction of luciferase (10 µg / mL), L-luciferin (200 µM), ATP (1.5 mM), coenzyme A (0.68 mM), in 80 mM HEPES pH 7.3, 150 mM NaCl, 20 mM $MgCl_2$.

This reaction was incubated at for 3 hours at room temperature (~23°C), with protection from light. Next, 10 μL of either purified LanABH1 (~10 mg / mL) or ddH_2O, was added to 190 μL of the luciferyl-CoA reaction mixture. No steps were taken to inactivate luciferase before the addition of the next component. At 5 and 40 minutes, 10 μL aliquots removed, and quenched 1:1 with 100% acetonitrile, and 5 μL was then injected for HPLC analysis. HPLC analyses were performed as previously described (Fallon et al., 2016), with the addition that reversed phase HPLC analysis of the luciferyl-CoA thioesters was performed using a modified buffer A containing H_2O with 10 mM triethylamine (TEA), 25 mM ammonium acetate (NH_4Ac), and 0.1% Formic acid. Without addition of ammonium acetate, luciferyl-CoA thioesters formed a poorly defined chromatographic peak. Luciferyl-CoA thioesters were detected using SIM analyses tuned to the $[M+H]^+$ ion of the compound in question.

Enzymology of recombinant LanSCR1

Dehydroluciferyl-CoA was synthesized in a 500 μL reaction of luciferase (10 μg / mL), dehydroluciferin (200 μM), ATP (1.5 mM), coenzyme A (0.68 mM), in 80 mM HEPES pH 7.3, 150 mM NaCl, 20 mM $MgCl_2$, and purified via peak collection off an analytical HPLC. The resulting fractions were concentrated by lyophilization, and redissolved in 50 μL 80 mM HEPES pH 7.3. 10 μL of the redissolved dehydroluciferyl-CoA solution added to 490 μL of coenzyme A (0.68 mM), in 80 mM HEPES pH 7.3, 150 mM NaCl, 20 mM $MgCl_2$. 190 μL of this dehydroluciferyl-CoA solution was then added to 10 μL of purified LanSCR1 (2 mg / mL). 2μL aliquots of either NADH (72 mM), NADPH (53 mM), or ddH_2O were then added to 28 μL aliquots of the LanSCR1 with dehydroluciferyl-CoA solution, producing the reaction mixture. The reaction mixtures was incubated at room temperature with protection from light, and at 15 and 31 hours after start of the reaction 14 μL aliquots were removed and quenched 1:1 acetonitrile, and 5 μL was then injected for HPLC analysis. HPLC analyses were performed as previously described (Fallon et al., 2016), with the addition that reversed phase HPLC analysis of the luciferyl-CoA thioesters was performed using a modified buffer A containing H_2O with 10 mM triethylamine (TEA), 25 mM ammonium acetate (NH_4Ac), and 0.1% Formic acid. Without addition of ammonium acetate, luciferyl-CoA thioesters would form a poorly defined chromatographic peak. Luciferyl-CoA thioesters were detected using SIM analyses tuned to the $[M+H]^+$ ion of the compound in question.

REFERENCES

Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, Robinson-Rechavi M. 2008. Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among SpeciesData Integration in the Life Sciences. Springer Berlin Heidelberg. pp. 124–131.

Das AK, Uhler MD, Hajra AK. 2000. Molecular Cloning and Expression of Mammalian Peroxisomaltrans-2-Enoyl-coenzyme A Reductase cDNAs. *J Biol Chem* **275**:24333–24340.

da Silva LP, da Silva JCGE. 2011. Kinetics of inhibition of firefly luciferase by dehydroluciferyl-coenzyme A, dehydroluciferin and L-luciferin. *Photochem Photobiol Sci* **10**:1039–1045.

De Nys K, Meyhi E, Mannaerts GP, Fransen M, Van Veldhoven PP. 2001. Characterisation of human peroxisomal 2,4-dienoyl-CoA reductase. *Biochim Biophys Acta* **1533**:66–72.

Fallon TR, Li F-S, Vicent MA, Weng J-K. 2016. Sulfoluciferin is Biosynthesized by a Specialized Luciferin Sulfotransferase in Fireflies. *Biochemistry* **55**:3341–3344.

Fallon TR, Lower SE, Chang C-H, Bessho-Uehara M, Martin GJ, Bewick AJ, Behringer M, Debat HJ, Wong I, Day JC, Suvorov A, Silva CJ, Stanger-Hall KF, Hall DW, Schmitz RJ, Nelson DR, Lewis SM, Shigenobu S, Bybee SM, Larracuente AM, Oba Y, Weng J-K. 2018. Firefly genomes illuminate parallel origins of bioluminescence in beetles. *Elife* **7**. doi:10.7554/eLife.36495

Ferdinandusse S, Denis S, van Berkel E, Dacremont G, Wanders RJ. 2000. Peroxisomal fatty acid oxidation disorders and 58 kDa sterol carrier protein X (SCPx). Activity measurements in liver and fibroblasts using a newly developed method. *J Lipid Res* **41**:336–342.

Ferdinandusse S, Kostopoulos P, Denis S, Rusch H, Overmars H, Dillmann U, Reith W, Haas D, Wanders RJA, Duran M, Marziniak M. 2006. Mutations in the gene encoding peroxisomal sterol carrier protein X (SCPx) cause leukencephalopathy with dystonia and motor neuropathy. *Am J Hum Genet* **78**:1046–1052.

Fontes R, Fernandes D, Peralta F, Fraga H, Maio I, Esteves da Silva JCG. 2008. Pyrophosphate and tripolyphosphate affect firefly luciferase luminescence because they act as substrates and not as allosteric effectors. *FEBS J* **275**:1500–1509.

Fraga H, Fernandes D, Fontes R, Esteves da Silva JCG. 2005. Coenzyme A affects firefly luciferase luminescence because it acts as a substrate and not as an allosteric effector. *FEBS J* **272**:5206–5216.

Fraga H, Fernandes D, Novotny J, Fontes R, Esteves da Silva JCG. 2006. Firefly luciferase produces hydrogen peroxide as a coproduct in dehydroluciferyl adenylate formation. *Chembiochem* **7**:929–935.

Gloerich J, Ruiter JPN, Van Den Brink DM, Ofman R, Ferdinandusse S, Wanders RJA. 2006. Peroxisomal trans-2-enoyl-CoA reductase is involved in phytol degradation. *FEBS Lett* **580**:2092–2096.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**:772–780.

Kowalik D, Haller F, Adamski J, Moeller G. 2009. In search for function of two human orphan SDR enzymes: hydroxysteroid dehydrogenase like 2 (HSDL2) and short-chain dehydrogenase/reductase-orphan (SDR-O). *J Steroid Biochem Mol Biol* **117**:117–124.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**:1870–1874.

Lloyd MD, Darley DJ, Wierzbicki AS, Threadgill MD. 2008. α-Methylacyl-CoA racemase--an “obscure”metabolic enzyme takes centre stage. *FEBS J* **275**:1089–1102.

Maeda J, Kato D-I, Okuda M, Takeo M, Negoro S, Arima K, Ito Y, Niwa K. 2017. Biosynthesis-inspired deracemizative production of d-luciferin by combining luciferase and thioesterase. *Biochim Biophys Acta Gen Subj* **1861**:2112–2118.

Nakamura M, Maki S, Amano Y, Ohkita Y, Niwa K, Hirano T, Ohmiya Y, Niwa H. 2005. Firefly luciferase exhibits bimodal action depending on the luciferin chirality. *Biochem Biophys Res Commun* **331**:471–475.

Niwa K, Nakamura M, Ohmiya Y. 2006. Stereoisomeric bio-inversion key to biosynthesis of firefly d-luciferin. *FEBS Lett* **580**:5283–5287.

Oba Y, Ojika M, Inouye S. 2003. Firefly luciferase is a bifunctional enzyme: ATP-dependent monooxygenase and a long chain fatty acyl-CoA synthetase. *FEBS Lett* **540**:251–254.

Okada K, Iio H, Kubota I, Goto T. 1974. Firefly bioluminescence III. Conversion of oxyluciferin to luciferin in firefly. *Tetrahedron Lett* **15**:2771–2774.

Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**:676–679.

Rattner A, Smallwood PM, Nathans J. 2000. Identification and Characterization of All -trans- retinol Dehydrogenase from Photoreceptor Outer Segments, the Visual Cycle Enzyme That Reduces All-trans -retinal to All- trans -retinol. *Journal of Biological Chemistry*. doi:10.1074/jbc.275.15.11034

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol* **32**:1342–1353.

Weng J-K, Noel JP. 2012. Structure--function analyses of plant type III polyketide synthasesMethods in Enzymology. Elsevier. pp. 317–335.

White EH, McCapra F, Field GF. 1963. The Structure and Synthesis of Firefly Luciferin. *J Am Chem Soc* **85**:337–343.

Wood KV, de Wet JR, Dewji N, DeLuca M. 1984. Synthesis of active firefly luciferase by in vitro translation of RNA obtained from adult lanterns. *Biochem Biophys Res Commun* **124**:592–596.

Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva EV. 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res* **45**:D744–D749.