bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Title:

Firefly genomes illuminate the origin and evolution of bioluminescence

Authors:

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

[1]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA.
[2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
[3]Department of Molecular Biology & Genetics, Cornell University, Ithaca, New York 14850, USA.
[4]Department of Biology, University of Rochester, Rochester, New York 14627, USA.
[5]Department of Environmental Biology, Chubu University, Kasugai, Aichi 487-8501, Japan.
[6]Graduate School of Bioagricultural Sciences, Nagoya University, Nagoya, Aichi 464-8601, Japan.
[7]Department of Biology, Brigham Young University, Provo, Utah 84602, USA.
[8]Department of Genetics, University of Georgia, Athens, Georgia 30602, USA.
[9]Biodesign Center for Mechanisms of Evolution, Arizona State University, Tempe, Arizona 85287, USA.
[10]Center of Agronomic Research National Institute of Agricultural Technology, Córdoba, Argentina.
[11]Centre for Ecology and Hydrology (CEH) Wallingford, Wallingford, Oxfordshire, UK.
[12]Department of Plant Sciences, University of California Davis, Davis, California, USA.
[13]Department of Microbiology Immunology and Biochemistry, University of Tennessee HSC, Memphis 38163, USA.
[14]Department of Biology, Tufts University, Medford, Massachusetts 02155, USA.
[15]NIBB Core Research Facilities, National Institute for Basic Biology, Okazaki 444-8585, Japan.

*These authors contributed equally to this work.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Summary

Fireflies are among the best-studied of the bioluminescent organisms. Despite long-term interest in the biochemistry, neurobiology, and evolution of firefly flash signals and the widespread biotechnological applications of firefly luciferase, only a limited set of genes related to this complex trait have been described. To investigate the genetic basis of firefly bioluminescence, we generated a high-quality reference genome for the Big Dipper firefly *Photinus pyralis*, from which the first laboratory luciferase was cloned, using long-read (PacBio), short-read (Illumina), and Hi-C sequencing technologies. To facilitate comparative genomics, we also generated short-read genome assemblies for a Japanese firefly *Aquatica lateralis* and a bioluminescent click beetle, *Ignelater luminosus*. Analyses of these genomic datasets supports at least two independent gains of luminescence in beetles, and provides new insights into the evolution of beetle bioluminescence and chemical defenses that likely co-evolved over their 100 million years of evolution.

Introduction

Fireflies are beetles (Family: Lampyridae) and consist of >2000 species present on every temperate continent. Also known as “lightning bugs”, they represent one of the most commonly observed and diverse lineages of bioluminescent organisms[1]. The characteristic flashes and glows of adult fireflies show intraspecific variation in duration, frequency, and color, enabling species-specific courtship dialogs in habitats that contain multiple simultaneously signaling firefly species (Supplementary Video 1)[2]. These luminous signals have long been studied for their function in both species recognition and mate choice[3], and for the evolutionary origins of their diversity[4]. Beyond the specialized luminous courtship behavior in the short-lived adult stage (~2–4 weeks), all fireflies are luminescent during the larval stage (1–2 years)[5]. Larval bioluminescence is thought to serve as an aposematic advertisement of their chemical defenses[6], including the potent cardiotoxic lucibufagins of *Photinus* fireflies[7]. The biochemical mechanism of firefly bioluminescence involves the oxidation of a specialized small molecule substrate, luciferin, by a specialized peroxisome-localized enzyme, luciferase, in an O_2, ATP and Mg^{2+}-dependent manner, forming the decarboxylated product oxyluciferin[8,9]. The cloning of firefly luciferase has enabled unique applications in medicine, industry, and research, through the use of firefly luciferase as a reporter gene[10], and through ATP quantification with the coupled luminescence reaction[11]. However, despite their global scientific and cultural relevance, fireflies remain an understudied group from a genetic perspective. In particular, the detailed molecular mechanisms supporting firefly bioluminescence, as well as their origins and evolution, await characterization.

To investigate the genetic basis, origins, and evolution of firefly bioluminescence, we generated high-quality reference genomes for two firefly species and a draft genome for a related bioluminescent click beetle (Family: Elateridae). The first species, the “Big Dipper Firefly”, *Photinus pyralis*, is a common, geographically widespread species in the United States (Figure 1). Due to its abundance and ease of identification, *P. pyralis* was used in classic studies of firefly physiology and bioluminescent biochemistry[12,13]. Additionally, *P. pyralis* has the smallest genome size of any measured lampyrid (422 ± 9

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mbp; XO male 2n=19)[14,15]. The second species, the "Heike-botaru", *Aquatica lateralis*, is a common Japanese firefly and is one of the few species that can be cultured in the laboratory. Finally, the "cucubano", *Ignelater luminosus*, is a bioluminescent click beetle found in the Caribbean. Notably, the terms "luciferase" and "luciferin" were first coined in experiments with bioluminescent click beetles in the late 1800's[16]. Elaterid and firefly luciferases are in the same enzyme superfamily, and their bioluminescent mechanisms and luciferins are chemically identical, suggesting a single origin of luminescence in this group. However, phylogenetic inference suggests independent origins of luminescence between these two families[17]. Together, these two lineages diverged from a common ancestor over 100 mya (Fig. 1c)[18,19]. Our comparative analysis of the genomic context of luciferases from these three species sheds light on the origin and evolution of bioluminescence in beetles. Our data also provide insight into the evolution of other traits, including chemical defenses and the viral and microbial holobiome associated with the unique lifestyle of bioluminescent beetles.

Results

Sequencing and assembly yield high-quality genomes

P. pyralis adult males were collected from the Great Smoky Mountains National Park, USA (GSMNP) and Mercer Meadows New Jersey, USA (MMNJ) (Fig. 1b), and sequenced using short-insert, mate-pair, Hi-C, and long read (Pacific Biosciences; PacBio) approaches (Extended Data Table 1). These were combined in a MaSuRCA[20] hybrid assembly (Supplementary Note 1). In contrast, the *A. lateralis* genome was derived from an ALL-PATHs[21] assembly of short insert and mate-pair reads from a single partially-inbred laboratory-reared adult female, whose lineage was first collected 25 years ago from a now extinct population in Yokohama, Japan (Supplementary Note 2). A single *Ignelater luminosus* adult male, collected in Mayagüez Puerto Rico, USA, was used to produce a high-coverage Supernova[22] linked-read draft genome (Supplementary Note 3). These genomes cover the two major firefly subfamilies: Lampyrinae (abundant in the Americas) and Luciolinae (abundant in the old world and Australia), and the Elateridae.

The *P. pyralis* (Ppyr1.3) and *A. lateralis* (Alat1.3) genome assemblies are high quality, as indicated by gene completeness and contiguity statistics comparable to the genome of the model beetle *Tribolium castaneum* (Supplementary Note 4.1). Remaining redundancy in the *P. pyralis* assembly, as indicated by duplicates of the Benchmarking Single Copy Orthologs (BUSCOs)[23] and the assembly size (Fig 2f; Supplementary Note 4.1), is likely due to the heterozygosity of the outbred input libraries (Supplementary Note 1). The *I. luminosus* genome assembly (Ilumi1.1) is less complete than that of Ppyr1.3 and Alat1.3, but is comparable to published insect genomes (Fig. 2f; Supplementary Note 4.1). We also assembled the complete mitochondrial genomes (mtDNA) for *P. pyralis and I. luminosus* (Supplementary Note 1.8; 3.10), to add to growing mtDNA resources in fireflies, including the recently-published mtDNA sequence of *A. lateralis*[24]. These mtDNA show high conservation of gene content and synteny as expected for metazoan mtDNA, with the exception of the ~1 kbp tandem repeat unit (TRU) found in the firefly mtDNAs.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To obtain a large-scale view of the *P. pyralis* genome, we super-scaffolded the assembly into 11 pseudo-chromosomal linkage groups using a Hi-C proximity-ligation linkage approach. These linkage groups contain 95% of the assembly (448.8 Mbp). A single linkage group, LG3a, corresponded to the X-chromosome based on read coverage and gene content (Supplementary Note 1.6.3). The size of LG3a (22.2 Mbp) was comparable to the expected X-chromosome size based on sex-specific genome size estimates using flow cytometry (~26 Mbp)[14]. We found that *T. castaneum* X-chromosome genes were relatively enriched on LG3a as compared to every other linkage group, suggesting that the X-chromosome content of these diverged beetles has been conserved for >200 MY (Supplementary Note 1.6.4). 42.6% of the *P. pyralis* assembly was found to be repetitive. 73% of this repeat content could not be annotated as any known repetitive sequence. However, we were able to identify the canonical telomeric repeat (TTAGG) in *P. pyralis* (Fig. 2b), as well as the telomerase (*TERT*) gene.

DNA methylation is common in eukaryotes, but variable in level within a genome, as well as general presence/absence, across insect taxa, in particular within Coleoptera[25]. Furthermore, the functions of DNA methylation across insects remains obscure[25,26]. Hence, we sought to shed light on cytosine methylation in fireflies by characterizing the methylation status of *P. pyralis* DNA through whole genome bisulfite sequencing (WGBS). Methylation at CpGs (mCG) was unambiguously detected in *P. pyralis* at ~20% in the genic regions, following a unique distribution as compared to methylation in other insects (Fig. 2d). Molecular evolution analyses of the DNA methyltransferases (DNMTs) show that orthologs of DNMT1 and DNMT3 were conserved in *P. pyralis* and *A. lateralis* (Supplementary Note 4.2), implying that many firefly lineages likely possess mCG.

Our coding gene annotation sets for *P. pyralis, A. lateralis,* and *I. luminosus* are comprised of 15,770, 14,285, and 27,552 genes containing 94.2%, 90.0%, and 91.8% of the Endopterygota BUSCOs, respectively. Analysis with the Orthofinder pipeline[27] indicated indicated a high similarity between the *P. pyralis* and *A. lateralis* genesets, whereas the *I. luminosus* geneset was diverged (Figure 2e; Supplementary Figure 4.1.3). Taken together, these results demonstrate that the genomes are a high-quality resource for genomic inquiry into the evolution of traits related to firefly bioluminescence.

Luciferase evolution in bioluminescent beetles

Fireflies are universally luminescent, suggesting a single origin of luminescence in their common ancestor. On the contrary, bioluminescence is rare within elaterid beetles (~200 out of 10,000 species), and previous studies of the origins of luminescence recovered a non-luminous elaterid ancestor[17]. However, the derived hypothesis that elaterid and firefly bioluminescence evolved independently remains controversial. To resolve the evolutionary history of beetle luciferases, we compared the genomic architecture of luciferases and their patterns of expression across the three sequenced taxa.

Firefly luciferase is hypothesized to have descended from an ancestral peroxisomal fatty acyl-CoA synthetase (PACS) because both it and its closely-related non-bioluminescent paralogs exhibit fatty acyl-CoA synthetase activity[28,29] (Figure 3a).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Here, two firefly luciferases, *Luc1* and *Luc2*, were unequivocally identified in both *P. pyralis* and *A. lateralis* genomes. Whereas *P. pyralis Luc1* was the first firefly luciferase identified and its orthologs have been widely cloned from other fireflies, the paralog *Luc2* was previously only known from a handful of divergent Asian taxa including *A. lateralis*[30,31]. In both species, *Luc1* is located within a cluster of related genes, including *PACSs* and *non-peroxisomal acyl-CoA synthetases* (*ACSs*), identified by their lack of the C-terminal peroxisomal targeting signal 1 (PTS1) (Figure 3d). *Luc2* is located on a different linkage-group from *Luc1* in *P. pyralis* and on a different scaffold from *Luc1* in *A. lateralis,* suggesting that *Luc1* and *Luc2* lie on different chromosomes in both taxa. No PACS or ACS gene was found in the vicinity of *Luc2* in either species.

The genomic architecture of firefly luciferases and closely related paralogs support a pervasive role of tandem duplication and neofunctionalization in the birth of this gene family. Both *Luc1* and *Luc2* are highly conserved at the level of gene structure—both are composed of seven exons with completely conserved exon/intron boundaries (Supplementary Figure 4.3.2). Phylogenetic analysis of luciferases, PACSs, and ACSs indicates that *Luc1* and *Luc2* represent two closely-related orthologous groups, and that the neighboring *PACS* and *ACS* genes near *Luc1* are co-orthologous, although the syntenic relationships are less clear, likely due to subsequent gene divergence and shuffling (Figure 3c,d). Altogether, the data support that tandem gene duplication gave rise to several ancestral *PACS* paralogs, one of which neofunctionalized to become the ancestral luciferase (*AncLuc*) (Figure 3b). While *AncLuc* evolved in-place to give rise to *Luc1*, *Luc2* likely derived from *AncLuc* via a long-range gene duplication event (e.g. transposon mobilization) prior to the divergence of the firefly subfamilies Lampyrinae and Luciolinae around 100 Mya.

Previous investigations of Asian taxa have shown that *Luc1* is responsible for light production from the lantern of adult, larvae, prepupa and pupa, whereas *Luc2* is responsible for the dim glow of egg, ovary, prepupa and pupal whole body[30]. Our RNA-Seq data show that *Luc1 and Luc2* in both *P. pyralis* and *A. lateralis* display expression patterns consistent with these previous reports. While *Luc2* is essentially not expressed in the adult male and larval lanterns, in both fireflies it is more highly expressed in eggs than larvae (Figure 3c). These results suggest that, since their divergence, *Luc1* and *Luc2* have established different but conserved roles in bioluminescence throughout firefly life cycle.

To resolve the independent origins of bioluminescence in fireflies and click beetles, we examined the genomic context of *I. luminosus* luciferase (*IlumiLuc*) and compared it to that of *P. pyralis* and *A. lateralis* (Figure 3d). Rather than the typical two elaterid luciferases responsible for prothorax and abdominal bioluminescence respectively, in *I. luminosus* we only identified a single luciferase, *IlumiLuc*, expressed in both the prothorax and abdominal lanterns. *IlumiLuc* was captured on an 87 kbp scaffold containing 8 other genes, one of which is a lowly expressed and partially assembled *PACS* gene (Figure 3d). However, this locus is not orthologous to either firefly luciferase locus based on genomic context and phylogenetic analysis: unlike the extant firefly *Luc1* locus or the hypothesized *AncLuc* loci, the most closely related *PACS*s to *IlumiLuc* based on phylogeny were not co-localized on the same scaffold genomically (Figure 3c,d). Instead, we identified a different scaffold that is likely orthologous to the *AncLuc* loci, containing both adjacent *ACS* and *PACSs*, as well as

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

orthologous exoribonuclease family and inositol monophosphatase family genes, which were found adjacent to the *A. lateralis Luc1* locus. Further, phylogenetic analysis suggested that elaterid luciferase co-orthologs formed a sister clade to non-luminous firefly PACS (Figure 3c, Clade c) indicating that elated luciferase did not evolve directly from firefly *AncLuc*. These genomic observations are consistent with the independent origins of firefly and elaterid luciferase.

To further substantiate this, we performed molecular adaptation analysis for *IlumiLuc* and its closely related homologs. Ancestral state reconstruction of luminescent activity on the luciferase paralog tree recovered two independent gains of luminescence as the most parsimonious scenario: once in elaterid beetles, and once in the common ancestor of firefly, phengodid, and rhagopthalmid beetles. 33% of the sites of the common ancestor of the elaterid luciferases showed a statistically significant signal of episodic positive selection with $d_N/d_S>1$ ($d_N/d_S=4.34$) as compared to the evolution of its paralogs using the aBSREL[31] branch-site selection test (Supplementary Note 4.3), implying that the common ancestor of elaterid luciferases underwent a period of accelerated directional evolution. Altogether, our results strongly support the independent neofunctionalization of luciferase in elaterid beetles and fireflies, and therefore at least two independent gains of luciferin-utilizing luminescence in beetles.

Evolution of firefly specialized physiology

Beyond luciferase, we sought to identify other specialized metabolic enzymes which might have co-evolved with firefly bioluminescence. In particular, the enzymes of the *de novo* biosynthetic pathway for firefly luciferin remain unresolved[32]. 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 exhibit high expression levels in the adult lantern, and would be differentially expressed when compared to non-luminescent tissues. To test this, we performed RNA-Seq and differential expression testing of the dissected *P. pyralis* and *A. lateralis* adult male lantern tissue as compared to a non-luminescent tissue. We identified set of candidate genes that were shared between *P. pyralis* and *A. lateralis*, annotated as enzymes, and both highly-expressed (HE) and differentially-expressed (DE) in the lantern (Figure 4e). Reassuringly, luciferase and luciferin sulfotransferase (LST), an enzyme recently implicated in luciferin storage in *P. pyralis*[33], were included in the candidate list using these four criteria. We identified the LST ortholog of *A. lateralis*, the first report of this gene in the Luciolinae, but not in *I. luminosus*. The lack of LST in the Elaterid suggests that LST-mediated luciferin storage was an ancestral trait unique to fireflies, consistent with independent evolutions of bioluminescence in Lampyridae and Elateridae[33].

In addition to these genes known to be important for bioluminescence, several other enzyme-encoding, HE, and DE lantern genes were identified that have implications for firefly lantern physiology. For instance, adenylate kinase may play a critical role in efficient recycling of AMP post-luminescence, whereas cystathionine gamma-lyase may be supportive of a key role of cysteine in luciferin biosynthesis[32] and recycling[34]. Highlighting the importance of luciferin storage as sulfoluciferin, we detected a combined adenylyl-sulfate kinase & sulfate adenylyltransferase, indicating that the full 3'-phosphoadenosine-5'-phosphosulfate (PAPS) biosynthetic pathway is present within

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the lantern. Furthermore, the firefly orthologs of this enzyme are the only members of their orthogroup to have a PTS1, suggesting specialized localization to the peroxisome, the location of the luminescence reaction (Supplementary Note 4.4.2). These findings suggest that, instead of sulfoluciferin being a passive storage compound, the levels of sulfoluciferin and luciferin may be actively regulated within the peroxisome of lantern cells in response to luminescence.

We also performed this analysis for genes not annotated as enzymes, yielding several genes with predicted lysosomal function (Supplementary Note 4.4.1). This could indicate that the abundant, but unidentified "differentiated zone granule" organelles of the firefly light organ[35] are, or are related to, lysosomes. We also found a highly and differentially expressed opsin, *Rh7*, in the light organ of *A. lateralis*, but not *P. pyralis.* Opsin expression has been associated with light perception in the light organ of other bioluminescent species, leading us to speculate that the role of *Rh7* is light perception (Supplementary Note 4.5).

Firefly bioluminescence has been postulated to have co-evolved as an aposematic warning of larval chemical defenses[36]. Lucibufagins are unpalatable defense steroids found in certain North American firefly species, most notably of the genera *Photinus*[7] and *Ellychnia*[37], and hence are a candidate for an ancestral firefly defense compound. To test whether lucibufagins were more widespread among fireflies, we surveyed the presence of lucibufagins in *P. pyralis* and *A. lateralis* by liquid-chromatography high-resolution accurate-mass mass-spectrometry (LC-HRAM-MS). Whereas lucibufagins were found in high abundance in *P. pyralis* , they were absent in *A. lateralis* (Figure 4b, Supplementary Note 4.6). This result suggests that lucibufagin biosynthesis evolved only in particular firefly taxa (e.g. subfamily: Lampyrinae), or that lucibufagin biosynthesis was an ancestral trait that has been lost in *A. lateralis*. Supporting the former hypothesis, the presence of lucibufagins in non-North-American Lampyrinae has been previously reported[38], but to date there are no reports of lucibufagins in the Luciolinae. Although lucibufagin metabolism is unresolved, their structures suggest a biosynthetic origin from cholesterol followed by a series of hydroxylations, -OH acetylations, and the side-chain oxidative pyrone formation[7] (Figure 4a). We hypothesized that cytochrome P450s, an enzyme family widely involved in specialized metabolism[39], could underlie several oxidative reactions in the proposed lucibufagin biosynthetic pathway. We therefore inferred the P450 phylogeny among bioluminescent beetles to identify any lineage-specific genes correlated to lucibufagin presence. Our analysis revealed a unique expansion of one P450 family, the CYP303 family, in *P. pyralis*. While all previously sequenced insect genomes, as well as the *A. lateralis* and *I. luminosus* genomes, contain only a single *CYP303* gene, the *P. pyralis* genome contains 11 *CYP303* genes and 2 pseudogenes (Figure 4d). The CYP303 ortholog of *D. melanogaster*, CYP303A1, was previously shown to play a role in the mechanosensory bristle development[40]. Although the exact biochemical function and substrate of *D. melanogaster* CYP303A1 is unknown, its closely related P450 families operate on the steroidal insect hormone ecdysone[40]. As ecdysone and lucibufagins are structurally similar, CYP303 may operate on steroid-like compounds, and the lineage-specific expansion of the CYP303 family in *P. pyralis* could have contributed to the metabolic evolution of lucibufagins.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Symbionts of bioluminescent beetles

Symbionts are another possible contributor to luciferin and or chemical defense metabolism. Whole genome sequencing of our wild-caught and laboratory reared fireflies revealed a rich holobiont. In addition to the nuclear and mitochondrial genomes of the host firefly species, we found various bacterial genomes, viral genomes, and the complete mtDNA for the phorid parasitoid fly, *Apocephalus antennatus*, the first mtDNA reported for genus *Apocephalus*. Inadvertently included in the *P. pyralis* PacBio library via undetected parasitization of the initial specimens, this mtDNA was assembled via a metagenomic approach (Supplementary Note 5.2). Independent collection of *A. antennatus* which emerged from *P. pyralis* adults and targeted COI sequencing later confirmed the taxonomic origin of this mtDNA (Supplementary Note 5.3). We also sequenced and metagenomically assembled the complete circular genome (1.29 Mbp, GC: 29.7%) for a *P. pyralis*-associated mollicute, *Entomoplasma luminosum* var. pyralis (Supplementary Note 5.1). *Entomoplasma* spp. were first isolated from the guts of North American fireflies[41] and our assembly provides the first complete genomic assembly of an *Entomoplasma* sp. Broad read coverage for the *E. luminosus* var. pyralis genome was detected in 5/6 of our *P. pyralis* DNA libraries, suggesting that *Entomplasma* is a highly prevalent symbiont of *P. pyralis*. It has been hypothesized that these mollicutes could play a role in firefly metabolism, specifically via contributing to cholesterol metabolism and lucibufagin biosynthesis[37].

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 (Supplementary Note 3.5.3), 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. Lastly, we detected two species of novel orthomyxoviridae-like ssRNA viruses, dubbed *Photinus pyralis* orthomyxo-like virus 1 and 2 (PpyrOMLV1/2), which were highly prevalent in our *P. pyralis* RNA-Seq dataset, and showed multi-generational transmission in the laboratory (Supplementary Note 5.4). Additionally, we found endogenous viral elements (EVEs) for PpyrOMLV1/2 in *P. pyralis* (Supplementary Note 5.4.1). These viruses are the first reported in any firefly species, and represent the second report of transgenerational transfer of an *Orthomyxoviridae* virus[42], and the second report of *Orthomyxoviridae* derived EVEs[43]. Together, these genomes from the firefly holobiont provide valuable resources for the continued inquiry of the symbiotic associates of fireflies and their ecological and biological significance.

Discussion

In this study, we generated high-quality genomic resources for three long-studied bioluminescent beetles, and through a series of comparative analyses, unveil molecular mechanisms underlying the origin and evolution of beetle bioluminescence. Examination of the genomic loci of several paralogous beetle luciferases, together with phylogenetic analysis, supports independent origins of luciferase in fireflies and elaterid beetles from common ancestral enzymes in fatty acid metabolism. In both lineages, such evolutionary development seems to be the end result of a series of tandem gene duplication events, which were subsequently sculpted by neofunctionalization and

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

subfunctionalization to give rise to extant luciferases (Figure 3b). The independent origins of firefly and elaterid luciferases therefore provides an exemplary system to study enzyme evolution via alternative mutational trajectories that link ancestral promiscuous activity to new adaptive catalytic function[44]. Additional genomic information from the other luminous beetle taxa (e.g. Phengodidae and Rhagophthalmidae), as well as non-luminous elateroid taxa (e.g. Cantharidae and Lycidae), will help to refine this evolutionary model with greater detail and confidence.

The parallel evolution of luciferases, however, does not indicate that the elaterid and firefly bioluminescent systems are entirely independent, as other subtraits of bioluminescence may have single or parallel origins. These subtraits include metabolic pathways for luciferin *de novo* biosynthesis and recycling, the developmental program and specialized physiology associated with the light organs, and the neurological networks that control bioluminescent signals. Elucidation of luciferin biosynthesis is of particular importance because the presence of luciferin is presumably a prerequisite to luciferase neofunctionalization. Through tissue-specific transcriptomics, we identified a list of conserved lantern-enriched firefly genes which encode enzymes likely involved in luciferin metabolism. Alternatively, microbial symbionts, such as the tenericutes detected in our *P. pyralis* and *A. lateralis* datasets, may contribute to luciferin biosynthesis. Recent reports have also shown that firefly luciferin is readily produced non-enzymatically by mixing benzoquinone and cysteine[45]. Benzoquinone is known to be a defense compound of divergent beetles and other arthropods (e.g. millipedes), therefore, promiscuous luciferin synthesis through spontaneous chemical reactions in non-bioluminescent taxa and dietary acquisition of such should be considered.

Currently, beetle bioluminescence is widely applied in biotechnology. However, the efficiency of the artificially reconstituted system pales in comparison to that of the native system evolved among fireflies and other beetles. The genomes and other related resources reported here provide a foundation for future inquiry of the molecular mechanisms that have evolved to support bioluminescence in these remarkable organisms, aiding development of improvements in the efficiency of the engineered bioluminescent systems. Furthermore, these resources provide the substrate for devising effect strategies to monitor and protect wild bioluminescent beetle populations in the face of rapidly changing climate and habitats due to human activities[46].

Acknowledgements

The authors would like to thank all contributors to the project via the Experiment.com crowdfunding platform (Supplementary Note 6). This work was supported by the Beckman Foundation (JKW), the Pew Scholar Program in the Biomedical Sciences (JKW), the Searle Scholars Program (JKW), the National Science Foundation (SES, Graduate Research Fellowship), and the Lars G. Ljungdahl Distinguished Investigator award (RJS). A part of work was supported by NIBB Cooperative Research Program (No. 12-202). We thank Fu-Shuang Li, Geoff Liou, Tomáš Pluskal and Aska Pluskal for their assistance with collection of *P. pyralis*, and D. Winston Bellot for his advice on genome assembly. Collection permits/specimens from the Mercer County Parks Commission, National Park Service, and Kathrin Stanger-Hall are gratefully acknowledged. The authors thank the Whitehead Institute high-performance-computing support, University of Rochester Center for Integrated Research Computing, and the

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Georgia Advanced Computing Resource Center (GACRC) for sequencing and computation resources. We thank Nick A. Rohr for MethylC-Seq library preparation and the Georgia Genomics Facility (GGF) for sequencing. We thank Beijing Genomics Institute for library prep and BGISEQ-500 RNA-Seq of *I. luminosus*. We thank Dr. Katsushi Yamaguchi for *A. lateralis* sequencing and data submission. We thank Mr. Haruyoshi Ikeya (Toin Gakuen High School, Yokohama, Japan) for providing *A. lateralis* specimens and advice on laboratory culture. We thank Dr. Kota Ogawa, Ryosuke Nakai, and Takahiro Bino for determining *A. lateralis* genome size, and R.N. for useful discussion.

Author contribution

T.R.F., S.E.S.L. and J.K.W. conceived the project. T.R.F. performed *P. pyralis* PacBio and Hi-C sequencing. S.E.S.L. performed *P. pyralis* Illumina sequencing. C.H.C. performed *P. pyralis* genome assembly. S.S. performed *A. lateralis* genome assembly. T.R.F. performed *I. luminosus*, mitochondrial, and non-viral symbiont genome assemblies. A.M.L. and C.J.S. performed repeat analysis. I.W. performed in situ hybridizations, A.J.B. performed methylation analysis. M.B. performed bacterial symbiont annotation and analysis. H.J.D. performed viral genome assembly and analysis. M.B.U. performed *A. lateralis* RNA-Seq, luciferase phylogenetic analysis, and Rh7 phylogenetic analysis. D.N. performed P450 manual annotation. T.R.F., S.E.S.L. and J.K.W. wrote the manuscript. All authors reviewed the manuscript and discussed the work.

Author Information

The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.K.W. (wengj@wi.mit.edu).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. Shimomura, O. *Bioluminescence: Chemical Principles and Methods*. (World Scientific, 2006).
2. Lloyd, J. E. Studies on the flash communication system in Photinus fireflies. (1966).
3. Lewis, S. M., Cratsley, C. K. & Demary, K. Mate recognition and choice in Photinus fireflies. *Ann. Zool. Fennici* **41,** 809–821 (2004).
4. Stanger-Hall, K. F. & Lloyd, J. E. Flash signal evolution in Photinus fireflies: character displacement and signal exploitation in a visual communication system. *Evolution* **69,** 666–682 (2015).
5. Buck, J. B. THE ANATOMY AND PHYSIOLOGY OF THE LIGHT ORGAN IN FIREFLIES. *Ann. N. Y. Acad. Sci.* **49,** 397–485 (1948).
6. De Cock, R. & Matthysen, E. Aposematism and Bioluminescence: Experimental evidence from Glow-worm Larvae(Coleoptera: Lampyridae). *Evol. Ecol.* **13,** 619–639 (1999).
7. Goetz, M. A., Meinwald, J. & Eisner, T. Lucibufagins, IV. New defensive steroids and a pterin from the firefly,Photinus pyralis (coleoptera: Lampyridae). *Experientia* **37,** 679–680 (1981).
8. White, E. H., McCapra, F. & Field, G. F. The Structure and Synthesis of Firefly Luciferin. *J. Am. Chem. Soc.* **85,** 337–343 (1963).
9. Suzuki, N. & Goto, T. 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 (1972).
10. Naylor, L. H. Reporter gene technology: the future looks bright. *Biochem. Pharmacol.* **58,** 749–757 (1999).
11. Shama, G. & Malik, D. J. The uses and abuses of rapid bioluminescence-based ATP assays. *Int. J. Hyg. Environ. Health* **216,** 115–125 (2013).
12. Bitler, B. & McELROY, W. D. The preparation and properties of crystalline firefly luciferin. *Arch. Biochem. Biophys.* **72,** 358–368 (1957).
13. de Wet, J. R., Wood, K. V., Helinski, D. R. & DeLuca, M. Cloning of firefly luciferase cDNA and the expression of active luciferase in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **82,** 7870–7873 (1985).
14. Lower, S. S. *et al.* Genome Size in North American Fireflies: Substantial Variation Likely Driven by Neutral Processes. *Genome Biol. Evol.* **9,** 1499–1512 (2017).
15. Wasserman, M. & Ehrman, L. Firefly Chromosomes, II (Lampyridae: Coleoptera). *Fla. Entomol.* **69,** 755–757 (1986).
16. Dubois, R. Note sur la physiologie des pyrophores. *CR Soc Biol* **2,** 559 (1885).
17. Sagegami-Oba, R., Oba, Y. & Ohira, H. 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 (2007).
18. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346,** 763–767 (2014).
19. Mckenna, D. D. *et al.* 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 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

20. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29,** 2669–2677 (2013).
21. Butler, J. *et al.* ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res.* **18,** 810–820 (2008).
22. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* **27,** 757–767 (2017).
23. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31,** 3210–3212 (2015).
24. Maeda, J. *et al.* 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 (2017).
25. Bewick, A. J., Vogel, K. J., Moore, A. J. & Schmitz, R. J. Evolution of DNA Methylation across Insects. *Mol. Biol. Evol.* **34,** 654–665 (2017).
26. Glastad, K. M. *et al.* Variation in DNA Methylation Is Not Consistently Reflected by Sociality in Hymenoptera. *Genome Biol. Evol.* **9,** 1687–1698 (2017).
27. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16,** 157 (2015).
28. Oba, Y., Ojika, M. & Inouye, S. Firefly luciferase is a bifunctional enzyme: ATP-dependent monooxygenase and a long chain fatty acyl-CoA synthetase. *FEBS Lett.* **540,** 251–254 (2003).
29. Oba, Y., Sato, M., Ohta, Y. & Inouye, S. Identification of paralogous genes of firefly luciferase in the Japanese firefly, Luciola cruciata. *Gene* **368,** 53–60 (2006).
30. Bessho-Uehara, M., Konishi, K. & Oba, Y. 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 (2017).
31. Smith, M. D. *et al.* Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* **32,** 1342–1353 (2015).
32. Oba, Y., Yoshida, N., Kanie, S., Ojika, M. & Inouye, S. 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 (2013).
33. Fallon, T. R., Li, F.-S., Vicent, M. A. & Weng, J.-K. Sulfoluciferin is Biosynthesized by a Specialized Luciferin Sulfotransferase in Fireflies. *Biochemistry* **55,** 3341–3344 (2016).
34. Okada, K., Iio, H., Kubota, I. & Goto, T. Firefly bioluminescence III. Conversion of oxyluciferin to luciferin in firefly. *Tetrahedron Lett.* **15,** 2771–2774 (1974).
35. Ghiradella, H. & Schmidt, J. T. Fireflies at one hundred plus: a new look at flash control. *Integr. Comp. Biol.* **44,** 203–212 (2004).
36. Branham, M. A. & Wenzel, J. W. The origin of photic behavior and the evolution of sexual communication in fireflies (Coleoptera: Lampyridae). *Cladistics* **19,** 1–22 (2003).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

37. Smedley, S. R. *et al.* Bufadienolides (lucibufagins) from an ecologically aberrant firefly (Ellychnia corrusca). *Chemoecology* **27,** 141–153 (2017).
38. Tyler, J., Mckinnon, W., Lord, G. A. & Hilton, P. J. A defensive steroidal pyrone in the Glow-worm Lampyris noctiluca L.(Coleoptera: Lampyridae). *Physiol. Entomol.* **33,** 167–170 (2008).
39. Hamberger, B. & Bak, S. Plant P450s as versatile drivers for evolution of species-specific chemical diversity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368,** 20120426 (2013).
40. Willingham, A. T. & Keil, T. A tissue specific cytochrome P450 required for the structure and function of Drosophila sensory organs. *Mech. Dev.* **121,** 1289–1297 (2004).
41. Hackett, K. J. *et al.* Lampyridae (Coleoptera): A plethora of mollicute associations. *Microb. Ecol.* **23,** 181–193 (1992).
42. Marshall, S. H., Ramírez, R., Labra, A., Carmona, M. & Muñoz, C. 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 (2014).
43. Katzourakis, A. & Gifford, R. J. Endogenous viral elements in animal genomes. *PLoS Genet.* **6,** e1001191 (2010).
44. Weng, J.-K. The evolutionary paths towards complexity: a metabolic perspective. *New Phytol.* **201,** 1141–1149 (2014).
45. Kanie, S., Nishikawa, T., Ojika, M. & Oba, Y. One-pot non-enzymatic formation of firefly luciferin in a neutral buffer from p-benzoquinone and cysteine. *Sci. Rep.* **6,** 24794 (2016).
46. Katoh, K., Sakai, S. & Takahashi, T. Factors maintaining species diversity in satoyama, a traditional agricultural landscape of Japan. *Biol. Conserv.* **142,** 1930–1936 (2009).
47. Letunic, I. & Bork, P. 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 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

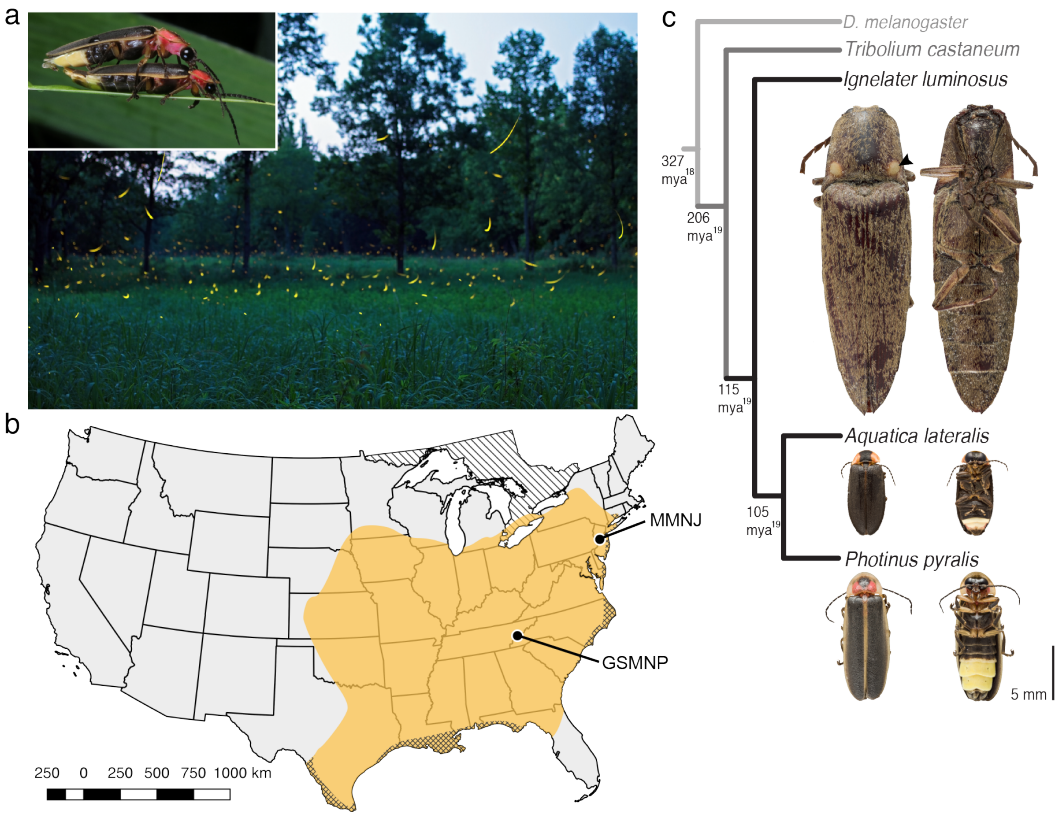
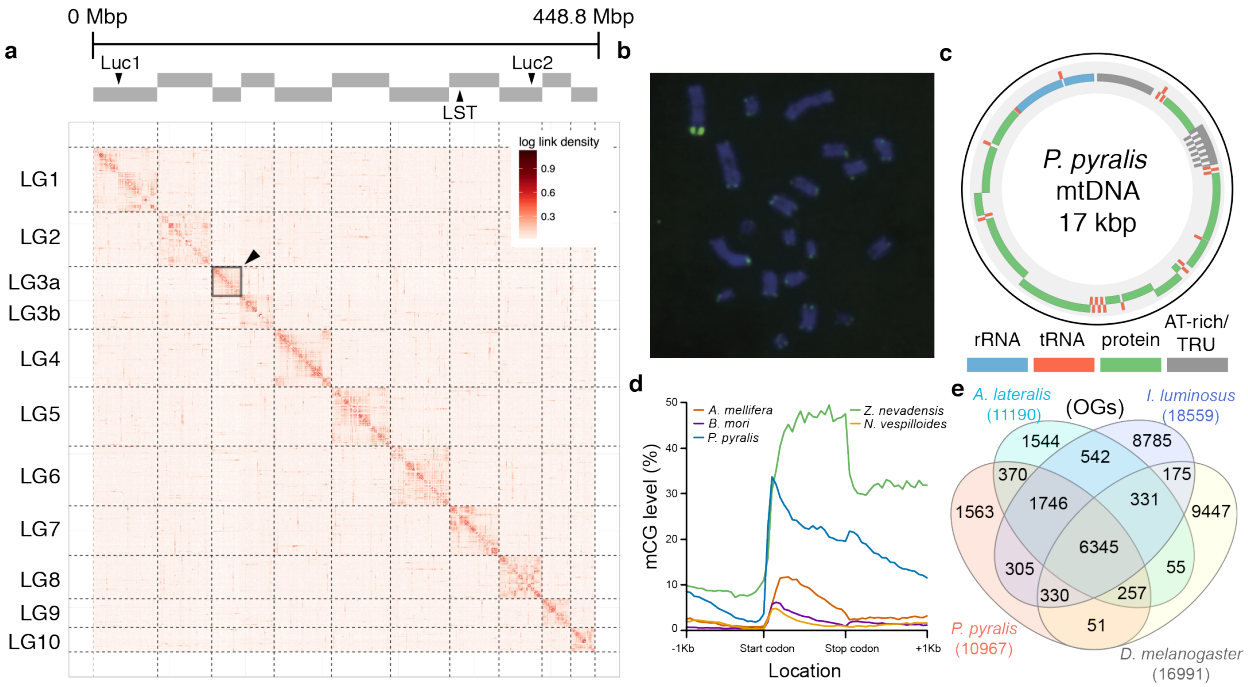


Figure 1 | Geographic and phylogenetic context of the Big Dipper firefly, *Photinus pyralis*.

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[2]. Photo credit: Alex Wild. Inset: male and female *P. pyralis* in early stages of mating. Photo credit: Terry Priest. **b,** Empirical range of *P. pyralis* in North America, extrapolated from 541 reported sightings (Supplementary Note 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. **c,** 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)[18,19]. Right: Dorsal and ventral photos of adult male specimens. Note the well-developed ventral light organs on the true abdominal segments 6 & 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).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



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	448 ± 7	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	Ilumi1.0	Chromium linked-reads	M	764 ± 7	845	91560	0.116	93.7	1.3

Figure 2 | Nuclear genome assembly of *Photinus pyralis*.

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 (black arrow) corresponds to the X chromosome. **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 female. **c,** Genome schematic of *P. pyralis* mitochondrial genome (mtDNA). Like other firefly mtDNAs it has a tandem repetitive unit (TRU). **d,** mCG is enriched across gene bodies of *P. pyralis* and shows a unique distribution as compared to other insects **e,** Orthogroup (OGs) clustering analysis of peptides with Orthofinder[27] shows a high degree of overlap of the *P. pyralis*, *A. lateralis,* and *I. luminosus* genesets with the geneset of *D. melanogaster*. **f,** Assembly statistics for presented genomes. *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[23] profile.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

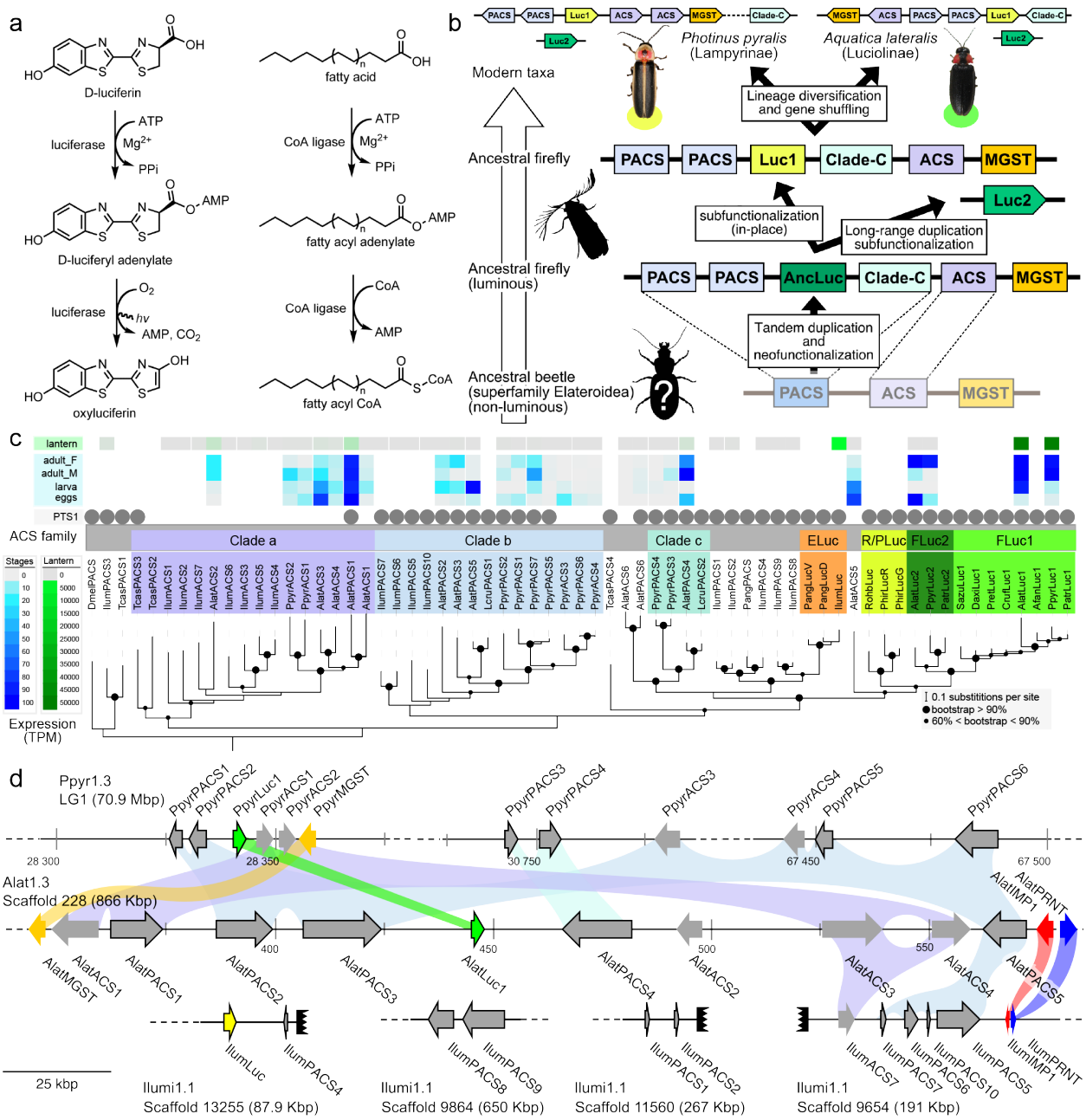


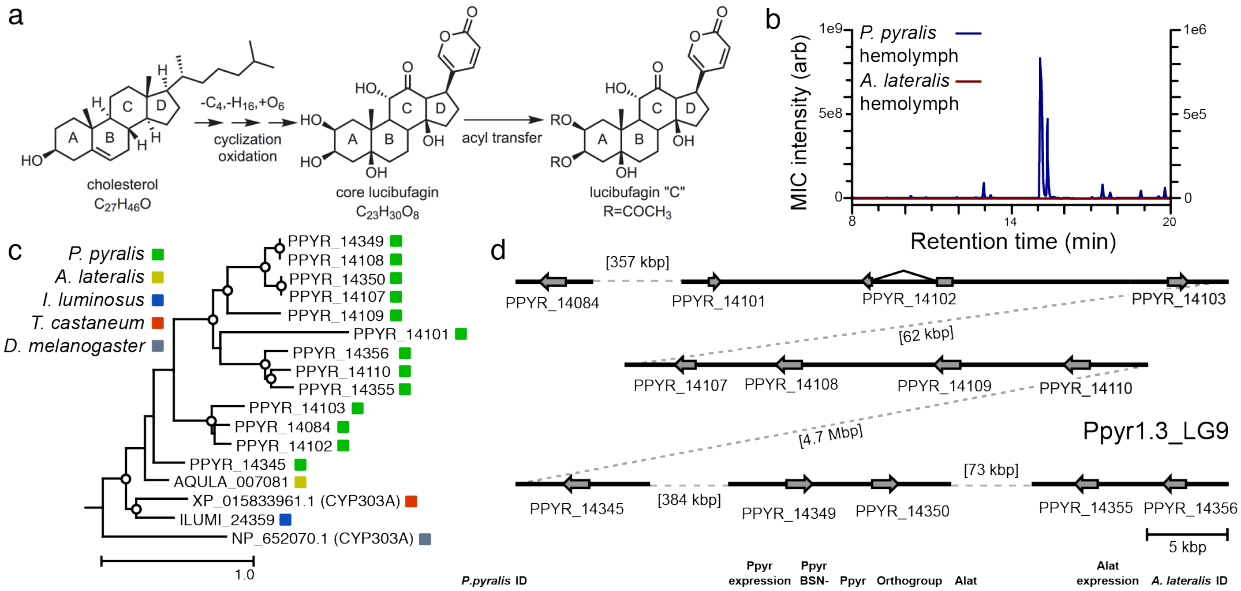
Figure 3 | Comparative genomic insights into luciferase evolution

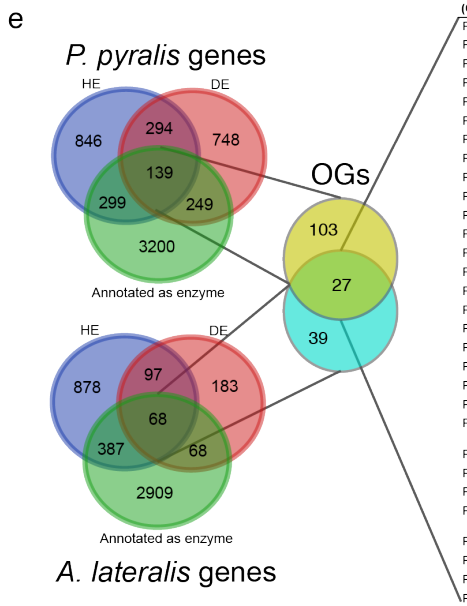
a, The catalytic mechanism of luciferase is related to that of fatty acyl-CoA synthetases. **b,** Model for genomic evolution of firefly luciferases. **c,** Maximum likelihood tree of luciferase homologs. Grey circles above gene names indicate the presence of peroxisomal targeting signal 1 (PST1). Color gradients indicate the TPM values of whole body in each sex/stage (grey to blue) and in the lantern (grey to green). Lantern values for *I. luminosus* are from whole prothorax. Tree and annotation visualized using iTOL[47]. **d**, Synteny analysis of beetle luciferases. About ten *PACS* and *ACS* genes flank the *Luc1* gene in both firefly genomes.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Although the *Luc1* loci in *P. pyralis* and *A. lateralis* are apparently derived from a common ancestor, the relative positions of the flanking *PACS* and *ACS* genes have diverged between the two species. *IlumiLuc* was captured a separate scaffold (Ilumi1.1_Scaffold13255) from its most most closely related *PACS*s (IlumPACS8,IlumPACS9) on Ilumi1.1_Scaffold9864), though one of the 8 other genes on the scaffold is a lowly expressed and partially assembled *PACS* gene (IlumiPACS4). In contrast, a different scaffold (Ilumi1.1_Scaffold9654) shows orthology to the firefly *Luc1* locus. Genes with *PTS1* indicated by a dark outline. Co-orthologous genes are labeled in the same color in the phylogenetic tree and are connected with corresponding color bands in synteny diagram. MGST: Gene accession numbers, annotation, and expression values are available on FigShare (10.6084/m9.figshare.5725690).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.





P.pyralis ID (OGS1.0)	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	66919	PTS1	OG0000168	PTS1	36044	1	AQULA_005067
PPYR_11147	Cystathionine gamma-lyase	3	38679		OG0002174		18096	3	AQULA_003032
PPYR_04899	Short chain dehydrogenase	4	28583	PTS1	OG0009894		9452	9	AQULA_008573
PPYR_09320	Saccharopine dehydrogenase-like	5	17563	PTS1	OG0000332	PTS1	6355	12	AQULA_012956
PPYR_10586	Alpha/beta hydrolase	6	17394		OG0001647	PTS1	1199	100	AQULA_007718
PPYR_06194	Alpha/beta hydrolase	7	13592	PTS1	OG0009284		850	161	AQULA_013805
PPYR_02512	Histidine Triad superfamily	8	11165		OG0007541		5575	13	AQULA_008871
PPYR_00996	Strictosidine synthase-like	12	4883		OG0004982		2529	35	AQULA_002761
PPYR_08520	Methionine-R-sulfoxide reductase	17	3960		OG0001893		2293	44	AQULA_008914
PPYR_08058	Acetyl-CoA hydrolase/transferase	20	3639		OG0004720		4981	17	AQULA_000701
PPYR_00003	Luciferin sulfotransferase*	24	3176	PTS1	OG0000048		2366	43	AQULA_012700
PPYR_08432	Adenylate kinase	36	2068		OG0001850	PTS1	3619	22	AQULA_007407
PPYR_14844	Malic oxidoreductase-like	55	1574	PTS1	OG0000471	PTS1	2441	41	AQULA_005495
PPYR_06564	Malic oxidoreductase-like	569	213		""	""	""	""	""
PPYR_04459	ABC transporter	75	1232		OG0000024		242	566	AQULA_007404
PPYR_06879	Metallo-beta-lactamase	78	1203		OG0006970		1880	51	AQULA_004381
PPYR_09240	CoA transferase	81	1174	PTS1	OG0004882	PTS1	690	203	AQULA_001958
PPYR_11151	Enolase	104	929		OG0008631		370	380	AQULA_003033
PPYR_01504	Alpha/beta hydrolase	155	677	PTS1	OG0000065	PTS1	904	148	AQULA_012908
PPYR_01505	Alpha/beta hydrolase	664	189		""	""	""	""	""
PPYR_10210	Methionine-S-sulfoxide reductase	174	639		OG0002286		640	227	AQULA_005939
PPYR_14372	Adenylyl-sulfate kinase & sulfate adenylyltransferase	214	539	PTS1	OG0000522	PTS1	4300	19	AQULA_001585
PPYR_05464	Peroxiredoxin	251	475		OG0000651		1434	72	AQULA_013952
PPYR_06980	Cytochrome P450	404	308		OG0001142		251	543	AQULA_002673
PPYR_10578	Short chain dehydrogenase	418	301		OG0003656		412	335	AQULA_002715
PPYR_09779	3'5'-cyclic nucleotide phosphodiesterase	441	286		OG0000969		104	1258	AQULA_002893
PPYR_01821	ABC transporter	477	260		OG0000024		647	223	AQULA_002548
PPYR_12812	Fatty acid hydroxylase	538	229		OG0001356		718	194	AQULA_001837
PPYR_01858	Enoyl-CoA hydratase/isomerase	675	188		OG0005842		652	221	AQULA_010152
PPYR_05219	DD-peptidase superfamily	1526	88		OG0005180		309	448	AQULA_004580

Figure 4 | Genomic insights into firefly specialized metabolism

a, Hypothesized lucibufagin biosynthetic pathway, starting from cholesterol. **b,** LC-HRAM-MS multi-ion-chromatograms (MIC) of hemolymph extracts from *P. pyralis* (left axis) and *A. lateralis* (right axis) exhibiting the exact masses for lucibufagins [M+H] ± 10 ppm (Supplementary Table 4.6.5.3). **c,** Maximum likelihood tree of CYP303 P450 enzymes from *P. pyralis*, *A. lateralis*, *T. castaneum*, and *D. melanogaster*. *P. pyralis* shows a notable *CYP303* family expansion, whereas the other species only have a single *CYP303*. Circles represent node bootstrap support >60%. Branch length measures substitutions per site. **d,** Genomic loci for *P. pyralis CYP303* family genes. These genes are found in multiple gene clusters on LG9. Introns >4 kbp are shown. **f,** Candidate enzymes of bioluminescent accessory metabolism. Enzymes which are highly expressed (HE), differentially expressed (DE), and annotated as enzymes via

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 with blank columns.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

***P. pyralis* genome sequencing and assembly.** A hybrid assembly was performed by combining previously published short-insert data[1] and newly-generated short-insert, mate-pair, long-read PacBio, and Hi-C sequencing data. *Photinus pyralis* males were captured in flight at the Great Smoky Mountains National Park, TN (GSMNP) and Mercer Meadows, NJ (MMNJ). Specimens for Illumina sequencing (GSMNP) were preserved in 95% ethanol at -80˚C until DNA extraction from thorax using phenol-chloroform extraction with RNAse digestion. DNA from Illumina TruSeq short insert (300 bp) and Nextera mate pair libraries (3-kb, 6-kb) were constructed from single individual (2 libraries) and pooled-specimens (1 library) extractions at the Georgia Genomics Facility (Athens, GA) and sequenced on 2 lanes of Illumina HiSeq2000 100 bp PE reads to a depth of 98x, 26x, and 14x, respectively, at BGI (Beijing, China). For PacBio sequencing, high-molecular-weight (HMW) DNA was isolated from 4 liquid-N_2-frozen, lyophilized, and -80˚C-stored MMNJ males using 100/G Genomic Tips with the Genomic Buffers kit (Qiagen, USA), and size-selected to 15 kbp+ using a BluePippin (Sage Science). The resulting library was sequenced on 61 SMRT cells using P6-C4 chemistry on a PacBio RSII instrument at the Broad Technology Labs (Cambridge, MA) to a depth of ~38x. Hi-C libraries were constructed and sequenced from two pooled MMNJ *P. pyralis* males at Phase Genomics (Seattle, WA) following established protocols[2,3,4]. See Supplementary Note 1.5 for more detail.

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)[5,6] pipeline that combined both Illumina-corrected PacBio reads (Mega-reads) and synthetic long reads constructed from short-insert reads alone (Super-reads) using a small overlap length (59 bp). The resultant assembly was then scaffolded and gap-filled with filtered Illumina mate-pair reads using Redundans (v 0.13a)[7] with default settings. Next, redundancy due to heterozygosity was removed using the built-in “deduplicate_contigs.sh” script in MaSuRCA. Subsequently, PacBio reads were used to scaffold and gap-fill the assembly using PBjelly (v15.8.24)[8] (parameters: -minMatch 8 -minPctIdentity 70 -bestn 1 -nCandidates 20 -maxScore -500 -noSplitSubreads), and the redundancy filter was applied again. To achieve linkage-group resolution of the genome, Hi-C reads were applied to the data following established protocols[2,3,4]. The scaffolding process was repeated 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, in agreement with previously reported chromosome counts of *P. pyralis*[9]. The Hi-C read linkage distribution was visualized by mapping the reads to the assembly using the juicer pipeline (v1.5.6)[10] and the juicebox (v1.5.2) Hi-C visualization software. Finally, LG3 was split manually to separate the X-chromosome, and the assembly was filtered for contamination using the blobtools toolset (v1.0.1)[11]. See Supplementary Note 1.6 for more detail.

The mitochondrial genome (mtDNA) was assembled separately from the nuclear genome using short reads that mapped to the mtDNA of the *Pyrocoelia rufa* (NC_003970.1[12]), a close relative, using SPAdes (v3.8.0)[13]. The full circular mtDNA was achieved by manually assembling the short-read results with a single high-quality

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

PacBio circular-consensus-sequencing (CCS) read that spanned a remaining unresolved repetitive region. See Supplementary Note 1.8 for more detail.

***A. lateralis* genome sequencing and assembly.** Genomic DNA was extracted from the whole body of a single laboratory-reared *A. lateralis* adult female using the QIAamp Kit (Qiagen). Individuals of this strain have been kept in a small population under laboratory conditions for over 25 generations. Purified DNA was fragmented with a Covaris S2 sonicator (Covaris, Woburn,MA, USA), size-selected with a Pippin Prep (Sage Science, Beverly, MA, USA), 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 HiSeq1500 using a 125x125 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 100x100 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)[14], with default parameters and the "HAPLOIDIFY = True" option. Scaffolds were filtered to remove non-firefly contaminant sequences using blobtools[11], resulting in the final assembly (Alat1.3). For further details see Supplementary Note 2.

***I. luminosus* genome sequencing and assembly.** Adults were collected in Mayagüez, Puerto Rico, frozen at -80 ˚C, lyophilized, and finally stored at -80 ˚C. HMW DNA was extracted from a single *I. luminosus* male using 100/G Genomic Tips with the Genomic buffers kit (Qiagen, USA), 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. Pulsed field gel electrophoresis (PFGE) quality control and 10x Chromium Genome v1 library construction was performed at the Hudson Alpha Institute of Biotechnology Genomic Services Lab (HAIB-GSL). The library was then sequenced on a single lane of 150-nt PE HiSeqX (153x). Reads were then assembled using Supernova (v1.1.1)[15]. For further details see Supplementary Note 3.

Transcriptome sequencing and assembly. *De novo* and reference-guided transcriptomes were constructed for *P. pyralis*, *A. lateralis*, and *I. luminous* using both newly-generated and previously published RNAseq data following previously described general workflow[16]. For newly-generated data, RNA was extracted using an RNeasy Mini Kit (Qiagen, USA) or Trizol (Invitrogen) with DNase treatment.

For *P. pyralis*, MMNJ strand-specific libraries (male whole body, female whole body, eggs, larvae) were prepared by the Whitehead Genome Technology Core (WI-GTC) using the TruSeq Stranded mRNA library prep kit (Illumina) and then sequenced on 2 lanes of HiSeq2500 PE 150 for an average of ~30M paired reads per library. These were combined with previously-published stranded data from SRA (SRR3521424)[1,17] for *de novo* transcriptome assembly and annotation using Trinity (v2.4.0)[18] and PASA (v2.1.0)[19]. These data were also used to construct a reference-guided transcriptome using HISAT2 (v2.0.5)[20] and StringTie (v1.3.3b)[21]. In addition,

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

non-strand specific libraries were prepared and sequenced from dissected photophores of MMNJ males (n = 3) simultaneously with recently published RNA-SEQ[17] for an average of ~10M reads per library (see [17] for complete methods). These were combined with published non-strand specific data (Supplementary Table 1.9.1.1)[1,17] for a reference-guided assembly following the same procedure as the stranded data. See Supplementary Note 1.9 for more detail.

For *A. lateralis*, non-strand specific libraries were prepared from male body (thorax-to-fifth segment), female head, male head, female lantern, male lantern, eggs, larvae, and pupae of both sexes using a TruSeq RNA Sample Preparation Kit v2 (Illumina) at half scale (Supplementary 2.7). These were multiplexed and sequenced in a single lane of Hiseq1500 101x101 bp paired-end reads, *de novo* assembled with Trinity, and reference-guided assembled using TopHat (v2.0.11)[22] and Cufflinks (v2.2.1)[23]. See Supplementary Note 2.7 for more detail.

For *I. luminosus*, libraries were prepared from head + prothorax, thorax, and abdomen using either TruSeq RNA Library Prep Kit v2 (Illumina) or BGISEQ-500 RNA sample prep protocol (Supplementary 3.6). TruSeq RNA of head + prothorax was sequenced on two lanes of an Illumina HiSeq 2500 using rapid mode 100x100 bp PE at the WI-GTC. The remaining libraries were sequenced on the BGISEQ-500 platform (PE 50). *De novo* and reference-guided assembly procedures followed those for *P. pyralis*. See Supplementary Note 3.7 for more detail.

Whole-genome bisulfite sequencing (WGBS). MethylC-seq libraries were prepared from HMW DNA prepared from four *P. pyralis* MMNJ males using a previously published protocol[24], and sequenced to ~36x expected depth of by Illumina NextSeq500. Sequencing data was aligned to Ppyr1.3 using methylpy[25]. 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).

Coding gene annotation. For each species, direct gene models from the reference-guided and aligned *de novo* transcripts were produced via ORF prediction of transcripts using Transdecoder (v5.0.2)[26]. These gene models were combined with *ab initio* gene predictions and protein alignments using the Evidence Modeler (EVM) pipeline (v1.1.1)[19]. Genesets for the homologs of the luciferase, P450, and LRE gene families were then manually curated to produce the official genesets (*P. pyralis*: Ppyr_OGS1.0, *A. lateralis*: Alat_OGS1.0, and *I. luminosus*: Ilumi_OGS1.0). The peptides of the genesets were functionally annotated using InterProScan (v5.26-65.0)[27]. FASTA and GFF files for the the reference genesets of *P. pyralis*, *A. lateralis*, and *I. luminosus* are available on www.fireflybase.org. Gene content and contiguity were assessed using BUSCO (v3)[28] and QUAST(v4.5)[29]. Details of annotation are available in the Supplementary Materials (*P. pyralis*: Note 1.10, *A. lateralis:* Note 2.8, *I. luminosus:* Note 3.8).

Repeat annotation. Repeat prediction for *P. pyralis* was performed *de novo* using RepeatModeler (v1.0.9)[30] and MITE-Hunter (v11-2011)[31]. RepeatModeler uses

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

RECON[32] and RepeatScout[33] 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. The RepeatModeler and MITE-Hunter libraries were combined and and classified using RepeatClassifier (RepeatModeler 1.0.9 distribution[30]) to create the final repeat library of 3047 interspersed repeats. *De novo* species-specific repeat libraries for *A. lateralis* and *I. luminous* were constructed using RepeatModeler (v1.0.9), for final library sizes of 2952 and 4251 interspersed repeats, respectively. We used RepeatMasker (v4.0.5)[34] and the final repeat libraries to identify and mask interspersed and tandem repeats in the genome assemblies using the MAKER pipeline (v3.0.0β)[35]. To confirm presence of canonical telomere sequence, 5' fluorescein-tagged $(TTAGG)_5$ oligo probes (FAM; Integrated DNA Technologies) were used in fluorescence *in situ* hybridization (FISH) on squashed larval tissues according to previously published methods[36] (Supplementary Note 1.13).

Protein orthogroup clustering. Orthologs were identified by clustering the *P. pyralis*, *A. lateralis*, and *I. luminosus* reference geneset peptides with the *D. melanogaste*r (NCBI GCF_000001215.4) and *T. castaneum* (NCBI GCF_000002335.3) reference protein genesets using the Orthofinder (v2.1.2)[37] pipeline with parameters "-A mafft -T fasttree -I 1.5". Results are available on FigShare (DOI:10.6084).

Transcript expression analysis. *P. pyralis, A. lateralis, and I. luminosus* RNA-Seq reads were pseudoaligned to Ppyr_OGS1.0, Alat_OGS1.0, and Ilumi_OGS1.0 geneset mRNAs using Kallisto (v0.43.1)[38] with 100 bootstraps (-b 100), producing transcripts-per-million reads (TPM). For differential expression testing (*P. pyralis* and *A. lateralis* only), Kallisto results were between sample normalized using Sleuth (v0.29.0)[39] 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) and *A. lateralis* (adult male thorax + abdominal segments 1-5 vs. adult male dissected lantern) were performed using the Wald test within Sleuth. Genes whose mean BSN-TPM across bioreplicates 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[40] and GOAtools[41], with the modification that the enzymatic activity GO term was manually added to select InterPro annotations: IPR029058, IPR036291, and IPR001279. HE/DE/E+NotE gene filtering and overlaps (Fig 4e) were performed using custom scripts available on FigShare (DOI:10.6084). Kallisto expression analysis results are also available on FigShare.

Luciferase analysis. To examine luciferase evolution, a gene tree was constructed for *P. pyralis*, *A. lateralis*, and *I. luminosus* luciferase homologs. Candidate homologs were first identified by protein blast of *DmelPACS* (CG6178), the most similar gene to firefly luciferase in *D. melanogaster,* against the genomes and transcriptomes (e-value < 1.0×10^{-60}). Additionally, peroxisomal targeting signal 1 (PST1) was predicted using the

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

regular expressions provided by the Eukaryotic Linear Motif database[42] and verified using the mendel PTS1 prediction server[43,44]. Then homologs 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)[45] using the BLOSUM62 matrix and filtered for spurious sequences and poorly aligned regions using trimAl[46] (parameters: gt = 0.5). The final alignment was 544 amino acids. Then, the best fit amino acid substitution model, LG+F Gamma, was estimated by Aminosan (v1.0.2016.11.07)[47] using the Akaike Information Criterion. Finally, a maximum likelihood gene phylogeny was estimated using RAxML (v8.2.9; 100 bootstrap replicates)[48]. Parsimony ancestral state reconstruction of luminescence activity of luciferase was performed in Mesquite (v3.31)[49]. Elaterid luciferase was tested for positive selection using the adaptive branch-site REL test for episodic diversification (aBSREL)[50] via the DataMonkey[51] server. For more information see Supplementary Note 4.

CYP303 evolutionary analysis. Candidate P450s were identified using BLASTP (e-value: $1x10^{-20}$) 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)[45], and a preliminary neighbor-joining (NJ) tree was generated using MEGA7[52]. 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 (5 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 4c was truncated in Dendroscope (v3.5.9)[53] to display only the *CYP303* clade. The multiple sequence alignment FASTA files and newick files of the full and truncated tree are available on FigShare (DOI: 10.6084).

Assessment of lucibufagin content. Hemolymph was extracted from *P. pyralis* and *A. lateralis* and assessed for lucibufagin content by liquid-chromatography high-resolution accurate-mass mass-spectrometry (LC-HRAM-MS). In brief, methanolic hemolymph extracts were separated via gradient reversed-phase chromatography on a 150 mm C18 column using a UltiMate 3000 (Dionex, USA) liquid chromatography system coupled to a Q-Exactive mass spectrometer (Thermo Scientific, USA). Solvent A was H_2O + 0.1% formic acid, whereas Solvent B was acetonitrile + 0.1% formic acid. The mass spectrometer was configured to perform MS1 scans with polarity switching plus data dependent fragmentation of the most abundant ion. The raw data in Thermo format was converted to mzML format using ProteoWizard MSConvert[54]. Data analysis was performed with Xcalibur (Thermo Scientific) and MZmine2 (v2.30)[55]. Known lucibufagins were detected via the exact mass of their $[M+H]^+$ ion. Unknown lucibufagins were

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

identified via the MS2 fragmentation similarity to known lucibufagins using the MS2 similarity search module of MZmine2. See Supplementary Note 4.6 for more detail. Raw LC-MS data is available on MetaboLights (Accession: XXXXX).

Holobiome. Identification and annotation of genomes belonging to organisms associated with *P. pyralis*, including endosymbionts and parasitoids, followed a long-read metagenomic sequencing and assembly approach. Briefly, circular contigs from a PacBio-only assembly with Canu (Ppyr0.1-PB; Supplementary Note 1.7) were compared to the NCBI nt database via BLASTN. This approach identified one mycoplasma genome with 99% 16S identity to *Entomoplasma luminosum* and one parasitoid mitochondrion, which were fully assembled and curated using a targeted approach (Supplementary Note 5). Contigs was polished using the Quiver algorithm through SMRTPortal v2.3.0.140893[56] with the "RS_Resequencing.1" protocol (default parameters). Gene annotations of the mycoplasmal and phorid mitochondrial genomes were performed by Prokka (v.1.12)[57] and MITOS2[58], respectively. The genome sequence and annotation of *Entomoplasma luminosum* var. pyralis is available on NCBI (Accession: XXXXX)

Viruses were discovered from published *P. pyralis* RNA-seq data (NCBI TSA: GEZM00000000.1) and confirmed in our new RNA-Seq datasets and the Ppyr1.2 genome. Briefly, 24 *P. pyralis* libraries were downloaded (SRA: taxid7054, accessed: 15th June 2017) and assembled with Trinity. Potential virus transcripts were identified with TBLASTN searches (E-value = $1x10^{-5}$) using the complete predicted non-redundant viral Refseq proteins as probe (accessed: 15th June 2017). Significant hits were confirmed using BLAST against NCBI nt and nr databases, manually curated, and annotated in a custom workflow (Supplement 1.10.2). Tentative virus detections, and the viral refseq collection were contrasted to the *P. pyralis* genome assembly Ppyr1.2 by BLASTX searches (E-value = $1x10^{-6}$) and inspected by hand to identify genome regions that could be associated to endogenous viral-like elements. 15 kbp genome flanking regions were retrieved and annotated. Transposable Elements (TE) 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 < $1x10^{-10}$).

Amino acid sequences of the predicted viral polymerases, specifically the PB1 subunit, were used for phylogenetic analyses. Briefly, multiple sequence alignments were created using MAFTT (v7.310)[45] and then unrooted maximum-likelihood phylogenetic trees generated in FastTree[59] 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 category by using a Bayesian approach with a gamma prior. Support for individual nodes was accessed using an approximate likelihood ratio test with the Shimodaira-Hasegawa-like procedure. Tree topology, support values and substitutions per site were based on 1,000 tree resamples.

Additional methods of identification and assembly of associated holobiome members in *P. pyralis* can be found in Supplementary Note 1.10.2.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Data availability. Genomic assemblies (Ppyr1.3, Alat1.3, and Ilumi1.1), associated derived data, BLAST server, and a 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. Mitochondrial genomes for *P. pyralis* and *I. luminosus* and *A. antennatus* are available on NCBI GenBank with accessions KY778696, MG242621, and MG546669. LC-MS data is available on MetaboLights (Accession: XXXXX). Other supporting datasets are available on FigShare (DOI:10.6084).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. Sander, S. E. & Hall, D. W. Variation in opsin genes correlates with signalling ecology in North American fireflies. *Mol. Ecol.* **24,** 4679–4696 (2015).
2. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
3. Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31,** 1119–1125 (2013).
4. Bickhart, D. M. *et al.* Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49,** 643–650 (2017).
5. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29,** 2669–2677 (2013).
6. Zimin, A. V. *et al.* 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 (2017).
7. Pryszcz, L. P. & Gabaldón, T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* **44,** e113 (2016).
8. English, A. C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7,** e47768 (2012).
9. Wasserman, M. & Ehrman, L. Firefly Chromosomes, II (Lampyridae: Coleoptera). *Fla. Entomol.* **69,** 755–757 (1986).
10. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3,** 95–98 (2016).
11. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. *F1000Res.* **6,** (2017).
12. Bae, J. S., Kim, I., Sohn, H. D. & Jin, B. R. 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 (2004).
13. Nurk, S. *et al.* Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. in *Lecture Notes in Computer Science* 158–170 (2013).
14. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 1513–1518 (2011).
15. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* **27,** 757–767 (2017).
16. Torrens-Spence, M. P., Fallon, T. R. & Weng, J. K. A Workflow for Studying Specialized Metabolism in Nonmodel Eukaryotic Organisms. *Methods Enzymol.* **576,** 69–97 (2016).
17. Al-Wathiqui, N., Fallon, T. R., South, A., Weng, J.-K. & Lewis, S. M. Molecular characterization of firefly nuptial gifts: a multi-omics approach sheds light on postcopulatory sexual selection. *Sci. Rep.* **6,** 38556 (2016).
18. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–652 (2011).
19. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9,** R7 (2008).
20. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12,** 357–360 (2015).
21. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33,** 290–295 (2015).
22. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
23. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).
24. Urich, M. A., Nery, J. R., Lister, R., Schmitz, R. J. & Ecker, J. R. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat. Protoc.* **10,** 475–483 (2015).
25. Schultz, M. D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523,** 212–216 (2015).
26. Haas, B. J. TransDecoder. Available at: https://github.com/TransDecoder/TransDecoder/wiki.
27. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30,** 1236–1240 (2014).
28. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31,** 3210–3212 (2015).
29. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29,** 1072–1075 (2013).
30. Smit, A. & Hubley, R. *RepeatModeler*. (2017).
31. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38,** e199 (2010).
32. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12,** 1269–1276 (2002).
33. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1,** i351–8 (2005).
34. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013--2015. *Institute for Systems Biology. http://repeatmasker. org* (2015).
35. Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18,** 188–196 (2008).
36. Larracuente, A. M. & Ferree, P. M. Simple method for fluorescence DNA in situ hybridization to squashed chromosomes. *J. Vis. Exp.* 52288 (2015).
37. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16,** 157 (2015).
38. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34,** 525–527 (2016).
39. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods* **14,** 687–690

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(2017).
40. Fallon, T. R. interproscan_to_enzyme_go.py. Available at: https://github.com/photocyte/interproscan_to_enzyme_go/tree/master.
41. Tang, H. *et al.* GOATOOLS: tools for gene ontology. *Zenodo.* (2015).
42. Dinkel, H. *et al.* ELM—the database of eukaryotic linear motifs. *Nucleic Acids Res.* **40,** D242–D251 (2012).
43. Neuberger, G., Maurer-Stroh, S., Eisenhaber, B., Hartig, A. & Eisenhaber, F. Motif refinement of the peroxisomal targeting signal 1 and evaluation of taxon-specific differences. *J. Mol. Biol.* **328,** 567–579 (2003).
44. Georg Neuberger, Sebastian Maurer-Stroh, Birgit Eisenhaber, Andreas Hartig and Frank Eisenhaber. The PTS1 Predictor. Available at: http://mendel.imp.ac.at/mendeljsp/sat/pts1/PTS1predictor.jsp.
45. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
46. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25,** 1972–1973 (2009).
47. Tanabe, A. S. 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 (2011).
48. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
49. Maddison, W. P. & Maddison, D. R. *Mesquite: a modular system for evolutionary analysis*. (2017).
50. Smith, M. D. *et al.* Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* **32,** 1342–1353 (2015).
51. Delport, W., Poon, A. F. Y., Frost, S. D. W. & Kosakovsky Pond, S. L. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26,** 2455–2457 (2010).
52. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **33,** 1870–1874 (2016).
53. Huson, D. H. & Scornavacca, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* **61,** 1061–1067 (2012).
54. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30,** 918–920 (2012).
55. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11,** 395 (2010).
56. Biosciences, P. SMRT Analysis Software. Available at: http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/.
57. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).
58. M. Bernt, A. Donath, F. Jühling, F. Externbrink, C. Florentz, G. Fritzsch, J. Pütz, M. Middendorf, P. F. Stadler. MITOS2 WebServer. Available at:

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

http://mitos2.bioinf.uni-leipzig.de.

59. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5,** e9490 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SUPPLEMENTARY MATERIALS

Fallon, Lower *et al.* 2017

TABLE OF CONTENTS

SUPPLEMENTARY NOTE 1: *Photinus pyralis* additional information 3
1.1 Taxonomy, biology, and life history 3
Supplementary Video 1: A *Photinus pyralis* courtship dialogue 5
1.2 Species distribution 5
1.2.1 Supplementary Figure: Detailed geographic distribution map for *P. pyralis* 7
1.3 Specimen collection and identification 7
1.3.1. Supplementary Figure: *P. pyralis* aedeagus (male genitalia) 8
1.4 Karyotype and genome size 8
1.5 Library preparation and sequencing 8
1.5.1 Illumina 8
1.5.2 PacBio 9
1.5.2.1 Supplementary Figure: PFGE of *P. pyralis* HMW DNA used for PacBio sequencing 10
1.5.2.2 Supplementary Figure: Subread length distribution for *P. pyralis* PacBio RSII sequencing. 11
1.5.3 Hi-C 11
1.6 Genome assembly 12
1.6.2 Ppyr1.1: MaSuRCA hybrid assembly 12
1.6.3 Ppyr1.2: Scaffolding with Hi-C 12
1.6.4 Ppyr1.3: Manual curation and taxonomic annotation filtering 13
1.6.4.1 Defining the X chromosome 13
1.6.4.2 Taxonomic annotation filtering 13
1.6.4.2.1 Supplementary Figure: Blobplot of Illumina short-insert reads aligned against Ppyr1.2 14
1.6.4.2.2 Supplementary Figure: Blobplot of *P. pyralis* PacBio reads aligned against Ppyr1.2 15
1.6.4.2.3 Supplementary Figure: Venn diagram representation of blobtools taxonomic annotation filtering approach for Ppyr1.2 scaffolds. 16
1.7 Ppyr0.1-PB: PacBio only genome assembly 16
1.8 Mitochondrial genome assembly and annotation 16
1.8.1 Supplementary Figure: Mitochondrial genome of *P. pyralis* 18
1.9 Transcriptome analysis 19
1.9.1 RNA-extraction, library preparation and sequencing 19
1.9.1.1 Supplementary Table: *P. pyralis* RNA sequencing libraries 20
1.9.2 De novo transcriptome assembly and genome alignment 20
1.9.3 Reference guided transcriptome assembly 21
1.10 Coding gene annotation 21

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1.10.1 P450 annotation 22
1.10.1.1 Supplementary Figure: *P. pyralis* P450 gene phylogenetic tree 24
1.10.2 Virus annotation and analysis 24
1.11 Repeat annotation 27
1.11.1 Supplementary Table: RepeatMasker annotation of *P. pyralis* repetitive elements 27
1.12 Methylation analysis 28
1.13 Telomere FISH analysis 28

SUPPLEMENTARY NOTE 2: *Aquatica lateralis* additional information 30
2.1 Taxonomy, biology, and life history 30
2.2 Species distribution 30
2.3 Specimen collection 30
2.4 Karyotype and genome size 31
2.5 Genomic sequencing and assembly 31
2.5.1 Taxonomic annotation filtering 31
2.5.1.1 Supplementary Figure: Blobplot of *A. lateralis* Illumina reads aligned against Alat1.2 33
2.6 RNA-extraction, library preparation and sequencing 33
2.6.1 Supplementary Table: *Aquatica lateralis* RNA sequencing 34
2.7 Transcriptome assembly 35
2.7.1 De novo transcriptome assembly and alignment 35
2.7.2 Reference guided transcriptome alignment and assembly 35
2.8 Coding gene annotation 35
2.9 Repeat annotation 36
2.9.1 Supplementary table: RepeatMasker annotation of *A. lateralis* repetitive elements 36

SUPPLEMENTARY NOTE 3: *Ignelater luminosus* additional information 38
3.1 Taxonomy, biology, and life history 38
3.2 Species distribution 39
3.3 Collection 39
3.3.1 Supplementary Figure: *I. luminosus* aedeagus (male genitalia) 40
3.4 Karyotype and genome size 40
3.5 Genomic sequencing and assembly 40
3.5.3 Taxonomic annotation filtering 41
3.5.3.1 Supplementary figure: Blobtools plot of Ilumi1.0 42
3.5.4 Indel polishing 42
3.6 RNA extraction, library prep, and sequencing 43
3.6.1 HiSeq2500 43
3.6.2 BGISEQ-500 43
3.6.3 Supplementary Table: *I. luminous* RNA-Seq libraries 44
3.7 Transcriptome assembly 44
3.7.1 De novo transcriptome assembly and alignment 44

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3.7.2 Reference guided transcriptome alignment and assembly 45
3.8 Coding gene annotation 45
3.9 Repeat annotation 46
3.10 Mitochondrial genome assembly and annotation 47
3.10.1 Supplementary Figure: Mitochondrial genome of *I. luminosus* 49

SUPPLEMENTARY NOTE 4: Comparative analyses 50
4.1 Assembly statistics and comparisons 50
4.1.1 Supplementary Table: Assembly statistics 50
4.1.2 Supplementary Table: Comparison of BUSCO conserved gene content with other insect genome assemblies 51
4.1.3 Supplementary Figure: Venn Diagram of *P.pyralis*, *A. lateralis*, *I.luminosus*, T. castaneum, and *D. melanogaster* orthogroup relationships. 53
4.2 Supplementary Figure: De novo methyltransferase gene phylogeny 54
4.3 Luciferase evolution 54
4.2.1 Supplementary Figure: Maximum likelihood phylogeny of luciferase homologs 56
4.2.2 Supplementary Figure: Conserved intron-exon structure of *P. pyralis* and *A. lateralis luciferases* 57
4.3.1 Testing for selection on Elaterid luciferase 58
4.3.1.1 Supplementary Figure: Adaptive branch-site REL test for episodic diversification (aBSREL) of elaterid luciferase homologs. 59
4.3.1.2 Supplementary Figure: Proportion of sites under positive selection 60
4.4 Non-enzyme highly and differentially expressed genes of the firefly lantern 60
4.4.1 Supplementary Table: 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 61
4.4.2 Supplementary Figure: Maximum likelihood phylogeny of the combined adenylyl-sulfate kinase & sulfate adenylyltransferase orthogroup. 63
4.5 Opsin analysis 63
4.5.1 Supplementary Figure: ML tree and gene expression levels of opsin genes. 65
4.6 LC-HRAM-MS of lucibufagin content in *P. pyralis* and *A. lateralis* 66
4.6.1 Supplementary Figure: Positive mode MS1 total-ion-chromatogram (TIC) of *P. pyralis* hemolymph LC-HRAM-MS data 68
4.6.2 Supplementary Figure: Negative mode MS1 total-ion-chromatogram (TIC) of *P. pyralis hemolymph LC-HRAM-MS data* 68
4.6.3 Supplementary Figure: Positive mode MS1 total-ion-chromatogram (TIC) of *A. lateralis* hemolymph LC-HRAM-MS data 69
4.6.4 Supplementary Figure: Negative mode MS1 total-ion-chromatogram (TIC) of *A. lateralis* hemolymph LC-HRAM-MS data 69
4.6.5 MS2 similarity search for *P. pyralis* lucibufagins 69
4.6.5.2 Supplementary Figure: MS2 spectral similarity network for *P. pyralis* hemolymph lucibufagins 71
4.6.5.3 Supplementary Table: Putative lucibufagin compounds from LC-HRAM-MS of *P. pyralis* hemolymph 71

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

4.6.7 MS2 similarity search for *A. lateralis* lucibufagins 72
4.6.7.1 Supplementary Table: Relative quantification of features identified by lucibufagin MS2 similarity search 73

SUPPLEMENTARY NOTE 5: Holobiont analyses 74
5.1 Assembly of the complete *Entomoplasma luminosum* var. pyralis genome 74
5.2 Assembly of Phorid mitochondrial genome 74
5.2.1 Supplementary Figure: Mitochondrial genome of *Apocephalus antennatus*. 76
5.3 Taxonomic identification of Phorid mitochondrial genome origin 76
5.4 Photinus pyralis orthomyxo-like viruses 77
5.4.1 Supplementary Figure: Photinus pyralis viruses and endogenous viral-like elements. 86
5.4.2 Supplementary Figure: Pairwise identity of OMLV viral proteins amongst identified OMLV viruses. 88
5.4.3 Supplementary Table: Best hits from BLASTP of PpyrOMLV proteins against the NCBI database 89
5.4.4 Supplementary Table: InterProScan domain annotation of PpyrOMLV proteins 90
5.4.5 Supplementary Table: Reads mapped to PpyrOMLV genome segments in available SRA and newly described P. pyralis RNA-Seq datasets 92
5.4.6 Supplementary Table: Reads mapped to PpyrOMLV genome segments in with FPKM values in available SRA and newly described P. pyralis RNA-Seq datasets 93
5.5 P. pyralis Endogenous virus-like Elements (EVEs) 94
5.5.1 Supplementary Table: FEVE hits from BLASTX of PpyrOMLV PB1 97
5.5.2 Supplementary Table: FEVE hits from BLASTX of PpyrOMLV PB2 99
5.5.3 Supplementary Table: FEVE hits from BLASTX of PpyrOMLV PA 99
5.5.4 Supplementary Table: FEVE hits from BLASTX of PpyrOMLV HA 99
5.5.5 Supplementary Table: FEVE hits from BLASTX of PpyrOMLV NP 100

SUPPLEMENTARY NOTE 6: Experiment.com donors 101

SUPPLEMENTARY NOTE 7: Data availability 103

Bibliography 104

SUPPLEMENTARY NOTE 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.[1,2] 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.[3] 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 minutes before sunset.[1] Male flashing is cued by

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ambient light levels, thus shaded or unshaded habitats can show up to a 30 minute difference in the initiation of male flashing[1]. Males can be cued to flash outside of true twilight if exposed to light intensities simulating twilight[4]. *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 minutes 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 second intervals[4]. Males emit their patrol flash while dipping down and then ascending vertically, creating a distinctive J-shaped flash gesture[1,4] (Fig. 1a). During courtship, females perch on vegetation and respond to a male patrol flash by twisting their abdomen towards the source of the flash and giving a single response flash given after a 2-3 sec delay (Supplemental Video 1). 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[5]; the molecular composition of this nuptial gift has recently been elucidated for *P. pyralis*[6]. In other *Photinus* species, male gift size decreases across sequential matings[7], and multiple matings are associated with increased female fecundity[8].

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*[9]. 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[10]. 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[11], which may be possibly be observed in the South as a "second wave" of signalling *P. pyralis* in September.

Anti-predator chemical defenses of male *P. pyralis* include several bufadienolides, known as lucibufagins, that circulate in the hemolymph[12]. Pterins have also been reported to be abundant in *P. pyralis*[13], 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 blood, and such reflex bleeding may also provide physical protection against small predators[14,15].

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Video 1: A Photinus pyralis courtship dialogue

A *P. pyralis* male (right) signals to a female on a leaf (center) as competitor males flash in the background. The male demonstrates the characteristic *P. pyralis* swooping "J" flash every ~5 seconds. The female demonstrates the characteristic abdominal rotation toward the male and delay in flash reply. Footage taken during summer 2012 in Jackson Co., Tennessee, USA by SESL.

1.2 Species distribution

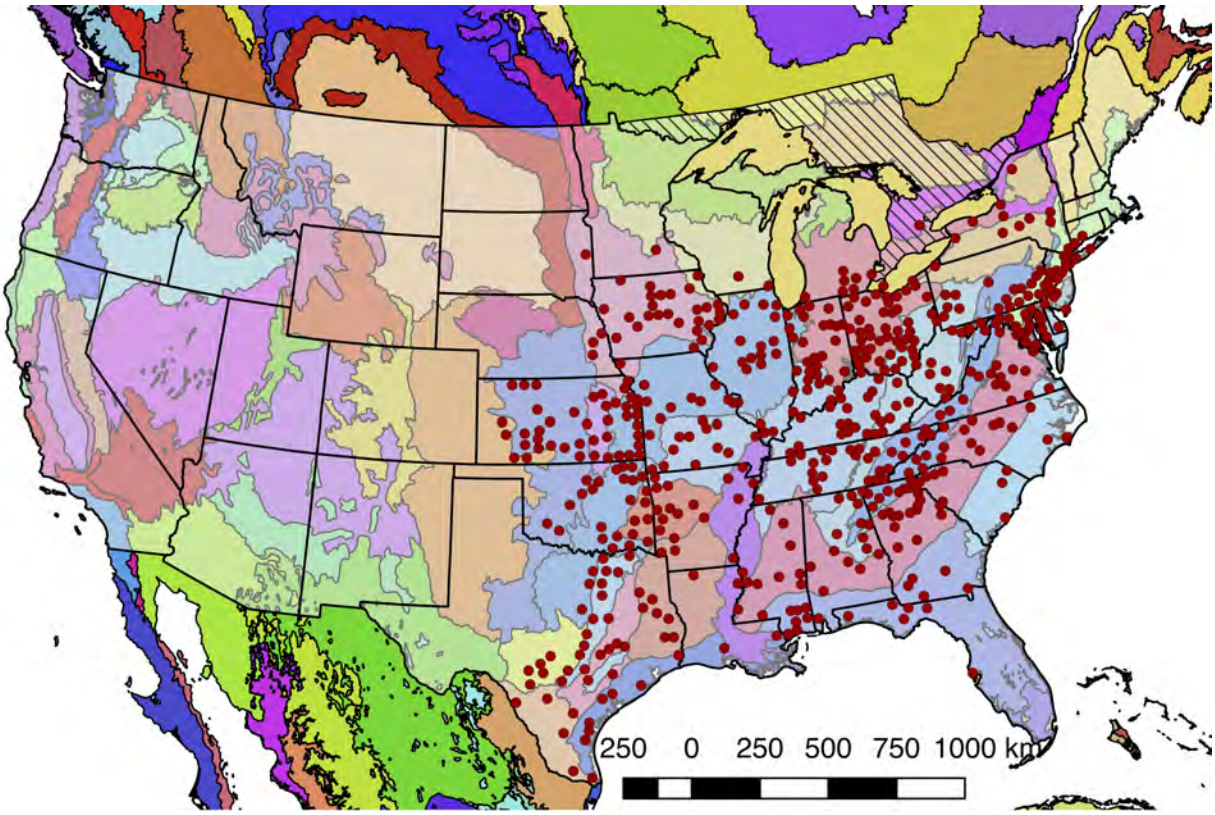
Although *Photinus pyralis* is widely distributed in the Eastern United States, published descriptions of its range are limited, with the notable exception of Lloyd's 1966 monograph[1] which addresses the range of many *Photinus* species. We therefore sought to characterize the current distribution of *P. pyralis* in order to provide an updated map to inform our experimental design and enable future population genetic studies. Three sources of data were used to produce the presented range map of *P. pyralis*: (i) Published[1,16] and unpublished sightings of *P. pyralis* at county level resolution, provided by Dr. J. Lloyd (University of Florida) and Dr. K. Stanger-Hall (University of Georgia), (ii) coordinates and dates of *P. pyralis* sightings, obtained

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

by targeted e-mail surveys to firefly field biologists, (iii) citizen scientist reports of *P. pyralis* through the iNaturalist platform[17]. 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)[18] was used for data viewing and figure creation. A custom Python script within QGIS[19] was used to link *P. pyralis* sightings to countries from the US census shapefile[20]. Outlying points that were located in Desert Ecoregions of the World Wildlife Fund (WWF) Terrestrial Ecoregions shapefile[21,22] or the westernmost edge of the range were manually removed, as they are likely not representative of the general 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 (Supplementary Figure 1.2.1), red circles indicate county-centroided presence records.

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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



1.2.1 Supplementary Figure: Detailed geographic distribution map for *P. pyralis*

P. pyralis sightings (red circles show county centroids) in the United States and Ontario, Canada (diagonal hashes). The World Wildlife Fund Terrestrial Ecoregions[21,22] are also shown (colored shapes). The dataset shown is identical to that used to prepare Figure 1b.

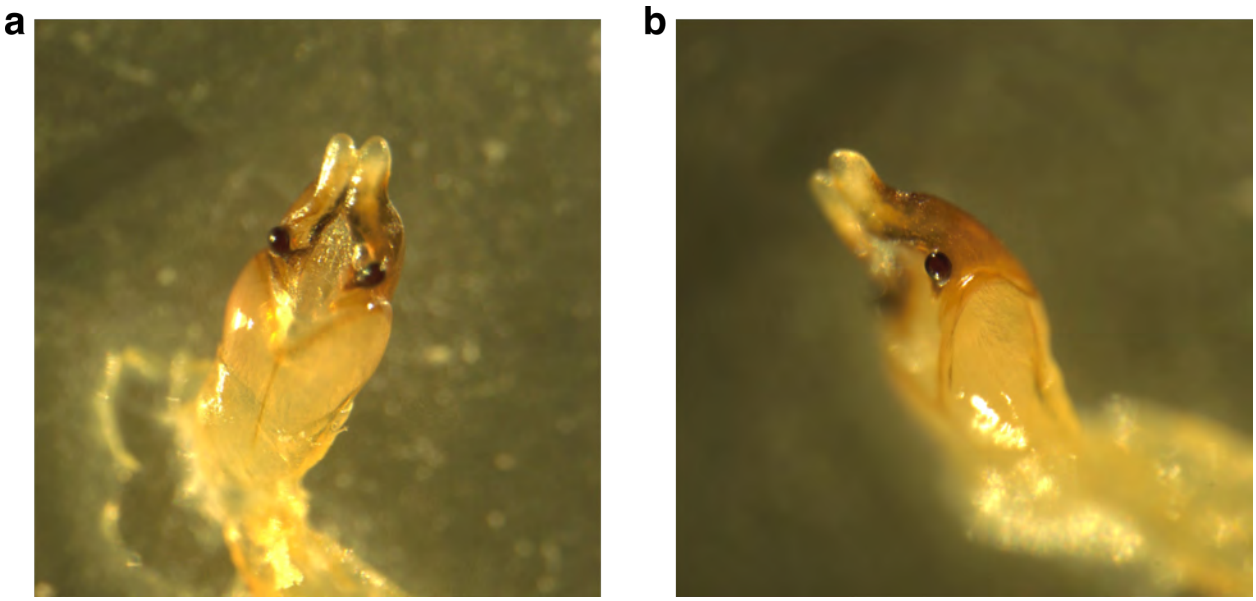
1.3 Specimen collection and identification

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,[23] flash pattern and behavior,[1] and *cytochrome-oxidase I* (*COI*) similarity (partial sequence: primers HCO, LCO[24]) 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.

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, verified to be likely *P. pyralis* by the presence of the margin of ventral unpigmented abdominal tissue anterior to the lanterns, flash frozen with liquid N_2, lyophilized, and stored at -80˚C until DNA extraction. A

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

single aedeagus (male genitalia) was dissected from the stored specimens and confirmed to match the *P. pyralis* taxonomic key[23] (Supplementary Figure 1.3.1).



1.3.1. Supplementary Figure: *P. pyralis* aedeagus (male genitalia)

a, Ventral and **b,** side view of a *P. pyralis* aedeagus dissected from specimens collected at the same time and place 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*[23].

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)[25]. 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[26]. From these analyses, the size of the X-chromosome is inferred to be ~26 Mbp.

1.5 Library preparation and sequencing

1.5.1 Illumina

DNA was extracted from sterile-water-washed thorax 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*

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

males are XO[27], 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 HiSeq2000 100x100 bp PE reads (University of Texas; Extended Data Table 1).

1.5.2 PacBio

High-molecular-weight DNA (HMW DNA) was extracted from lyophilized adult male *P. pyralis* (dry mass 90.8 mg), which were first washed using 95% ethanol, using 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 4 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 greatly 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.

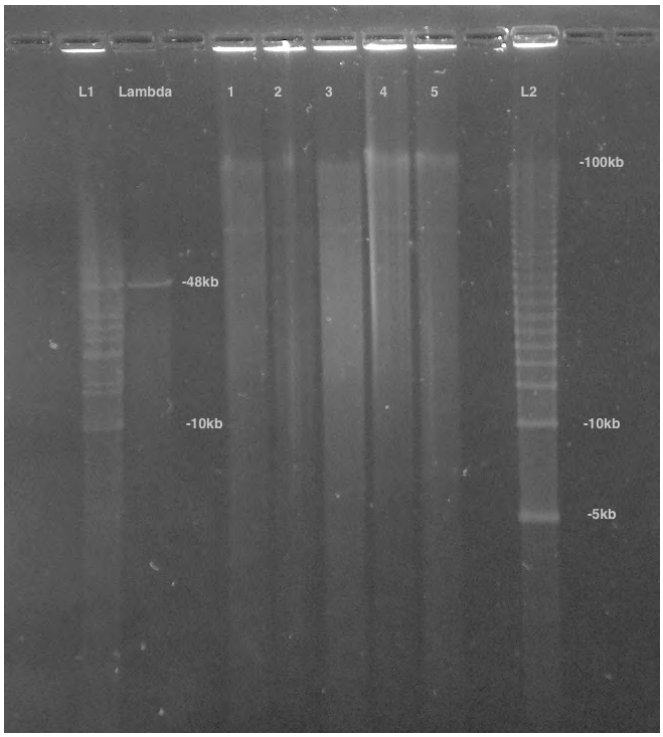
Precipitated HMW DNA was redissolved in 80 µL Qiagen QLE buffer (10 mM Tris-Cl, 0.1 mM EDTA, pH 8.5) yielding 17.1 µg of DNA (214 ng/µL) and glycogen (500 ng/µL). Final DNA concentration was measured with a Qubit fluorometer (Thermo Scientific) using the Qubit Broad Range kit. Manipulations hereafter, 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, USA).

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 BioRad CHEF DRIII system. The sample was run out for 16 hours at 6 volts/cm with an angle of 120 degrees with a running temperature of 14˚C. The gel was stained with SYBRgreen dye (Thermo Scientific - Part No. S75683). 1 µg of 5 Kbp ladder (BioRad, part no 170-3624) was used as a standard. These results demonstrated the HMW DNA had a mean size of >48 Kbp (Supplementary Figure 1.5.2.1). 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 minutes. 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, USA). Two separate cassettes were run. In each cassette, 2 lanes were used in which there

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

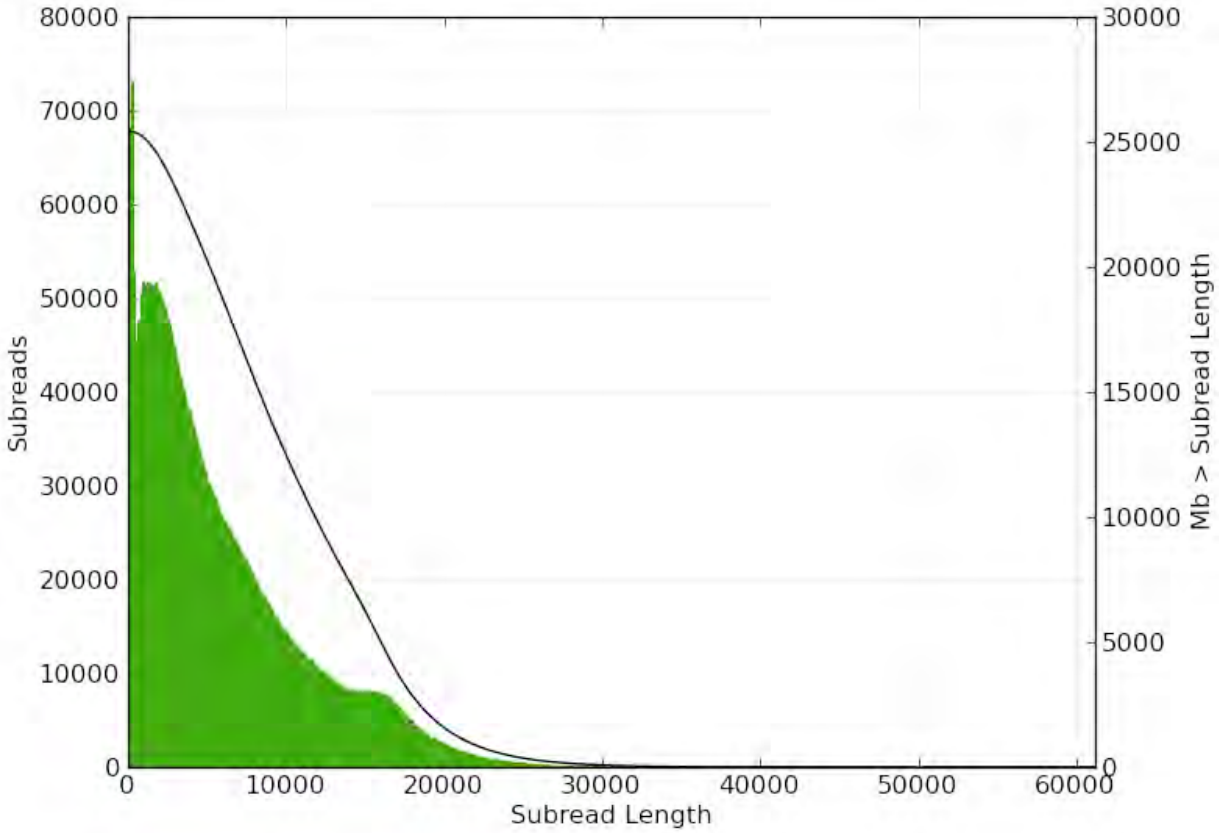
was 1362 ng/lane (PAC20kb kit). Constructs 15 Kbp and above were eluted over a period of four hours. An additional damage repair step was carried out post size-selection. Insert size range for the final library was determined using the Fragment Analyzer System (Advanced Analytical, Ankeney IA, USA). The size-selected SMRTbell library was then sequenced over 61 SMRT cells on the PacBio RSII instrument of the Broad Technology Labs (Cambridge, MA), using the P6 v.2 polymerase and the v.4 DNA Sequencing Reagent (part numbers 100-372-700, 100-612-400). PacBio sequencing data is available on the NCBI Sequence Read Archive (Bioproject PRJNA378805).



1.5.2.1 Supplementary Figure: 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.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



1.5.2.2 Supplementary Figure: Subread length distribution for *P. pyralis* PacBio RSII sequencing.

Figure produced with SMRTPortal (v2.3.0.140936)[28] by aligning all PacBio reads from data from the 61 SMRT cells against Ppyr1.3 using the RS_Resequencing.1 protocol with default parameters.

1.5.3 Hi-C

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[29–31]. Briefly, a streamlined version of the standard Hi-C protocol[29] 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.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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, (2) Ppyr1.1 scaffolded using Hi-C data, producing assembly Ppyr1.2, and (3) Ppyr1.2 manually curated for proper X-chromosome assembly and removal of putative non-firefly sequences, producing Ppyr1.3.

1.6.2 Ppyr1.1: MaSuRCA hybrid assembly

We first applied MaSuRCA (v3.2.1_01032017)[32,33] to correct short and long reads. MaSuRCA includes a step to remove Illumina adaptors within the pipeline. We modified the pipeline to assemble the genome using both corrected long reads (Mega-reads) and synthetic long reads (Super-reads) with smaller overlap length (59 bp). Then we used all reads, including short reads, to call the genomic consensus.

To scaffold the contigs, we first manually integrated the MaSuRCA assembly by replacing the mitochondrial contigs with complete mitochondrial assemblies from *P. pyralis* and *Apocephalus antennatus* (Supplementary Note 5.2). We scaffolded and gap-filled assembly by Redundans (v0.13a)[34] with default settings using Illumina reads. The Illumina short-reads from mate-end libraries were filtered by Nxtrim (v0.4.1)[35] with parameters "--separate --rf --justmp" so that only the mate-end reads were retained. After scaffolding by Illumina, the redundant sequences were removed by the built-in MaSuRCA script. Then we applied PBjelly (v15.8.24)[36] and PacBio reads to scaffold and gap-fill the assembly. Again, the redundant sequences from the output assembly were removed by the MaSuRCA built-in script. Finally, we replaced mitochondrial sequences which had been artifically extended by the scaffolding, gap-filling and sequence extension process with the proper sequences .

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[30] and later expanded upon[31]. Briefly, Hi-C reads were mapped to Ppyr1.1 with BWA (v1.7.13)[37], 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[30,31]. 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[30,31], 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*[25]. However, to verify the correctness of this assumption, scaffolds were created for haploid chromosome numbers ranging from 5 to 15. A

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

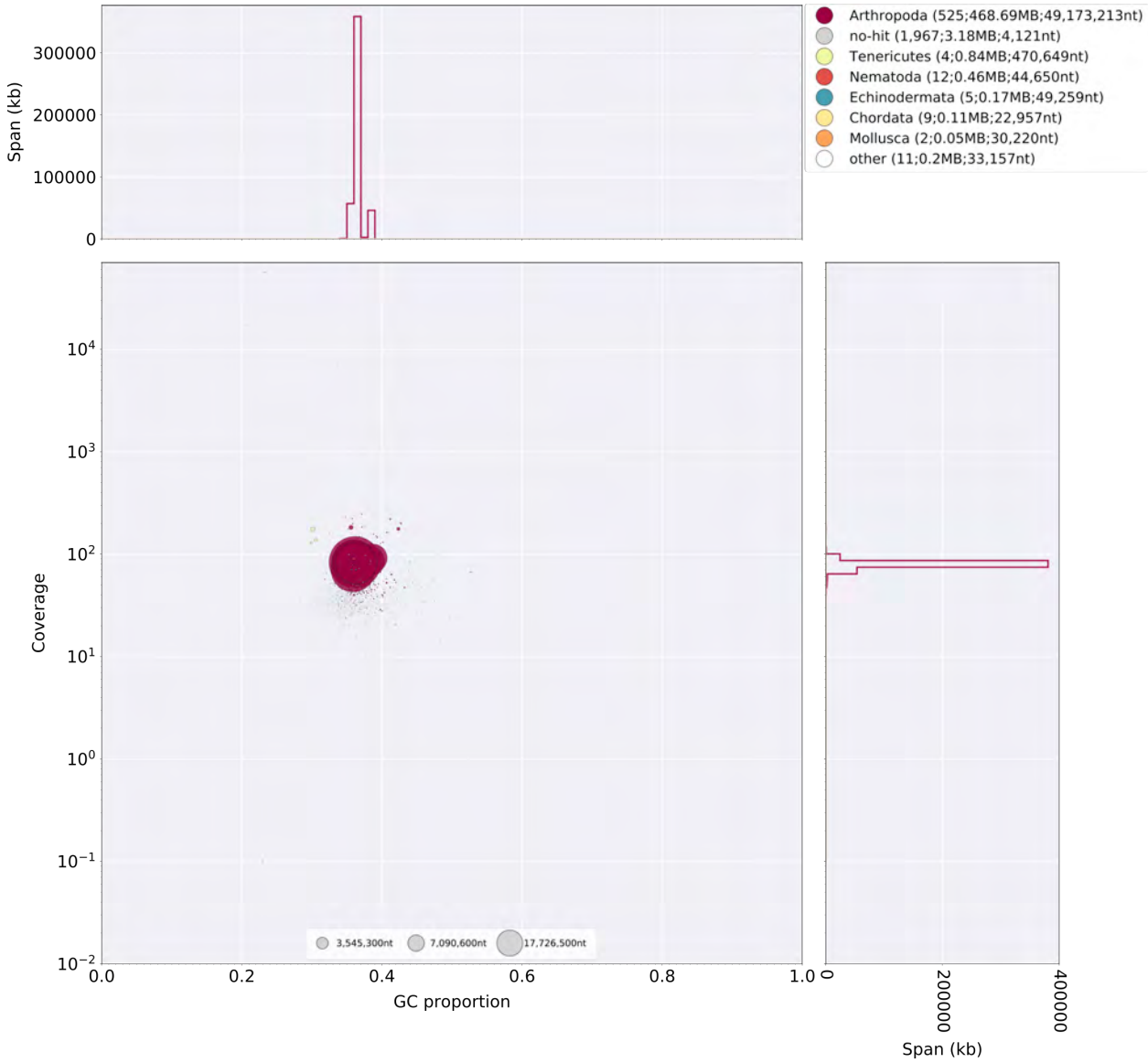
Hi-C data was mapped and converted to .hic format with the juicer pipeline (v1.5.6)[38], and then visualized using juicebox (v1.5.2)[39]. This visualization revealed a clear breakpoint in Hi-C linkage density on LG3 at ~22,220,000 bp. Mapping of Illumina short-insert and PacBio reads with Bowtie2 (v2.3.1)[40] and SMRTPortal (v2.3.0.140893) with the "RS_Resequencing.1" protocol, followed by visualization with Qualimap (v2.2.1)[41], 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)[42] and Exonerate(v2.2.0)[43] 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[44], we sought to systematically remove assembled non-firefly contaminant sequence from Ppyr1.2. Using the blobtools toolset (v1.0.1)[45], 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)[46] 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 (Supplementary Figure 1.6.3.2.1) and PacBio libraries (Supplementary Figure 1.6.3.2.2), and if the taxonomic phylum was annotated as "Arthropod" or "no-hit" by the blobtools pipeline. 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 (Supplementary Note 5) were removed. Furthermore we removed two contigs representing the mitochondrial genome of *P. pyralis* (complete mtDNA

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

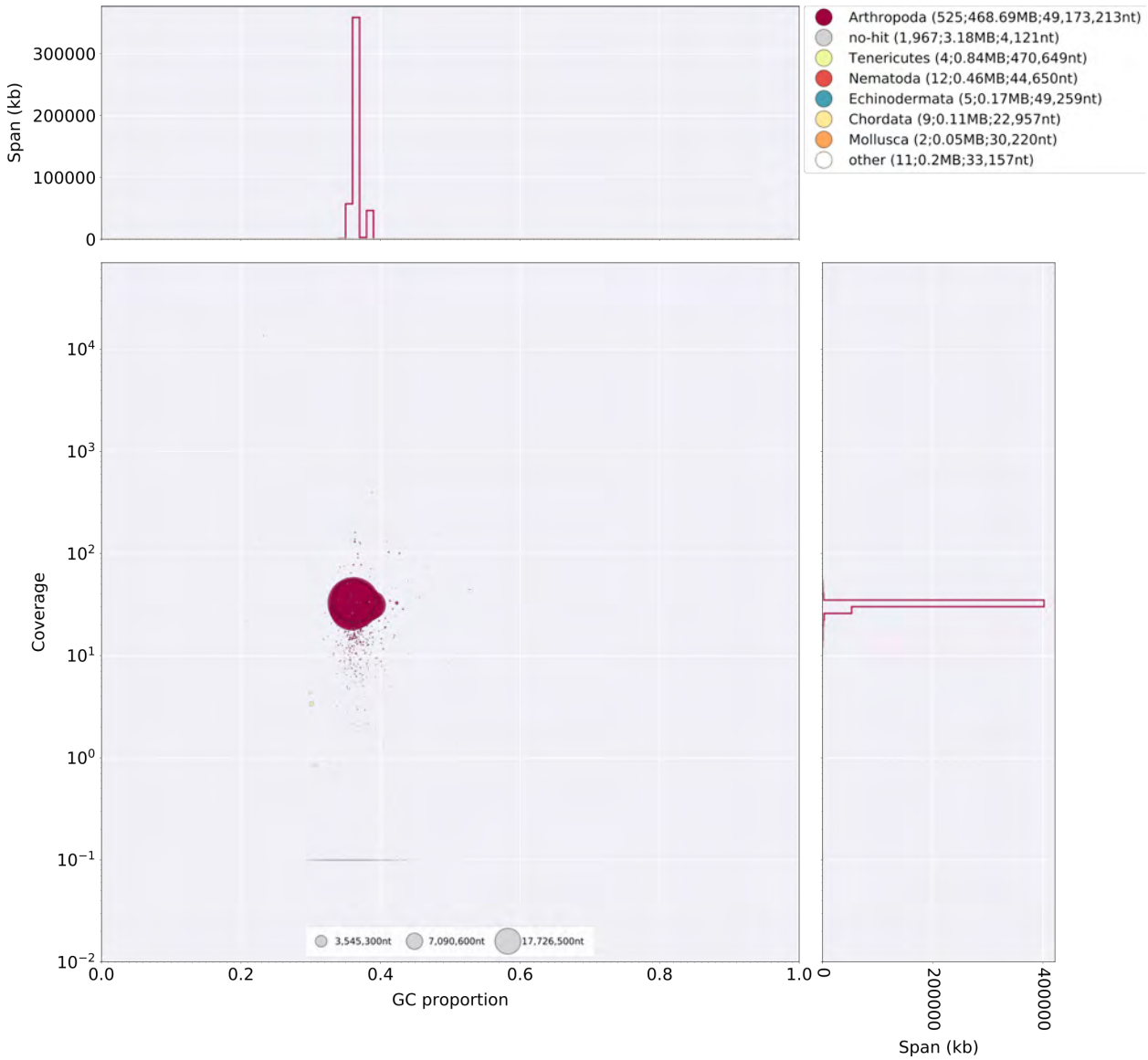
available via Genbank: KY778696). The final filtered assembly, Ppyr1.3, is available at www.fireflybase.org.



1.6.4.2.1 Supplementary Figure: Blobplot of Illumina short-insert reads aligned against Ppyr1.2

Coverage shown represents mean coverage of reads from the Illumina short-insert library (Sample name 8369; Supplementary Table 1.5.4), aligned against Ppyr1.2 using Bowtie2 with parameters (--local). Scaffolds were taxonomically annotated as described in Supplementary Note 1.6.3.2.

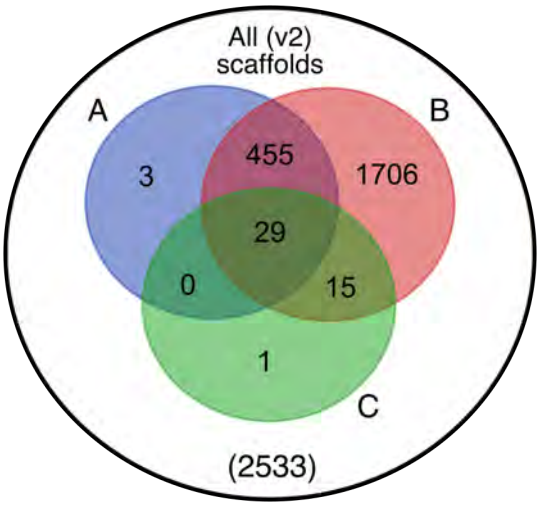
bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



1.6.4.2.2 Supplementary Figure: Blobplot of *P. pyralis* PacBio reads aligned against Ppyr1.2

Coverage shows represents mean coverage of reads from the PacBio library (Sample name 1611; Supplementary Table 1.5.4). 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 Supplementary Note 1.6.3.2.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



1.6.4.2.3 Supplementary Figure: 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 celsl to .FASTQ format subreads using SMRTPortal (v2.3.0.140893)[28] with the "RS_Subreads.1" protocol with default parameters. These subreads were then input into Canu (Github commit 28ecea5 / v1.6)[47] with parameters "genomeSize=450m corOutCoverage=200 ovlErrorRate=0.15 obtErrorRate=0.15 -pacbio-raw". The unpolished contigs from this produced genome assembly are dubbed Ppyr0.1-PB.

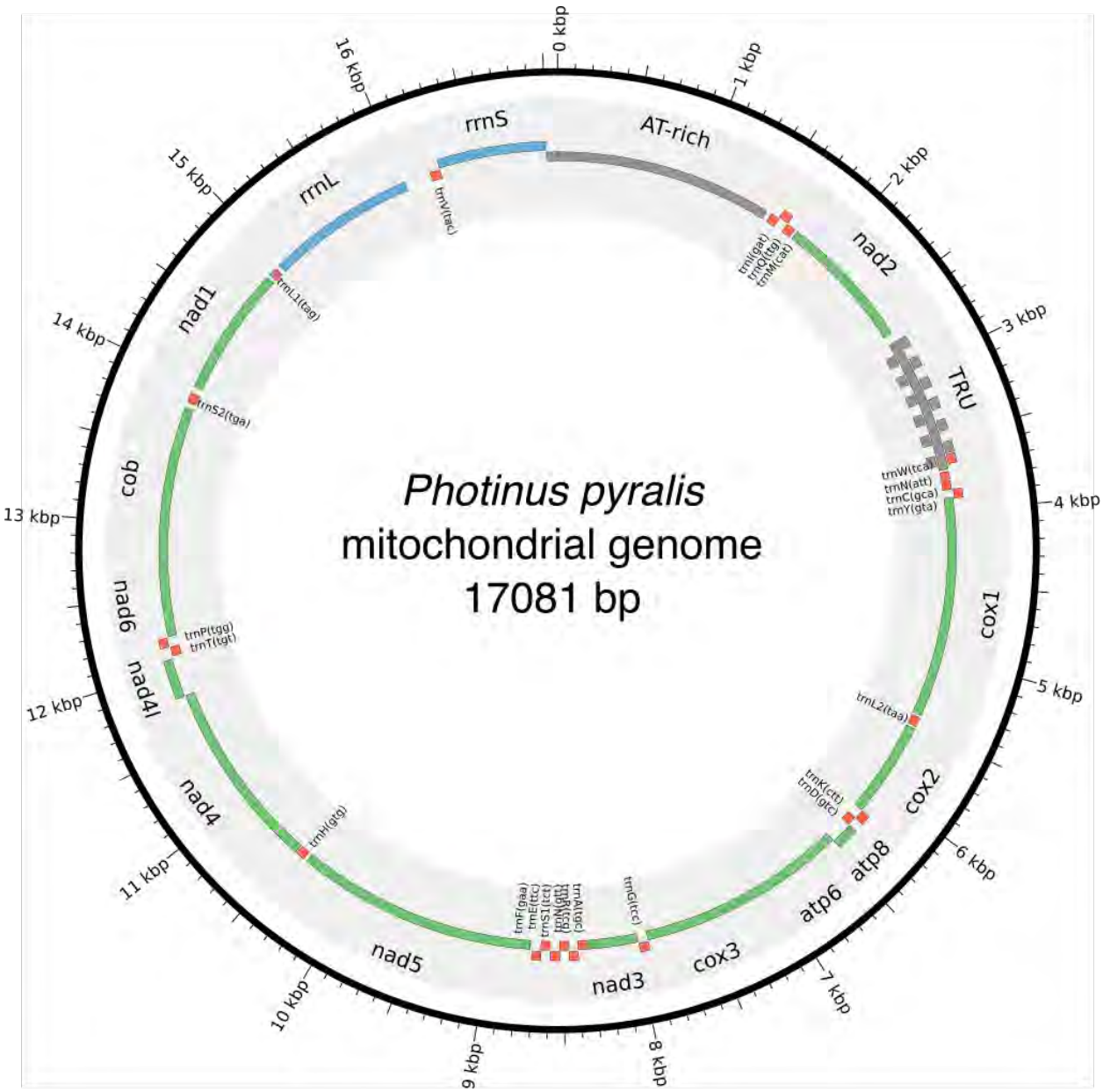
1.8 Mitochondrial genome assembly and annotation

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; Extended Data Table 1) were mapped to the known mtDNA of the closest available relative, *Pyrocoelia rufa* (NC_003970.1[48]) using bowtie2 v2.3.1 (parameters: --very-sensitive-local). All concordant read pairs were input to SPAdes

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(v3.8.0)[49] (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)[48], 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[50]. 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; Extended Data Table 1). The *P. pyralis* mtDNA was then "restarted" using seqkit[51], such that the FASTA record break occurred in the AT-rich region, and annotated using the MITOS2 annotation server[52]. Low confidence and duplicate gene predictions were manually removed from the MITOS2 annotation. The final *P. pyralis* mtDNA plus annotations is available on GenBank (KY778696).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



1.8.1 Supplementary Figure: 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[53].

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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, 1st instar larvae, adult male, and adult female; Supplementary Table 1.9.1.1). 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 N_2, and stored at -80˚C for RNA extraction. Resulting eggs were washed 3x with dilute bleach/ H_2O 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 (1st 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 150x150 nt on one rapid mode flowcell (2 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 (Supplementary Table 1.9.1.1). 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 3 individuals per biological replicate. These tissues and libraries were co-prepared and sequenced with other libraries (full library preparation and sequencing details here[6]) at a depth of ~10M paired reads per library.

All of these data were combined with previously published tissue, sex, and stage-specific libraries (Supplementary Table 1.9.1.1) for reference-guided transcriptome assembly. Strand-specific data was used for *de novo* transcriptome assembly.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1.9.1.1 Supplementary Table: *P. pyralis* RNA sequencing libraries

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

Library name	Source[a]	N	Sex/stage	Tissue	Library type
8175 Photinus pyralis male head (adult) transcriptome	SRA1	1	M/A	H	
8176 Photinus pyralis male light organ (adult) transcriptome	SRA1	1	M/A	PA	
8819 Photinus pyralis light organ (larval) transcriptome	SRA1	1	L	PA	
9_Photinus_sp_1_lantern	SRA2	1	M/A	PA	Strand-specific. Ribo-zero
Ppyr_FatBody_1	SRA3	6	M/A	FB	
Ppyr_FatBody_2	SRA3	6	M/A	FB	
Ppyr_FatBody_3	SRA3	6	M/A	FB	
Ppyr_FatBody_Mated	SRA3	4	M/A	FB	
Ppyr_FThorax	SRA3	3	F/A	T	
Ppyr_MThorax_1	SRA3	6	M/A	T	
Ppyr_MThorax_2	SRA3	6	M/A	T	
Ppyr_MThorax_3	SRA3	6	M/A	T	
Ppyr_OAG_1A	SRA3	6	M/A	AG	
Ppyr_OAG_1B	SRA3	6	M/A	AG	
Ppyr_OAG_2	SRA3	6	M/A	AG	
Ppyr_OAG_Mated	SRA3	4	M/A	AG	
Ppyr_SDGBursa	SRA3	3	F/A	SD	
Ppyr_SG_Mated	SRA3	4	M/A	SG	
Ppyr_Spermatheca	SRA3	3	F/A	SC	
Ppyr_SpiralGland_1	SRA3	6	M/A	SG	
Ppyr_SpiralGland_2	SRA3	6	M/A	SG	
Ppyr_SpiralGland_3	SRA3	6	M/A	SG	
Ppyr_Lantern_1A	**	6	M/A	P	
Ppyr_Lantern_2	**	6	M/A	P	
Ppyr_Lantern_3	**	6	M/A	P	
Ppyr_Eggs	**	7	E13	E	Strand-specific
Ppyr_Larvae	**	4	L1	WB	Strand-specific
Ppyr_wholeFemale*	**	1	F/A	WB	Strand-specific
Ppyr_wholeMale	**	1	M/A	WB	Strand-specific

[a] SRA1= NCBI BioProject PRJNA289908; SRA2= NCBI BioProject PRJNA321737; SRA3= NCBI BioProject PRJNA328865

* Parent of eggs and larvae with data from this study

** This study

1.9.2 *De novo* transcriptome assembly and genome alignment

MMNJ strand-specific reads (WholeMale, WholeFemale, eggs, larvae) and strand-specific reads from SRA (SRR3521424) were pooled (158.6M paired-reads) as input for *de novo* transcriptome assembly. Transcripts were assembled using Trinity (v2.4.0)[54] with default parameters except the following: (--SS_lib_type RF --trimmomatic --min_glue 2 --min_kmer_cov 2 --jaccard_clip --no_normalize_reads). Then gene structures were predicted from alignment of the de novo transcripts to the Ppyr1.3 genome using the PASA pipeline (v2.1.0)[55] 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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 blat (v. 36x2)[56] and gmap (v2017-09-11)[57] 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. Subsequently, potentially spurious transcripts were filtered out using a custom script[58] 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)[59] with default parameters. Direct gene models 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.

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 (Supplementary Table 1.9.1.1) using HISAT2 (v2.0.5)[60] and StringTie (v1.3.3b)[61]. For each library, reads were first mapped to the Ppyr1.3 draft genome 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. 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 gene models 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.

1.10 Coding gene annotation

We annotated the coding gene structure of *P. pyralis* by integrating direct gene models produced from the *de novo* transcriptome (Supplementary Note 1.9.2) and reference guided transcriptome (Supplementary Node 1.9.3), with a lower weighted contribution of *ab initio* gene predictions, using the Evidence Modeler (EVM) algorithm (v1.1.1)[55]. First, Augustus (v3.2.2)[62] 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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(v3.0.0β)[63] in "est2genome" mode. Preliminary gene models were used to train SNAP (v2006-07-28)[64] following the MAKER instructions[65]. 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.

We then integrated the *ab initio* predictions with our *de novo* and reference guided direct 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, 6 sources of evidence were used for EVM: de novo transcriptome direct gene models (Ppyr1.3_Trinity_stranded; 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 gene models (Ppyr1.3_Stringtie_stranded; weight = 3), Augustus and SNAP *ab initio* gene models (via MAKER; weight = 2). A custom script[66] was necessary to convert MAKER GFF format to an EVM compatible GFF format. Lastly, gene models for luciferase homologs 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 gene models.

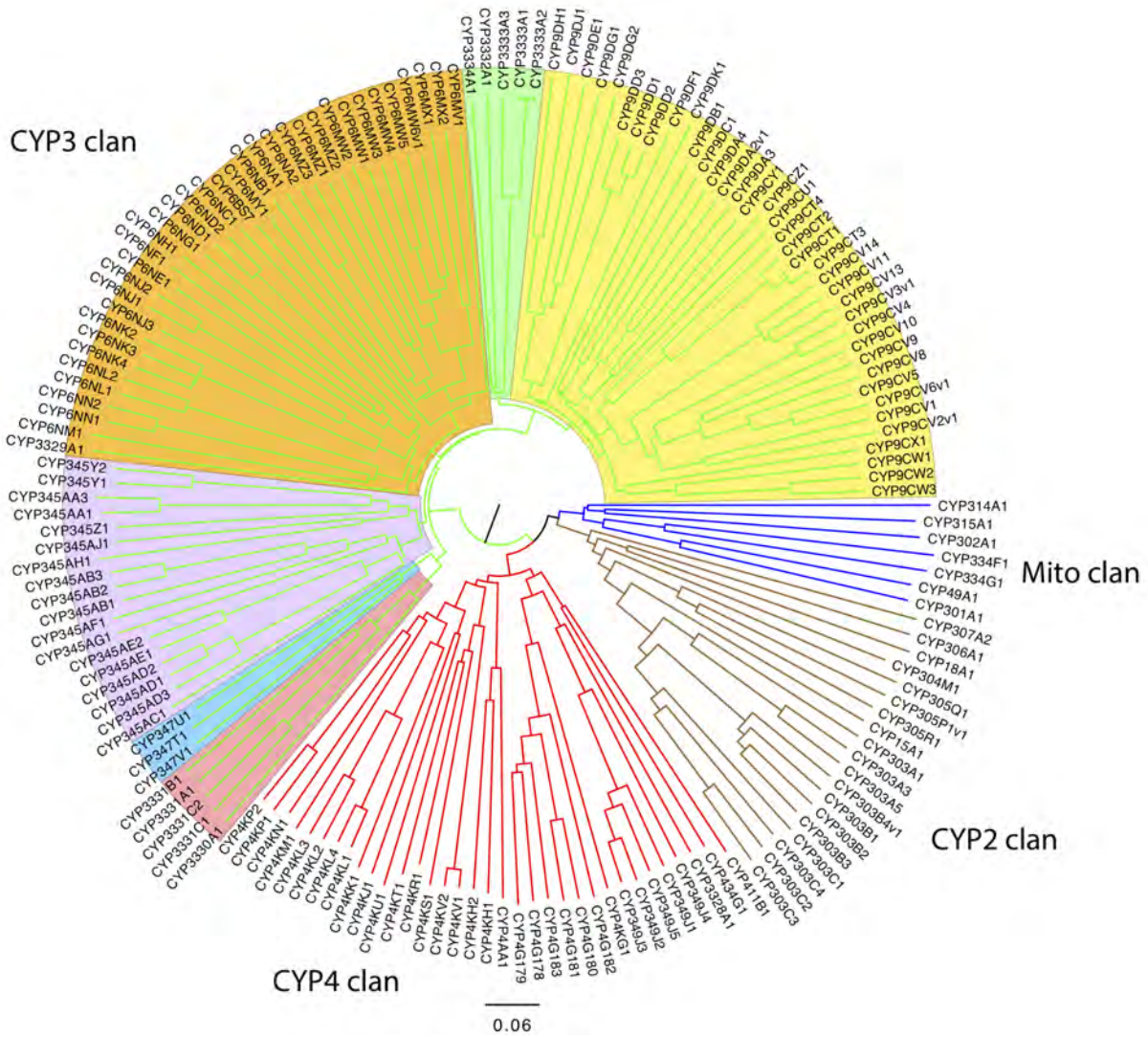
1.10.1 P450 annotation

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 5,837 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 DNXXXX IDs. All peptides with these IDs were retrieved and clustered with CD-Hit (v4.5.4)[67] to 99% percent 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 PpyrOGS1.0. Some hybrid sequences were corrected. The final set contains 170 named cytochrome P450 sequences (166 genes, 4 pseudogenes).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The cytochrome P450s in insects belong to four established clans CYP2, CYP3, CYP4 and Mito (Supplementary Figure 1.10.1.1). *P. pyralis* has about twice as many P450s as *Drosophila melanogaster* (86 genes, 4 pseudogenes) and slightly more than the red flour beetle *Tribolium castaneum* (137 genes, 10 pseudogenes). 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[68] in the CYP2 and Mito clans. The CYP4G subfamily makes a hydrocarbon waterproof coating for the exoskeleton[69]. Additional conserved P450s are CYP15A1 (juvenile hormone[69]) and CYP18A1 (20-hydroxyecdysone degradation[70]) in the CYP2 clan. Most of the other P450s are limited to a narrower phylogenetic range. Many are unique to a single genus, though this may change as more sampling is done. It is common for P450s to expand into gene blooms[71].

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



1.10.1.1 Supplementary Figure: *P. pyralis* P450 gene phylogenetic tree

Neighbour-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[72] 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. 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[54] with default parameters. Resulting transcriptomes were assessed for similarity to known viral sequences by

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

TBLASTN searches (max e-value = 1 x 10^{-5}) using as 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)[40]. 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[73]. Further, conserved UTR sequences were identified using ClustalW2[74] (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[75] 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)[76] and integrated with SMART[77], Pfam[78], and PROSITE[79] results to characterize the functional domains. Secondary structure was predicted with Garnier as implemented in EMBOSS (v6.6)[80], signal and membrane cues were assessed with SignalP (v4.1)[81], and transmembrane topology and signal peptides were predicted by Phobius[82]. 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 x 10^{-5}) searches of the complete 2,754 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) [83] and unrooted maximum-likelihood phylogenetic trees were constructed using FastTree [84] 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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 1,000 tree resamples.

To facilitate taxonomic identification we complemented BLASTP data with 2 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[85] (e-value = 1e10-2). 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[86], and contrasted with related literature. Heterotrimeric viral polymerase 3D structure prediction was generated with the SWISS-MODEL automated protein structure homology-modelling server[87] with the best fit template 4WSB: the crystal structure of Influenza A virus 4WSB. Predicted structures were visualized in UCSF Chimera[88] 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[40] 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.

All curation, phylogeny construction, and visualization were conducted in Geneious 8.1.9 (Biomatters, Ltd.). Animal silhouettes in Supplementary Figure 5.4.1 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[89].

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).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1.11 Repeat annotation

See Main Text: Methods.

1.11.1 Supplementary Table: RepeatMasker annotation of *P. pyralis* repetitive elements

Repetitive element class	Subclass	# of elements	Length occupied (bp)	Percentage of sequence
SINEs:	(All)	1231	129753	0.03%
	ALUs	0	0	0.00%
	MIRs	0	0	0.00%
LINEs:	(All)	42257	16154151	3.41%
	LINE1	842	366809	0.08%
	LINE2	8221	3624800	0.77%
	L3/CR1	3912	1412208	0.30%
LTR elements:	(All)	24416	11143713	2.35%
	ERVL	60	13699	0.00%
	ERVL-MaLRs	0	0	0.00%
	ERV_classI	606	130557	0.03%
	ERV_classII	1926	406670	0.09%
DNA elements:	(All)	73491	22983364	4.85%
	hAT-Charlie	933	205696	0.04%

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

	TcMar-Tigger	91	14141	0.00%
Unclassified:	(All)	690206	148520332	31.35%
Total interspersed repeats:	(All)	N.R.	198931313	41.99%
Small RNA:	(All)	604	183519	0.04%
Satellites:	(All)	1498	571536	0.12%
Simple repeats:	(All)	48605	2467925	0.52%
Low complexity:	(All)	7762	405465	0.09%

1.12 Methylation analysis

Methylation analysis was performed using methylpy[90] (see main text Methods for details). Methylpy calls programs for read processing and aligning: (i) reads were trimmed of sequencing adapters using Cutadapt[91], (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 [92]), and (iii) PCR duplicates were removed using Picard[93].

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[94], with some modification. Briefly, we dissected larvae in 1X PBS and treated tissues with a hypotonic solution (0.5% Sodium citrate) for 7 minutes. We transferred treated larval tissues to 45% acetic acid for 30 seconds, fixed in 2.5% paraformaldehyde in 45% acetic acid for 10 minutes, 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 minutes per wash. Slides were mounted in VectaShield with DAPI (Vector Laboratories), visualized on a Leica DM5500 upright

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

fluorescence microscope at 100X, imaged with a Hamamatsu Orca R2 CCD camera. Images were captured and analyzed using Leica's LAX software.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SUPPLEMENTARY NOTE 2: *Aquatica lateralis* additional information

2.1 Taxonomy, biology, and life history

Aquatica lateralis (Motschulsky, 1860) (Japanese name, Heike-botaru) 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[95].

The life cycle of *A. lateralis* is usually one 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[96].

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 7 in females[96–98]. The adult is dusk active. Male adults flash yellow-green for about 1.0 second in duration every 0.5-1.0 seconds while flying ~1 m above the ground. Female adults, located on low grass, respond to the male signal with flashes of 1-2 seconds in duration every 3-6 sec. Males immediately approach females and copulate on the grass[96,99]. 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[100]. *A. lateralis* larvae have eversible glands on each of the 8 abdominal segments[95]. The contents of the eversible glands is perhaps similar to that reported for *A. leii*[101].

2.2 Species distribution

The geographical range of *A. lateralis* includes Siberia, Northeast China, Kuril Isls, Korea, and Japan (Hokkaido, Honshu, Shikoku, Kyushu, Tsushima Isls.)[102]. 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[103]. For example, in 2017, Japanese Ministry of Environment began efforts to protect the population of *A. lateralis* in the Imperial Palace, Tokyo, where 3,000 larvae cultured in an aquarium were released in the pond beside the Palace[104].

2.3 Specimen collection

Individuals used for genome and RNA sequencing 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[105] by Mr. Haruyoshi Ikeya, a highschool teacher in Yokohama, Japan. Mr. Ikeya collected adult *A. lateralis* specimens from

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

their natural habitat in Yokohama and 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 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 (*Physa 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)[106]. The Y chromosome is much smaller than X chromosome, and the typical behaviors of XY chromosomes, such as partial conjugation of X/Y at first meiotic metaphase and separation delay of X/Y at first meiotic anaphase, were observed in testis cells[106].

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).

2.5 Genomic sequencing and assembly

See Main Text: Methods for details. The final assembly (Alat1.3) consists of 5,388 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. Genome sequencing library statistics are available in Extended Data Table 1.

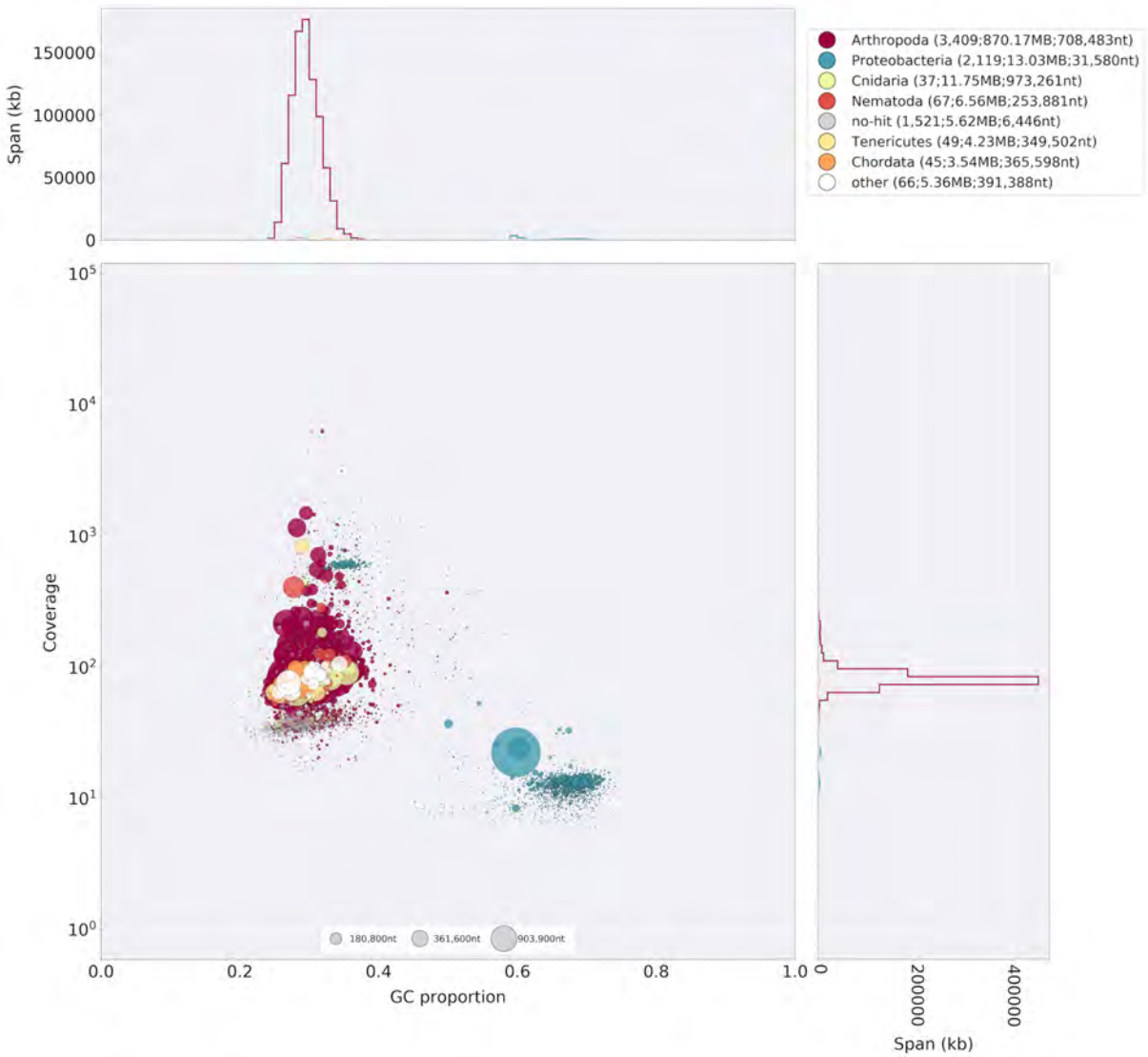
2.5.1 Taxonomic annotation filtering

Potential contaminants in Alat1.2 were identified using the blobtools toolset (v1.0)[45]. First, scaffolds were compared to known sequences by performing a blastn (v2.5.0+) nucleotide

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sequence similarity search against the NCBI nt database and a diamond (v0.9.10)[46] 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 (<30x) (see Supplementary Figure 2.5.1.1). 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.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



2.5.1.1 Supplementary Figure: 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; Extended Data Table 1), aligned against Alat1.2 using Bowtie2. Scaffolds were taxonomically annotated as described in Supplementary Note 2.5.1.

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, 5th (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 (Supplementary Table 2.6.1). Live specimens were anesthetized on ice and dissected during the day. The lantern tissue was

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 6 cycles. A subset of nine libraries (BdM1, HeF1, HeM1, LtF1, LtM1, Egg1, Lrv1, PpEF, PpLM; Supplementary Table 2.6.1) were multiplexed and sequenced in a single lane of Hiseq1500 101x101 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 Hiseq1500 66 bp single-end reads. Sequence quality was inspected by FastQC[107].

2.6.1 Supplementary Table: *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 = 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	Biosample ID	N	Sex/ Stage	Tissue	Library type
R102L6_idx13	BdM1	SAMD00106518	1	M/A	B	Illumina paired-end, non-stranded specific, PolyA
R128L1_idx25	BdM2	SAMD00106519	1	M/A	B	Illumina single-end, non-stranded specific, PolyA
R128L2_idx27	BdM3	SAMD00106520	1	M/A	B	Illumina single-end, non-stranded specific, PolyA
R102L6_idx15	HeF1	SAMD00106521	3	F/A	H	Illumina paired-end, non-stranded specific, PolyA
R128L1_idx22	HeF2	SAMD00106522	3	F/A	H	Illumina single-end, non-stranded specific, PolyA
R128L2_idx23	HeF3	SAMD00106523	3	F/A	H	Illumina single-end, non-stranded specific, PolyA
R102L6_idx12	HeM1	SAMD00106524	2	M/A	H	Illumina paired-end, non-stranded specific, PolyA
R128L1_idx20	HeM2	SAMD00106525	2	M/A	H	Illumina single-end, non-stranded specific, PolyA
R128L2_idx21	HeM3	SAMD00106526	2	M/A	H	Illumina single-end, non-stranded specific, PolyA
R102L6_idx16	LtF1	SAMD00106527	5	F/A	La	Illumina paired-end, non-stranded specific, PolyA
R128L1_idx06	LtF2	SAMD00106528	5	F/A	La	Illumina single-end, non-stranded specific, PolyA
R128L2_idx12	LtF3	SAMD00106529	5	F/A	La	Illumina single-end, non-stranded specific, PolyA
R102L6_idx14	LtM1	SAMD00106530	5	M/A	La	Illumina paired-end, non-stranded specific, PolyA
R128L1_idx05	LtM2	SAMD00106531	5	M/A	La	Illumina single-end, non-stranded specific, PolyA
R128L2_idx19	LtM3	SAMD00106532	5	M/A	La	Illumina single-end, non-stranded specific, PolyA
R128L2_idx15	WAF1	SAMD00106533"	1	F/A	W	Illumina single-end, non-stranded specific, PolyA
R128L1_idx16	WAF2	SAMD00106534	1	F/A	W	Illumina single-end, non-stranded specific, PolyA
R128L2_idx18	WAF3	SAMD00106535	1	F/A	W	Illumina single-end, non-stranded specific, PolyA
R128L1_idx11	WAM1	SAMD00106536	1	M/A	W	Illumina single-end, non-stranded specific, PolyA
R128L2_idx13	WAM2	SAMD00106537	1	M/A	W	Illumina single-end, non-stranded specific, PolyA
R128L1_idx14	WAM3	SAMD00106538	1	M/A	W	Illumina single-end, non-stranded specific, PolyA
R102L6_idx4	Egg1	SAMD00106539	19.6 mg (~30-50)	E ~6h AEL	W	Illumina paired-end, non-stranded specific, PolyA

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

R128L1_idx01	Egg2	SAMD00106540	21.6 mg (~30-50)	E ~7d AEL	W	Illumina single-end, non-stranded specific, PolyA
R102L6_idx5	Lrv1	SAMD00106541	1	L	W	Illumina paired-end, non-stranded specific, PolyA
R128L1_idx03	Lrv2	SAMD00106542	1	L	W	Illumina single-end, non-stranded specific, PolyA
R128L2_idx04	Lrv3	SAMD00106543	1	L	W	Illumina single-end, non-stranded specific, PolyA
R128L1_idx07	PpEM	SAMD00106544	1	M/P-E	W	Illumina single-end, non-stranded specific, PolyA
R128L2_idx10	PpLF	SAMD00106545	1	F/P-L	W	Illumina single-end, non-stranded specific, PolyA
R128L1_idx09	PpMF	SAMD00106546	1	F/P-M	W	Illumina single-end, non-stranded specific, PolyA
R128L2_idx08	PpMM	SAMD00106547	1	M/P-M	W	Illumina single-end, non-stranded specific, PolyA
R102L6_idx7	PpEF	SAMD00106548	1	F/P-E	W	Illumina paired-end, non-stranded specific, PolyA
R102L6_idx6	PpLM	SAMD00106549	1	M/P-L	W	Illumina paired-end, non-stranded specific, PolyA

2.7 Transcriptome assembly

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; Supplementary Table 2.6.1) were pooled. These represent RNA prepared from various tissues (head, thorax+abdomen, lantern) and stages (egg, pupae, adult) of both sexes. The reads were cleaned with Cutadapt[91]: Low-quality ends (<QV30) and adapter sequences were trimmed, while reads shorter than 50 bp were discarded. The cleaned reads were assembled *de novo* with Trinity (vr20140413p1)[54] in the paired-end mode (parameters: --min_kmer_cov 2, --normalize_max_read_cov 160).

2.7.2 Reference guided transcriptome alignment and assembly

The same set of the RNA-Seq reads used in the de novo transcriptome assembly (see above) were mapped to the genome with TopHat[108] using default parameters except adding an option "--min-intron-length 30" to allow short introns, because our preliminary analysis showed that the *A. lateralis* genome contains a substantial number of genes with short introns, including luciferase. In addition, TopHat with default parameters produced incorrect alignments of the reads spanning short introns. The mapped reads were assembled into transcripts with Cufflinks[109] (parameters: --min-intron-length 30).

2.8 Coding gene annotation

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 (Supplementary Note 2.7.1) and de-novo transcriptome (Supplementary Note 2.7.2) approaches. In the reference-guided approach, RNA-Seq reads were mapped to the genome assembly with TopHat and assembled into transcripts with Cufflinks. The Cufflinks transcripts were subjected to the TransDecoder program to extract ORFs. In the *de novo* transcriptome approach, RNA-seq reads were assembled de novo by Trinity and ORFs were predicted using TransDecoder. We used CD-HIT-EST[67] to reduce the redundancy of the

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

2.9 Repeat annotation

See Main Text: Methods.

2.9.1 Supplementary table: RepeatMasker annotation of *A. lateralis* repetitive elements

Repetitive element class	Subclass	# of elements	Length occupied (bp)	Percentage of sequence
SINEs:	(All)	0	0	0.00%
	ALUs	0	0	0.00%
	MIRs	0	0	0.00%
LINEs:	(All)	133598	49227182	5.42
	LINE1	0	0	0.00%
	LINE2	4696	2735009	0.30%
	L3/CR1	648	430251	0.05
LTR elements:	(All)	48201	23019597	2.53%
	ERVL	0	0	0.00%

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

	ERVL-MaLRs	0	0	0.00%
	ERV_classI	568	132882	0.01%
	ERV_classII	0	0	0.00%
DNA elements:	(All)	207311	71380866	7.86%
	hAT-Charlie	8416	3283510	0.36%
	TcMar-Tigger	1361	598443	0.07%
Unclassified:	(All)	414260	95508648	10.51%
Total interspersed repeats:	(All)	N.R.	239136293	26.32%
Small RNA:	(All)	0	0	0.00%
Satellites:	(All)	0	0	0.00%
Simple repeats:	(All)	154302	6790679	0.75%
Low complexity:	(All)	54324	2690707	0.30%

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SUPPLEMENTARY NOTE 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, that includes about 10,000 species[110] (17 subfamilies[111]) 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 from tropical and subtropical regions of Americas and some small Melanesian islands[111,112]. For instance, the tropical American *Pyrophorus noctilucus* is considered the largest (~30 mm) and brightest insect[113,114]. All luminous species are closely related - luminous click beetles belong to the tribes Pyrophorini and Euplinthini[112,115] of the subfamily Agrypninae, with the single exception of *Campyloxenus pyrothorax* (Chili) in the related subfamily Campyloxeninae[116]. The luminescence of a pair of pronotal 'light organs' of the adult *Balgus schnusei*[117], a species that has now been assigned to the Thylacosterninae of the Elateridae[111], has not been confirmed by later observation. This near-monophyly of bioluminescent elaterid taxa is supported by both morphological[118] and molecular phylogenetic analysis[119,120,121], though early morphological phylogenies were inconsistent[116,122–125]. 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[112]. Often this species is called *Pyrophorus luminosus* as an 'auctorum', a name used to describe a variety of taxa[126]. This use of "Pyrophorus" an auctorum may be due to the heightened difficulty of classifying Elateridae[112]. The genus *Ignelater* is characterized by the presence of both dorsal and ventral photophores[112,127]. An unreviewed report suggested that the adult *I. luminosus* has a ventral light organ only in males[128]. 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[127]. The earliest fossil of an Elateridae species was recorded from the Middle Jurassic of Inner Mongolia, China[129]. McKenna and Farrell suggested that, based on molecular analyses, the family Elateridae originated in the Early Cretaceous (130 Mya)[130]. It is expected that many recent genera in Elateroidea were established by the Early Tertiary[131].

The exact function of bioluminescence across different life stages remains unknown for many luminous elaterid species. 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[132]. However, it is unknown if *I. luminosus* larvae produce light. Adult *I. luminosus* are luminescent and bioluminescent courtship behavior was described in an unreviewed study[133]. 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 second 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. It is unclear if the male

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ventral lantern response represents a direct control of light production from the ventral lantern, or simply the beetle exposing a constitutively luminescent ventral lantern which is normally obscured from view.

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 flight[134].

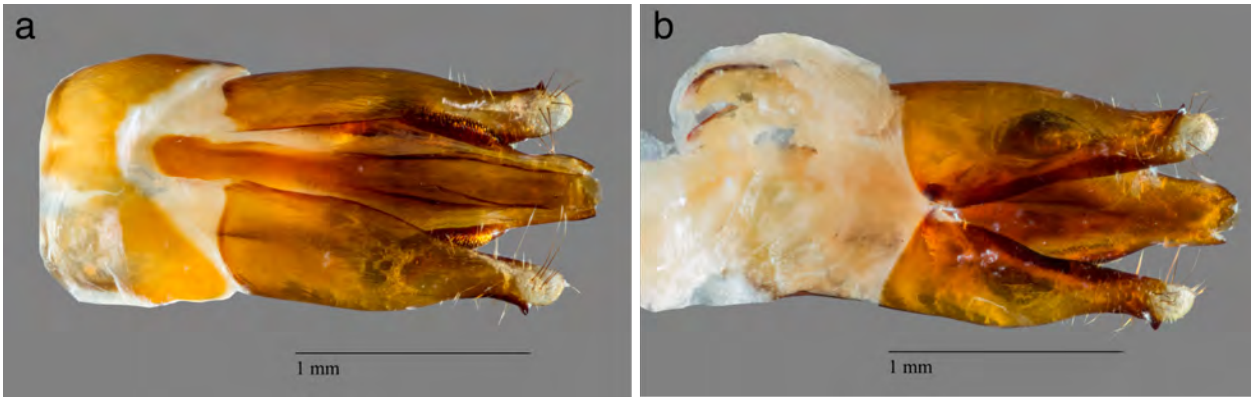
3.2 Species distribution

I. luminosus is often considered to be endemic to Puerto Rico[135], however the genus *Igenater* 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[112]. Similarly, *I. luminosus* itself has been reported on the island of Hispaniola[133,136], indicating *I. luminosus* is not restricted to Puerto Rico. This geographic distribution of *Ignelater* suggests that Puerto Rico likely contains multiple *Ignelater* species and, given the difficulty of distinguishing different species of bioluminescent elaterids by morphological characters, a definitive species distribution for *I. luminosus* cannot be stated, other than this species is seemingly not 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 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[112,127,137] (Supplementary Figure 3.3.1). All inspected specimens were male (3/3). Separate specimens 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.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



3.3.1 Supplementary Figure: *I. 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[112,127,137].

3.4 Karyotype and genome size

The karyotype of Puerto Rican *I. luminosus* (as *Pyrophorus luminosus*) was reported as 2n=14A + X_1X_2Y[135]. 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).

3.5 Genomic sequencing and assembly

HMW DNA was extracted from a single male specimen of *I. luminosus* which was first washed with 95% ethanol, using a 100/G Genomic Tip with the Genomic buffers kit (Qiagen, USA). 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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

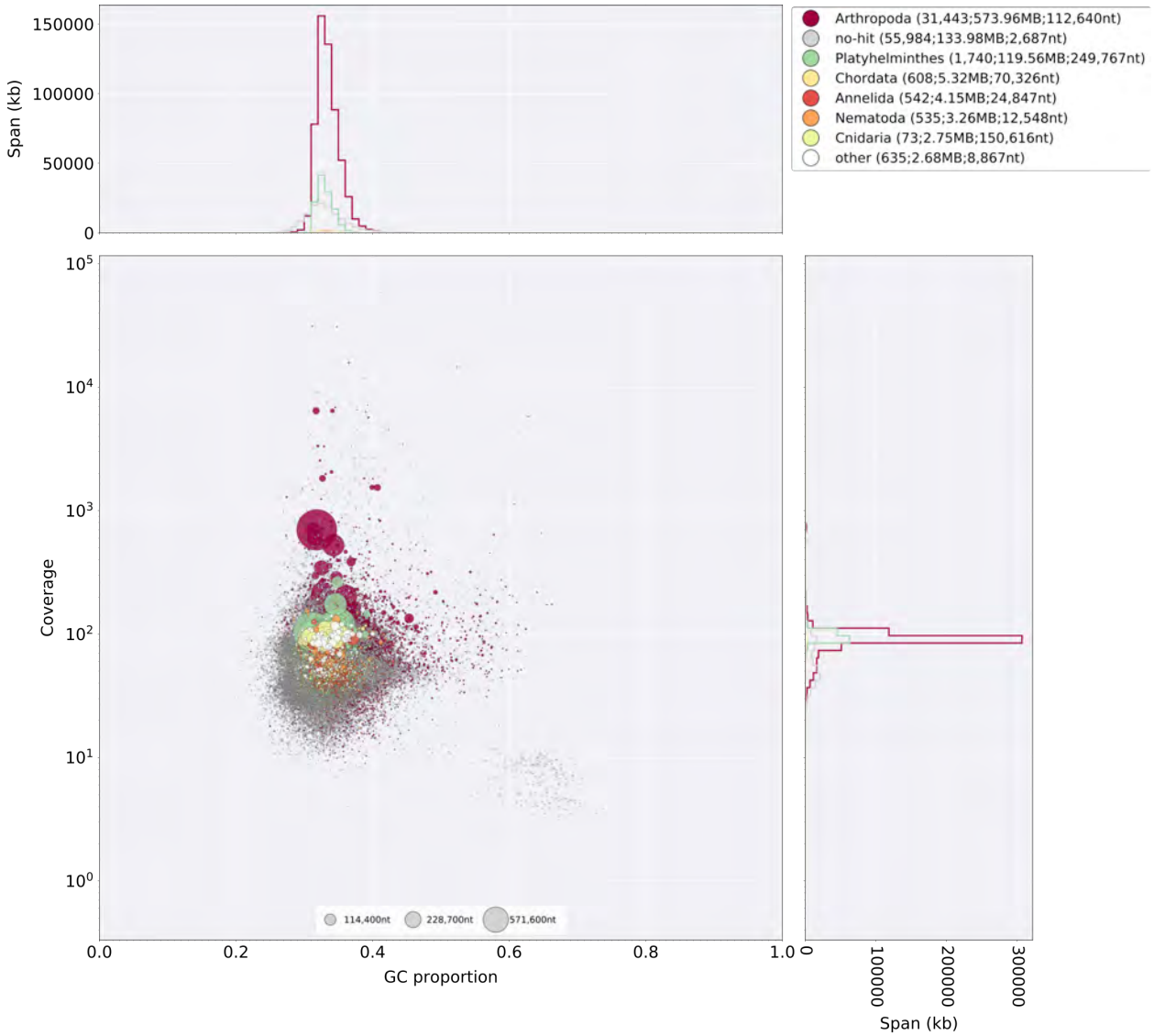
10x Genomics Chromium Genome v1 library construction was performed. The library was then sequenced on one HiSeqX lane. 408,838,927 paired reads (150x150 PE) were produced, corresponding to a genomic coverage of 153x. The library was also sequenced on one HiSeq2500 lane, where 145,250,480 reads (150x150 PE) were produced, corresponding to a genomic coverage of 54x. A summary of the library statistics for the genomic sequencing is available in Extended Data 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)[138], on a 40 core 1 TB RAM server at the Whitehead Institute for Biomedical Research. The assembly was exported to FASTA format using Supernova mkoutput (parameters: --style=pseudohap). A Supernova (v1.2.2) assembly was also produced from the HiSeq2500 reads, but on a brief inspection the quality was equivalent to Ilumi1.0, so the new assembly was not used for further analyses.

3.5.3 Taxonomic annotation filtering

We sought to systematically remove assembled non-elaterid contaminant sequence from Ilumi1.0. Using the blobtools toolset (v1.0.1),[45] 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)[46] 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" (Supplementary Figure 3.5.3.1). A tab delimited text file containing the results of this blobtools annotation are 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", and polishing indels and gap-filling with Pilon (Supplementary Note 3.5.4). 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[139,140]. For a beetle like *I. luminosus*, which has a extended predatory larval stage,

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the accidental ingestion and harboring of a *Hymenolepis sp.* Is plausible, potentially enabling HGT between worm and host over evolutionary timescales.



3.5.3.1 Supplementary figure: Blobtools plot of Ilumi1.0

Coverage shown represents mean coverage of reads from the HiSeqX Chromium library sequencing (Sample name 1610_IlumiHiSeqX; Supplementary Table 3.5.1, 3.5.2), aligned against Ilumi1.0 using Bowtie2 with parameters (--local). Scaffolds were taxonomically annotated as described in Supplementary Note 3.5.3.

3.5.4 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 runs. Mapping of the 1610_IlumiHiSeqX and 1706_IlumiHiSeq2500 reads (Supplementary Table 3.5.1, 3.5.2) with Bowtie2 using parameters (--local), revealed that this indel was not supported by the majority of

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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[138], we therefore sought to correct these errors globally through the use of Pilon (v1.2.2)[50]. In order to run Pilon efficiently, we split the taxonomically filtered Ilumi1.0 reference (Ilumi1.0b; Supplementary Note 3.5.3) using Kirill Kryukov's fasta_splitter.pl script (v0.2.6)[141], 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 final 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 (C:93.3% to C:94.8%), suggesting that indel polishing and gap filling had a net positive effect.

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 100x100 bp PE. This library was multiplexed with the *P. pyralis* RNA-Seq libraries of Al-Wathiqui and colleagues[6], 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, thorax, and abdomen of 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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(RCA) was performed to produce DNA Nanoballs (DNBs). The qualified DNBs were loaded into the patterned nanoarrays and the libraries were sequenced as 50x50 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.

3.6.3 Supplementary Table: *I. luminous* RNA-Seq libraries

Library name	Source	N	Sex	Tissue	Notes
Pyrophorus_luminosus_head	This report	1	M*	Prothorax and head (lantern containing)	Illumina RNA-Seq
Prothorax_A3	This report	1	M*	Prothorax and head (lantern containing)	BGISEQ-500 RNA-Seq
Thorax_A3	This report	1	M*	Thorax	BGISEQ-500 RNA-Seq
Abdomen_A3	This report	1	M*	Abdomen (lantern containing)	BGISEQ-500 RNA-Seq
Prothorax_A4	This report	1	M*	Prothorax and head (lantern containing)	BGISEQ-500 RNA-Seq
Thorax_A4	This report	1	M*	Thorax	BGISEQ-500 RNA-Seq
Abdomen_A4	This report	1	M*	Abdomen (lantern containing)	BGISEQ-500 RNA-Seq

* Gender inferred. See Supplementary Note 3.3 for a discussion on this inference.

3.7 Transcriptome assembly

Both *de novo* and reference guided transcriptome assembly approaches using Trinity and Stringtie respectively were used.

3.7.1 *De novo* transcriptome assembly and alignment

For the *de novo* transcriptome approach, all available *I. luminosus* RNA-Seq reads (head+prothorax,thorax, 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)[54] using default parameters exception the following: (--min_glue 2 --min_kmer_cov 2 --jaccard_clip --no_normalize_reads). 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 using the PASA pipeline with blat (v36x2) and gmap (v2017-09-11) (--

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 gene models 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.

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,thorax, abdomen - both Illumina and BGISEQ-500) using HISAT2 (v2.0.5)[60] and StringTie (v1.3.3b)[61]. 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 gene models 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.

3.8 Coding gene annotation

We annotated the coding gene structure of *I. luminosus* by integrating direct gene models produced from the *de novo* transcriptome (Supplementary Note 3.7.1) and reference guided transcriptome (Supplementary Node 3.7.2), with a lower weighted contribution of *ab initio* gene predictions, using the Evidence Modeler (EVM) algorithm (v1.1.1)[55]. First, Augustus (v3.2.2)[62] 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), *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 gene models, using EVM. In the final version, eight sources of evidence were used for EVM: *de novo* transcriptome direct gene models (Ilumi1.1_Trinity_unstranded; weight=8), reference guided transcriptome direct gene models (Ilumi1.1_Stringtie_unstranded; weight = 4),

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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[66] was used to convert MAKER GFF format to an EVM compatible GFF format.

3.9 Repeat annotation

See Main Text: Methods.

3.9.1 Supplementary Table: RepeatMasker annotation of *I. luminosus* repetitive elements

Repetitive element class	Subclass	# of elements	Length occupied (bp)	Percentage of sequence
SINEs:	(All)	0	0	0.00%
	ALUs	0	0	0.00%
	MIRs	0	0	0.00%
LINEs:	(All)	82257	39170650	4.65%
	LINE1	0	0	0.00%
	LINE2	6646	3639251	0.43%
	L3/CR1	3047	2088463	0.25%
LTR elements:	(All)	19489	11170336	1.33%
	ERVL	0	0	0.00%
	ERVL-MaLRs	0	0	0.00%
	ERV_classI	0	0	0.00%
	ERV_classII	0	0	0.00%

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DNA elements:	(All)	149429	70171758	8.33%
	hAT-Charlie	3178	1374510	0.16%
	TcMar-Tigger	0	0	0.00%
Unclassified:	(All)	696477	152541426	18.10%
Total interspersed repeats:	(All)	N.R.	273054170	32.39%
Small RNA:	(All)	0	0	0.00%
Satellites:	(All)	0	0	0.00%
Simple repeats:	(All)	112969	5342894	0.63%
Low complexity:	(All)	29128	1397433	0.17%

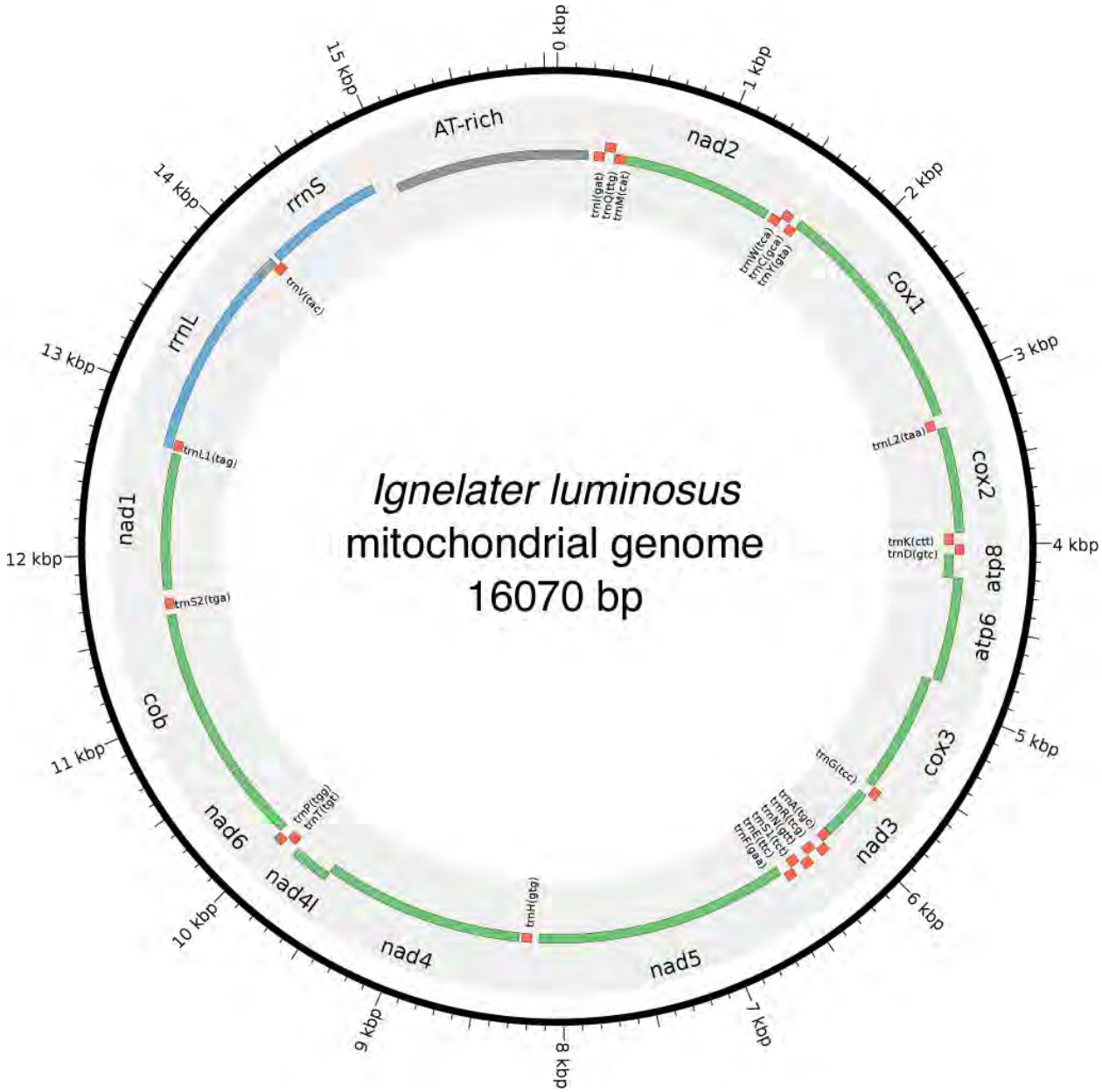
3.10 Mitochondrial genome assembly and annotation

The mitochondrial genome sequence of *I. luminosus* was assembled by a targeted sub-assembly approach. First, Chromium reads were mapped to the previously sequenced mitochondrial genome of the Brazilian elaterid beetle *Pyrophorus divergens* (NCBI ID: NC_009964.1)[142], using Bowtie2 (v2.3.1; parameters: --very-sensitive-local)[92]. Although these reads still contain the 16bp 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 & 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[143] without error correction and with the plasmidSPAdes module[144] enabled (parameters: -t 16 --plasmid -k55,127 --cov-cutoff 1000 --only-assembler).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The resulting "assembly_graph.fastg" file was viewed in Bandage[145], 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 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[146].

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



3.10.1 Supplementary Figure: Mitochondrial genome of *I. luminosus*

The mitochondrial genome of *I. luminosus* was assembled and annotated as described. in the supplementary methods. Figure produced with Circos[53].

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SUPPLEMENTARY NOTE 4: 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[147] using a dust-masked minikraken database to eliminate comparison with repetitive sequences. Overall contamination levels were low, in agreement with a low level of contamination in our final assembly (see Supplementary 1.6.3.2 above). On average, contamination was 3.5% in the PacBio reads (whole body) and 1.6% in the Illumina reads (only thorax) (Extended Data 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[148], was used to calculate genome quality statistics for comparison and optimization of assembly methods (Supplementary Table 4.1.1). BUSCO (v3.0.2)[149] 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 (Supplementary Table 4.1.2). 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 non-beetle insect genome predictions, "--species=fly" was used.

4.1.1 Supplementary Table: 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 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	Ppyr1.2 + Blobtools +	472/422	2160/ 7533	192.5	49,173	C:97.2% [S:88.8%,

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

	PacBio	manual filtering					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%,
Ilumi1.1	Linked read	Ilumi1.0 + Blobtools + Pilon indel & gap polishing	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 Supplementary Note 1.4, 2.4, 3.4.
*** Calculated with QUAST (v4.5)[148], parameters "-e --scaffolds --est-ref-size X --min-contig 0" and the measured genome size for "est-ref-size"

4.1.2 Supplementary Table: Comparison of BUSCO conserved gene content with other insect genome assemblies

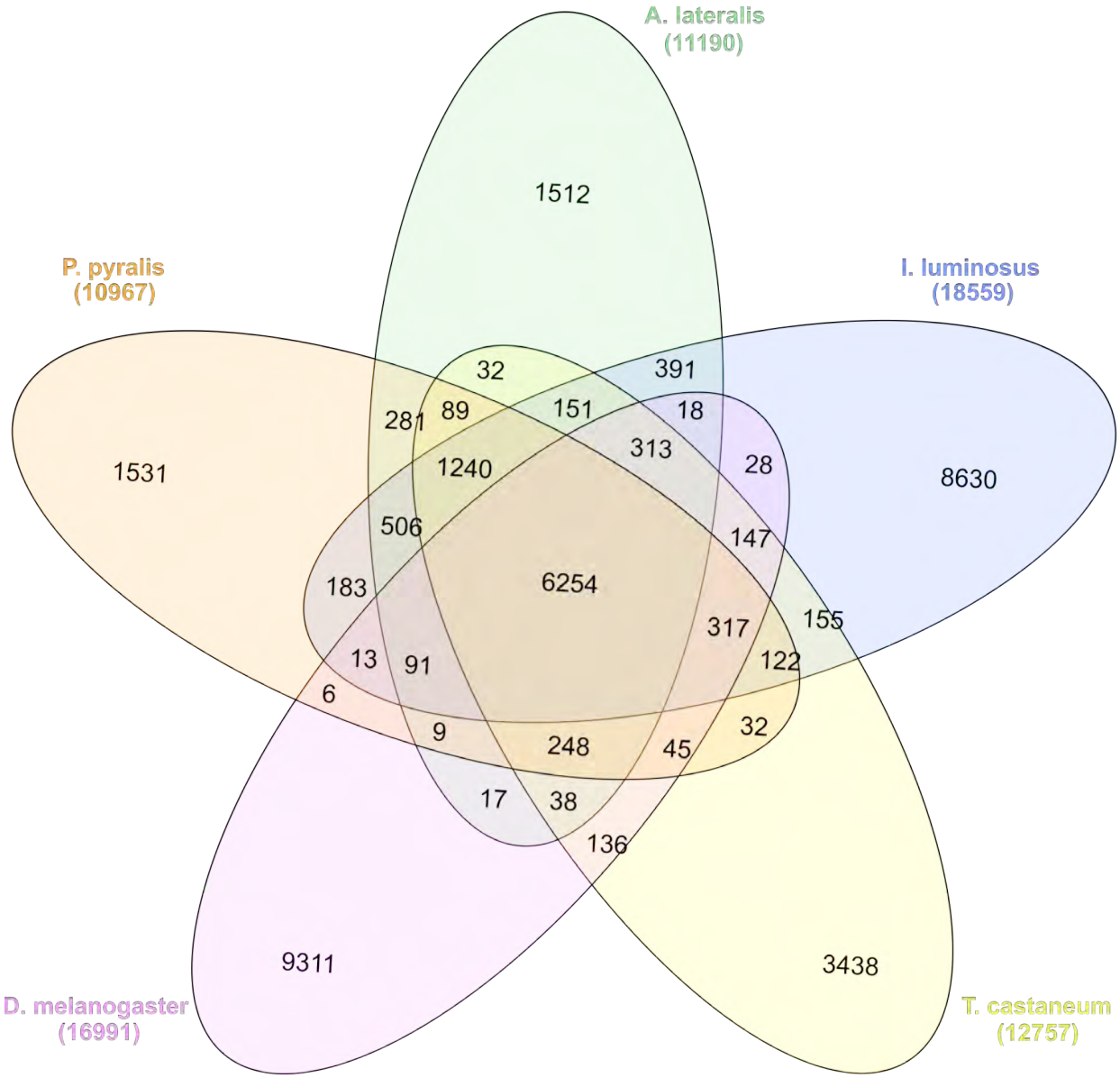
Species	Genome version (NCBI, or this report)	Note	Genome BUSCO (endopterygota_odb9)	Protein geneset BUSCO (endopterygota_odb9)
Drosophila melanogaster	GCA_000001215.4 Release 6	Model insect	C:99.4%[S:98.7%,D:0.7%], F:0.4%,M:0.2%,n:2442	C:99.8%[S:43.0%,D:56.8%], F:0.2%,M:0.0%,n:2442
Tribolium castaneum	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:99.0%[S:69.2%,D:29.8%], F:0.5%,M:0.5%,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%	C:90.0%[S:89.1%,D:0.9%], F:3.2%,M:6.8%,n:2442
Nicrophorus vespilloides[150]	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[151]	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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Igenlater luminosus	Ilumi1.1*	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

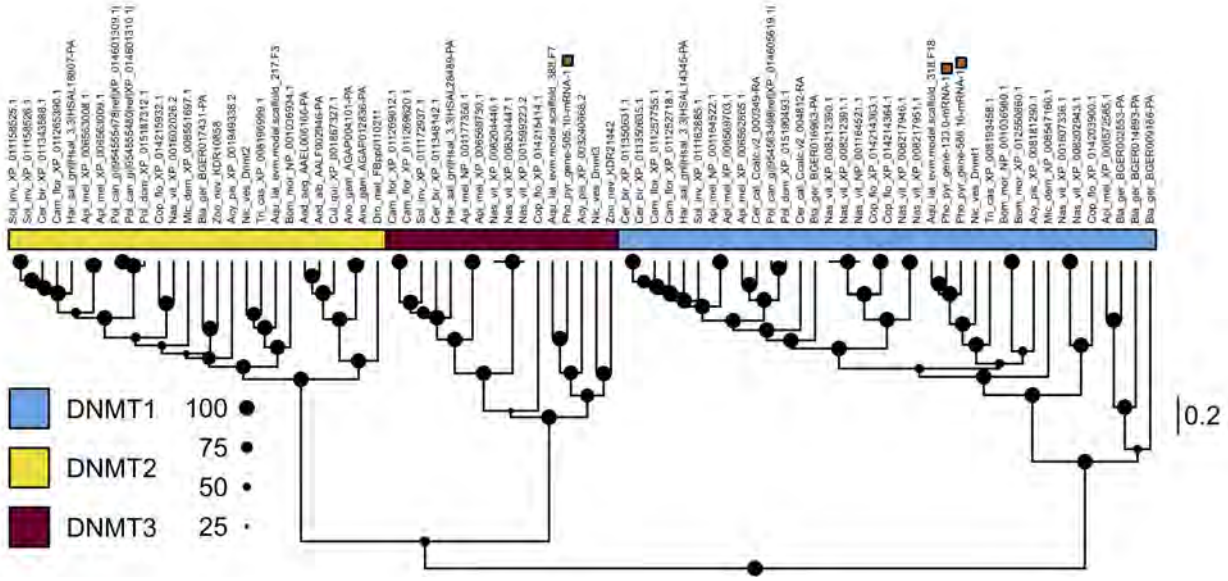
bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



4.1.3 Supplementary Figure: Venn Diagram of *P. pyralis*, *A. lateralis*, *I. luminosus*, *T. castaneum*, and *D. melanogaster* orthogroup relationships.

Orthogroups were calculated between the Ppyr_OGS1.0, Alat_OGS1.0, Ilumi_OGS1.0, genesets, and the T. casteneum and D. melanogaster NCBI reference genesets using OrthoFinder[152]. See maintext methods for full description of methods.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



4.2 Supplementary Figure: *De novo* 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, which is in contrast to a single copy of DNMT1 in the firefly *Aquatica lateralis*. Size of circles at nodes corresponds to posterior probability support. Branch lengths are in amino acid substitutions per site.

4.3 Luciferase evolution

The gene for firefly luciferase was first isolated from the North American firefly *P. pyralis*[153,3,154] and then identified from the Japanese fireflies *Luciola cruciata* [155] and *Aquatica lateralis*[156]. 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*[157] (Photurinae), *L. cruciata*[158] (Luciolinae), *A. lateralis*[159] (Luciolinae), *Luciola parvula*[160] (Luciolinae), and *Pyrocoelia atripennis*[161] (Lampyrinae).

Luciferase genes have also been isolated from members of the other luminous beetles families: Phengodidae, Rhagophthalmidae, and Elateridae[162,163,164,165] with amino acid identities to firefly luciferases at >48%[166]. 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[167 168], suggesting the monophyly of Elateroidea and a single origin of the luminescence in the ancestor of these three lineages[166]. However, ambiguity in

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the evolutionary relationships among luminous beetles, including luminous Elaterids, does not yet exclude multiple origins.

From our reference genesets, protein blast search detected 30 luciferase homologs (E-value < $1x10^{-60}$) to *P. pyralis* luciferase (PpyrLuc1; Genbank accession AAA29795) from *P. pyralis, A. lateralis,* and *I. luminosus.* We defined the luciferase co-orthology as followings; (1) shows an E-value lower than $1.0x10^{-60}$, (2) phylogenetically sister to *DmelPACS* (CG6178), which is the most similar gene to firefly luciferase in *D. melanogaster,* based on the maximum likelihood (ML) phylogenetic reconstruction (Supplementary Figure 4.2.1). Multiple alignment, gap filtering, and reconstruction of ML tree were performed as described in the Main Text: Methods. The genes co-orthologous to *DmelPACS* were used for further analysis (see the main text and Figure 3).

As an alternative to the duplication-neofunctionalization model, the high copy number of PACS, ACS, and luciferase and dual enzymatic functions of ancestral luciferase could indicate an innovation-amplification-divergence (IAD) model[169] of gene family evolution, where multiple functions in a gene appear prior to gene duplication and duplicates then diverge under selection for separate functions.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

4.2.1 Supplementary Figure: Maximum likelihood phylogeny of luciferase homologs

The maximum likelihood tree was reconstructed from a 541 amino acid multiple sequence alignment, generated via a BLASTP search using *P. pyralis* luciferase as query (e-value: 1.0 x 10^-60). Members of the clade that includes both known firefly luciferase and CG6178 of *D. melanogaster* are defined as luciferase co-orthologous genes (highlighted in gray). Branch length represents substitutions per site.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

4.2.2 Supplementary Figure: Conserved intron-exon structure of P. pyralis and A. lateralis luciferases

a, Intron-exon structure of P.pyralis & A. lateralis Luc1 & Luc2 is globally similar with 7 exons and a similar intron between homologous introns **b,** MAFFT L-INS-i multiple sequence

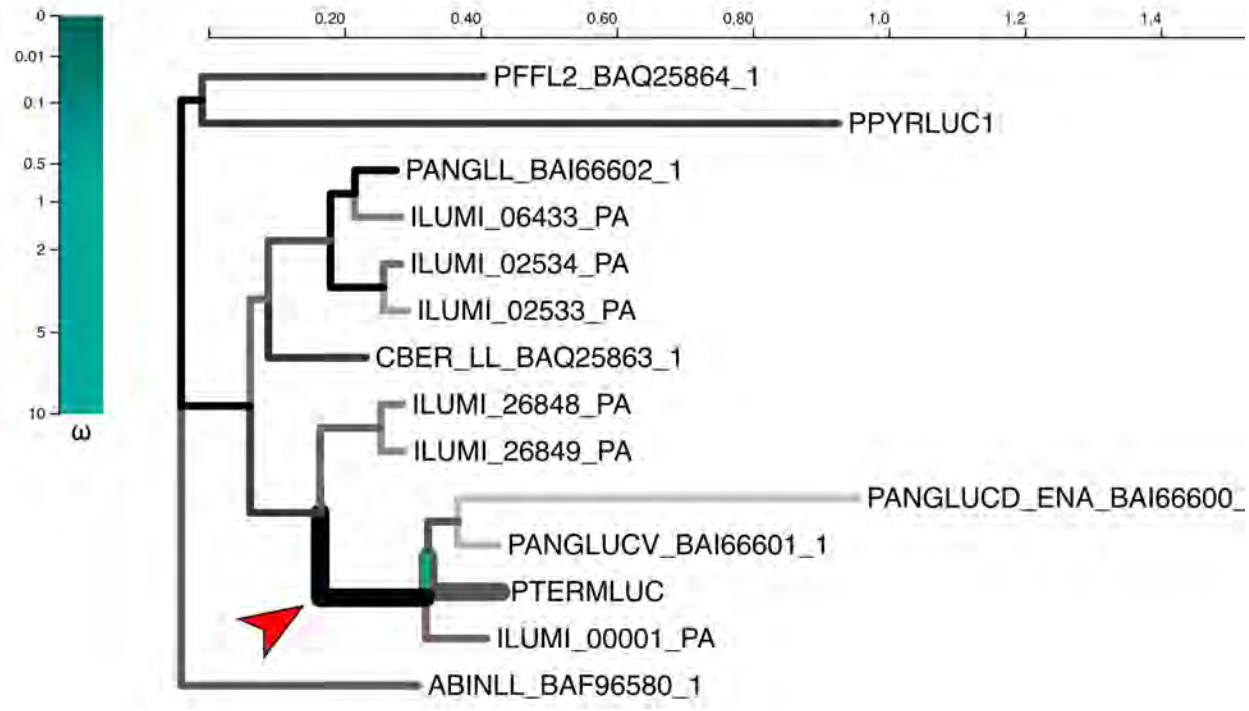
bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

alignment of luciferase gene nucleotide sequences demonstrates the location of intron-exon junctions (bolded blue text) is completely conserved amongst the 4 luciferases.

4.3.1 Testing for selection on Elaterid luciferase

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). This includes the dorsal (PangLucD; ENA ID=BAI66600.1) and ventral (PangLucV; ENA ID = BAI66601.1) luciferases, and a luciferase-like homolog without luciferase-activity (PangLL; ENA ID=BAI66602.1) from *Pyrophorus angustus*[170], and two unpublished but database deposited luciferase homologs without luciferase-activity (data not shown) from *Cryptalaus berus* (CberLL; ENA ID = BAQ25863.1) and *Pectocera fortunei fortunei* (PffL2; ENA ID=BAQ25864.1). The peptide and CDS sequence of the *Pyrearinus termitilluminans* were manually transcribed from the literature[164], as these sequences were seemingly never deposited in a publically accessible sequence database. The CDS sequence of complete luciferase of *I. luminosus* (IlumiLuc; ILUMI_00001-PA), two related paraogs (IlumiLL1; ILUMI_26849-PA, IlumiLL2; ILUMI_26848-PA), and 2 other paralogs (ILUMI_02534-PA; ILUMI_06433-PA) The nucleotide sequences were multiple-sequenced-aligned with MUSCLE[171] in “codon” mode within MEGA7[172], using parameters (Gap Open = -.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). This MSA is available on FigShare (DOI: 10.6084/m9.figshare.5691277). An adaptive branch-site REL test for episodic diversification was performed on the DataMonkey server[173] using the adaptive branch-site REL test for episodic diversification (aBSREL) method[174]. The previously mentioned nucleotide MSA, was input and a gene phylogeny was produced internally by DataMonkey without specifying parameters. The input was 13 sequences with 553 sites (codons). The aBSREL analysis found evidence of episodic diversifying selection on 1 out of 23 branches in the phylogeny. A total of 23 branches were formally tested for diversifying selection. Significance was assessed using the Likelihood Ratio Test at a threshold of $p \leq 0.005$, after correcting for multiple hypothesis testing.

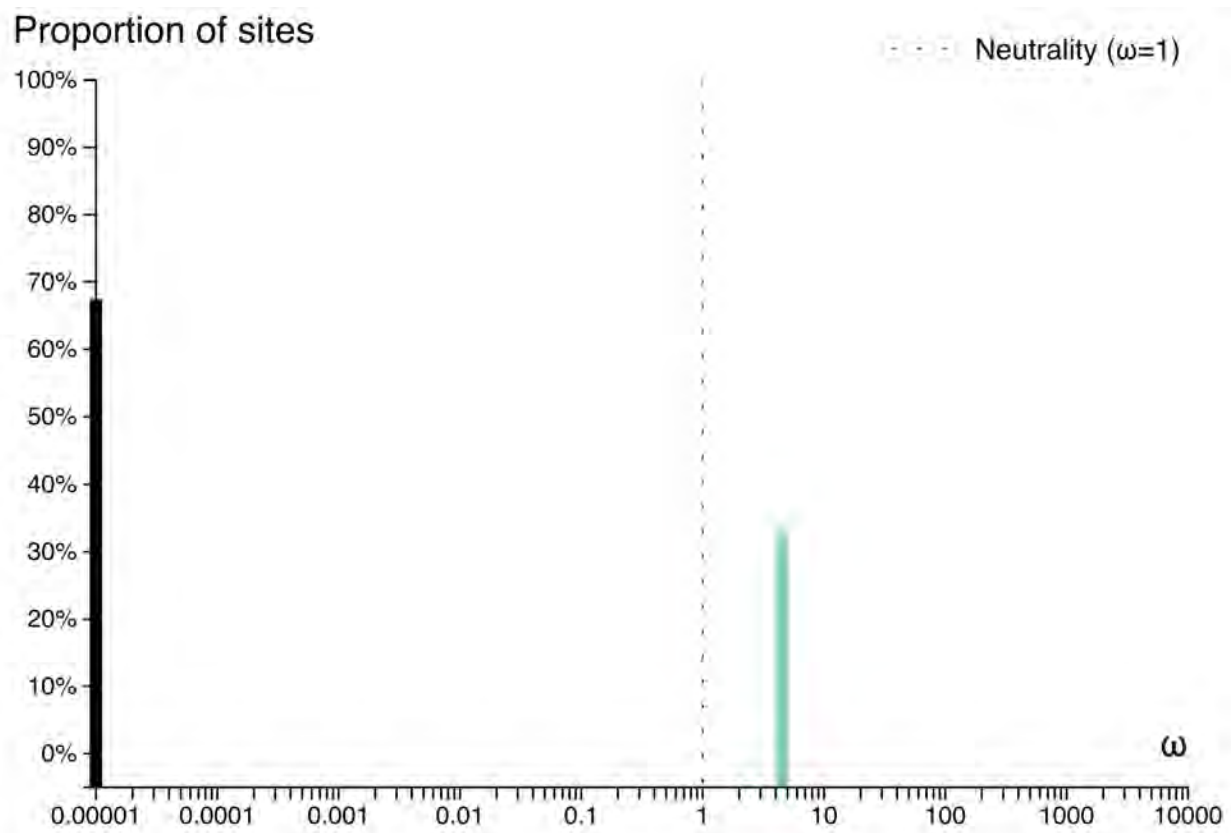
bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



4.3.1.1 Supplementary Figure: Adaptive branch-site REL test for episodic diversification (aBSREL) of elaterid luciferase homologs.

The branch leading to the common ancestor of elaterid luciferases (red arrow) was the only branch recovered with significant ($p<0.005$) evidence of positive selection, with 33% of sites showing strong directional selection (d_N/d_S or $\omega = 4.34$), and 67% of sites showing strong purifying selection (d_N/d_S or $\omega = 0.00$). The tree and results from the full adaptive model are shown. Increased thickness of the bars indicates a smaller p-value, color of the branches indicates the value of ω, and branch length indicates the number of substitutions per site.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



4.3.1.2 Supplementary Figure: Proportion of sites under positive selection

Values for d_N/d_S or ω for the two divisions of sites in the branch of the elaterid luciferases from the aBSREL analysis. 33% of the sites show strong directional selection whereas 67% of the sites show strong purifying selection.

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*[175]. This fatty acid binding protein was previously reported to bind strongly to fatty acids, and weakly to luciferin. Three G-coupled protein receptors (GCPRs) 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[176].

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; Supplementary Table 4.4.1) previously identified

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

from *P. pyralis* light organ extracts via photoaffinity labeling with an octopamine analog and partial N-terminal Edman degradation[176]. 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 light reaction, such a release of O_2 could enhance flash intensity or accelerate flash kinetics. Further research is required to test this hypothesis.

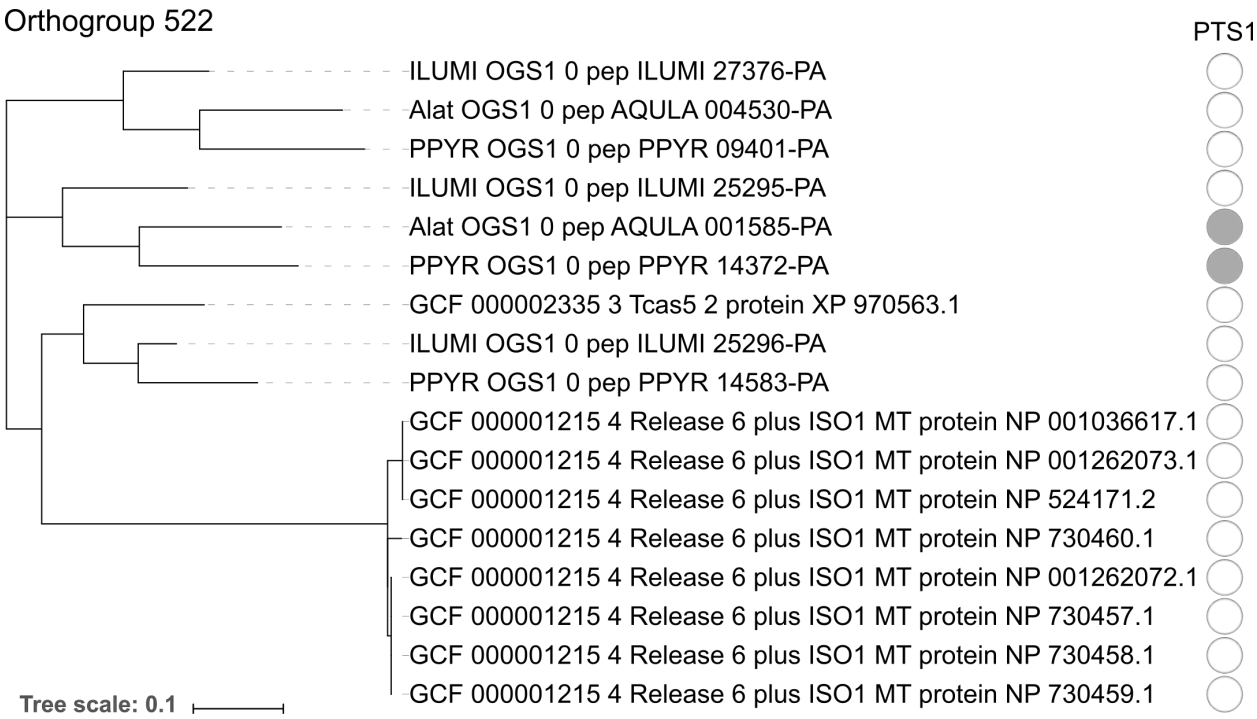
4.4.1 Supplementary Table: Highly expressed (HE), differentially expressed (DE), non-enzyme annotated (NotE), lantern genes whose closest relative in the opposite species is also HE, DE, NotE. BSN-TPM = between sample normalized TPM

P.pyralis ID (OGS1.0)	Predicted function	Ppyr BSN-TPM	Ppyr expression rank	Alat expression rank	Alat BSN-TPM	A. lateralis ID (OGS1.0)
PPYR_04589	Fatty-acid binding protein	71119	1	8	10464	AQULA_005257
PPYR_05098	Peroxisomal biogenesis factor 11 (PEX11)	4016	14	26	3294	AQULA_005466
PPYR_14966	Octopamine binding secreted hemocyanin	2095	35	21	3658	AQULA_008529
PPYR_11733	MFS transporter superfamily	1859	42	84	1335	AQULA_012209
PPYR_07633	Reticulon	1561	56	109	1123	AQULA_005090
PPYR_09394	lysosomal Cystine Transporter	1101	87	69	1494	AQULA_009474
PPYR_05852	Vacuolar ATP synthase 16kDa subunit	839	118	287	475	AQULA_001418
PPYR_11443	RNA-binding domain superfamily	784	134	1221	108	AQULA_003174
PPYR_11300	Mitochondrial outer membrane translocase complex	511	232	402	349	AQULA_004355
PPYR_04602	Leucine-rich repeat domain superfamily	460	262	378	373	AQULA_004134
PPYR_13497	Mitochondrial substrate/solute carrier	440	285	379	372	AQULA_003680

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

PPYR_04424	Domain of unknown function (DUF4782)	380	332	1296	101	AQULA_013946
PPYR_08278	Protein of unknown function DUF1151	366	348	430	325	AQULA_000628
PPYR_13261	Major facilitator superfamily	310	403	158	862	AQULA_007558
PPYR_05702	Sulfate permease family	226	543	396	357	AQULA_013064
PPYR_05993	V-type ATPase, V0 complex, 116kDa subunit family	211	579	541	251	AQULA_000400
PPYR_04179	Haemolymph juvenile hormone binding protein	202	606	879	152	AQULA_011187
PPYR_08298	Peroxisomal membrane protein (Pex16)	199	623	395	358	AQULA_013536
PPYR_06294	Homeobox-like domain superfamily - Abdominal-B-like	198	627	737	186	AQULA_000483
PPYR_01677	MFS transporter superfamily	108	1234	455	302	AQULA_002485

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



4.4.2 Supplementary Figure: Maximum likelihood phylogeny of the combined adenylyl-sulfate kinase & sulfate adenylyltransferase orthogroup.

Peptide sequences from P.pyralis, A. lateralis, I.luminosus, T. castaneum, and D. melanogaster were clustered (orthogroup # 522), multiple sequence aligned, and refactored into a species rooted maximum likelihood tree, via the OrthoFinder pipeline (maintext methods). PTS1 sequences were predicted from the peptide sequence using the PTS1 predictor server[177]. Figure produced with iTOL[178].

4.5 Opsin analysis

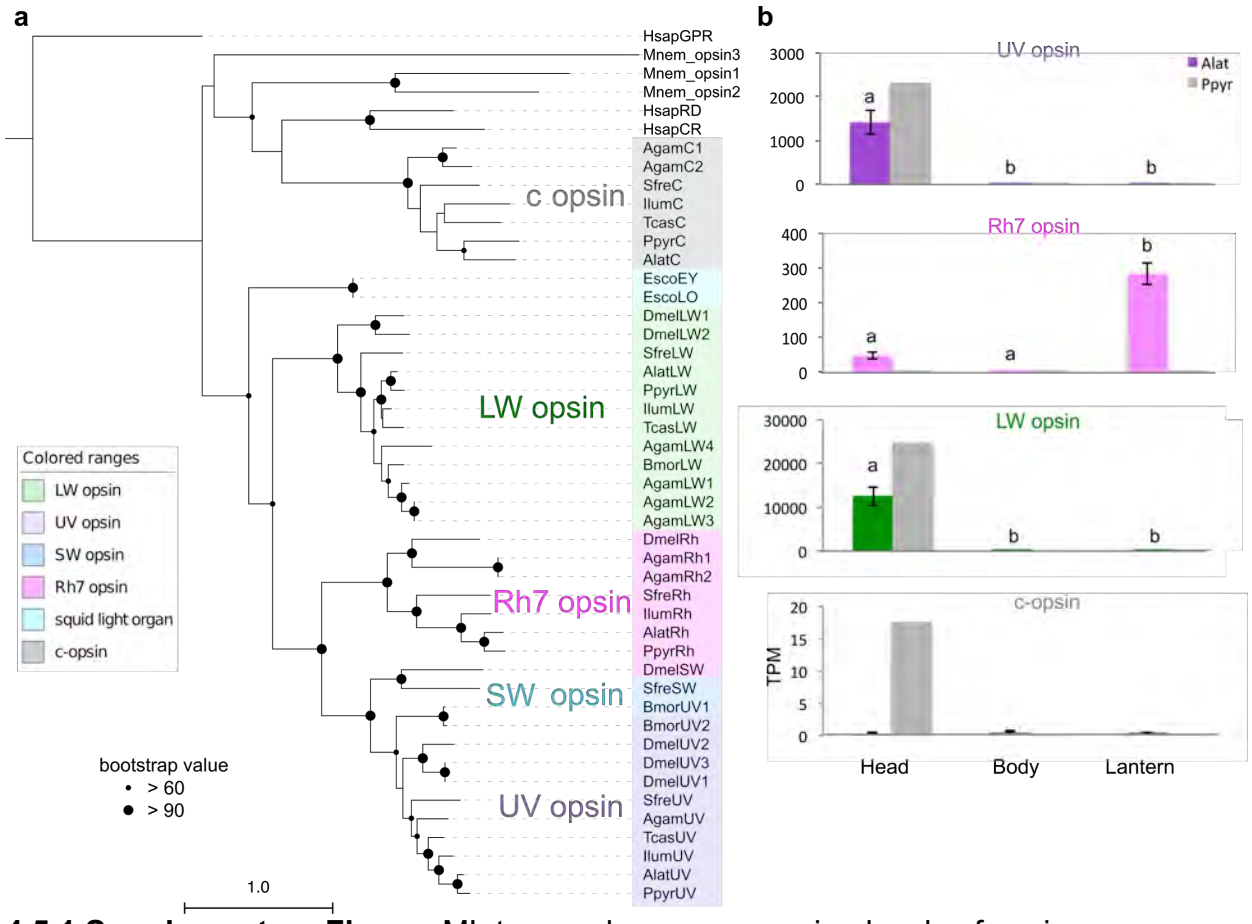
Opsins are G-protein-coupled receptors that, together with a bound chromophore, form visual pigments that detect light (reviewed in[179]). While opsin genes are known for their expression in photoreceptors and function in vision, they have also been found to be expressed in other tissues, suggesting 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[179,180]. Both insects and mammals may retain the alternate opsin type, generally in a non-visual capacity. The ancestral insect is hypothesized

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

to have 3 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[181,182]. A likely non-visual c-opsin was also detected, though not highly expressed[181,182].

To confirm the previously documented opsin presence and expression patterns, we collected candidate opsin genes via BLASTP searches (e-value: $1x10^{-20}$) of the Ppyr_OGS1.0, Alat_OGS1.0 and Ilumo_OGS1.0 reference genesets, as well as collected sequences via literature searches, followed by maximum likelihood phylogenetic reconstruction (Supplementary Figure 4.5.1a), and expression analyses of the opsins (Supplementary Figure 4.5.1.b). 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 differentially, but lowly expressed, in *P. pyralis* head tissue only (Supplemental Figure 4.5.1b). In contrast, *Rh7* was not expressed in the *P. pyralis* light organ, but was differentially expressed in the light organ of *A. lateralis* (Supplementary Figure 4.5.1b). The detection of *Rh7* in our genomes is unusual in beetles[183], though emerging genomic resources across the order have detected it in two taxa: *Anoplophora glabripennis*[184] and *Leptinotarsa decemlineata*[185]. *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[186]. 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)[187]. In the bioluminescent Ctenophore *Mnemiopsis leidyi,* three c-opsins are co-expressed with the luminous photoprotein in the photophores[188]. In the bobtail squid, *Euprymna scolopes,* one of the c-opsin isoforms is expressed in the bacterial symbiotic light organ[189]. Thus, it is possible that *Rh7* has a photo sensory function in the lantern of *A. lateralis*, though 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.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



4.5.1 Supplementary Figure: ML tree and gene expression levels of opsin genes.

a, Opsin Maximum likelihood (ML) tree. Collected opsin sequences were multiple sequence aligned with MAFFT L-INS-i[83] with default parameters. Gaps and ambiguous sequences were filtered with trimAL software[190] (parameter: -gt =0.5), and the ML tree reconstructed with MEGA7[172] with LG+G (5 gamma categories (+G, parameter = 1.3856) substitution model using 362 aa of multiple amino acid alignment. 100 bootstrap replicates were performed. HsapGPR was used as the outgroup sequence. Black circles on each node indicate bootstrap values. Scale bar equals substitutions per site. Taxon abbreviation: **Hsap:** *Homo sapiens*, **Mnem:** *Mnemiopsis leidyi* **Agam**: *Anopheles gambiae*, **Sfre:** *Sympetrum frequens*, **Ilum:** *Ignelater luminosus,* **Bmor**: *Bombyx mori*, **Ppyr:** *Photinus pyralis*, **Tcas:** *Tribolium castaneum*, **Dmel:** *Drosophila melanogaster.* The tree in Newick format, multiple sequence alignment files, and an excel document linking the provided gene names to the original sequence accession IDs and species name is available on FigShare (DOI: 10.6084/m9.figshare.5723005) **b,** Bar graphs indicate the gene expression levels in each body parts of averaged both male and female adult. The gene expressions in *A. lateralis* are tested with Tukey-Kramer method (three experimental replicates). UV and LW opsins are significantly highly expressed in the head ($p < 0.005$). On the other hand, *Rh7* was significantly highly expressed in the lantern ($p < 0.001$). Error bar represents standard error.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

We chose to analyze extracted hemolymph from both *P. pyralis*, and *A. lateralis* for lucibufagin content, by liquid-chromatography high-resolution accurate-mass mass-spectrometry (LC-HRAM-MS), as lucibufagins are known to accumulate in the hemolymph. 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 (~20,000xg) for 30 seconds 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 50µL methanol to precipitate proteins and other macromolecules. For *A. lateralis*, 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 minute 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 PFTE 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.

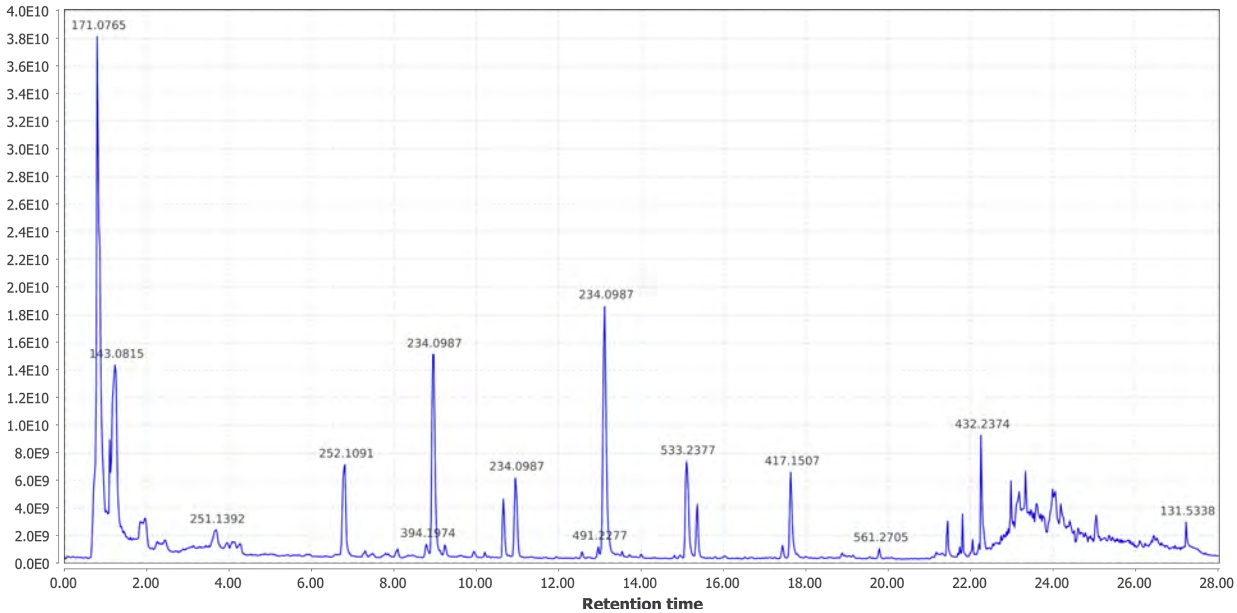
Injections of these filtered extracts (*P. pyralis* 10 µL; *A. lateralis* 5 µL) was separated on an UltiMate 3000 liquid chromatography system (Thermo Scientific) equipped with a 150 mm C18 Column (Kinetex 2.6 µm silica core shell C18 100Å pore, P/No. 00F-4462-Y0, Phenomenex) coupled to a Q-Exactive mass spectrometer (Thermo Scientific). For P. pyralis, compounds were separated by reversed-phase chromatography on the C18 column by 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 minutes, 95% B for 4 min, and 5% B for 5 min; flow rate 0.8 mL/min. For *A. lateralis*, compounds were separated by reversed-phase chromatography on the C18 column by 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, 5-80% 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 1 MS1 scan from *m/z* 120-1250 followed by 1 data-dependent MS2 scan using HCD fragmentation with a stepped collision energy of 10, 15, 25 normalized collision energy (NCE). Positive mode and negative mode MS1 and MS2 data was 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 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[191]. Data analysis was performed with Xcalibur (Thermo Scientific) and MZmine2 (v2.30)[192].

Within MZmine2, data was first cropped to 20 minutes in order to compare the A. lateralis and P.pyralis data which was obtained with the same LC gradient parameters. Profile MS1 data was then converted to centroid mode with the Mass detection module(Parameters:

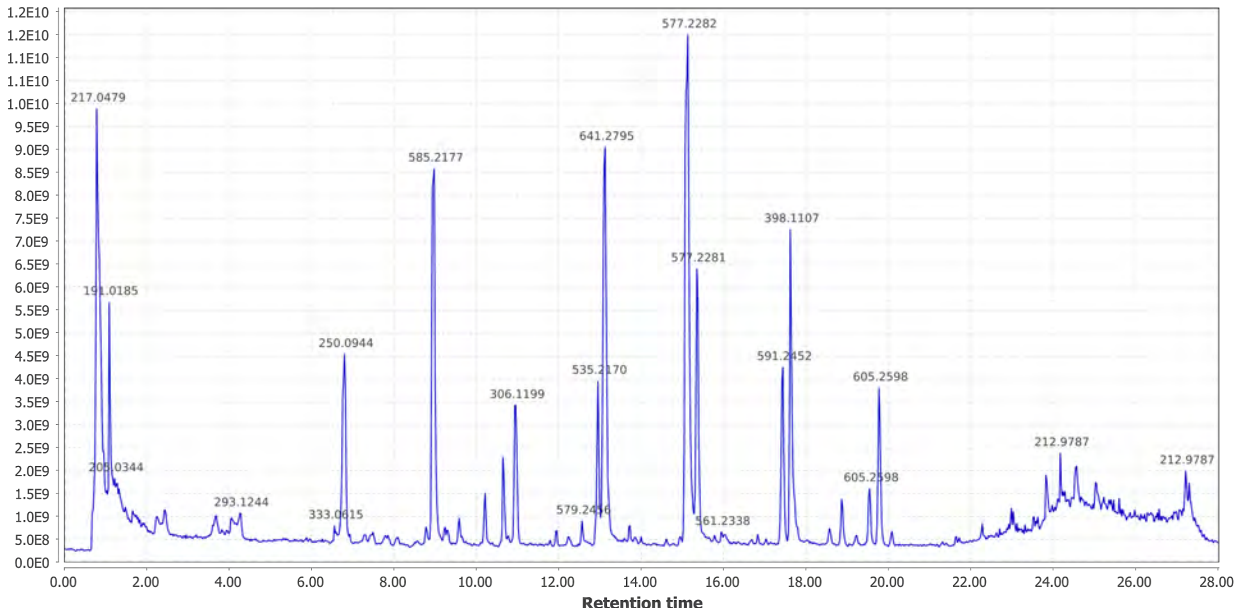
bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mass Detector = Exact mass, Noise level = 1.0E4), whereas MS2 data was converted to centroid mode with (Noise level=1.0E1). Ions were built into chromtograms using the Chromatogram Builder module with parameters (min_time_span = 0.10,min_height = 1.0E4, *m/z* tolerance = 0.001 *m/z* or 5 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 5 ppm, Retention time tolerance = 0.2 min, Monotonic shape=yes, Maximum charge = 2, Representative isotope=Most intense). The two peaklists (*P. pyralis*, *A. lateralis*) 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. There was no detected signal within the exact mass and retention time window defined by the 17 lucibufagin isomers of *P. pyralis*, allowing us to conclude that these lucibufagin isomers are likely absent from hemolymph of adult male *A. lateralis*.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

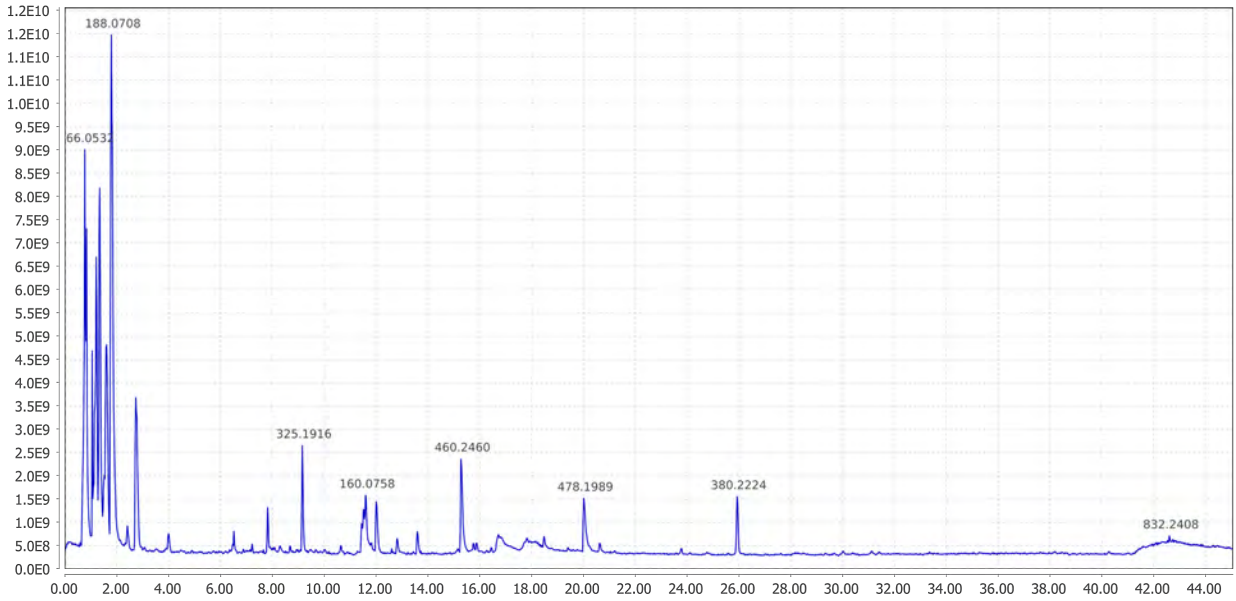


4.6.1 Supplementary Figure: Positive mode MS1 total-ion-chromatogram (TIC) of *P. pyralis* hemolymph LC-HRAM-MS data

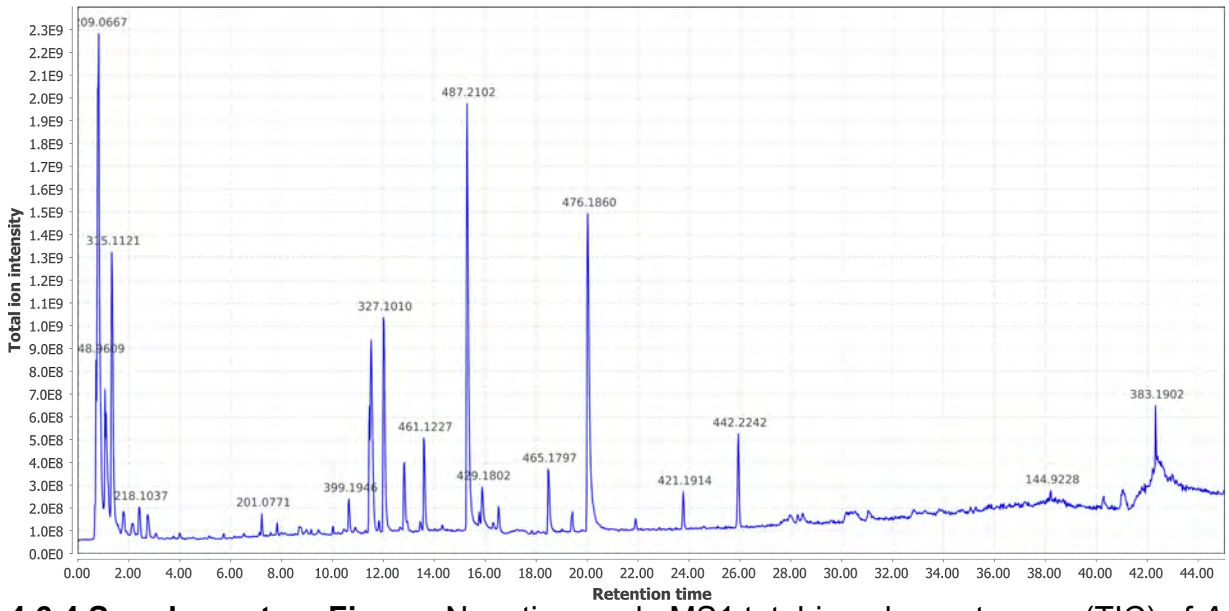


4.6.2 Supplementary Figure: Negative mode MS1 total-ion-chromatogram (TIC) of *P. pyralis* hemolymph LC-HRAM-MS data

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



4.6.3 Supplementary Figure: Positive mode MS1 total-ion-chromatogram (TIC) of *A. lateralis* hemolymph LC-HRAM-MS data

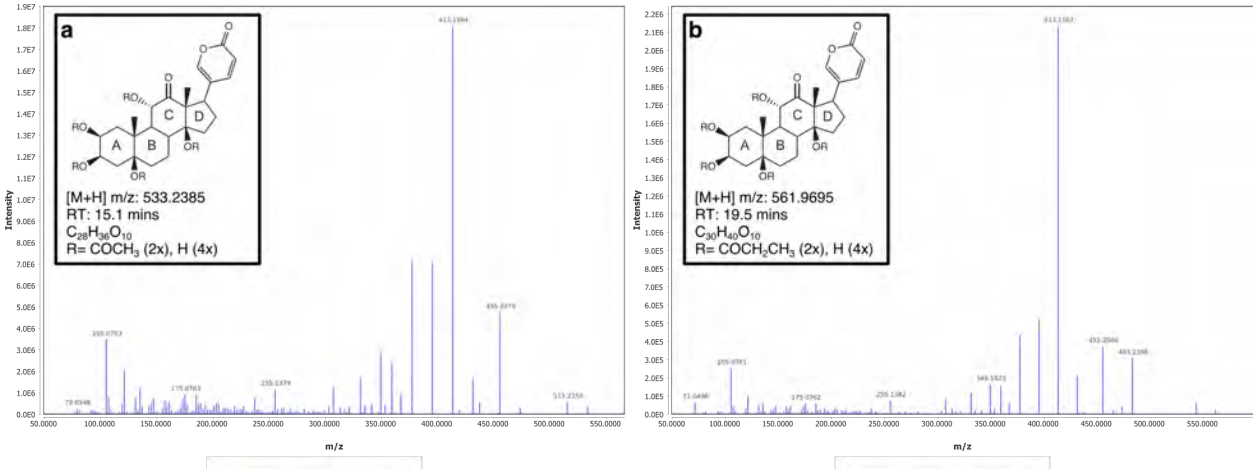


4.6.4 Supplementary Figure: Negative mode MS1 total-ion-chromatogram (TIC) of *A. lateralis* hemolymph LC-HRAM-MS data

4.6.5 MS2 similarity search for *P. pyralis* lucibufagins

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

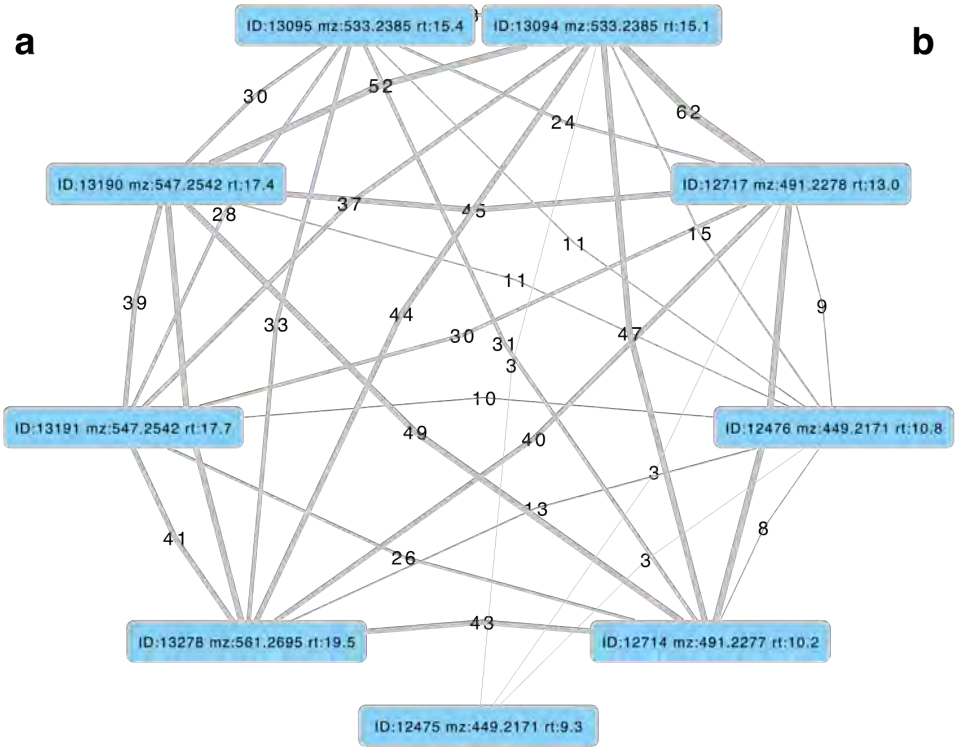
We first performed a MS2 similarity search for ions that showed a similar MS2 spectra to that of diacetylated lucibufagin ([M+H] *m/z* 533.2377, RT = 15.10 mins) (Supplementary Figure 4.6.5.1). This search was performed through the MS2 similarity search module of MZmine2 (v2.30) with parameters (*m/z* tolerance: 0.0004 *m/z* or 1 PPM; minimum # of ions to report: 3). This MS2 similarity search revealed 9 putative lucibufagin isomers with highly similar MS2 spectra (Supplementary Figure 4.6.5.2), which expanded to 17 putative lucibufagin isomers when considering features without MS2 spectra, but with identical exact masses and close retention times (ΔRT < 2 min) to the previously identified 9 (Supplementary Table 4.6.5.3). Chemical formula prediction was assigned to each precursor ion using SIRUIS (v3.5.1)[193]. The structural identity of these ions was easily interpreted in light that the different chemical formula represented the core lucibufagins that had undergone acetylation ($COCH_3$) or propylation ($COCH2CH_3$), 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 minutes), thus it may be possible that more highly substituted lucibufagins with a longer retention times are present, but not detected in the current analysis.



4.6.5.1 Supplementary Figure: Positive mode MS2 spectra of **a,** diacetylated lucibufagin and **b,** dipropylated lucibufagin

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a



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	$C_{24}H_{29}O_6$

4.6.5.2 Supplementary Figure: MS2 spectral similarity network for *P. pyralis* hemolymph lucibufagins

a) MS2 similarity network produced with the MZmine2 MS2 similarity search module. Nodes represent MS2 spectra from the initial dataset, whereas edges represent an MS2 similarity match between two MS2 spectra. Thickness / label of the edge represents the number of ions matched between the two MS2 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 Adducts and complexes of the presented ions were manually removed.

4.6.5.3 Supplementary Table: Putative lucibufagin compounds from LC-HRAM-MS of *P. pyralis* hemolymph

Assigned ion identity	Ion type	*m/z*	Retention time (mins)	Feature area (arb)
Core lucibufagin isomer 1	[M+H]	449.2171	7.9	6.7E+05
Core lucibufagin isomer 2	""	""	9.3	1.1E+07
Monoacetylated lucibufagin isomer 1	""	491.2277	10.2	4.2E+07
Core lucibufagin isomer 3	""	449.2171	10.8	1.7E+07
Monoacetylated lucibufagin isomer 2	""	491.2277	11.4	1.1E+06
Monoacetylated lucibufagin isomer 3	""	""	11.9	1.8E+07

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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	""	533.2385	15.1	4.0E+09
Diacetylated lucibufagin isomer 2	""	""	15.4	1.9E+09
Monoacetylated, mono propylated lucibufagin isomer 1	""	547.2542	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	""	561.2695	18.9	1.4E+08
Dipropylated lucibufagin isomer 2	""	""	19.5	3.9E+07
Dipropylated lucibufagin isomer 3	""	""	19.8	1.8E+08

4.6.7 MS2 similarity search for *A. lateralis* lucibufagins

Although our earlier LC-HRAM-MS analysis (Supplementary Note 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 not present in *P .pyralis* that are present in *A. lateralis*. To address this, we performed a MS2 similarity search against the *A. lateralis* adult male hemolymph MS2 spectra, with the MS2 spectra of lucibufagin "C" (*m/z* 533.2385, RT=15.1) as bait, using the MZmine2 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 MS2 spectra were matched (Supplementary Table 4.6.7.1). None of these features were detected in *P. pyralis*. 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 are derived from lucibufagins. The notable degree of MS2 similarity may be due to the *A. lateralis* compounds being also 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* hemolymph, is absent from the *P. pyralis* hemolymph, as is possibly a steroidal compound.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

4.6.7.1 Supplementary Table: Relative quantification of features identified by lucibufagin MS2 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	C22H38NO7P*; C25H29N7O2*	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	C23H28N2O3*	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 (MS2 analysis), and MZmine2 (isotope pattern analysis).
N.D., Not determined

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SUPPLEMENTARY NOTE 5: Holobiont analyses

5.1 Assembly of the complete *Entomoplasma luminosum* var. pyralis genome

The complete genome of *E. luminosum* var. 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 assembly graph with Bandage v0.8.1[145], 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 resulting consensus sequence was restarted with seqkit[51] 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 3 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 using Prokka v.1.12[194] with the mycoplasma/spiroplasma genetic code parameter (code 4). 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 RNA gene sequence to the NCBI 16S RNA gene database indicated that this gene had 99% identity to the *E. luminosum* 16S sequence (ATCC 49195 - formerly *Mycoplasma luminosum;* NCBI Assembly ID ASM52685v1)[195,196], leading to our description of this genome as *Entomoplasma luminosum* var. pyralis. Protein overlap comparisons using the Orthofinder pipeline (v1.1.10)[152] 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* var. pyralis.

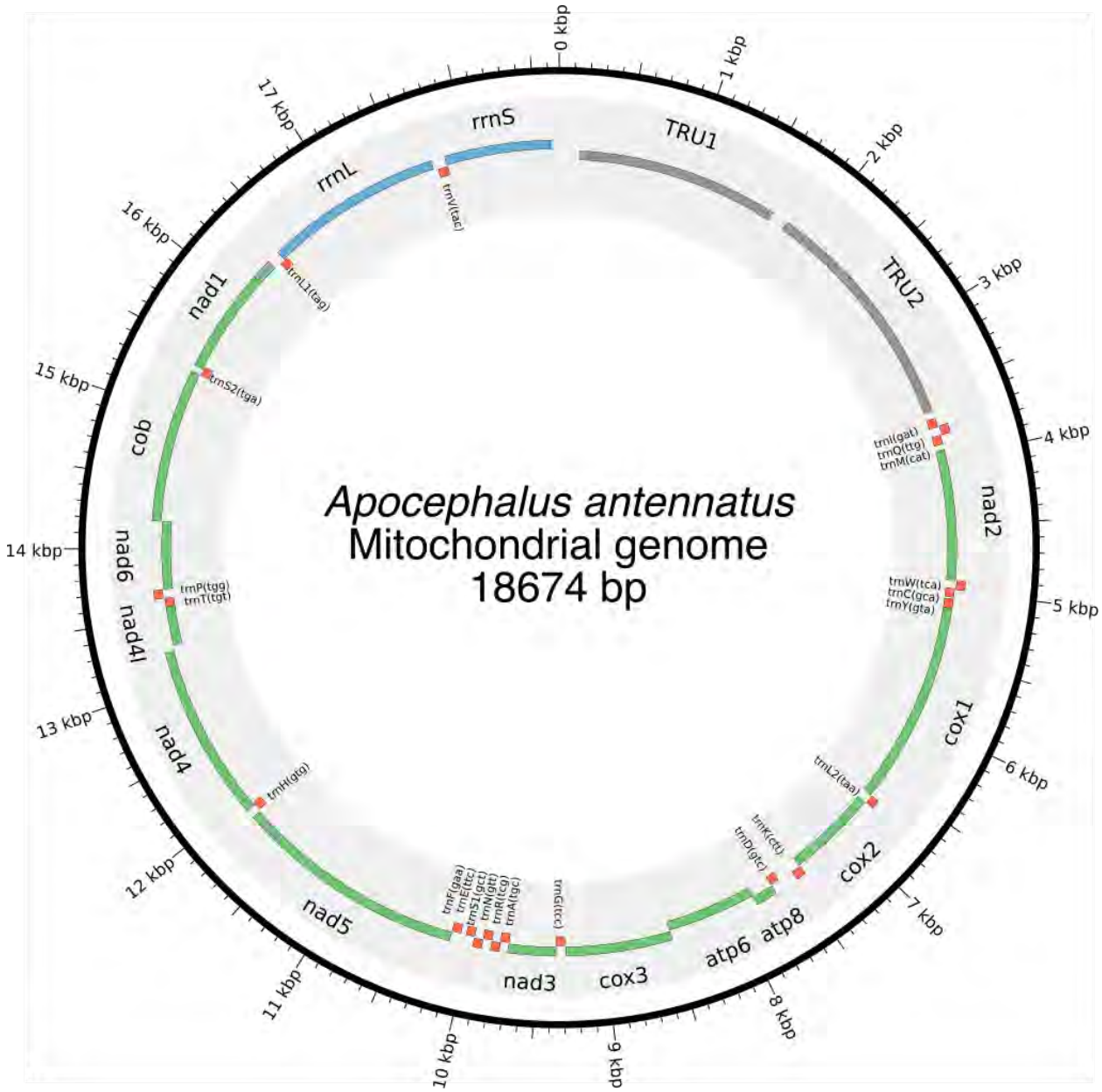
5.2 Assembly of Phorid mitochondrial genome

The complete mitochondrial genome of the dipteran parasatoid *Apocephalus antennatus,* first detected 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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(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 3 partitions: *P. pyralis* mapping, *M. scalaris* mapping, and other, and input into Canu v1.6+44 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 *P. pyralis* DNA. 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)[197], 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; Supplementary Figure 5.2.1). This mtDNA was taxonomically identified to originate from *A. antennatus* (Supplementary Note 5.3)

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



5.2.1 Supplementary Figure: Mitochondrial genome of *Apocephalus antennatus*.

The mitochondrial genome of *A. antennatus* was assembled and annotated as described in the Supplementary Note 5.2, and taxonomically identified as described in Supplementary Note 5.3. Figure produced with Circos[53]

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 (Supplementary Note 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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mitochondrial genome sequence to our 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 5 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 these specimens and a COI fragment was PCR-amplified and Sanger sequenced. The forward primer was forward primer 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[198] 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 & 2 (PpyrOMLV1 & 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). 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 & 2. Genomic and phylogenetic studies suggest that the detected viruses correspond to a new lineage within the *Orthomyxoviridae* family (ssRNA(-)) (Supp. Fig. 5.4.1 A-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 (Supplementary Note 5.5, Supp. Fig. 5.4.1j). 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 & PpyrOMLV2) share their genomic architecture and evolutionary clustering (Supp. Fig. 5.4.1a-h, Supp Fig. 5.4.2). 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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

protein - PA (PpyrOMLV1: 754 aa, 86.6 kDA; PpyrOMLV2: 762 aa, 87.9 kDA). PpyrOMLV1 & 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 & 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 (Supp Table 2). PpyrOMLV1 & 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[199]. Genome segment four (1.6 Kbp size) encodes a Hemaglutinin 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 (Supp. Fig 5.4.1 E). Despite sharing genome architecture and structural and functional domains of their predicted proteins, PpyrOMLV1 & 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 (Supp. Fig. 5.4.2). 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 & PpyrOMLV2 genome segments present an overall high identity in their respective RNA segments ends (Supp. Fig 5.4.1 F). 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[200]. 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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 & PpyrOMLV2 to the OMV virus family. These are the first viruses that have been associated with the *Lampyridae* beetle family, which includes over 2,000 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[201]. 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[201]. As described previously, every OMV RNA segment possess conserved and partially complementary 5′- and 3′-end sequences with promoter activity[202]. OMV structural proteins are tentatively common to all genera involving the three polypeptides subunits that form the viral RdRP (PA, PB1, PB2)[203]; 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[204]. The HA protein of FLUCV, besides its hemagglutinating and envelope fusion function, has an esterase activity that induces host receptor enzymatic destruction[201]. In contrast, the HA of THOV is divergent to influenzavirus HA proteins, and presents high sequence similarity to a baculovirus surface glycoprotein[205]. 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[201]. 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,

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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)[201]. 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[206]. 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[207]. 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 (Supp. Figure 5.4.1.H) 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[203]. 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 (Supp. Figure 5.4.1.G-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[200,208]. 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[201]. In thogotoviruses, no requirement for HA glycoprotein cleavage have been demonstrated[205]. Integral membrane

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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[209]. 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[210]. 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[211]. 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. Influenzaviruses 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, though rare, is well documented. Influenza B virus infect humans and cause epidemics, and have been rarely found in seals. Influenzaviruses C cause limited outbreaks in humans and have been occasionally found on dogs. Influenza spreads globaly in a yearly outbreak, resulting in about three to five million cases of severe illness and about 250,000 to 500,000 human deaths[212]. Influenzavirus D has been recently reported and accepted and infects cows and swine[213]. 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 dromedarri*, a species of camel ticks[214,215]. 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[216]. Quaranfil and Johnston Atoll are transmitted by ticks and infect avian species[217].

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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 (Supplementary Fig. 5.4.1.b). When the genetic distances of firefly viruses proteins and ICTV accepted OMV species were computed, a strong similarity was evident (Supplementary Fig.5.4.1.b-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 (Supplementary Fig. 5.4.2). 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 (Supplementary Figure 5.4.2). 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 & 2 were assessed by the generation and analyses of 29 specific RNA-Seq libraries of *P. pyralis* (refer to Specimens/libraries Table). 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* (Supplementary Figure 5.4.1 J, Supplementary Table 2). Wild caught individuals were collected in period spanning six years, and locations separated as much as 900 miles (New Jersey – Georgia, USA). Interestingly PpyrOMLV1 & 2 were found in individuals of both location, and the corresponding assembled isolate virus sequences presented negligible differences, with an inter-individual

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 & 2, which appears to be significantly lower than of Influenzaviruses, specifically *Influenza A virus,* which are characterized by high mutational rate (ca. 1 mutation per genome replication) associated to the absence of RNA proofreading enzymes [218]. In addition we evaluated the presence of PpyrOMLV1 & 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 [201]. 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 (Supplementary Table 5.4.5, 5.4.6), indicating a widespread tissue/organ tropism of PpyrOMLV1 & 2. This tentatively pervasive tropism of PpyrOMLV1 & 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[219]. 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[220]. The presence of PpyrOMLV1 & 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 & 2 in eight

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 OMV viral genome and may occur during mixed infections. Reassortment generates viral diversity and has been associated to host gain of Influenzavirus[221]. 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[206]. 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 & 2.

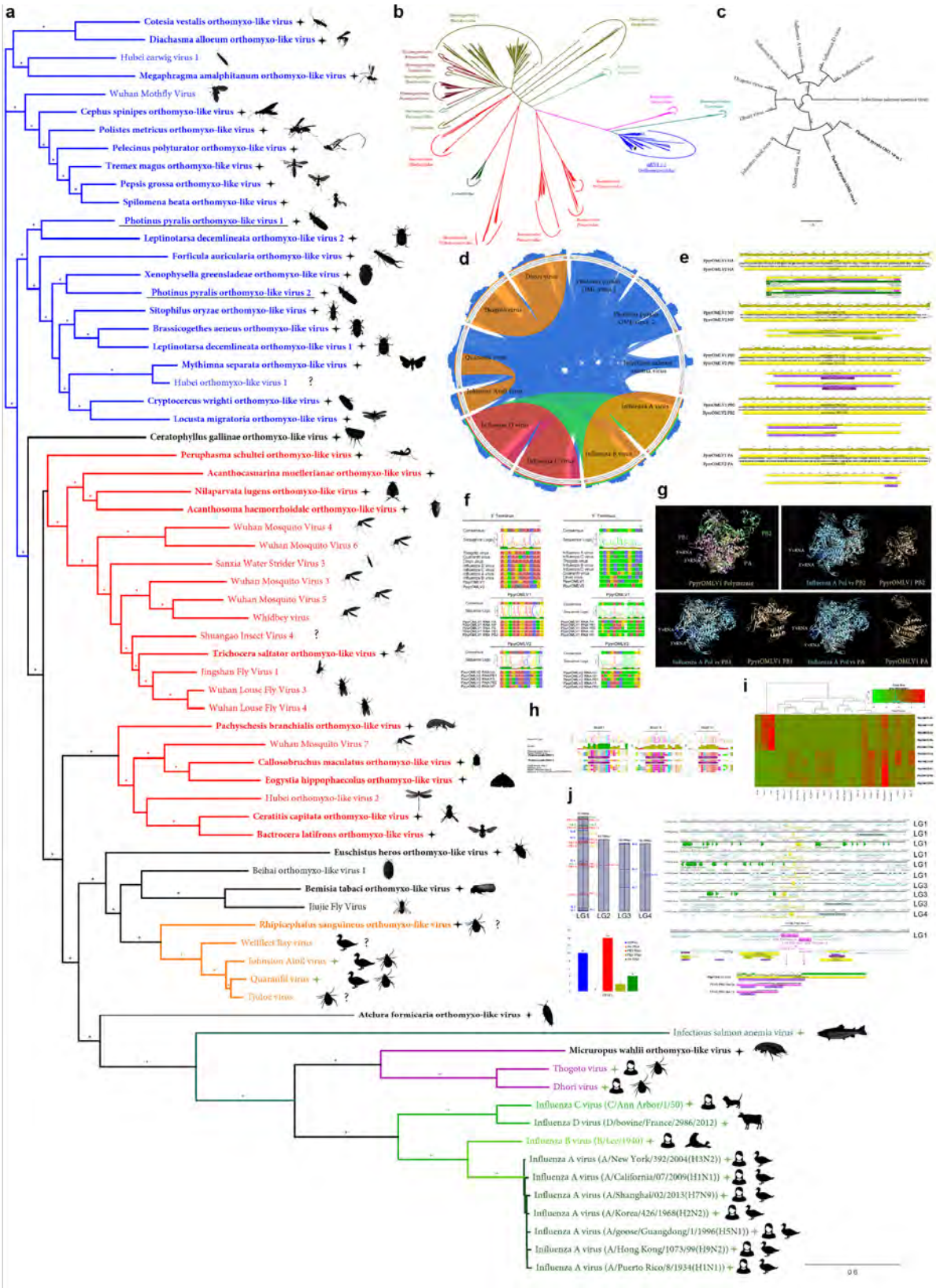
In light of the presence of virus RNA in reproductive glands, we further explored the potential life style of PpyrOMLV1 & 2 related to eventual vertical transmission. Vertical transmission is extremely exceptional for OMV, and has only been conclusively described for the *Infectious salmon anemia virus* (*Isavirus)*[222]. 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 1st 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 1st instar larvae, in the sequenced library average virus RNA levels were of 10.42 FPKM. Future experiments should focus on PpyrOMLV1 & 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 & 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 & 2 are associated with any potential alteration of phenotype of the infected host. We failed to unveil

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

any specific effect of the presence of PpyrOMLV1 & 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 & 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[223].

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

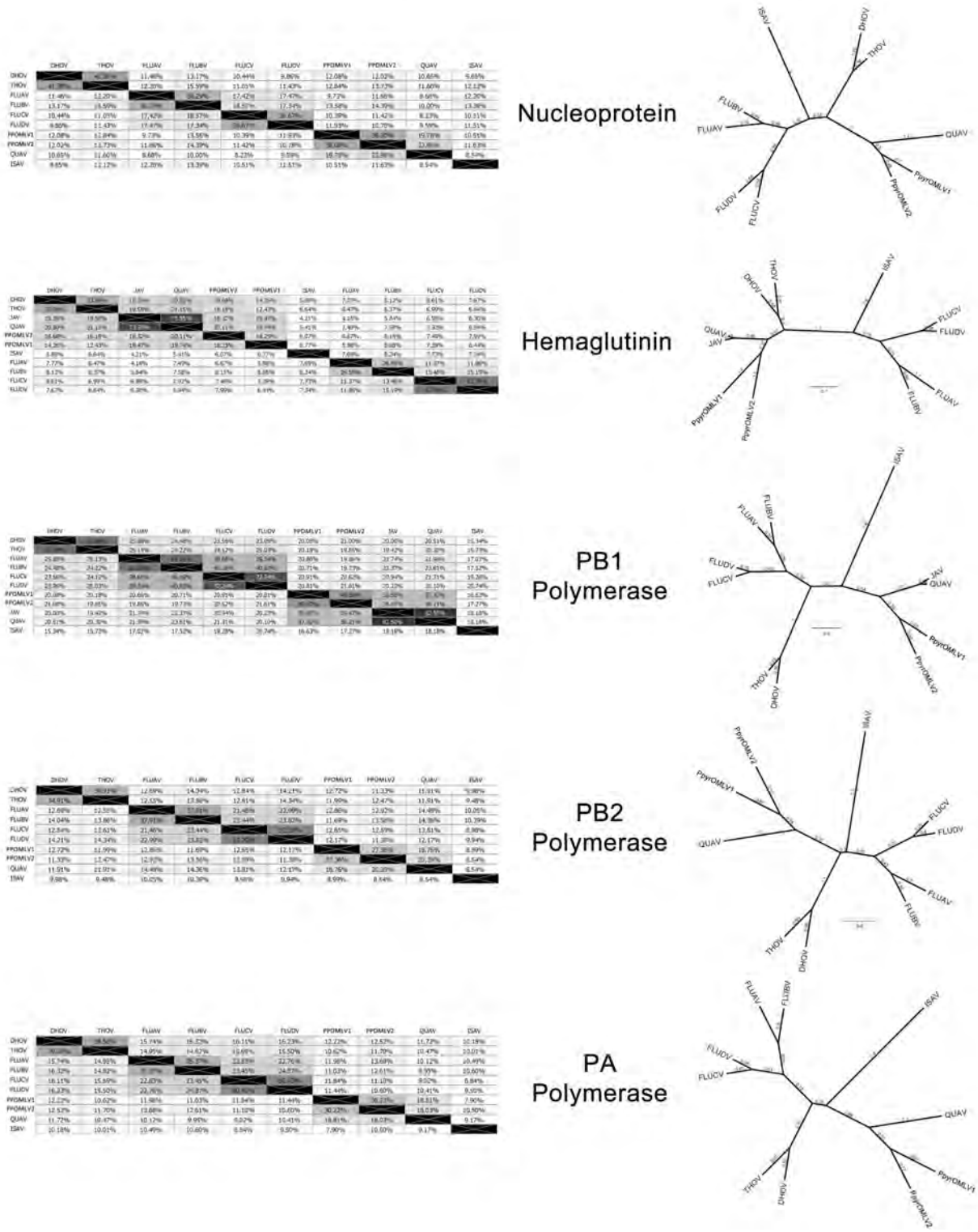


5.4.1 Supplementary Figure: *Photinus pyralis* viruses and endogenous viral-like elements.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(**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 & 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 & 2. (**F**) Virus genomic noncoding termini analyses of PpyrOLMV1 & 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 & 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 & 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.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



5.4.2 Supplementary Figure: Pairwise identity of OMLV viral proteins amongst identified OMLV viruses.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

5.4.3 Supplementary Table: 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%**

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

5.4.4 Supplementary Table: 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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

5.4.5 Supplementary Table: Reads mapped to PpyrOMLV genome segments in available SRA and newly described *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 OMLV 1 HA	11	541	2	160	0	4	881	2	0	2	199	2848
Ppyr OMLV 1 NP	0	321	0	141	0	0	523	0	0	0	120	1460
Ppyr OMLV 1 PA	3	256	0	95	0	0	306	1	0	5	100	660
Ppyr OMLV 1 PB1	2	364	2	208	0	4	820	0	0	0	669	1464
Ppyr OMLV 1 PB2	5	194	0	152	2	0	319	2	0	0	106	696
Ppyr OMLV 2 HA	12	444	266	124	54	247	549	38	22	10	232	710
Ppyr OMLV 2 NP	29	526	275	144	66	299	653	24	205	57	274	1067
Ppyr OMLV 2 PA	12	88	216	72	40	204	97	18	15	8	50	838
Ppyr OMLV 2 PB1	9	115	75	72	26	78	76	8	74	57	146	493
Ppyr OMLV 2 PB2	5	50	57	67	47	131	110	22	85	72	173	728

	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	578	2	6	867	0	0	0	0	1664	7826	15586
Ppyr OMLV 1 NP	0	289	0	3	647	0	2	0	0	644	5216	6562
Ppyr OMLV 1 PA	0	124	0	2	626	0	0	0	0	1264	3692	9564
Ppyr OMLV 1 PB1	2	460	0	3	1607	2	0	0	0	2824	7144	15952

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ppyr OMLV 1 PB2	0	188	0	2	848	0	0	0	0	648	2562	10568
Ppyr OMLV 2 HA	13	236	23	546	337	286	43	190	415	0	0	0
Ppyr OMLV 2 NP	32	248	22	501	482	196	51	127	432	0	0	0
Ppyr OMLV 2 PA	14	93	6	234	222	131	75	54	97	0	0	0
Ppyr OMLV 2 PB1	29	90	4	168	180	63	22	96	190	0	0	0
Ppyr OMLV 2 PB2	49	90	6	256	230	94	22	57	96	0	0	0

5.4.6 Supplementary Table: Reads mapped to PpyrOMLV genome segments in with FPKM values in available SRA and newly described *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	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 OMLV1 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 OMLV1 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 OMLV1 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 OMLV1 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 OMLV2 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 OMLV2 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 OMLV2 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 OMLV2 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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ppyr OMLV2 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_egs
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 OMLV 1 PB2	0.00	3.91	0.00	0.05	15.05	0.00	0.00	0.00	0.00	4.06	15.96	46.51
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 & 2, we examined our assembly of *P. pyralis* for putative signatures or paleovirological traces[224–226] that would indicate ancestral integration of virus related sequences into the firefly host. Remarkably, we found Endogenous virus-like Elements (EVEs)[227], sharing significant sequence identity with most PpyrOMLV1 & 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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

depend on integration into the genome of the host cell as an obligate step in their replication strategy[228]. 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[229]. 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 & 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[230,231]. Interestingly, when we screened our *P. pyralis* genome assembly Ppyr1.2 by BLASTX searches (E-value $<1e10^{-6}$) of PpyrOMLV1 & 2 genome segments, we identified several genome regions that could be defined as Firefly EVEs, which we termed FEVEs (Supplementary Fig. 5.1 J; Supplementary Table 5.5.1-5.5.5). 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 & 2 (ca. 46 % of FEVEs), followed by N encoding genome segment five (ca. 33 % of detected FEVEs). In addition we identified four FEVEs related to genome segment three (PA region) and two FEVEs associated to genome segment two (PB2 encoding region). We found no evidence of FEVEs related to the hemagglutinin coding genome segment four (HA). 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 (Supplementary Table 5.5.1-5.5.5), 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) (Supplementary Fig. 5.4.1 J) 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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

case of *Ixodes scapularis*, the putative vector of *Quaranfil virus* and *Johnston Atoll virus* corresponding to genus *Quaranjavirus* [227]. 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 & 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 & Gifford[227] 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[232]. 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[233]. Recently, in mosquitoes, a new mechanism of antiviral immunity against RNA viruses has been proposed, relying in the production and expression of EVEs DNA[234]. 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[235]. In addition, integration and eventual modulation in the host genome may be associated with an interaction between viral RNA and the mosquito RNAi machinery[236]. The piRNA pathway mediates through small RNAs and Piwi-Argonaut proteins the repression of TE

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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[231,237]. 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[234]. 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.

5.5.1 Supplementary Table: 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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ppyr1.2_LG2	34546169	34545801	(-)	52.80%	1.16E-31	34.10%	EVE PB1 like-14

5.5.2 Supplementary Table: 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

5.5.3 Supplementary Table: 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

5.5.4 Supplementary Table: FEVE hits from BLASTX of PpyrOMLV HA

(None detected)

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

5.5.5 Supplementary Table: 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

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SUPPLEMENTARY NOTE 6: Experiment.com donors

Liliana Bachrach	Doug Fambrough	Benjamin Lower	Luis Cunha	Joshua Guerriero
Atsuko Fish	Tom Alar	Noreen Huefner	David Esopi	John Skarha
Rutong Xie	Richard Hall	Zachary Michel	Jack Hynes	Keith Guerin
Nathan Shaner	Joe Doggett	Joe T. Bamberg	Michael McGurk	Pureum Kim
Sara Lewis	Mark Lewis	Lauren Solomon	Peter Berx	Milo Grika
Jing-Ke Weng	Sarah Sander	Dr. Husni Elbahesh	Matt Grommes	Daniel Zinshteyn
Peter Rodenbeck	Daniel Bear	Kathryn Larracuente	Colette Dedyn	Tom Brekke
Larry Fish	Don Salvatore	Matthew Cichocki	Florencia Schlamp	Edoardo Gianni
Amanda Larracuente	Emily Davenport	Marcel Bruchez	Marie Lower	Cindy Wu
Hunter Lower	Ted Sharpe	Robert Unckless	Michael R. McKain	Christina Tran
Allan Kleinman	David Plunkett	Arvid Ågren	Ben Pfeiffer	Eric Damon Walters
Misha Koksharov	Tim Fallon	Margaret S Butler	Kathryn Keho	Geoffrey Giller
Sarah Shekher	Edward Garrity	Yasir Ahmed-Braimah	Jenny Wayfarer	Fahd Butt
Jared Lee	Huaping Mo	Ruth Ann Grissom	Darby Thomas	Christophe Mandy
Raphael De Cock	TimG	Tomáš Pluskal	Emily Hatas	

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Linds Fallon	Jan Thys	Genome Galaxy	Richard Casey	
Grace Li	Francisco Martinez Gasco	Dustin Greiner	William Nicholls	

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SUPPLEMENTARY NOTE 7: Data availability

Files on FigShare:
(1) *Photinus pyralis* sighting records (Excel spreadsheet)
(2) Ppyr1.2 Blobtools results
(3) Alat1.2 Blobtools results
(4) Ilumi1.0 Blobtools results
(5) Nucleotide multiple sequence alignment for Elaterid luciferase homolog branch selection test (Supplementary Note 4.3)
(6) Protein multiple sequence alignment for P450 tree - Supplementary Fig 1.10.1.1
(7) Extended Data Table 1 - Genomic sequencing library statistics
(8) Photinus pyralis orthomyxo-like virus 1 sequence and annotation
(9) *Photinus pyralis orthomyxo-like virus 2 sequence and annotation*
(10) Orthofinder protein clustering analysis
(11) Ppyr_OGS1.0 kallisto RNA-Seq expression quantification (TPM)
(12) Alat_OGS1.0 kallisto RNA-Seq expression quantification (TPM)
(13) Fig 4e. Ppyr_OGS1.0 Sleuth / differential expression Venn diagram analysis (BSN-TPM)
(14) Fig 4e. Alat_OGS1.0 Sleuth / differential expression Venn diagram analysis (BSN-TPM)
(15) Ilumi_OGS1.0 kallisto kallisto RNA-Seq expression quantification (TPM)
(16) Supplementary Video 1: A Photinus pyralis courtship dialogue
(17) Figure 4c. CYP303 maximum likelihood gene tree
(18) Supplementary Figure 4.5.1a Opsin gene tree
(19) Figure 3c Maximum likelihood tree of luciferase homologs.

Files on www.fireflybase.org:
(1) *P. pyralis/A. lateralis/I.luminosus* geneset GFF files
(2) *P. pyralis/A. lateralis/I.luminosus* mRNA nucleotide FASTA files

Tracks on www.fireflybase.org JBrowse genome browser:
For each genome:
(1) Gaps
(2) Repeats
(3) Direct gene-models (Stringtie)
(4) Direct gene-models (Trinity)
(5) Official geneset gene-models (OGS1.0)

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Bibliography

1. Lloyd, J. E. Studies on the flash communication system in Photinus fireflies. (1966).
2. Lloyd, J. E. Fireflies (Coleoptera: Lampyridae). in *Encyclopedia of Entomology* (ed. Capinera, J. L.) 1429–1452 (Springer Netherlands, 2008).
3. de Wet, J. R., Wood, K. V., Helinski, D. R. & DeLuca, M. Cloning of firefly luciferase cDNA and the expression of active luciferase in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **82,** 7870–7873 (1985).
4. Case, J. F. Flight studies on photic communication by the firefly Photinus pyralis. *Integr. Comp. Biol.* **44,** 250–258 (2004).
5. Reijden, E. D. van der, Monchamp, J. D. & Lewis, S. M. The formation, transfer, and fate of spermatophores in Photinus fireflies (Coleoptera: Lampyridae). *Can. J. Zool.* **75,** 1202–1207 (1997).
6. Al-Wathiqui, N., Fallon, T. R., South, A., Weng, J.-K. & Lewis, S. M. Molecular characterization of firefly nuptial gifts: a multi-omics approach sheds light on postcopulatory sexual selection. *Sci. Rep.* **6,** 38556 (2016).
7. Cratsley, C. K., Rooney, J. A. & Lewis, S. M. Limits to Nuptial Gift Production by Male Fireflies, Photinus ignitus. *J. Insect Behav.* **16,** 361–370 (2003).
8. Rooney, J. & Lewis, S. M. Fitness advantage from nuptial gifts in female fireflies. *Ecol. Entomol.* **27,** 373–377 (2002).
9. Faust, L. & Faust, H. The Occurrence and Behaviors of North American Fireflies (Coleoptera: Lampyridae) on Milkweed, Asclepias syriaca L. *Coleopt. Bull.* **68,** 283–291 (2014).
10. Hess, W. N. Notes on the biology of some common Lampyridae. *Biol. Bull.* **38,** 39–76 (1920).
11. Faust, L. F. *Fireflies, Glow-worms, and Lightning Bugs: Identification and Natural History of*

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the Fireflies of the Eastern and Central United States and Canada. (University of Georgia Press, 2017).

12. Meinwald, J., Wiemer, D. F. & Eisner, T. 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 (1979).
13. Goetz, M. A., Meinwald, J. & Eisner, T. Lucibufagins, IV. New defensive steroids and a pterin from the firefly,Photinus pyralis (coleoptera: Lampyridae). *Experientia* **37,** 679–680 (1981).
14. Blum, M. S. & Sannasi, A. Reflex bleeding in the lampyrid Photinus pyralis: Defensive function. *J. Insect Physiol.* **20,** 451–460 (1974).
15. Faust, L., De Cock, R. & Lewis, S. Thieves in the Night: Kleptoparasitism by Fireflies in the Genus Photuris Dejean (Coleoptera: Lampyridae). *Coleopt. Bull.* **66,** 1–6 (2012).
16. Luk, S. P. L., Marshall, S. A. & Branham, M. A. The fireflies of Ontario (Coleoptera: Lampyridae). *Can. J. Arthropod Identif.* **16,** 1–105 (2011).
17. Common Eastern Firefly (Photinus pyralis). *iNaturalist* Available at: https://www.inaturalist.org/taxa/129350-Photinus-pyralis.
18. Foundation, O. S. G. *QGIS Geographic Information System*. (2017).
19. Fallon, T. R. Ppyralis_QGIS_sighting_to_centroided_county.py. Available at: https://github.com/photocyte/2017_misc_scripts/blob/master/Ppyralis_QGIS_sighting_to_centroided_county.py.
20. United States Census Bureau. Cartographic Boundary Shapefiles - Counties. Available at: https://www.census.gov/geo/maps-data/data/cbf/cbf_counties.html. (Accessed: 2017)
21. Olson, D. M. *et al.* 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 (2001).
22. Fund, W. W. Terrestrial Ecoregions of the World. Available at:

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

https://www.worldwildlife.org/publications/terrestrial-ecoregions-of-the-world. (Accessed: 2017)

23. Green, J. W. Revision of the nearctic species of Photinus (Coleoptera: Lampyridae). *Proc. Calif. Acad. Sci.* **28,** 561–613 (1956).
24. Stanger-Hall, K. F. & Lloyd, J. E. Flash signal evolution in Photinus fireflies: character displacement and signal exploitation in a visual communication system. *Evolution* **69,** 666–682 (2015).
25. Wasserman, M. & Ehrman, L. Firefly Chromosomes, II. (Lampyridae: Coleoptera). *Fla. Entomol.* **69,** 755–757 (1986).
26. Lower, S. S. *et al.* Genome Size in North American Fireflies: Substantial Variation Likely Driven by Neutral Processes. *Genome Biol. Evol.* **9,** 1499–1512 (2017).
27. Dias, C. M., Schneider, M. C., Rosa, S. P., Costa, C. & Cella, D. M. The first cytogenetic report of fireflies (Coleoptera, Lampyridae) from Brazilian fauna. *Acta Zool.* **88,** 309–316 (2007).
28. Biosciences, P. SMRT Analysis Software. Available at: http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/.
29. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
30. Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31,** 1119–1125 (2013).
31. Bickhart, D. M. *et al.* Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49,** 643–650 (2017).
32. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29,** 2669–2677 (2013).
33. Zimin, A. V. *et al.* Hybrid assembly of the large and highly repetitive genome of Aegilops

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* **27,** 787–792 (2017).

34. Pryszcz, L. P. & Gabaldón, T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* **44,** e113 (2016).
35. O'Connell, J. *et al.* NxTrim: optimized trimming of Illumina mate pair reads. *Bioinformatics* **31,** 2035–2037 (2015).
36. English, A. C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7,** e47768 (2012).
37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
38. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3,** 95–98 (2016).
39. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3,** 99–101 (2016).
40. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
41. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32,** 292–294 (2016).
42. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
43. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6,** 31 (2005).
44. Koutsovoulos, G. *et al.* 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 (2016).
45. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. *F1000Res.* **6,** (2017).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

46. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12,** 59–60 (2015).

47. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27,** 722–736 (2017).

48. Bae, J. S., Kim, I., Sohn, H. D. & Jin, B. R. 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 (2004).

49. Nurk, S. *et al.* Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. in *Lecture Notes in Computer Science* 158–170 (2013).

50. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9,** e112963 (2014).

51. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11,** e0163962 (2016).

52. M. Bernt, A. Donath, F. Jühling, F. Externbrink, C. Florentz, G. Fritzsch, J. Pütz, M. Middendorf, P. F. Stadler. MITOS2 WebServer. Available at: http://mitos2.bioinf.uni-leipzig.de.

53. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19,** 1639–1645 (2009).

54. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–652 (2011).

55. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9,** R7 (2008).

56. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).

57. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21,** 1859–1875 (2005).

58. Fallon, T. R. PASA_expression_filter_2017.py. Available at:

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

https://github.com/photocyte/PASA_expression_filter_2017.

59. Haas, B. J. TransDecoder. Available at: https://github.com/TransDecoder/TransDecoder/wiki.
60. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12,** 357–360 (2015).
61. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33,** 290–295 (2015).
62. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7,** 62 (2006).
63. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12,** 491 (2011).
64. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5,** 59 (2004).
65. MAKER Tutorial for GMOD Online Training 2014. Available at: http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial_for_GMOD_Online_Training_2014.
66. Fallon, T. R. maker_gff_to_evm_gff_2017.py. Available at: https://github.com/photocyte/maker_gff_to_evm_gff_2017.
67. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).
68. Rewitz, K. F., O'Connor, M. B. & Gilbert, L. I. Molecular evolution of the insect Halloween family of cytochrome P450s: phylogeny, gene organization and functional conservation. *Insect Biochem. Mol. Biol.* **37,** 741–753 (2007).
69. Helvig, C., Koener, J. F., Unnithan, G. C. & Feyereisen, R. CYP15A1, the cytochrome P450 that catalyzes epoxidation of methyl farnesoate to juvenile hormone III in cockroach corpora

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

allata. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 4024–4029 (2004).

70. Guittard, E. *et al.* CYP18A1, a key enzyme of Drosophila steroid hormone inactivation, is essential for metamorphosis. *Dev. Biol.* **349,** 35–45 (2011).
71. Sezutsu, H., Le Goff, G. & Feyereisen, R. Origins of P450 diversity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368,** 20120428 (2013).
72. Clustal Omega. Available at: https://www.ebi.ac.uk/Tools/msa/clustalo/.
73. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31,** 3406–3415 (2003).
74. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
75. Wheeler, D. L. Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* **31,** 28–33 (2003).
76. Marchler-Bauer, A. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45,** D200–D203 (2017).
77. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx922
78. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44,** D279–85 (2016).
79. Sigrist, C. J. A. *et al.* PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* **3,** 265–274 (2002).
80. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277 (2000).
81. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8,** 785–786 (2011).
82. Käll, L., Krogh, A. & Sonnhammer, E. L. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338,** 1027–1036 (2004).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

83. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).

84. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* **5,** e9490 (2010).

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

86. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190 (2004).

87. Biasini, M. *et al.* SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42,** W252–8 (2014).

88. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).

89. Khomtchouk, B. B., Hennessy, J. R. & Wahlestedt, C. shinyheatmap: Ultra fast low memory heatmap web interface for big data genomics. *PLoS One* **12,** e0176334 (2017).

90. Schultz, M. D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523,** 212–216 (2015).

91. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).

92. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

93. Picard Tools - By Broad Institute. Available at: http://broadinstitute.github.io/picard. (Accessed: 18th December 2017)

94. Larracuente, A. M. & Ferree, P. M. Simple method for fluorescence DNA in situ hybridization to squashed chromosomes. *J. Vis. Exp.* 52288 (2015).

95. Fu, X. H., Ballantyne, L. A. & Lambkin, C. L. Aquatica gen. nov. from mainland China with a description of Aquatica wuhana sp. nov.(Coleoptera: Lampyridae: Luciolinae). *Zootaxa*

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2530, 1–18 (2010).

96. Ohba, N. Mystery of Fireflies. Yokosuka City Mus. Yokosuka, Japan (In Japanese). (2004).
97. Branham, M. A. & Wenzel, J. W. The origin of photic behavior and the evolution of sexual communication in fireflies (Coleoptera: Lampyridae). *Cladistics* **19,** 1–22 (2003).
98. Kanda, S. Firefly. (1935).
99. Ohba, N. Studies on the communication system of Japanese fireflies. *Sci. Rept. Yokosuka City Mus.* **30,** 1–62 (1983).
100. Ohba, N. & Hidaka, T. Reflex bleeding of fireflies and prey-predator relationship. *Science Report of the Yokosuka City Museum* **49,** 1–12 (2002).
101. Fu, X. *et al.* Structure and function of the eversible glands of the aquatic firefly Luciola leii (Coleoptera: Lampyridae). *Chemoecology* **17,** 117–124 (2007).
102. KAWASHIMA & I. A check-list of Japanese fireflies (Coleoptera, Lampyridae and Rhagophthalmidae). *Jpn. J. syst. Ent., Matsuyama* **9,** 241–261 (2003).
103. Conservation of Ecosystem. in *Conservation Biology* (ed. Higuchi, H.) 71–102 (Univ. Tokyo Press, Tokyo, 1996).
104. Environment, J. M. of. Press Release. (2017).
105. Ikeya, H. Melanic strain of Luciola lateralis. *Bull. Firefly Mus. Toyota Town* **8,** 175–177 (2016).
106. Inoue, M. & Yamamoto, H. Cytological studies of family Lampyridae I. Karyotypes of Luciola lateralis and L. cruciata. *La Kromosomo* **II-45,** 1440–1443 (1987).
107. Andrews, S. A quality control tool for high throughput sequence data. *FastQC* Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
108. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
109. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(2010).

110. Slipinski, S. A., Leschen, R. A. B. & Lawrence, J. F. Order Coleoptera Linnaeus, 1758. in *Animal Biodiversity: An Outline of Higher-level Classification and Survey of Taxonomic Richness* (ed. Zhang, Z. –Q) 203–208 (Magnolia Press, Auckland., 2011).

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

112. Costa, C. Systematics and evolution of the tribes Pyrophorini and Heligmini, with description of Campyloxeninae, new subfamily (Coleoptera, Elateridae). *Arq. Zool.* **26,** 49–190 (1975).

113. Harvey, E. N. & Stevens, K. P. THE BRIGHTNESS OF THE LIGHT OF THE WEST INDIAN ELATERID BEETLE, PYROPHORUS. *J. Gen. Physiol.* **12,** 269–272 (1928).

114. Levy, H. C. Greatest bioluminescence. in *Book of Insect Records* (ed. Walker, T. J.) 72–73 (Univ. Florida, Florida., 1998).

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

116. Stibick, J. N. L. Classification of the Elateridae (Coleoptera). *Relationships and classification of the subfamilies and tribes. Pacific Insects* **20,** 145–186 (1979).

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

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

119. Oba, Y. & Sagegami-Oba, R. Phylogeny of Elateridae inferred from molecular analysis. *Nat Insects* **42,** 30–42 (2007).

120. Sagegami-Oba, R., Oba, Y. & Ohira, H. Phylogenetic relationships of click beetles

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Coleoptera: Elateridae) inferred from 28S ribosomal DNA: insights into the evolution of bioluminescence in Elateridae. *Mol. Phylogenet. Evol.* **42,** 410–421 (2007).

121. Kundrata, R. & Bocak, L. 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 (2011).

122. Hyslop, J. A. The phylogeny of the Elateridae based on larval characters. *Ann. Entomol. Soc. Am.* **10,** 241–263 (1917).

123. Ôhira, H. Morphological and taxonomic study on the larvae of Elateridae in Japan (Coleoptera). *H. Ohira, Okazaki City, Japan* **61,** (1962).

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

125. Ôhira, H. 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 (2013).

126. Johnson, P. J. 58. Elateridae Leach 1815. *American beetles* **2,** 160–173 (2002).

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

128. Reyes, N. & Lee, V. Behavioral and morphological observations of Ignelater luminosus in Dominica. (2010).

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

130. McKenna, D. D. & Farrell, B. D. Beetles (Coleoptera). *The timetree of life* **278,** 289 (2009).

131. Grimaldi, D. & Engel, M. S. *Evolution of the Insects*. (Cambridge University Press, 2005).

132. Costa, C. & Vanin, S. A. Coleoptera Larval Fauna Associated with Termite Nests (Isoptera) with Emphasis on the 'Bioluminescent Termite Nests' from Central Brazil. *Psyche* **2010,** (2010).

133. Kretsch, E. Courtship Behavior of Ignelater luminosus. (2000).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

134. Vélez, S. Biogeographic and Genetic Approaches to the Natural History of the Bioluminescent Jamaican Click Beetle, Pyrophorus plagiophthalamus (Coleoptera: Elateridae). (University of Notre Dame, 2006).

135. Virkki, N., Flores, M. & Escudero, J. Structure, orientation, and segregation of the sex trivalent in Pyrophorus luminosus III. (Coleoptera, Elateridae). *Can. J. Genet. Cytol.* **26,** 326–330 (1984).

136. Perez-Gelabert, D. E. *Arthropods of Hispaniola (Dominican Republic and Haiti): A checklist and bibliography*. (Magnolia Press, 2008).

137. Rosa, S. P. New species of Ignelater Costa (Coleoptera, Elateridae, Pyrophorini). *Pap. Avulsos Zool.* **50,** 445–449 (2010).

138. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* **27,** 757–767 (2017).

139. Center for Disease Control and Prevention (CDC). Hymenolepiasis. Available at: https://www.cdc.gov/dpdx/hymenolepiasis/index.html#lifeCycle.

140. Sheiman, I. M., Shkutin, M. F., Terenina, N. B. & Gustafsson, M. K. S. A behavioral study of the beetle Tenebrio molitor infected with cysticercoids of the rat tapeworm Hymenolepis diminuta. *Naturwissenschaften* **93,** 305–308 (2006).

141. Kryukov, K. FASTA Splitter. Available at: http://kirill-kryukov.com/study/tools/fasta-splitter/.

142. Arnoldi, F., Ogoh, K., Ohmiya, Y. & Viviani, V. R. 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 (2007).

143. Nurk, S. *et al.* Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. in *Lecture Notes in Computer Science* 158–170 (2013).

144. Antipov, D. *et al.* plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics* **32,** 3380–3387 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

145. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31,** 3350–3352 (2015).

146. Bae, J. S., Kim, I., Sohn, H. D. & Jin, B. R. 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 (2004).

147. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15,** R46 (2014).

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

149. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31,** 3210–3212 (2015).

150. Cunningham, C. B. *et al.* The Genome and Methylome of a Beetle with Complex Social Behavior, Nicrophorus vespilloides (Coleoptera: Silphidae). *Genome Biol. Evol.* **7,** 3383–3396 (2015).

151. Poelchau, M. *et al.* The i5k Workspace@NAL--enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic Acids Res.* **43,** D714–9 (2015).

152. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16,** 157 (2015).

153. Wood, K. V., de Wet, J. R., Dewji, N. & DeLuca, M. Synthesis of active firefly luciferase by in vitro translation of RNA obtained from adult lanterns. *Biochem. Biophys. Res. Commun.* **124,** 592–596 (1984).

154. De Wet, J. R., Wood, K. V., DeLuca, M., Helinski, D. R. & Subramani, S. Firefly luciferase gene: structure and expression in mammalian cells. *Mol. Cell. Biol.* **7,** 725–737 (1987).

155. Masuda, T., Tatsumi, H. & Nakano, E. Cloning and sequence analysis of cDNA for

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

luciferase of a Japanese firefly, Luciola cruciata. *Gene* **77,** 265–270 (1989).

156. Tatsumi, H., Kajiyama, N. & Nakano, E. Molecular cloning and expression in Escherichia coli of a cDNA clone encoding luciferase of a firefly, Luciola lateralis. *Biochim. Biophys. Acta* **1131,** 161–165 (1992).

157. Ye, L., Buck, L. M., Schaeffer, H. J. & Leach, F. R. Cloning and sequencing of a cDNA for firefly luciferase from Photuris pennsylvanica. *Biochim. Biophys. Acta* **1339,** 39–52 (1997).

158. Oba, Y., Mori, N., Yoshida, M. & Inouye, S. 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 (2010).

159. Oba, Y. *et al.* 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 (2013).

160. Bessho-Uehara, M. & Oba, Y. 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 (2017).

161. Bessho-Uehara, M., Konishi, K. & Oba, Y. 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 (2017).

162. Wood, K. V., Lam, Y. A., Seliger, H. H. & McElroy, W. D. Complementary DNA coding click beetle luciferases can elicit bioluminescence of different colors. *Science* **244,** 700–702 (1989).

163. Viviani, V. R., Bechara, E. J. & Ohmiya, Y. Cloning, sequence analysis, and expression of active Phrixothrix railroad-worms luciferases: relationship between bioluminescence spectra and primary structures. *Biochemistry* **38,** 8271–8279 (1999).

164. Viviani, V. R. *et al.* Cloning and molecular characterization of the cDNA for the Brazilian

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

larval click-beetle Pyrearinus termitilluminans luciferase. *Photochem. Photobiol.* **70,** 254–260 (1999).

165. Ohmiya, Y., Sumiya, M., Viviani, V. R. & Ohba, N. Comparative aspects of a luciferase molecule from the Japanese luminous beetle, Ragophthalmus ohbai. *Sci. Rep. Yokosuka City Mus* **47,** 31–38 (2000).

166. Oba, Y. & Hoffmann, K. H. Insect Bioluminescence in the Post-Molecular Biology Era. *Insect Molecular Biology and Ecology* 94–120 (2014).

167. Timmermans, M. J. T. N. *et al.* Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Res.* **38,** e197 (2010).

168. Timmermans, M. J. T. N. & Vogler, A. P. Phylogenetically informative rearrangements in mitochondrial genomes of Coleoptera, and monophyly of aquatic elateriform beetles (Dryopoidea). *Mol. Phylogenet. Evol.* **63,** 299–304 (2012).

169. Näsvall, J., Sun, L., Roth, J. R. & Andersson, D. I. Real-time evolution of new genes by innovation, amplification, and divergence. *Science* **338,** 384–387 (2012).

170. Oba, Y., Kumazaki, M. & Inouye, S. 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 (2010).

171. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).

172. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **33,** 1870–1874 (2016).

173. Delport, W., Poon, A. F. Y., Frost, S. D. W. & Kosakovsky Pond, S. L. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26,** 2455–2457 (2010).

174. Smith, M. D. *et al.* Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* **32,** 1342–1353 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

176. Nathanson, J. A., Kantham, L. & Hunnicutt, E. J. Isolation and N-terminal amino acid sequence of an octopamine ligand binding protein. *FEBS Lett.* **259,** 117–120 (1989).

177. Georg Neuberger, Sebastian Maurer-Stroh, Birgit Eisenhaber, Andreas Hartig and Frank Eisenhaber. The PTS1 Predictor. Available at: http://mendel.imp.ac.at/mendeljsp/sat/pts1/PTS1predictor.jsp.

178. Letunic, I. & Bork, P. 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 (2016).

179. Briscoe, A. D. & Chittka, L. THEEVOLUTION OFCOLORVISION ININSECTS. *Annu. Rev. Entomol.* **46,** 471–510 (2001).

180. Porter, M. L. *et al.* Shedding new light on opsin evolution. *Proc. Biol. Sci.* **279,** 3–14 (2012).

181. Sander, S. E. & Hall, D. W. Variation in opsin genes correlates with signalling ecology in North American fireflies. *Mol. Ecol.* **24,** 4679–4696 (2015).

182. Martin, G. J., Lord, N. P., Branham, M. A. & Bybee, S. M. Review of the firefly visual system (Coleoptera: Lampyridae) and evolution of the opsin genes underlying color vision. *Org. Divers. Evol.* **15,** 513–526 (2015).

183. Feuda, R., Marlétaz, F., Bentley, M. A. & Holland, P. W. H. Conservation, Duplication, and Divergence of Five Opsin Genes in Insect Evolution. *Genome Biol. Evol.* **8,** 579–587 (2016).

184. McKenna, D. D. *et al.* 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 (2016).

185. Schoville, S. D. *et al.* A model species for agricultural pest genomics: the genome of the Colorado potato beetle, Leptinotarsa decemlineata (Coleoptera: Chrysomelidae). *bioRxiv* 192641 (2017). doi:10.1101/192641

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

186. Sakai, K. *et al.* Drosophila melanogaster rhodopsin Rh7 is a UV-to-visible light sensor with an extraordinarily broad absorption spectrum. *Sci. Rep.* **7,** 7349 (2017).

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

188. Schnitzler, C. E. *et al.* Genomic organization, evolution, and expression of photoprotein and opsin genes in Mnemiopsis leidyi: a new view of ctenophore photocytes. *BMC Biol.* **10,** 107 (2012).

189. Tong, D. *et al.* Evidence for light perception in a bioluminescent organ. *Proceedings of the National Academy of Sciences* **106,** 9836–9841 (2009).

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

191. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30,** 918–920 (2012).

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

193. Böcker, S., Letzel, M. C., Lipták, Z. & Pervukhin, A. SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics* **25,** 218–224 (2009).

194. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).

195. Kyrpides, N. C. *et al.* Genomic Encyclopedia of Type Strains, Phase I: The one thousand microbial genomes (KMG-I) project. *Stand. Genomic Sci.* **9,** 1278–1284 (2014).

196. Williamson, D. L. *et al.* 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 (1990).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

197. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23,** 1026–1028 (2007).

198. Lloyd, J. E. Firefly Parasites and Predators. *Coleopt. Bull.* **27,** 91–106 (1973).

199. Guilligay, D. *et al.* Comparative structural and functional analysis of orthomyxovirus polymerase cap-snatching domains. *PLoS One* **9,** e84973 (2014).

200. Reich, S., Guilligay, D. & Cusack, S. An in vitro fluorescence based study of initiation of RNA synthesis by influenza B polymerase. *Nucleic Acids Res.* **45,** 3353–3368 (2017).

201. King, A. M. Q., Lefkowitz, E., Adams, M. J. & Carstens, E. B. *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*. (Elsevier, 2011).

202. Hsu, M. T., Parvin, J. D., Gupta, S., Krystal, M. & Palese, P. 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 (1987).

203. Pflug, A., Lukarska, M., Resa-Infante, P., Reich, S. & Cusack, S. Structural insights into RNA synthesis by the influenza virus transcription-replication machine. 2017. *Virus Res.* 30782–30781

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

205. Leahy, M. B., Dessens, J. T., Weber, F., Kochs, G. & Nuttall, P. A. The fourth genus in the Orthomyxoviridae: sequence analyses of two Thogoto virus polymerase proteins and comparison with influenza viruses. *Virus Res.* **50,** 215–224 (1997).

206. Kimble, J. B. Zoonotic Transmission of Influenza H9 subtype through Reassortment. (University of Maryland, College Park, 2013).

207. Te Velthuis, A. J. W. & Fodor, E. Influenza virus RNA polymerase: insights into the mechanisms of viral RNA synthesis. *Nat. Rev. Microbiol.* **14,** 479–493 (2016).

208. Hengrung, N. *et al.* Crystal structure of the RNA-dependent RNA polymerase from influenza C virus. *Nature* **527,** 114–117 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

209. Hara, K., Kashiwagi, T., Hamada, N. & Watanabe, H. 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 (2017).

210. Wulan, W. N., Heydet, D., Walker, E. J., Gahan, M. E. & Ghildyal, R. Nucleocytoplasmic transport of nucleocapsid proteins of enveloped RNA viruses. *Front. Microbiol.* **6,** 553 (2015).

211. Sikora, D., Rocheleau, L., Brown, E. G. & Pelchat, M. Influenza A virus cap-snatches host RNAs based on their abundance early after infection. *Virology* **509,** 167–177 (2017).

212. Thompson, W. W. *et al.* Estimates of US influenza-associated deaths made using four different methods. *Influenza Other Respi. Viruses* **3,** 37–49 (2009).

213. Hause, B. M. *et al.* Isolation of a novel swine influenza virus from Oklahoma in 2011 which is distantly related to human influenza C viruses. *PLoS Pathog.* **9,** e1003176 (2013).

214. Anderson, C. R. & Casals, J. Dhori virus, a new agent isolated from Hyalomma dromedarii in India. *Indian J. Med. Res.* **61,** 1416–1420 (1973).

215. Haig, D. A., Woodall, J. P. & Danskin, D. THOGOTO VIRUS: A HITHERTO UNDERSCRIBED AGENT ISOLATED FROM TICKS IN KENYA. *J. Gen. Microbiol.* **38,** 389–394 (1965).

216. Mjaaland, S., Rimstad, E., Falk, K. & Dannevig, B. H. 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 (1997).

217. Presti, R. M. *et al.* Quaranfil, Johnston Atoll, and Lake Chad viruses are novel members of the family Orthomyxoviridae. *J. Virol.* **83,** 11599–11606 (2009).

218. Pauly, M. D., Procario, M. & Lauring, A. S. The mutation rates and mutational bias of influenza A virus. *bioRxiv* 110197 (2017). doi:10.1101/110197

219. Zeng, H. *et al.* Tropism and infectivity of influenza virus, including highly pathogenic avian H5N1 virus, in ferret tracheal differentiated primary epithelial cell cultures. *J. Virol.* **87,**

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2597–2607 (2013).

220. Mansfield, K. G. Viral tropism and the pathogenesis of influenza in the Mammalian host. *Am. J. Pathol.* **171,** 1089 (2007).

221. Steel, J. & Lowen, A. C. Influenza A virus reassortment. *Curr. Top. Microbiol. Immunol.* **385,** 377–401 (2014).

222. Marshall, S. H., Ramírez, R., Labra, A., Carmona, M. & Muñoz, C. 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 (2014).

223. Hall, R. A. *et al.* Commensal Viruses of Mosquitoes: Host Restriction, Transmission, and Interaction with Arboviral Pathogens. *Evol. Bioinform. Online* **12,** 35–44 (2016).

224. Ballinger, M. J., Bruenn, J. A., Hay, J., Czechowski, D. & Taylor, D. J. Discovery and evolution of bunyavirids in arctic phantom midges and ancient bunyavirid-like sequences in insect genomes. *J. Virol.* **88,** 8783–8794 (2014).

225. Metegnier, G. *et al.* Comparative paleovirological analysis of crustaceans identifies multiple widespread viral groups. *Mob. DNA* **6,** 16 (2015).

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

227. Katzourakis, A. & Gifford, R. J. Endogenous viral elements in animal genomes. *PLoS Genet.* **6,** e1001191 (2010).

228. Temin, H. M. Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. *Mol. Biol. Evol.* **2,** 455–468 (1985).

229. Bushman, F. *et al.* Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Microbiol.* **3,** 848–858 (2005).

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

231. Palatini, U. *et al.* Comparative genomics shows that viral integrations are abundant and

bioRxiv preprint doi: https://doi.org/10.1101/237586; this version posted December 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

express piRNAs in the arboviral vectors Aedes aegypti and Aedes albopictus. *BMC Genomics* **18,** 512 (2017).

232. Olson, K. E. & Bonizzoni, M. Nonretroviral integrated RNA viruses in arthropod vectors: an occasional event or something more? *Curr Opin Insect Sci* **22,** 45–53 (2017).

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

234. Goic, B. *et al.* Virus-derived DNA drives mosquito vector tolerance to arboviral infection. *Nat. Commun.* **7,** 12410 (2016).

235. Aaskov, J., Buzacott, K., Thu, H. M., Lowry, K. & Holmes, E. C. Long-term transmission of defective RNA viruses in humans and Aedes mosquitoes. *Science* **311,** 236–238 (2006).

236. Goic, B. *et al.* RNA-mediated interference and reverse transcription control the persistence of RNA viruses in the insect model Drosophila. *Nat. Immunol.* **14,** 396–403 (2013).

237. Miesen, P., Joosten, J. & van Rij, R. P. PIWIs Go Viral: Arbovirus-Derived piRNAs in Vector Mosquitoes. *PLoS Pathog.* **12,** e1006017 (2016).