THE EVOLUTION OF FIREFLY VISUAL PIGMENTS

by

SARAH EMILY SANDER

(Under the Direction of DAVID HALL)

ABSTRACT

Animals use signals in order to elicit reactions from conspecific or heterospecific individuals that will increase their own fitness. Signals involved in mating are perhaps the most important conspecific signals as they directly relate to an organism's fitness. How and why new mating signals arise and how they spread through a population remain open fields of research in evolutionary biology. However it has been difficult to study signaling traits from a genetic perspective since signals, and the receptors that enable their detection, can be complex morphological and/or behavioral traits based on many genes. Fireflies (Coleoptera: Lampyridae) offer an ideal system to study signal evolution because of their conspicuous and highly variable sexual signals. With over 2000 species worldwide, fireflies exhibit lighted signals ranging from simple glows to complex flashes, as well as unlighted long-distance pheromone signals. Aside from differing in pattern, lighted signals also differ in color and range from blue/green to orange. Genes are known that govern light emission color, luciferase, and visual receptor sensitivity, opsins. Here, I conduct foundational research on the variation and molecular evolution of the opsins of North American fireflies between and within species to better understand how signalers and receivers have evolved in this system. I find evidence for only two opsins, one

longwavelength-detecting and one ultraviolet, across 38 firefly species. Both opsins show molecular changes associated with evolutionary transitions from nocturnal to diurnal behavior. In contrast, within one widespread species, *Photinus pyralis*, I find little variation in opsins and luciferase across populations. Finally, in using developing genomic resources to identify opsins and luciferases, I discovered substantial variation in genome size across the family. Investigation into proximate and ultimate causes of genome size variation showed a dynamic repeat landscape and little evidence for selective explanations of genome size evolution. These studies highlight the utility of fireflies as a system to study both the genetics underlying signal evolution and genome evolution.

INDEX WORDS: Lampyridae; opsins; population genetics; molecular evolution; spectral tuning; transposable elements; transcriptome; RADseq; 454 sequencing

THE EVOLUTION OF FIREFLY VISUAL PIGMENTS

by

SARAH EMILY SANDER

B.A., Amherst College, 2006

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2015

THE EVOLUTION OF FIREFLY VISUAL PIGMENTS

by

SARAH EMILY SANDER

Major Professor:	David Hall
Committee:	Sonia Altizer
	Jeremy DeBarry
	Kelly Dyer
	Kathrin Stanger-Hall
	John Wares

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2015

DEDICATION

To the child in all of us with a jar waiting to be filled with lightning bugs.

And to my family.

ACKNOWLEDGEMENTS

I would especially like to thank my committee: Jeremy deBarry, for helping me through repetitive DNA analysis; Sonia Altizer, for introducing me into the wonderful world of insect field ecology; John Wares for pop gen advice and cheerleading; Kelly Dyer as my go to PAML (and manuscript rejection) troubleshooter; Kathrin for getting me started in the firefly system, especially that first field trip, and for helpful lab meetings; and, of course, Dave Hall for BBQ, fieldwork, hashing things out, and taking a chance on me as his first graduate student. I would also like to thank my other firefly colleagues and mentors: Lynn Faust, Raphael de Cock, Sara Lewis, and Zach Marion for their support in the field. Thanks as well to James Lloyd at the University of Florida for inspiring many of these studies.

This work would not have been possible if it weren't for the people and places that helped me with my epic field odysseys, especially: Allegheny National Forest, Gina Baucomb, the Jim Bever Family, the Bieker Family, Tom Brightman (Longwood Gardens), Ashley Brown, Ken and Peggy Butler of the Black Caddis Bed and Breakfast, Cincinnati Center for Field studies (Cincinnati University), the Entomological Society of Pennsylvania, the David Fisk Family, Gordon and Doris Fisk, the Friehauf Family, Great Smoky Mountains National Park, Megan and Blane Hollingsworth, Illinois State Parks, Paul Kisel (Eisenhower State Park, Texas), the Jared Lee Family, Joel Martin (Dixon Center, Auburn University), Mike Marsh (Whitehall Experimental Forest, University of Georgia), Jerry McCollum (Wharton Conservation Center), Dave McNaughton (Fort Indiantown Gap), Mississippi State Parks, The Nature Conservancy, Cheryl Pinzone, Mike Quinn, David Riskind (Texas State Parks), Willem Roosenburg, the David

Sander Family, Kevin Smith (Tyson Research Center, Washington University in St. Louis), Smithfield Farm (Berryville, VA), Tennessee State Parks, Bill and Ann Thorpe, Tonya Saint John, Joan Strassman and David Queller, and Dorset Trapnell.

Accomplishing this work would have been impossible without the support network here at UGA. Thanks to the GACRC staff and the folks at GGF (especially Magdy Alabady, Roger Nilsen, Estefania Olivar, and Jeff Wagner). Special thanks to Travis Glenn and Troy Kieran for 3RAD guidance and to Dave Brown for keeping my computer running, even when I tried my hardest to break it. Thanks to Willie and James for making me feel safer during my late nights in lab. Thanks to Susan, Michelle, and Tina for keeping things running smoothly. Thanks to Darlene for her cheerful "Good Morning!."

Thank you to my ever-supportive family: Mom, Dad, Jake, and Kaylee. Thank you to my Aunt Becky for getting me through my orals. Thanks to my UGA family, especially: Jen (thank goodness you turned out to be the right Jen), Jared, Dave, Milton, Mark, Megan, Emily, Louisa, and the Ecotones. Thanks to the undergraduates who have helped me out, especially Jenna Pallansch, Katharin Korunes, Allison Hall, John Chamberlain, and James Workman.

This work was supported by the National Science Foundation (Graduate Research Fellowship to Sarah E. Sander and Doctoral Dissertation Improvement Grant DEB-1311315 to David W. Hall and Sarah E. Sander) and the National Institute of General Medical Sciences of the National Institute of Health (award number T32GM007103 to Sarah E.Sander). The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health. I would also like to acknowledge the generous support of the Bishop Award (Genetics), an Innovative and Interdisciplinary Research Grant (University of

Georgia Graduate School), and the Interdisciplinary Life Sciences program at the University of Georgia.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Evolutionary biologists are particularly interested in traits involved in speciation, such as those that are involved in reproductive isolation between closely related species. Mating signals are especially interesting because they are an example of a pre-zygotic barrier to reproduction and can be a result of, or lead to, reproductive isolation. Animals exhibit a diverse array of mating signals, from acoustic to chemical to visual signals. How this diversity is generated and maintained is a central question in evolutionary biology and animal behavior.

Signals may experience conflicting selection pressures from a variety of sources. Conspecifics may exert stabilizing selection on signals for species recognition, or select for divergent signals when in contact with another species (reproductive character displacement) (Lofstedt 1993). On the other hand, sexual selection by conspecifics may lead to directional selection, shifting the signaling norm. Selection on signals may also be mediated by environmental factors, outside eavesdroppers, or constrained by evolutionary history or energetic costs (Romer 1993; Ryan & Rand 1993).

Questions in signal evolution remain: given that mating signals may be under stabilizing, disruptive, or directional selection pressures, what are the major factors in how and why have they diversified? Are signalers and receivers coevolving or does previous bias exist? What is the relative contribution of the sources of selection on signal evolution and how does that affect the speciation process? Do new signals arise from mutation or standing variation, single genes or many?

The genetics of signaling have been particularly difficult to disentangle given that signals and signal reception are often complex traits under the control of many genes. However, genes essential to signal production and reception have been identified in some systems. Using fireflies as a system to study signal evolution I ask: how do genes governing reception evolve with respect to signal variation?

Variation in firefly light signals offers a prime opportunity to study signal evolution

Fireflies, in the beetle family Lampyridae, are a diverse, globe-spanning group that includes many species known for their enthralling light displays. These displays may serve several purposes: to discourage predators, to attract mates, or to attract prey (Lloyd 1966). With over 2000 species worldwide, fireflies show extensive variation in light displays, from "dark" fireflies with no adult light production, to species where only adult females produce light, to species in which both sexes produce light. In addition to pair-wise light-signal interactions between individuals, some species gather in large breeding colonies and flash in synchrony (Copeland & Moiseff 2004).

All fireflies are luminous as larvae and larval luminescence is proposed to function as an aposematic signal (DeCock & Matthysen 1999). After at least a year as semi-fossorial larvae predating on worms and snails, fireflies pupate and emerge as adults (McLean *et al.* 1972). Adult fireflies live only a few weeks, during which time they must find a mate and reproduce. It is thought that most species of firefly do not eat as adults, thus most of their active time as adults is spent searching for mates (Lloyd 1997). Exceptions to the rule are predatory *Photuris* fireflies that prey upon other firefly species. These predatory fireflies mimic female flash signals of other species to attract males and eat them (Lloyd 1984).

In the general case, mating consists of flying males that engage in flash dialogues with perched females (Lloyd 1983). Patrolling males produce a species-specific flash pattern of light pulses separated by a longer "flash interval". Females respond to the appropriate male signal based on the flash interval and respond after a species-specific delay (Lloyd 1966). After several rounds of back-and-forth dialogue with a female, the male will land nearby, "scramble" up the vegetation to approach the female, and mating will commence (Vencl & Carlson 1998). Flash patterns are thought to function primarily in species recognition and sex identification (Carlson & Copeland 1985).

Aside from differing in pattern, lighted signals also differ in color, though little attention has been brought to bear on this aspect of signal variation. While color seems to have little to do with male-female recognition (Ohba 2004), there is demonstrated variation in signal color across firefly species as well as anecdotal evidence of more dramatic color variation among genera (Lall *et al*. 1980; personal observation).

Primary hypotheses for divergent signal colors in fireflies include ecological selection, sexual selection, and predation. A correlation study showed that fireflies that were active at dusk had light emissions that were significantly more yellow in color versus later flashers that had green emissions (Lall *et al*. 1980). It is possible that luminescent signals and corresponding visual receptors have coevolved to better detect signals in different ambient light conditions during evening activity periods. While sexual selection is another hypothesis for evolution of signal color, to date there has been no work on signal color and mate preference (Lloyd 1979). Shifts in signal color may also be advantageous in avoiding predation by moving to a signaling channel that is less well detected by predators (Lloyd 1973).

Despite the lack of consensus on why different signal colors have evolved, general predictions can be made about the evolution of signalers and receivers. Because light signals are essential in finding and choosing mates, it is expected that natural selection should favor a greater ability to perceive flashes over distance and identify species-specific messages. Thus, receptors in the eye may be "tuned" to a species' emission color and have evolved to detect light emissions of a species-specific wavelength.

Light signal detection is dependent on opsins

Vision in insects is mediated by light-detecting visual pigments in the retinula cells of the compound eye. These pigments consist of a light-absorbing retinal-based chromophore bound to a G-protein-coupled seven transmembrane receptor protein, opsin (Palczewski *et al.* 2000). Collectively, these two molecules are referred to as rhodopsin and interact with one another to respond to light. While the chromophore is synthesized in a complex biochemical pathway, opsins are coded for by genes of 1 to 18 kilobases (Chou *et al.* 1996; Briscoe 1999; Townson *et al.* 1998; Yokoyama & Yokoyama 2000). The resultant protein is typically 280-380 amino acids in length (Gartner 2000).

Both the chromophore and the opsin show a characteristic peak of absorbance at a particular wavelength of light, or λ_{max}. Previous work in vertebrates and some insects has shown that changes in the amino acid sequence of opsin are correlated with changes in λ_{max}. These changes, at "spectral tuning sites," may correlate with habitat (gobies: Larmuseau *et al.* 2011; cichlids: Terai *et al.* 2006) or the use of UV-pigmentation as a sexual signal (butterflies: Briscoe *et al.* 2010).

Changes in the chromophore can also cause changes in λ_{max}. Most insect species use one of two types of chromophore, either A1, with a λ_{max} of 383 nm, or A3 with a λ_{max} of 379 nm. However, fireflies use both types (Gleadall *et al.* 1989). Since there is little variability in structure and λ_{max} between chromophores, it is thought unlikely that differences between types are responsible for spectral tuning of visual pigments (Briscoe & Chittka 2001). However, few studies have examined the effect of *in vivo* chromophore substitution, especially in insects.

The opsin interacts with the chromophore to extend the λ_{max} of the visual pigment, ranging from 300 to 600 nm. Mutagenesis experiments with bovine rhodopsin have identified 30 sites involved in modifying the spectral sensitivity of the visual pigment (i.e. spectral tuning) in vertebrates (Yokoyama 2008). Ancestral reconstruction of opsin sequence in butterflies suggests only a handful of amino acid sites are responsible for spectral tuning (Briscoe 2001). These sites usually occur in the chromophore binding pocket and single amino acid changes can shift absorption spectra by over 15 nm (Nathans 1990).

In insects, opsins can be divided into 3 spectral classes: UV-opsins that have a λ_{max} between 300 and 400 nm in the ultraviolet range, B-opsins that peak in the blue range from 400 nm to 500 nm, and LW-opsins that absorb at long wavelengths of 480 to 600 nm (Jackowska *et al.* 2007). Ancestral reconstructions suggest that primitive insects had opsins of each spectral class (Briscoe & Chittka 2001). Recent work has revealed independent opsin duplications within several derived insect lineages such as flies, bees, and butterflies, and subsequent functional diversification of duplicates (Spaethe & Briscoe 2004). However, these diversifications are usually associated with the tuning of color vision in diurnal insects. Investigation of the *Tribolium castaneum* genome revealed the presence of only two opsins—one UV and one LW (Jackowska *et al.* 2007). This was hypothesized to be linked with its nocturnal/light-avoiding

behavior. However, other nocturnal insect lineages, such as *Manduca* moths and augochlorine bees, retain greater numbers of opsins (Tierney *et al*. 2011).

Evidence for spectral tuning of opsins

Many vertebrate lineages show evidence of opsin adaptation to dim environments. These adaptations can be in opsin number or sequence. For example, many nocturnal primates have lost their short-wavelength opsin (Ankel-Simons & Rasmussen 2008). Work on cichlids links habitat characteristics, such as turbidity and water depth, to positive selection on opsin tuning sites (Terai *et al.* 2006; Seehausen *et al.* 2008). These characteristics are correlated with ambient light levels and mating signal color. In fact, convergent evolution in spectral tuning sites in vertebrate lineages across environments suggests that selection can act on ~5 nm difference in λ_{max} (Yokoyama *et al.* 2008).

Work on insects, especially butterflies, has revealed links between mating signal color or foraging habits and spectral tuning of opsins (Briscoe & Chittka 2001; Briscoe *et al.* 2010). These studies have been generally limited to diurnal species, probably because most nocturnal insects do not use the visual signaling channel when light levels are low. Fireflies and other bioluminescent beetles provide one of the few examples of nocturnal organisms that rely primarily on visual signaling (Greenfield 2002).

Previous research on spectral tuning of visual receptors in fireflies used electroretinogram techniques to record nerve impulses originating from visual cells in response to exposure to light. These studies show that spectral sensitivity of firefly eyes closely matches the conspecific signal (Eguchi *et al.* 1984; Cronin *et al.* 2000). Using degenerate primers, Oba identified two opsins in the dusk-active Little Asian Firefly*, Luciola cruciata*, one UV and one long-wavelength (Oba &

Kainuma 2009). Though electroretinograms suggest that fireflies also have a blue opsin, Oba was unable to amplify additional opsins from this species. Spectral sensitivity is most likely under the control of LW-opsins, while UV-opsins may be used to detect daylight and determine the onset of flashing (Lall *et al.* 1980).

Here, I characterize the variation of opsins across (**Chapter 2**) and within (**Chapter 4**) firefly species and investigate ultimate explanations for that variation. The central hypothesis is that selection has acted to "tune" opsins to a specific visual sensitivity. The sources of this selection may be signal color and/or environmental light conditions. I predict that at least three opsins will be found across North American fireflies: one LW, one UV, and one B. Since firefly light signals are in the long-wavelength range, I expect amino acid variation in LW opsin to (a) show signatures of selection and (b) correlate with differences in signal color both across and within species. To test these predictions, I use high-throughput sequencing of genomes (four species) and expressed genes (six additional species) to identify all opsins in North American fireflies. I then use these data to amplify opsins from 28 other species, and, within one species, 12 different populations, and examine opsin sequence variation with respect to both species- and population-level variation in signal color and environmental light conditions.

Within species, I also examine variation in luciferase, the gene that codes for the enzyme responsible for light production (**Chapter 4**). Since mutations in the luciferase amino acid sequence cause changes in light color, I expect sequence variation to correlate with differences in signal color across populations. If selection is responsible, then this sequence variation should show greater differentiation among populations with different signal colors than neutrally evolving sequences throughout the genome. To test these predictions, I use reduced-representation high-throughput sequencing to generate neutral loci across the genomes of 154

individuals from 12 populations. I then examine the variation in luciferase with respect to variation at neutral loci and differences in signal color and light environment.

In developing genomic resources for opsin discovery and identification of neutral loci, I identify significant variation in genome size across species. I explore proximate and ultimate explanations for this evolution in **Chapter 3**. I test both selective (physiological-body size, metabolism) and neutral (phylogenetic, effective population size) explanations, as well as investigate the contribution of repetitive DNA to the genomes of species with different genome sizes. I predict that species with larger genomes will have more repetitive DNA. To test this, I perform low coverage genomic sequencing of 21 specimens to identify common repeats that may substantially contribute to genome size.

Significance

There is a paucity of genetic work on signal evolution because signals are frequently complex traits and candidate genes for signal generation and reception are not known (Greenfield 2002). Particularly, there is a lack of information about visual signaling in arthropods. Lighted fireflies offer a prime example of organisms relying primarily on visual sexual signals (Greenfield 2002). Candidate genes governing signal color (luciferase) and signal reception (opsins) are known, though they have not been investigated thoroughly, especially in a natural context.

This work has implications for the broader study of molecular evolution, diversification of gene families, the evolution of sexual signals, and genome evolution.
Establishing the evolutionary history of visual receptors in this signaling system is essential for future study on the genetics of signaling. The knowledge gained from this project informs future

studies on signal evolution: how do new signals arise, persist, and spread through a population? What are primary mechanisms of selection on signals- predation, sexual selection, ecological selection? While answering the above questions is beyond the scope of this work, this project provides a strong foundation on which future studies can build.

References

Ankel-Simons F, Rasmussen DT (2008) Diurnality, nocturnality, and the evolution of primate visual systems. *American Journal of Physical Anthropology*, **137**, 100-117.

Briscoe AD (1999) Intron splice sites of *Papilio glaucus* PglRh3 corroborate insect opsin phylogeny. *Gene*, **230**, 101-109.

Briscoe AD (2001) Functional diversification of lepidopteran opsins following gene duplication. *Molecular Biology and Evolution*, **18**, 2270-2279.

Briscoe AD, Bybee SM, Bernard GD, Yuan F, Sison-Mangus MP, *et al.* (2010) Positive selection of a duplicated UV-sensitive visual pigment coincides with wing pigment evolution in Heliconius butterflies. *Proceedings of the National Academy of Sciences***,** **107**, 3628-3633.

Briscoe AD, Chittka L (2001) The evolution of color vision in insects. *Annual Review of Entomology*, **46**, 471-510.

Carlson AD, Copeland J (1985) Flash communication in fireflies. *The Quarterly Review of Biology*, **60**, 415-436.

Chou WH, Hall KJ, Wilson DB, Wideman CL, Townson SM, *et al.* (1996) Identification of a novel Drosophila opsin reveals specific patterning of the R7 and R8 photoreceptor cells. *Neuron*, **17**, 1101-1115.

Copeland J, Moiseff A (2004) Flash precision at the start of synchrony in *Photuris frontalis*. *Integrative Comparative Biology*, **44**, 259-263.

Cronin TW, Jarvilehto M, Weckstrom M, Lall AB (2000) Tuning of photoreceptor spectral sensitivity in fireflies (Coleoptera: Lampyridae). *Journal of Comparative Physiology A***:** 1-12.

DeCock R, Matthysen E (1999) Aposematism and bioluminescence: experimental evidence from glow-worm larvae (Coleoptera: Lampyridae). *Evolutionary Ecology*, **13**, 619-639.

Eguchi E, Nemoto A, Meyer-Rochow VB, Ohba N (1984) A comparative study of spectral sensitivity curves in three diurnal and eight nocturnal species of Japanese fireflies. *Journal of Insect Physiology*, **30**, 607-612.

Gartner W (2000) Invertebrate visual pigments. In. *Molecular Mechanisms in Visual Transduction*, edited by DG Stavenga, WJ DeGrip and EN Pugh Jr. Amsterdam: Elsevier, pp. 297-388

Gleadall IG, Hariyama T, Tsukahara Y (1989) The visual pigment chromophores in the retina of insect compound eyes, with special reference to the Coleoptera. *Journal of Insect Physiology*, **35**, 787-795.

Greenfield MD (2002) *Signalers and Receivers: Mechanisms and Evolution of Arthropod Communication*. Oxford University Press, New York.

Jackowska M, Bao R, Liu Z, McDonald EC, Cook TA, *et al.* (2007) Genomic and gene regulartory signatures of cryptozoic adaptation: loss of blue sensitive photoreceptors through exapansion of long wavelength-opsin expression in the red flour beetle *Tribolium castaneum*. *Frontiers in Zoology* **4**.

Lall AB, Seliger HH, Biggley WH, Lloyd JE (1980) Ecology of colors of firefly bioluminescence. *Science*, **210**, 560-562.

Larmuseau MHD, Vanhove MPM, Huyse T, Volckaert FAM, Decorte R (2011) Signature of selection on the rhodopsin gene in the marine radiation of American Seven-spined gobies (Gobiidae, Gobiosomatini). *Journal of Evolutionary Biology*.

Lloyd JE (1966) Studies on the flash communication system in Photinus fireflies. In. *Miscellaneous Publications*. Ann Arbor, Michigan: Museum of Zoology, University of Michigan.

Lloyd JE (1973) Firefly parasites and predators. *The Coleopterists Bulletin,* **27**, 91-106.

Lloyd JE (1979) Sexual selection in luminescent beetles. In. *Sexual Selection and Reproductive Competition in Insects*, edited by MS Blum and NA Blum. New York: Academic, pp. 293-342.

Lloyd JE (1983) Bioluminescence and communication in insects. *Annual Review of Entomology*, **28**, 131-160.

Lloyd JE (1984) Occurrence of aggressive mimicry in fireflies. *The Florida Entomologist*, **67**, 368-376.

Lloyd JE (1997) Firefly mating ecology, selection, and evolution. In. *The Evolution of Mating Systems in Insects and Arachnids*, edited by JC Choe and BJ Crespi. Cambridge, UK: Cambridge University Press, pp. 184-192.

Lofstedt C (1993) Moth pheromone genetics and evolution. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, **340**, 167-177.

McLean M, Buck J, Hanson FE (1972) Culture and larval behavior of Photurid fireflies. *American Midland Naturalist*, **87**, 133-145.

Nathans J (1990) Determinants of visual pigment absorbance: role of charged amino acids in the putative transmembrane segments. *Biochemistry*, **29**, 937-942.

Oba Y, Kainuma T (2009) Diel changes in the expression of long-wavelength sensitive and ultraviolet-sensitive opsin genes in the Japanese firefly, *Luciola cruciata*. *Gene*, **436**, 66-70.

Ohba N (2004) Flash communication systems of Japanese fireflies. *Integrative and Comparative Biology*, **44**, 225-233.

Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, *et al.* (2000) Crystal structure of rhodopsin: a G protein-coupled receptor. *Science*, **289**, 739-745.

Romer H (1993) Environmental and biological constraints for the evolution of long-range signalling and hearing in acoustic insects. *Philosophical Transactions: Biological Sciences*, **340**, 179-185.

Ryan MJ, Rand AS (1993) Sexual selection and signal evolution: the ghost of biases past. *Philosophical Transactions: Biological Sciences*, **340**, 187-195.

Seehausen O, Terai Y, Magalhaes IS, Carleton KL, Mrosso HDJ, *et al.* (2008) Speciation through sensory drive in cichlid fish. *Nature*, **455**, 620-626.

Spaethe J, Briscoe AD (2004) Early duplication and functional diversification of the opsin gene family in insects. *Molecular Biology and Evolution*, **21**, 1583-1594.

Terai Y, Seehausen O, Sasaki T, Takahashi K, Mizoiri S, *et al.*n (2006) Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids. *PLoS Biology*, **4**, e433.

Tierney SM, Sanjur O, Grajales GG, Santos LM, Bermingham E, *et al.* (2011) Photic niche invasions: phylogenetic history of the dim-light foraging augochlorine bees (Halictidae). *Proceedings of the Royal Society B: Biological Sciences*, **279**, 794-803.

Townson SM, Chang BSW, Salcedo E, Chadwell LV, Pierce NE, *et al.* (1998) Honeybee blue- and ultraviolet-sensitive opsins: cloning, heterologous expression in Drosophila, and physiological characterization. *The Journal of Neuroscience*, **18**, 2412-2422.

Vencl FV, Carlson AD. (1998) proximate mechanisms of sexual selection in the firefly *Photinus pyralis* (Coleoptera: Lampyridae). *Journal of Insect Behavior*, **11**, 191-207.

Yokoyama S (2008) Evolution of dim-light and color vision pigments. *Annual Review of Genomics and Human Genetics*, **9**, 259-282.

Yokoyama S, Yokoyama R (2000) Comparative molecular biology of visual pigments. In. *Molecular Mechanisms in Visual Transduction*, edited by DG Stavenga, WJ DeGrip and EN Pugh Jr. Amsterdam: Elsevier, pp. 257-296.

CHAPTER 2

VARIATION IN OPSIN GENES CORRELATES WITH SIGNALING ECOLOGY IN NORTH AMERICAN FIREFLIES[1]

[1] Sander, S.E. and D.W. Hall. Accepted by *Molecular Ecology*.
Reprinted here with permission of the publisher.

Abstract

Genes underlying signal reception should evolve to maximize signal detection in a particular environment. In animals, opsins, the protein component of visual pigments, are predicted to evolve according to this expectation. Fireflies are known for their bioluminescent mating signals. The eyes of nocturnal species are expected to maximize detection of conspecific signal colors emitted in the typical low-light environment. This is not expected for species that have transitioned to diurnal activity in bright daytime environments. Here we test the hypothesis that opsin gene sequence plays a role in modifying firefly eye spectral sensitivity. We use genome and transcriptome sequencing in four firefly species, transcriptome sequencing in six additional species, and targeted gene sequencing in 28 other species to identify all opsin genes present in fireflies and to elucidate amino acid sites under positive selection. We also determine whether amino acid substitutions in opsins are linked to evolutionary changes in signal mode, signal color, and light environment. We find only two opsins, one long wavelength and one ultraviolet, in all firefly species and identify 25 candidate sites that may be involved in determining spectral sensitivity. In addition, we find elevated rates of evolution at transitions to diurnal activity, and changes in selective constraint on LW opsin associated with changes in light environment. Our results suggest that changes in eye spectral sensitivity are at least partially due to opsin sequence. Fireflies continue to be a promising system in which to investigate the evolution of signals, receptors, and signaling environments.

Introduction

The diversity of visual signals in nature is a long-standing enigma in evolutionary biology. Natural selection favors signals and receptors that maximize the detection of signals against environmental "noise", leading to the expectation that signals, receptors, and the environments in which signals are displayed will be evolutionarily linked, a process known as "sensory drive" (Endler 1992). One prediction of this framework, as applied to visual signals, is that visual receptors will be "tuned" to best detect signals in a specific light environment. This "spectral tuning" hypothesis can be extended to the genes underlying reception, generating the prediction that genes involved in tuning vision should evolve in response to selection pressure from signals, light environments, or both.

Vision in most animals is mediated by opsin proteins in the eye. Opsins are signaling proteins that contain a conserved lysine residue which binds a vitamin A-derived chromophore that is required for light absorption (Palczewski *et al.* 2000). An opsin with a bound chromophore, collectively termed a visual pigment, maximally absorbs light at a particular wavelength, λ_{max} (Yokoyama 2008). Differences in opsin amino acid sequence and/or chromophore type can change λ_{max} (Yokoyama 2000). Amino acid sites in the opsin at which changes in sequence alter λ_{max} are generally located in regions that interact with the chromophore, termed the chromophore binding pocket (Wilkie *et al.* 2000).

The range of wavelengths that an organism can detect is affected by opsin copy number. Many animal lineages have several visual opsin paralogs that are classified based on the wavelengths of light they detect. For example, the common ancestor of all insects is thought to have contained three visual opsins, one sensitive to long wavelengths (LW; λ_{max} from 600 to 480 nm), one sensitive to blue wavelengths (B; λ_{max} from 480 nm to 400 nm), and one sensitive to

ultraviolet wavelengths (UV; λ_{max} from 400 to 300 nm)(Briscoe & Chittka 2001). Selection on visual spectral sensitivity may affect both the number of opsin paralogs and their peak sensitivities. For example, nocturnal, cave-dwelling, and low-light-living species tend to show reduced selective constraint on or the loss-of-function of one or more opsin copies (e.g. beetles: Jackowska *et al.* 2007; flying squirrels: Carvalho *et al.* 2006).

Fireflies are a diverse, globe-spanning family of beetles that allow for testing of the associations between signals, light environments, and visual receptor evolution (e.g. Biggley *et al.* 1967; Lall, Seliger *et al.* 1980). This family includes many species renowned for their nocturnal lighted mating displays. Generally, flying males flash to stationary females, who respond with flashes of their own (Lloyd 1966). While fireflies are perhaps best known for their species-specific variation in flash pattern (e.g. Lloyd 1966), their flashes also differ in color, with peak emission wavelengths ranging from green (554 nm) to yellow (580 nm) across North American species (Seliger *et al.* 1964; Biggley *et al.* 1967; Lall, Seliger, *et al.* 1980). Nocturnal taxa are also active at different times and in different habitats (Lloyd 1966). These conditions alter the light environment in which signals are displayed (considered in Endler 1993; Théry *et al.* 2008). In addition, there are several diurnal lineages that have independently lost adult light signals and instead use long-distance pheromones to identify and locate mates (Stanger-Hall *et al.* 2007; Stanger-Hall & Lloyd 2015). Thus, fireflies provide the opportunity to examine the evolution of signals and signal reception in response to changes in signal color, light environment, and signal mode (nocturnal, light signals vs. diurnal, no light signals).

Fireflies also vary in visual reception. Previous work provided physiological evidence that firefly vision is tuned to detect conspecific light signals with the recorded peak spectral sensitivity closely matching the peak emission wavelength of the conspecific light signal (Lall,

Chapman, *et al.* 1980; Lall 1981; Lall *et al.* 1982; Eguchi *et al.* 1984; Cronin *et al.* 2000). These data suggested that there are three expressed opsins in fireflies: one LW, one B, and one UV, similar to the hypothesized situation in the ancestral insect. However, this prediction has not been supported to date—only two expressed opsins, one LW and one UV, were found in the dusk-active Little Asian Firefly*, Luciola cruciata* (Oba & Kainuma 2009).

Here we examine whether there is evidence that opsin sequence and copy number contribute to the reported spectral sensitivity of the firefly eye. To accomplish this goal we first elucidate the molecular evolution of visual opsins across 38 species of North American fireflies. We then test for relationships between selective constraint, amino acid sequence, and signaling ecology in a phylogenetic context. We predict that at least three opsins, one LW, one UV, and one B, determine the spectral sensitivity of firefly eyes. Further, we gather evidence for the potential role of opsins in spectral tuning in several ways. (1) We develop a set of candidate functional sites by testing for selection and constraint across the opsin molecule. We expect to find evidence of positive selection at a subset of amino acids that may influence λ_{max}. (2) We examine the location of candidate sites, which we predict will be in positions that change λ_{max}, specifically the chromophore binding pocket. (3) We examine both selective constraint and amino acid sequence variation in relationship to potential sources of selection. Positive selection and amino acid substitutions at candidate sites should correlate with changes in signaling ecology, including ambient light, conspecific light emissions, or both. These patterns should be most apparent between nocturnal and diurnal taxa since these species differ the most in both signal (light signal vs. no light signal) and light environment (night vs. day). To generate our dataset and test our predictions, we used high-throughput RNA and genomic sequencing to

identify putative opsins, and then employed tests of selection and ancestral state reconstruction to investigate their molecular evolutionary history.

Materials and Methods

Specimen and Data collection

Firefly specimens were caught by hand and the date, time, temperature, locality, habitat type, and, when possible, flash pattern were recorded. Light emission spectra from 1-5 individuals per population per species were measured using a portable spectrophotometer (Appendix A, Text S1). Specimens were initially identified to species by a combination of ecological, morphological, and behavioral characteristics (Green 1956, 1957; Lloyd 1966, 1969; Fender 1966; Luk *et al.* 2011). To confirm species identification, genomic DNA was extracted from legs or thorax using a DNeasy Blood and Tissue kit (QIAGEN), and 647 bp of *cytochrome c oxidase I* (COI) were amplified and sequenced (Stanger-Hall *et al.* 2007; Stanger-Hall & Lloyd 2015).

RNA Sequencing

We selected 10 firefly species for RNA sequencing based on taxonomic divergence, divergence in signaling mode and emission wavelength, and the availability of specimens (Figure 2.1; Appendix A, Table S1). Males were used since females of many species are difficult to locate in the field. For each of the 10 species, RNA was isolated from 1-6 heads harvested during the active period and immediately frozen in liquid nitrogen or stored in RNAlater (LifeTechnologies). Bodies of specimens selected for RNA and genomic sequencing, and all other specimens used in PCR amplification, were preserved in 95% ethanol at -80°C until

extraction. RNA was also isolated from the adult light organs of *Photinus pyralis* and *Photinus macdermotti,* and from the larval light organ of *Pn. pyralis* based on specimen availability.

Total RNA from all samples was extracted using Trizol (LifeTechnologies) and treated with DNase prior to library construction. TruSeq RNA libraries (Illumina) were constructed at the Georgia Genomics Facility (Athens, GA) with an average insert size of 150 bp. Samples were individually barcoded and pooled before submission to BGI (Hong Kong) for 100 bp, paired-end sequencing in one lane of Illumina HiSeq2000 v3.

Genomic Sequencing

Sequencing of genomic DNA was performed on a taxonomically diverse subset of species (*Pn. pyralis, Pn. scintillans, Pyractomena borealis,* and *Phausis reticulata*) in order to determine opsin copy number. Genomic DNA was phenol-chloroform extracted (Sambrook *et al.* 1989) from thorax (*Pn. pyralis*, *Py. borealis*) or whole body (*Pn. scintillans*, *Pa. reticulata*) and treated with RNAseA before library preparation. Truseq DNA libraries (Illumina) with an average insert size of 300 bp were constructed at the Georgia Genomics Facility before submission to BGI for sequencing. All four species were individually barcoded, pooled, and sequenced in one lane of Illumina HiSeq2000 v3 100 bp paired-end reads for 10x (*Pn. scintillans*, *Py. borealis*, *Pa. reticulata*) and 25x (*Pn. pyralis*) coverage.

Opsin Identification

Following sequencing, Illumina reads were assessed for quality using FastQC v0.10.1 (Babraham Bioinformatics 2012) and then trimmed, adapters removed, and filtered for quality using the fastqmcf program in the eautils package v1.1.2 (Aronesty 2011; parameters: -m 13, -C 1000000, -x 0.01, -q 20, -w 4). Transcriptomes were assembled *de novo* using the Trinity

pipeline v20121005 with default parameters (Grabherr *et al.* 2011). Candidate opsin transcripts were identified by querying the previously published amino acid sequences of *L. cruciata* opsins (GenBank: LW, AB300328; UV, AB300329; Oba & Kainuma 2009) against the assembled transcriptomes using tBLASTn (Altschul *et al.* 1990; default parameters) and then querying the results against the NCBI nucleotide database using BLASTn (evalue: 1e-06). Transcriptome components with the greatest identity to insect opsins were then aligned in Geneious R6 (Biomatters Ltd 2013) using Muscle (Edgar 2004). Alignments were manually reviewed for sequence similarity and the presence of an open reading frame (ORF) longer than 300 bp (100 amino acids). Expression levels were obtained by aligning trimmed mRNA sequencing reads to the assembled transcriptome components using bowtie (Langmead *et al.* 2009) and quantifying expression with RSEM (Li & Dewey 2011). Putative opsin wavelength sensitivity was inferred from homology to experimentally validated insect LW and UV opsin sequences and confirmed using a neighbor-joining phylogeny (Appendix A, Figure S4).

Finally, to verify the presence of putative opsins and examine opsin copy number variation across the firefly genomes, *L. cruciata* mRNA sequence for UV and LW opsins were queried against the four libraries of genomic DNA sequences using the dc-megablast program for blastn. These sequencing reads were matched to their paired-ends using faSomeRecords (UCSC Genome Browser) and assembled into contigs in Sequencher (Gene Codes Corp. 2011) using the clean data assembly algorithm (minimum match percentage = 85, minimum overlap = 10). Following assembly, each contig was BLASTed against the NCBI nt database using the megablast program for blastn (default parameters) and contigs with top hits to opsin were retained. Contigs were then "walked out" to encompass ~1 kb on either side of the coding

sequence (CDS) using the process described above substituting the ends of assembled contigs as queries.

To corroborate opsin sequences identified bioinformatically, degenerate primers were designed from transcriptomic and genomic sequences and used to amplify LW and UV opsin from genomic DNA. After putative LW and UV opsins were verified in the genomes of the 10 individuals used to obtain transcriptome data, the primers were then used to amplify opsins from the genomic DNA of 28 additional species for which habitat, activity period, and spectra data were available (Figure 2.1; Appendix A, Table S1). Information on primer sequences, PCR cycling conditions, and Sanger sequencing is given in Appendix A, Table S2. Sequences were assembled *de novo*, annotated for intron-exon boundaries using homology to transcriptome sequences, and introns removed to obtain the CDS in Geneious. All CDS were then aligned using Muscle in Geneious and manually reviewed. The final alignment for LW opsin was 1,032 bp in length and represented amino acids 18-378, ending 18 amino acids upstream of the stop codon. The final alignment for UV opsin was 1,137 bp in length, included the start site, and ended 6 amino acids upstream of the stop codon. Amino acids at either end of the opsins are not known to be involved with the chromophore and their exclusion from the final alignments should not affect tests of our hypotheses. Homology models were created in SwissModel (Arnold *et al.* 2006) using squid rhodopsin (*Todarodes pacificus*, PDB: 2z73A; Murakami & Kouyama 2008) as a template (Appendix A, Text S2). Chromophore binding pocket sites were identified by visual inspection of Van der Waals forces in each model. Amino acid sites are numbered in reference to the full-length *Pn. pyralis* amino acid sequences.

Species phylogeny

Analysis of opsin sequences was performed on species tree topologies achieved by extending the three-locus dataset described in Stanger-Hall and Lloyd (2015) to include taxa outside *Photinus*. Briefly, *wingless* (*WG*, 420 bp), *rudimentary* (*CAD*, 594 bp), and *COI* (1272 bp) sequences were obtained from nine additional taxa (Appendix A, Table S3) and tree construction procedures from Stanger-Hall and Lloyd (2015) applied to the extended dataset. This analysis resulted in two slightly different, highly supported tree topologies, depending on taxon sampling. Subsequent analyses of positive selection were performed on both topologies and resulted in similar findings. Figures presented in the main text display the topology consistent with that presented in Stanger-Hall and Lloyd (2015). Details of the phylogenetic methods, comparison of the species phylogeny to a phylogeny generated from the opsin sequences, and the robustness of the phylogeny to the model of sequence evolution, number of taxa, and the method of construction are discussed in the Supporting Information (Appendix A, Text S3).

Identifying positively selected sites

The two tree topologies with branch lengths were used to examine rates of molecular evolution across branches and sites using PAML4 (Yang 2007). *Phausis reticulata*, a North American species shown to be basal to all other North American taxa (Stanger-Hall *et al.* 2007), was used as the outgroup in the opsin analyses since a complete dataset containing all of the loci used in our phylogeny and the opsins was not available from any other beetle taxon. We used the metric ω, the ratio of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site, dN/dS, as a measure of positive selection. An ω value that is

less than, equal to, or greater than 1 is indicative of purifying selection, no selection (neutrality), or positive selection, respectively. The branch models tested included M0 (one ω over all branches and sites), two-rate (one ω for branches with transitions to diurnal and one for branches without transitions), and a free-ratio model (each branch has its own ω). The site models tested were: M1a (neutral) with a class of sites evolving under purifying selection and a class of sites evolving neutrally; M2a (selection) with 3 classes of sites, one under purifying selection, one evolving neutrally, and one evolving under positive selection; M3 (discrete) with 3 classes of sites, similar to M2a, except that the sites classes are constrained to have successively greater ω values rather than a specific ω; M7 (beta) with 10 classes of sites evolving with $0 < \omega < 1$, sampled according to a beta distribution; and M8 (beta), similar to the M7 model, with the addition of a class of sites with ω greater than 1. Nested models were compared using likelihood ratio tests (LRTs). In models that included a positively selected class of sites, Bayes Empirical Bayes analysis (Yang *et al.* 2005) was performed to identify sites under selection.

Fitmodel v20140407 (Guindon *et al.* 2004) was used to evaluate sites under positive selection along branches without defining a branch of interest *a priori*. Fitmodel accomplishes this by allowing sites to switch between ω classes. The M2a and M3 selection models can then be tested while incorporating different switching models. Equal switching models estimate a single switching rate among the ω classes, while biased switching models estimate a separate switching rate between each pair of classes. Fitmodel has been shown to outperform standard methods for detecting selection at sites along branches, especially when the foreground branches defined in a standard analyses do not represent the branches along which there has been selection (Lu & Guindon 2014).

Testing for correlations with signal and ecological traits

We tested for correlations between measures of opsin evolution for each lineage and four explanatory signal and ecological traits: 1) signaling mode (nocturnal/light or diurnal/no light), 2) spectra (mean male peak emission wavelength), 3) habitat (open, mixed, or closed) based on the amount of canopy cover in the signaling environment, and 4) activity start time (early or late) based on when the first individual was observed signaling relative to sunset (Lall, Seliger, *et al.* 1980). All 38 North American taxa from which we were able to obtain LW and UV opsin sequences were used to test signal mode. Only nocturnal taxa were used to test emission spectra, habitat, and activity time because the spectral tuning hypothesis for fireflies is predicted for light-signaling taxa only and adults of diurnal species do not produce light. Data for signal and ecological traits was gathered from both the literature (Seliger *et al.* 1964; Biggley *et al.* 1967; Otte & Smiley 1977; Cicero 1983; Stanger-Hall & Lloyd 2015) and field measurements. Where possible, the spectra, habitat, and activity time values for the population where the specimen(s) used for opsin sequencing was caught were used in the final dataset (Text S1). Investigation of spectra was limited to the 28 species for which we were able to obtain emission data. Male average peak emission wavelengths were used because female data were not available in either the literature or in our collected dataset across a sufficient number of species.

Signaling and ecological traits were tested for phylogenetic signal using Blomberg's K (Blomberg *et al.* 2003) in the picante package in R (Kembel *et al.* 2010; Appendix A, Table S4). All traits except for signaling mode showed significant or nearly significant phylogenetic signal; therefore, analyses were performed on values calculated using phylogenetic independent contrasts (Felsenstein 1985) using the ape package (Paradis *et al.* 2004) in R. In contrast to the other traits, nocturnal/diurnal activity was treated as a discrete variable, reconstructed using

maximum parsimony, and measures of opsin evolution compared between branches with different signal modes.

The measures of opsin evolution examined included: number of amino acid substitutions, number of parallel (same amino acid to same amino acid) and convergent (different amino acid to same amino acid) substitutions, dN, whole protein ω values, and site-specific ω values. The number and types (parallel or convergent) of amino acid substitutions between nodes was determined using ancestral sequence reconstruction. Amino acid alignments were first tested for an appropriate model of evolution using ProtTest v3.3 (best models, LW: LG; UV: JTT; Darriba *et al.* 2011), then ancestral states at each site were reconstructed and visualized using the phangorn package in R (Schliep 2011; Kenaley *et al.* 2014). Absolute numbers and types of substitutions were parsed from phangorn output using a custom R script. Distributions for numbers of parallel and convergent amino acid substitutions occurring in nocturnal lineages were generated by bootstrapping the parsed dataset in R for nocturnal branches only (1000 reps). Estimates of selective constraint along branches and at sites along branches were taken from PAML free-ratio and fitmodel M2aS1 (LW) or M3S2 (UV) model output, respectively. These models were shown to be the best fit for the data using LRTs.

We explored the relationships between several response and explanatory variables using partial correlations to control for divergence (branch length) in SPSS v22 (IBM Corp. 2013). The response variables examined were the number of amino acid substitutions along lineages, difference in dN, cumulative nonsynonymous substitution rate, and difference in selective constraint (ω) between taxa for each opsin. The explanatory variables were change in peak emission wavelength, habitat, and activity time. Before analysis, variables were assessed for normality in JMP Pro 10 (SAS Institute Inc.). Change in peak emission wavelength was normally

distributed, while all other variables did not conform to normality even after transformation. Accordingly, we proceeded with the analysis under the assumption that partial correlation analysis is robust to departures from normality (Voortman & Druzdzel 2008).

Results

Transcriptome and genome sequencing reveal two opsin genes

RNAseq of heads from 10 species yielded 172 million reads totaling 34.5 Gb (Table 2.1). Several transcripts showed homology to known insect opsins using BLAST. However, only two transcripts, one with homology to LW opsin and one with homology to UV opsin, had additional evidence to suggest that they encode *bona fide* opsin proteins: both had complete open reading frames (ORFs) and could be recapitulated from whole-genome shotgun Illumina reads. Comparing opsins assembled from genomic reads and the putative opsin transcripts revealed that firefly LW opsin genes are 1247-1935 bp in length (including introns), encode 344-381 amino acids, and contain 5 exons (Appendix A, Figure S1a). The firefly UV opsins are 1404-1525 bp in length, encode 385-386 amino acids, and have 6 exons (Appendix A, Figure S1b). The lengths of both opsins are within the range of other sequenced insect opsins. Both candidate opsins also have the structural characteristics of other described opsins, including 7 transmembrane domains and the conserved lysine at K324 (LW) and K323 (UV) where the chromophore is bound (Appendix A, Figure S2). In addition, the full-length LW and UV opsin candidates are highly expressed in heads; in all 10 species they are in the top 5% of expressed genes (e.g. Appendix A, Figure S3). In contrast, opsins are expressed at 10,000-fold lower levels in adult and larval light organs compared to heads in the two species for which we have data, *Pn. pyralis* and *Pn. macdermotti* (Figure 2.2). No other putative visual opsin transcripts identified in the RNAseq

data could be reconstructed from genomic reads. However, we did detect a distantly related c-opsin in several species (data not shown). Arthropod c-opsins diverged from arthropod visual opsins in an ancient split and are thought to function in circadian rhythm regulation (Arendt 2003; Porter *et al.* 2012). In species where we did not detect full-length ORFs of c-opsin, we were able to detect members of the extended G protein-coupled receptor gene family, indicating that our analysis would have detected additional visual opsins if they were present. A neighbor-joining phylogeny of insect opsin sequences showed that the putative LW and UV opsins fall in clades with LW and UV, but not B, opsins from other insect species (Appendix A, Figure S4).

LW and UV opsin survey in fireflies

Using transcriptome and genome sequences, we designed degenerate primers and used them to amplify and sequence genomic LW and UV opsins from a total of 38 species representing a diversity of ecological and signaling traits. Across all taxa in our dataset, within both LW and UV opsins, there was 86-94% identity at the amino acid level with the previously published *L. cruciata* opsin sequences, and ~70 % identity with other beetle opsins. There was no evidence for recent gene duplications of LW or UV opsin in any of the lineages examined: coverage of opsins in genomic sequences was within that expected for single-copy genes, there was no evidence of multiple flanking regions in either the transcriptome or genomic datasets, and double peaks in Sanger-sequence chromatograms were consistent with heterozygosity rather than duplication.

LW and UV opsins show evidence of positive selection at specific sites

Using PAML (Yang 2007), we found evidence of strong purifying selection ($\omega < 1$) over the entire gene for both LW and UV opsins (M0, one ratio model; LW: $\omega = 0.07$, UV: $\omega = 0.05$; Table 2.2). In both cases, models that included a small proportion of sites under positive selection over the entire phylogeny (M2a, M8) were a better fit than the M0 model (Table 2.2). Bayes Empirical Bayes analysis under the M8 model identified seven sites with $\omega > 1$ in LW opsin (Table 2.3); of these, the ω values were statistically greater than neutrality ($\omega = 1$) for three sites, 181, 235, and 188, indicating positive selection. In UV opsin, Bayes Empirical Bayes analysis under the M8 model identified seven sites with $\omega > 1$, but none were statistically greater than 1.

In order to examine selection on sites across branches of the phylogeny, we employed a fitmodel analysis (Guindon *et al.* 2004). The best models for LW opsin based on LRTs were M2aS1 (selection model with equal switching; $\omega_0 = 0.00$, $\omega_1 = 1$, $\omega_2 = 2.75$), and M3S2 (discrete model with biased switching; $\omega_0 = 0.000$, $\omega_1 = 0.003$, $\omega_2 = 9.79$; Table 2.4), with one exception based on the specific tree topology that was used (Appendix A, Text S3). The four sites detected by the M2aS1 models included all three of the positively selected sites identified in the PAML site-model analysis, though the posterior probabilities did not exceed 0.95. The M3S2 model identified most of the 104 variable sites in LW as under positive selection on at least one branch of the phylogeny. However, this model was too complex for our data because the posterior probabilities for sites being in the positively selected class were below the recommended cut-off of 0.9 (maximum: 0.84), the estimates of ω_0 and ω_1 were nearly equal, and the switching parameter was zero. Thus, we further investigated only the eight sites with evidence for $\omega > 1$ across the PAML M2a, PAML M8, and fitmodel M2aS1 analyses (Table 2.3).

For UV opsin, fitmodel supported models M2aS1 ($\omega_0 = 0.01$, $\omega_1 = 0.3$, $\omega_2 = 1$) and M3S2 (discrete with biased switching; $\omega_0 = 0.01$, $\omega_1 = 0.5$, $\omega_2 = 20$; Table 2.4), again with one exception based on the specific topology used (Appendix A, Text S3). Even though M2aS1 is a selection model, it did not estimate any sites with $\omega > 1$. In contrast, the M3S2 model identified one site, 133, as being in the ω_2 class ($\omega = 20$) on three of the 73 branches in the phylogeny. Though the posterior probabilities of site 133 being in the ω_2 class on these branches were low (range: 0.52-0.58), investigation of nonsynonymous and synonymous nucleotide changes showed that this site had multiple nonsynonymous hits in these lineages. Consequently, we considered this site, along with the seven sites identified in the PAML analysis, in further analysis (Table 2.3).

Most positively selected sites are outside the chromophore binding pocket

We used homology modeling to estimate the tertiary structure of firefly opsins and identify sites in the chromophore binding pocket. To do this, we compared the sequence of firefly LW and UV opsins separately to Japanese Common Squid (*T. pacificus*) opsin, the closest species with a known protein tertiary structure (Murakami & Kouyama 2008). In total, we identified 24 sites in each opsin protein that are likely to interact with or have potential long-range effects on the chromophore. Twenty-two of the 24 binding pocket sites identified in LW opsin were invariant across the 38 species (Appendix A, Table S5). None of the eight sites with $\omega > 1$ in LW opsins were identified as a binding pocket site (Figure 2.3). In UV opsin, 19 of 24 binding pocket sites were invariant and two of the eight positively selected sites, 133 and 299, were identified as binding pocket sites (Figure 2.3, Appendix A, Table S5).

Amino acid changes are linked to signaling mode

We used ancestral state reconstruction to identify branches where evolutionary transitions between signaling modes (nocturnal, light/diurnal, no light) occurred. Maximum parsimony reconstruction supported a model with a nocturnal ancestor and four independent transitions to diurnal activity (Figure 2.4). PAML branch analysis supported models with elevated whole-molecule ω along branches shifting to diurnal activity relative to branches remaining nocturnal in both LW and UV opsins (Table 2.2; LW: 0.10 vs 0.07; UV: 0.09 vs 0.05) indicating positive or relaxed purifying selection. In LW opsin, all four sites identified in fitmodel were under positive selection along at least one of the branches with a transition from nocturnal to diurnal activity. However, these sites were also positively selected along some nocturnal branches. In UV opsin, the single site identified by fitmodel was positively selected along the branch leading to the most recent common ancestor (MRCA) of *Ellychnia*, a diurnal clade of fireflies, as well as on the two branches within the group.

To better understand the relationship between signal mode and amino acid substitutions at sites of interest (positively selected, binding pocket, and other variable sites), we reconstructed ancestral amino acid sequences across the phylogeny using maximum likelihood. Significantly more changes occurred on diurnal branches than expected based on their phylogenetic representation (Fisher's exact text (one-tailed); LW: diurnal: 53 changes on 6 branches, nocturnal: 177 changes on 68 branches, p=0.0024; UV: diurnal: 67 changes, nocturnal: 137 changes, p = 0.0001). In addition, there were significantly greater numbers of parallel and convergent amino acid substitutions between pairs of diurnal lineages than pairs of nocturnal lineages for LW opsin (parallel: diurnal: 1.6 changes, nocturnal: 0.35 changes, p=0.004; convergent: diurnal: 0.13 changes, nocturnal: 0.006 changes, p=0.015), but not for UV opsin

(parallel: diurnal: 0.47 changes, nocturnal: 0.04 changes, p=0.34; convergent: diurnal: 0 changes, nocturnal: 0.001 changes).

LW opsin: One parallel (A248V) and one convergent (S/T309A) change were noted at two of the eight positively selected sites in LW opsin in diurnal lineages. The number of changes was significantly higher in diurnal than nocturnal lineages for convergent changes (diurnal: 0.067 changes, nocturnal: 0.004 changes, p=0.046), but not for parallel changes (diurnal: 0.067 changes, nocturnal: 0.012 changes, p=0.14). However, the specific amino acid changes that occurred were not unique to diurnal lineages. There were no parallel or convergent amino acid changes in the LW opsin binding pocket; however, two other sites had amino acid changes exclusive to diurnal lineages (L59I in *Ld. atra* and *El. corrusca*; T/I301V in *Ld. atra* and MRCA Ellychnia).

UV opsin: Parallel changes in the diurnal lineages occurred at two of the eight positively selected sites in UV opsin (V59L and V218A), significantly higher than the number of parallel changes on nocturnal branches (parallel: diurnal: 0.13, nocturnal: 0.02, p=0.029), but the specific amino acid changes that occurred were not unique to diurnal lineages. There was one parallel change in the UV binding pocket that was unique to diurnal lineages (F128Y), but this was not significantly different from the number of parallel changes that occurred in nocturnal lineages. There were no convergent changes at positively selected or binding pocket sites in UV opsin on diurnal lineages; however two binding pocket sites showed single amino acid changes exclusive to diurnal lineages (G129A *El. bivulneris*, S203T *Ld. atra*). Two additional sites showed parallel changes unique to diurnal lineages (I68V and D214N in *Ld. atra* and *El. corrusca*).

In summary, the transition from nocturnal to diurnal activity is associated with an increase whole-molecule ω in both LW and UV opsins. A total of 25 positively selected, binding

pocket, parallel, or convergent amino acid substitutions are excellent candidates for functional studies. Descriptions of specific candidate sites and figures showing ancestral sequence reconstructions are given in the Supporting Information (Appendix A, Text S4; Figures: S5-S7 (LW) and S8-S10 (UV)).

Changes in selective constraint linked to transitions in ecological and signaling traits

For each of the nocturnal species we used data on male peak emission wavelength, habitat in which signals are displayed, and activity start time to examine whether molecular evolution of opsins is correlated with changes in signal color or light environment. We found a significant positive correlation between the change in selective constraint (ω) in LW opsin and change in activity time along branches after accounting for divergence between taxa (pr = 0.54, p=0.01). This relationship was not statistically significant for UV opsin (Table 2.5). No significant correlations were found between response variables (number of amino acid substitutions, dN, whole molecule ω values, site-specific ω values) and signal emission color or habitat.

Discussion

The visual opsins of North American fireflies

Vision in North American fireflies is due to two opsins, one LW and one UV opsin. It is surprising that we did not detect a B opsin among the 10 transcriptomes and 4 genomes sequenced, since this opsin was present in the common ancestor of insects and previous studies on firefly eye sensitivity suggest the presence of blue-sensitive photoreceptors (Lall, Chapman, *et al.* 1980; Lall 1981; Lall *et al.* 1982; Eguchi *et al.* 1984; Lall *et al.* 1988; Booth *et al.* 2004).

Similarly, soldier beetles in the family Cantharidae, a sister family to Lampyridae, also exhibit sensitivity to blue wavelengths (Horridge 1979).

We are confident that we would have detected a B opsin if present given our ability to detect c-opsins as well as other divergent gene family transcripts. It is possible that we did not detect B opsins because we sequenced only species that lack them, because they are exclusively expressed at times that we did not sample, or because they are expressed at very low levels. However, we are confident that none of these possibilities apply for the following reasons: First, we sequenced the transcriptome of *Photuris frontalis*, the species with the strongest evidence for blue-sensitivity (Lall *et al.* 1988) and found no B opsin. Second, we found no evidence for any other visual opsins, including a B opsin, in the four genomic data sets. Third, the sequencing depth of the head tissue transcriptomes was sufficient to identify very rare transcripts. For example, we identified luciferase transcripts, the light-producing enzyme putatively expressed only in the light organ, at a level that was three orders of magnitude lower than the LW and UV opsins. Fourth, a recently-published study did not find a B opsin in the transcriptomes of nine firefly species (Martin *et al.* 2015). Based on our data, we conclude that fireflies have lost the hypothesized ancestral insect B opsin paralog.

In other beetle species there is evidence for both the presence (Scarabidae: Théry *et al.* 2008; Curculionidae: Groberman & Borden 1982; Coccinelidae: Lin 1993) and absence (Elateridae: Lall *et al.* 2000; Lall *et al.* 2010; Tenebrionidae: Yinon 1970) of blue sensitivity. In some cases, loss of blue sensitivity indicates a loss of B opsin (e.g. *Tribolium castaneum*: Jackowska *et al*. 2007); however, it is unknown to what extent loss and retention of blue sensitivity is coupled to loss or retention of the B opsin across beetle taxa. Without a B opsin, sensitivity to blue wavelengths in *Photuris* and other species may be explained by additional

photosensitizing pigments that extend the range of wavelengths either LW or UV opsin are able to detect (Lall *et al.* 1982), similar to "antenna" pigments in deep-sea dragonfish (Douglas *et al.* 2000). Direct measures of λ_{max} are needed to confirm the phylogenetically inferred spectral absorbances of each of the firefly opsins.

The functions of LW and UV opsins

The presence of only two opsins in fireflies and their inferred absorbances (LW and UV) has implications for possible sources of selection on their amino acid sequences. Vision in adult fireflies is likely used for the detection of conspecific and heterospecific light signals, avoidance of obstacles during locomotion, and timing the onset of activity. In contrast, vision probably plays a minor role for larvae, which are generally active at night and are often below leaves or in the soil. All firefly larvae are bioluminescent, but they appear to use their light an aposematic signal to predators rather than for communication with conspecifics (McLean *et al.* 1972). The evolution of visual pigments is thus likely to be primarily driven by selection in adults.

Unlike in some other insect lineages (e.g. Briscoe & Chittka 2001), there has been no lineage-specific diversification of either LW or UV opsins in fireflies. Since there is no known UV component of firefly light emissions (Eguchi *et al.* 1984), these results are concordant with monochromatic detection of firefly light signals using the LW opsin (Lall *et al.* 2000; Lall & Worthy 2000). UV opsin may aid in navigation during flight, especially in species active at twilight in open habitats because UV and blue wavelengths are enriched in these conditions (Cronin *et al.* 2000), or it may be important for detecting polarized light (Dacke *et al.* 2004). UV opsin may also be involved in detecting the threshold of ambient light that cues initiation of evening flashing activity (Lall 1993, 1994).

Spectral tuning and the molecular evolution of LW and UV opsins

We examined the evolution of LW and UV opsins by testing for positive selection within each opsin across 38 North American species. To investigate the spectral tuning hypothesis, we further determined whether positively selected sites in each opsin occurred in locations that are functionally relevant to tuning (the chromophore binding pocket). Visual opsins generally exhibit purifying selection across the entire molecule suggesting functional constraints (e.g. Terai *et al.* 2006; Briscoe *et al.* 2010; Shen *et al.* 2010; Sivasundar & Palumbi 2010; Audzijonyte *et al.* 2012; Weadick & Chang 2012; Meredith *et al.* 2013; Kenaley *et al.* 2014). For both LW and UV opsins the entire molecule showed evidence of selective constraint. However, 16 sites across both opsins exhibited $\omega > 1$ in at least one of our tests of selection. Three of these sites (LW: 181, 188, and 235) were significantly elevated, and seven others were either in or near the chromophore binding pocket and/or have been identified in other insect studies (Table 2.3). Contrary to our expectations, only two of these 16 sites were predicted to alter λ_{max} based on their location in the binding pocket in the homology model. The low number of λ_{max}–altering sites was not caused by fewer identified binding pocket sites; we predicted 24 sites, similar to other insect studies (Wakakuwa *et al.* 2010).

The lack of overlap between positively selected sites and binding pocket sites suggests that either (a) the homology model is inaccurate or, if the homology model is correct, that (b) sites may be under selection for spectral tuning through long-range effects or (c) selection targets other opsin functions. Z-scores indicated that our homology model was under-performing relative to models of other genes generated in SwissModel, likely due to low identity between firefly opsins and the most similar available template, squid opsin (Appendix A, Text S2). However, six of the positively selected sites that we identified are common to studies of selection

and spectral tuning in other insect opsins (Briscoe 2001; Briscoe *et al.* 2010; Wakakuwa *et al.* 2010; Tierney *et al.* 2011). The presence of elevated rates of evolution at these sites across butterflies, bees, and fireflies suggests that, even if our homology model is inaccurate, there may be similar selective pressures at work across these diverse lineages. Besides spectral tuning of λ_{max}, it is possible that selection for other functions includes breadth of sensitivity (Lall, Seliger, *et al.* 1980), thermal stability (Endler 1992, Frentiu *et al.* 2015), protein folding, chromophore uptake, or interactions with downstream molecules, some of which may affect λ_{max} (Mackin *et al.* 2014; Schott *et al.* 2014). For example, site 356 in LW opsin, one of the positively selected sites, is likely not involved in spectral tuning since it is located in helix 8, parallel to the membrane, and instead probably interacts with the downstream G protein.

These functional hypotheses require a better crystal structure for homology modeling, preferably from a firefly species, to corroborate the position of amino acid sites relative to the chromophore and assess the potential effects of identified mutations on opsin structure and function. These will then need to be functionally verified by empirically measuring the effects of amino acid substitutions at the sites identified in this study.

The influence of light signaling on opsin evolution

If the predictions of sensory drive, specifically tuning of genes underlying visual reception, apply to fireflies, we expect to see changes at amino acid sites in opsins associated with adult light signaling, especially at sites predicted to alter λ_{max}. If these changes are driven purely by selection for conspecific signal detection, we expect to find these patterns specifically in LW opsin.

We find several lines of evidence supporting an association between changes in opsin sequence and changes in signal mode. (1) ω values were higher, indicating faster evolution, in lineages of diurnal species relative to nocturnal species. Since ω for both LW and UV opsin in diurnal lineages is less than 1, this could be due to either positive or relaxed purifying selection (Bielawski & Yang 2003). Surprisingly, UV opsin and LW opsin showed similar patterns, suggesting that both opsins have important functions to fireflies transitioning to a diurnal lifestyle. (2) There were more parallel and convergent amino acid changes among diurnal lineages than nocturnal lineages in LW opsins, though not in UV opsins. This was primarily driven by comparisons between *Ld. atra* and *Ellychnia* lineages. The lack of amino acid substitutions in the other two diurnal taxa can be explained by recent divergence (branch lengths) from their nocturnal sister taxa (Stanger-Hall & Lloyd 2015). (3) A binding pocket site in UV opsin, site 133, was under positive selection in diurnal *Ellychnia* lineages. The S133A substitution is of particular interest because it is a polar to nonpolar change and such changes at binding pocket sites are known to affect λ_{max} (Briscoe 2008). In addition, site-directed mutagenesis found that a S133A substitution blue shifted absorbance in a butterfly B opsin (Wakakuwa *et al.* 2010).

While this evidence supports that some variable sites in LW and at least one binding pocket site in UV opsin are involved in spectral tuning associated with a change in signal mode (i.e. the transition to diurnal, unlighted, pheromone signals), the other sites identified in LW and UV opsin may or may not be involved in spectral tuning. The fact that we did not find a correlation for UV opsin suggests that sources of selection may be different for LW vs UV opsin, as expected based on their presumed different functions. Activity time rather than habitat may best reflect the quality and amount of light available in the environment and future work should

seek to quantify light environments in both spectrum and intensity to examine the relationship with LW opsin more closely.

Contrary to our expectations, we did not find evidence of spectral tuning of LW opsin to male emission wavelength. There was a positive correlation between number of amino acid changes in LW opsin and change in emission spectra; however, both variables were correlated with branch length and the relationship became nonsignificant when branch length was taken into account. Interestingly, specific substitutions at two of the candidate sites are associated with shifts in opsin absorbance in other taxa that are in the same direction as the emission spectrum shifts identified in the present study (Appendix A, Table S6). All three nocturnal lineages that had T108M substitutions were red-shifted in emission relative to their reconstructed ancestor. This site, analogous to V63M in Heliconius, is associated with a red shift in opsin absorbance in butterflies (Briscoe 2001). In addition, V188I, associated with a red shift in emission in fireflies, is analogous to V143I in Heliconius, also associated with a red shift in opsin absorbance. Neither of these sites are identified as in the binding pocket based on homology modeling, yet are associated with shifts in opsin absorbance.

There are several possible explanations for our overall negative finding while individual substitutions have some evidence for functional effects on absorbance. First, there may not be enough large phylogenetically-independent transitions in spectra to give us sufficient power to detect a relationship. Sampling more taxa would help to alleviate this problem. Second, the presence of selective constraint across the entire molecule may make single amino acid substitutions difficult to detect as driven by selection. Third, the homology model may not identify all the amino acid sites that are candidates for changing λ_{max}, or it may be combinations of changes at different sites that tune opsins (Yokoyama *et al.* 2008). Finally, a lack of

concordance between opsin amino acid sequence, spectral sensitivity, and light environment has been observed in other species (e.g. Mysis: Audzijonyte *et al*. 2012), suggesting that the paradigm of opsin spectral tuning does not hold universally and that other mechanisms should be examined.

Other mechanisms of spectral tuning

Our data suggest that changes in opsin sequence, some with potential to affect λ_{max}, occur at large transitions in signaling characteristics, especially the transitions from nocturnal lighted signaling to diurnal pheromone signaling. However, they do not explain the observed tight correlation between firefly visual sensitivities and male emission wavelengths. Fine-tuning of spectral sensitivity may instead be due to other molecules that interact with the opsins or with photons of light.

One candidate molecule that may be involved is the chromophore. While most insect lineages use a single chromophore in their visual pigment, Asian firefly species possess both A1 and A3 chromophore types (Gleadall *et al.* 1989). These two types are similar in chemical structure; however, alternate chromophores are known to change the spectral absorbance of their paired opsin (Briscoe & Chittka 2001). The difference in λ_{max} with different chromophores is typically large (35-40 nm), much greater than the difference in LW spectral sensitivity observed across all firefly species (26 nm), suggesting that the use of alternate chromophores may not explain fine-tuning. It is also possible that the observed blue sensitivity without a B opsin may be caused by the use of different chromophores.

Alternatively, fine-tuning may involve screening pigments that absorb specific wavelengths, thus modifying the spectrum of light that reaches the opsins (Seliger *et al.* 1982a,

b; Cronin *et al*. 2000). Our data give ambivalent support for predictions of LW opsin spectral absorbance based on this hypothesis (Appendix A, Text S5). Nevertheless, screening pigments have been found in 12 North American firefly species, with the exception of *Ld. atra*, the only diurnal species examined to date (Seliger *et al.* 1982b; Lall *et al*. 1988; Cronin *et al.* 2000). Screening pigments may be less constrained than opsins, allowing rapid evolution of eye sensitivity. It is also possible that screening pigments may be responsible for blue sensitivity. If screening pigments in the rhabdom mask light in the UV spectrum they could effectively convert UV receptors into blue receptors, much like the "sunscreen" pigments in mantis shrimp (Bok, et al. 2014). Future work will develop the link between opsin sequence and absorbance, chromophore usage, and screening pigments in relation to both blue sensitivity and the spectral tuning hypothesis.

Conclusion

We have demonstrated that across firefly species, there is a single LW opsin, a single UV opsin, and no B opsin. Within both LW and UV, there is evidence for parallel and convergent amino acid changes at transitions from the use of nocturnal light signals to diurnal pheromone signals, and within LW opsin, evidence for greater changes in selective constraint with greater changes in activity time. In addition, there is evidence for positive selection at six sites that have been identified in other insect orders as under positive selection and/or of functional importance (LW sites 108, 181, 188, 356; UV sites 59, 133; Table 2.3, S6). This study represents a first step in testing the molecular basis for spectral tuning in fireflies with a comparative dataset. These data provide candidate sites and mutations for future functional testing. Recent advances in insect opsin expression systems (e.g. Frentiu *et al.* 2015) will aid in this effort. Given the tight

correspondence between signal phenotypes and spectral sensitivity (e.g. Lall 1981), the diversity of species in both signaling and ecological traits, the existing and expandable phylogeny, and emerging genomic resources, fireflies continue to be a rich study system for investigating the evolution of signaling and signal reception in a comparative context.

Acknowledgements

This work was supported by the National Science Foundation (GRF to S.E.S and DDIG DEB-1311315 to D.W.H. and S.E.S.) and the NIGMS of the National Institute of Health (award number T32GM007103 to S.E.S.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to thank: Allegheny National Forest (permit to Lynn Faust), Great Smoky Mountains National Park (permit to Kathrin Stanger-Hall), Raphael de Cock, Lynn Faust, Sara Lewis, Michael Marsh (University of Georgia), Jerry McCollum (Charles H. Wharton Conservation Center), David McNaughton (Fort Indiantown Gap), Jenna Pallansch, the Sander Family, and Dorset Trapnell for collection assistance; Megan Behringer, Zachary Wood, and Kelly Dyer for bioinformatics support, homology-modeling advice, and manuscript comments respectively; and Kathrin Stanger-Hall for collection assistance, manuscript comments, and inspiring the study.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403-410.

Arendt D (2003) Evolution of eyes and photoreceptor cell types. *International Journal of Developmental Biology*, **47**, 563-571.

Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, **22**, 195-201.

Aronesty E (2011) *ea-utils*: Command-line tools for processing biological sequencing data. http://code.google.com/p/ea-utils.

Audzijonyte A, Pahlberg J, Viljanen M, Donner K, Väinölä R (2012) Opsin gene sequence variation across phylogenetic and population histories in *Mysis* (Crustacea: Mysida) does not match current light environments or visual-pigment absorbance spectra. *Molecular Ecology*, **21**, 2176-2196.

Biggley WH, Lloyd JE, Seliger HH (1967) The spectral distribution of firefly light II. *Journal of General Physiology*, **50**, 1681-1692.

Bielawski JP, Yang Z (2003) Maximum likelihood methods for detecting adaptive evolution after gene duplication. *Journal of Structural and Functional Genomics,* **3,** 201-212.

Blomberg SP, Garland T, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, **57**, 717-745.

Bok MJ, Porter ML, Place AR, Cronin TW (2014) Biological sunscreens tune polychromatic ultraviolet vision in mantis shrimp. *Current Biology,* **24**, 1636-1642.

Booth D, Stewart AJA, Osorio D (2004) Colour vision in the glow-worm *Lampyris noctiluca* (L.) (Coleoptera: Lampyridae): evidence for a green-blue chromatic mechanism. *Journal of Experimental Biology*, **207**, 2373-2378.

Briscoe AD (2001) Functional diversification of Lepidopteran opsins following gene duplication. *Molecular Biology and Evolution*, **18**, 2270-2279.

Briscoe AD (2008) Reconstructing the ancestral butterfly eye: focus on the opsins. *Journal of Experimental Biology*, **211**, 1805-1813.

Briscoe AD, Bybee SM, Bernard GD, Yuan F, Sison-Mangus MP, Reed RD, Warren AD, Llorente-Bousquets J, Chiao C-C (2010) Positive selection of a duplicated UV-sensitive visual pigment coincides with wing pigment evolution in Heliconius butterflies. *Proceedings of the National Academy of Sciences*, **107**, 3628-3633.

Briscoe AD, Chittka L (2001) The evolution of color vision in insects. *Annual Reviews of Entomology*, **46**, 471-510.

Carvalho L, Cowing JA, Wilkie SE, Bowmaker JK, Hunt DM (2006) Shortwave visual sensitivity in tree and flying squirrels reflects changes in lifestyle. *Current Biology*, **16**, R81-R83.

Cicero JM (1983) Lek assembly and flash synchrony in the Arizona firefly *Photinus knulli* Green (Coleoptera: Lampyridae). *Coleopterists Bulletin*, **37**, 318-342.

Cronin TW, Jarvilehto M, Weckstrom M, Lall AB (2000) Tuning of photoreceptor spectral sensitivity in fireflies (Coleoptera: Lampyridae). *Journal of Comparative Physiology A*, **186**, 1-12.

Dacke M, Byrne MJ, Scholtz CH, Warrant EJ (2004) Lunar orientation in a beetle. *Proceedings of the Royal Society B: Biological Sciences*, **271**, 361-365.

Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, **27**, 1164-1165.

Douglas RH, Mullineaux CW, Partridge JC (2000) Long-wave sensitivity in deep-sea stomiid dragonfish with far-red bioluminescence: evidence for a dietary origin of the chlorophyll-derived retinal photosensitizer of *Malacosteus niger*. *Philosophical Transactions of the Royal Society B: Biological Sciences,* **355**, 1269-1272.

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Eguchi E, Nemoto A, Meyer-Rochow VB, Ohba N (1984) A comparative study of spectral sensitivity curves in three diurnal and eight nocturnal species of Japanese fireflies. *Journal of Insect Physiology*, **30**, 607-612.

Endler JA (1992) Signals, signal conditions, and the direction of evolution. *American Naturalist,* **139**, S125-S153.

Endler JA (1993) The color of light in forests and Its implications. *Ecological Monographs*, **63**, 2-27.

Felsenstein J (1985) Phylogenies and the comparative method. *American Naturalist*, **125**, 1-15.

Fender KM (1966) The genus *Phausis* in America north of Mexico (Coleoptera-Lampyridae). *Northwest Science*, **40**, 83-95.

Frentiu FD, Yuan F, Savage WK, Bernard GD, Mullen SP, Briscoe AD (2015) Opsin clines in butterflies suggest novel roles for insect photopigments. *Molecular Biology and Evolution*, **32**, 368-379.

Gleadall IG, Hariyama T, Tsukahara Y (1989) The visual pigment chromophores in the retina of insect compound eyes, with special reference to the Coleoptera. *Journal of Insect Physiology,* **35**, 787-795.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology,* **29**, 644-652.

Green J (1957) Revision of the nearctic species of *Pyractomena* (Coleoptera: Lampyridae). *Wasmann Journal of Biology*, **15**, 237-284.

Green J (1956) Revision of the Nearctic species of *Photinus* (Lampyridae: Coleoptera). *Proceedings of the California Academy of Sciences*, **28**, 561-613.

Groberman LJ, Borden JH (1982) Electrophysiological response of *Dendroctonus pseudotsugae* and *Ips paraconfusus* (Coleoptera: Scolytidae) to selected wavelength regions of the visible spectrum. *Canadian Journal of Zoology*, **60**, 2180-2189.

Guindon Sp, Rodrigo AG, Dyer KA, Huelsenbeck JP (2004) Modeling the site-specific variation of selection patterns along lineages. *Proceedings of the National Academy of Sciences*, **101**, 12957-12962.

Horridge GA (1979) The eye of the soldier beetle *Chauliognathus pulchellus* (Cantharidae). *Proceedings of the Royal Society B: Biological Sciences,* **203**, 361-378.

Jackowska M, Bao R, Liu Z, McDonald EC, Cook TA, Friedrich M (2007) Genomic and gene regulatory signatures of cryptozoic adaptation: Loss of blue sensitive photoreceptors through exapansion of long wavelength-opsin expression in the red flour beetle *Tribolium castaneum*. *Frontiers in Zoology*, **4,** 24

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics,* **26**, 1463-1464.

Kenaley CP, Devaney SC, Fjeran TT (2014) The complex evolutionary history of seeing red: molecular phylogeny and the evolution of an adaptive visual system in deep-sea dragonfishes (Stomiiformes: Stomiidae). *Evolution,* **68**, 996-1013.

Lall A, Chapman R, Trouth CO, Holloway J (1980) Spectral mechanisms of the compound eye in the firefly *Photinus pyralis* (Coleoptera: Lampyridae). *Journal of Comparative Physiology,* **135**, 21-27.

Lall A, Lord E, Trouth CO (1982) Vision in the firefly *Photuris lucicrescens* (Coleoptera: Lampyridae): Spectral sensitivity and selective adaptation in the compound eye. *Journal of Comparative Physiology,* **147**, 195-200.

Lall AB (1981) Electroretinogram and the spectral sensitivity of the compound eyes in the firefly *Photuris versicolor* (Coleoptera-Lampyridae): A correspondence between green sensitivity and species bioluminescence emission. *Journal of Insect Physiology,* **27**, 461-468.

Lall AB (1993) Action spectra for the initiation of bioluminescent flashing activity in males of twilight-active firefly *Photinus scintillans* (Coleoptera: Lampyridae). *Journal of Insect Physiology,* **39**, 123-127.

Lall AB (1994) Spectral cues for the regulation of bioluminescent flashing activity in the males of twilight-active firefly *Photinus scintillans* (Coleoptera: Lampyridae) in nature. *Journal of Insect Physiology,* **40**, 359-363.

Lall AB, Cronin TW, Carvalho AA, de Souza JM, Barros MP, Stevani CV, Bechara EJ, Ventura DF, Viviani VR, Hill AA (2010) Vision in click beetles (Coleoptera: Elateridae): pigments and spectral correspondence between visual sensitivity and species bioluminescence emission. *Journal of Comparative Physiology A,* **196**, 629-638.

Lall AB, Seliger HH, Biggley WH, Lloyd JE (1980) Ecology of colors of firefly bioluminescence. *Science,* **210**, 560-562.

Lall AB, Strother GK, Cronin TW, Seliger HH (1988) Modification of spectral sensitivities by screening pigments in the compound eyes of twilight-active fireflies (Coleoptera: Lampyridae). *Journal of Comparative Physiology A,* **162**, 23-33.

Lall AB, Ventura DSF, Bechara EJH, de Souza JM, Colepicolo-Neto P, Viviani VR (2000) Spectral correspondence between visual spectral sensitivity and bioluminescence emission spectra in the click beetle *Pyrophorus punctatissimus* (Coleoptera: Elateridae). *Journal of Insect Physiology,* **46**, 1137-1141.

Lall AB, Worthy KM (2000) Action spectra of the female's response in the firefly *Photinus pyralis* (Coleoptera: Lampyridae): evidence for an achromatic detection of the bioluminescent optical signal. *Journal of Insect Physiology,* **46**, 965-968.

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

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

Lin J (1993) Identification of photoreceptor locations in the compound eye of *Coccinella septempunctata* Linnaeus (Coleoptera, Coccinellidae). *Journal of Insect Physiology,* **39**, 555-562.

Lloyd JE (1966) Studies on the flash communication system in *Photinus* fireflies. In. Miscellaneous Publications. Ann Arbor, Michigan: Museum of Zoology, University of Michigan.

Lloyd JE (1969) Flashes, behavior and additional species of Nearctic *Photinus* fireflies (Coleoptera: Lampyridae). *Coleopterists Bulletin,* **23**, 29-40.

Lu A, Guindon S (2014) Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences. *Molecular Biology and Evolution,* **31**, 484-495.

Luk S, Marshall SA, Branham MA (2011) The fireflies of Ontario (Coleoptera: Lampyridae). *Canadian Journal of Arthropod Identification,* **16**, 1-105.

Mackin KA, Roy RA, Theobald DL (2014) An empirical test of convergent evolution in rhodopsins. *Molecular Biology and Evolution,* **31**, 85-95.

Martin G, Lord N, Branham M, Bybee S (2015) Review of the firefly visual system (Coleoptera: Lampyridae) and evolution of the opsin genes underlying color vision. *Organisms Diversity & Evolution*, preprint, 1-14.

McLean M, Buck J, Hanson FE (1972) Culture and larval behavior of Photurid fireflies. *American Midland Naturalist,* **87**, 133-145.

Meredith RW, Gatesy J, Emerling CA, York VM, Springer MS (2013) Rod monochromacy and the coevolution of Cetacean retinal opsins. *PLoS Genetics*, **9**: e1003432.

Murakami M, Kouyama T (2008) Crystal structure of squid rhodopsin. *Nature,* **453**, 363-367.

Oba Y, Kainuma T (2009) Diel changes in the expression of long-wavelength sensitive and ultraviolet-sensitive opsin genes in the Japanese firefly, *Luciola cruciata*. *Gene,* **436**, 66-70.

Otte D, Smiley J (1977) Synchrony in Texas fireflies with a consideration of male interaction models. *Biology of Behavior,* **2**, 143-158.

Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Trong IL, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. *Science*, **289**, 739-745.

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics,* **20**, 289-290.

Porter ML, Blasic JR, Bok MJ, Cameron EG, Pringle T, Cronin TW, Robinson PR (2012) Shedding new light on opsin evolution. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 3-14.

Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Schliep KP (2011) phangorn: phylogenetic analysis in R. *Bioinformatics,* **27**, 592-593.

Schott RK, Refvik S, Hauser FE, López-Fernández H, Chang BSW (2014) Divergent positive selection in rhodopsin from lake and riverine cichlid fishes. *Molecular Biology and Evolution,* **31**, 1149-1165.

Seliger HH, Buck JB, Fastie WG, McElroy WD (1964) The spectral distribution of firefly light. *Journal of General Physiology*, **48**, 95-104.

Seliger HH, Lall AB, Lloyd JE, Biggley WH (1982a) The colors of firefly bioluminescence—I. Optimization model. *Photochemistry and Photobiology,* **36,** 673-680.

Seliger HH, Lall AB, Lloyd JE, Biggley WH (1982b) The colors of firefly bioluminescence—II. Experimental evidence for the optimization model. *Photochemistry and Photobiology,* **36**, 681-688.

Shen YY, Liu J, Irwin DM, Zhang YP (2010) Parallel and convergent evolution of the dim-light vision gene *RH1* in bats (Order: Chiroptera). *PLoS ONE*, **5**, e8838.

Sivasundar A, Palumbi SR (2010) Parallel amino acid replacements in the rhodopsins of the rockfishes (*Sebastes spp.)* associated with shifts in habitat depth. *Journal of Evolutionary Biology*, **23**, 1159-1169.

Stanger-Hall KF, Lloyd JE, Hillis DM (2007) Phylogeny of North American fireflies (Coleoptera: Lampyridae): Implications for the evolution of light signals. *Molecular Phylogenetics and Evolution,* **45**, 33-49.

Stanger-Hall KF, Lloyd JE (2015) Flash signal evolution in *Photinus* fireflies: Character displacement and signal exploitation in a visual communication system. *Evolution*, **69**, 666-682.

Terai Y, Seehausen O, Sasaki T, Takahashi K, Mizoiri S, Sugawara T, Sato T, Watanabe M, Konijnendijk N, Mrosso HDJ, *et al.* (2006) Divergent selection on opsins drives incipient speciation in lake victoria cichlids. *PLoS Biology,* **4,** e433.

Théry M, Pincebourde S, Feer F (2008) Dusk light environment optimizes visual perception of conspecifics in a crepuscular horned beetle. *Behavioral Ecology,* **19**, 627-634.

Tierney SM, Sanjur O, Grajales GG, Santos LM, Bermingham E, Wcislo WT (2011) Photic niche invasions: phylogenetic history of the dim-light foraging augochlorine bees (Halictidae). *Proceedings of the Royal Society B: Biological Sciences*, **279**, 794-803.

Voortman M, Druzdzel MJ (2008) Insensitivity of constraint-based causal discovery algorithms to violations of the assumption of multivariate normality. In. David Wilson & H. Chad Lane, ed., 'FLAIRS Conference', AAAI Press, pp. 690-695.

Wakakuwa M, Terakita A, Koyanagi M, Stavenga DG, Shichida Y, Arikawa K (2010) Evolution and mechanism of spectral tuning of blue-absorbing visual pigments in butterflies. *PLoS ONE*, **5**, e15015.

Wang B, Xiao J-H, Bian S-N, Niu L-M, Murphy RW, Huang D-W (2013) Evolution and expression plasticity of opsin genes in a fig pollinator, *Ceratosolen solmsi*. *PLoS ONE*, **8,** e53907.

Weadick CJ, Chang BS (2012) Complex patterns of divergence among green-sensitive (RH2a) African cichlid opsins revealed by clade model analyses. *BMC Evolutionary Biology,* **12**, 206.

Wilkie SE, Robinson PR, Cronin TW, Poopalasundaram S, Bowmaker JK, Hunt DM. (2000) Spectral tuning of avian violet- and ultraviolet-sensitive visual pigments. *Biochemistry,* **39**, 7895-7901.

Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution,* **24**,1586-1591.

Yang Z, Wong WSW, Nielsen R (2005) Bayes Empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution,* **22**,1107-1118.

Yinon U (1970) The visual mechanisms of *Tenebrio molitor*: some aspects of the spectral response. *Journal of Experimental Biology,* **53**, 221-229.

Yokoyama S (2000) Molecular evolution of vertebrate visual pigments. *Progress in Retinal and Eye Research,* **19,** 385-419.

Yokoyama S (2008) Evolution of dim-light and color vision pigments. *Annual Review of Genomics and Human Genetics,* **9,** 259-282.

Yokoyama S, Yang H, Starmer WT (2008) Molecular basis of spectral tuning in the red- and green-sensitive (M/LWS) pigments in vertebrates. *Genetics,* **179**, 2037-2043.

Table 2.1. Transcriptome assembly metrics

			Trinity Assemblies					
Species	*Number of Reads*	*% lost to QC*	*Total # of components (# comps 1 – 4 kb)*	*Mean length*	*Median length*	*# similar to opsin*	*# BLAST to opsin*	*# putative opsins highly expressed*
Photinus pyralis	39,069,471	1.7	77,811 (27,766)	1715	881	107	10	2
Photuris sp.	12,359,044	1.4	54,415 (17,746)	1395	653	34	6	2
Photinus scintillans	15,456,064	1.8	63,440 (20,865)	1440	666	39	5	2
Lucidota atra	11,194,970	1.8	56,584 (19,259)	1440	718	73	10	2
Photinus carolinus	7,472,450	1.6	43,254 (13,950)	1338	615	25	2	2
Photinus macdermotti	41,783,634	4.9	76,743 (30,730)	1701	1022	80	12	2
Photinus australis	13,145,094	1.7	43,963 (14,649)	1140	606	32	4	2
Photuris frontalis	11,373,006	1.3	56,427 (17,855)	1278	589	34	4	2
Pyractomena borealis	11,431,732	1.6	52,416 (18,463)	1544	792	60	5	2
Phausis reticulata	9,245,174	0.02	67,843 (21,132)	1297	581	47	12	2

Table 2.2. Results of PAML analysis for LW and UV opsin genes

			Parameters[b]					
Gene	Model[a]	lnL	ω_0/p (p_0)	ω_1/q (p_1)	ω_2/ ω_P (p_2)	Null	LRT	df
LW opsin	M0 (one rate)	-6550.53	0.07					
Branch	H1 (N vs D)[c]	-6548.44	0.07	0.10		M0	4.17*	1
	Free	**-6432.33**	**see Appendix A, Figure S11**			**M0**	**236.40***	**37**
Sites	M1a (neutral)	-6348.90	0.04 (0.75)	1 (0.25)		M0	403.26*	1
	M2a (selection)	**-6302.40**	**0.04 (0.91)**	**1 (0.09)**	**999[d] (0.0003)**	**M1a**	**92.99***	**2**
	M3 (discrete)	**-6285.76**	**0.02 (0.84)**	**0.49 (0.16)**	**999[d] (0.0003)**	**M0**	**529.54***	**4**
	M7 (beta)	-6329.70	0.14	0.48				
	M8 (beta, selection)	**-6277.28**	**0.15**	**1.22**	**243.12 (0.001)**	**M7**	**104.85***	**2**
UV opsin	M0 (one rate)	-7241.40	0.05					
Branch	H1 (N vs D)[c]	-7236.91	0.05	0.09		M0	8.98*	1
	Free	**-7151.25**	**see Appendix A, Figure S11**			**M0**	**180.31***	**37**
Sites	M1a (neutral)	-7111.53	0.04 (0.80)	1 (0.20)		M0	259.74*	1
	M2a (selection)	**-7060.18**	**0.04 (0.94)**	**1 (0.06)**	**999[d] (0.0004)**	**M1a**	**102.71***	**2**
	M3 (discrete)	**-7031.77**	**0.02 (0.83)**	**0.34 (0.17)**	**999[d] (0.0003)**	**M0**	**419.26***	**4**
	M7 (beta)	-7093.36	0.14	0.62				
	M8 (beta, selection)	**-7028.47**	**0.18**	**2.08**	**443.45 (0.0006)**	**M7**	**129.78***	**2**

[a] Best models within a nested set of models (branch, site: selection, discrete, beta) are shown in bold

[b] For branch models, ω_0 (background branches) and ω_1 (branches of interest); for site models M1a-M3, ω_x and p_x (proportion of sites) in each class (0, 1, 2); for site models M7 and M8, p and q describe the shape of the beta distribution, ω_P is the value of ω for the positively selected site class, and p_2 gives the proportion of sites in this class.

[c] Nocturnal versus diurnal branches

[d] A value of 999 indicates an ω_2 greater than 1, but unable to be precisely estimated

*$p < 0.05$

Table 2.3. Sixteen candidate sites with evidence for positive selection and/or function

					Change at homologous site
Gene	Site[a]	Identified in Model(s)	Location[b]	Evidence for Phenotypic Effect	Site, Organism, Opsin
LW	T108	M2a, M8	NC	Shift from 530 to 550 nm	63, *Heliconius*, LW[c]
				Cline with latitude	112, *Liminitis* LW[d]
	A181 *	M2a, M8, M2aS1	H4	Shift between 530 and 510 nm	136, *Vanessa*, LW[c]
				Shift from 530 to 550 nm	136, *Heliconius*, LW[c]
				Shift between 530 and 575 nm	136, *Papilio,* LW[c]
				Positively selected in parasitic wasps	198, Fig wasp, LW[e]
	L188 *	M2a, M8, M2aS1	NC	Shift between 530 and 575 nm	143, *Papilio*, LW[c]
				Shift from 530 to 550 nm	143, *Heliconius*, LW[c]
				Faster evolution after duplication	143, *Bombus*, LW[f]
	V235 *	M2a, M8, M2aS1	H5		
	A248	M8	H5		
	F307	M2aS1	H6		
	T309	M8	EC		
	A356	M8	T	Positively selected in dim-light foragers	229, Halictidae, LW[g]
UV	L59	M2a, M8	NC	Positively selected after duplication	60, *Heliconius*, UV[h]
	S133	M3S2	BP	S to A shifts absorption (450-437 nm)	116, *Pieris rapae*, B and V[i]
	V218	M8	NC		
	A267	M8	IC		
	A268	M8	IC		
	S299	M8	BP		
	T321	M2a, M8	NC		
	T373	M8	T		

[a] Site number and amino acid in reference to *Pn. pyralis*

[b] NC: near chromophore, but not in contact; BP: binding pocket; H: helix, far away from binding pocket; EC: extracellular loop; IC: intracellular loop; T: tail

[c] Ancestral state reconstruction (Briscoe 2001)

[d] Within-species cline with latitude. Hypothesized involvement in thermal stability. (Frentiu et al. 2015)
[e] Tests of positive selection (Wang et al. 2013)
[f] Tests of positive selection (Spaethe and Briscoe 2004)
[g] Tests of positive selection (Tierney et al. 2011)
[h] Tests of positive selection (Briscoe et al. 2010)
[i] Site-directed mutagenesis (Wakakuwa et al. 2010)
* significant at p=0.05, BEB analysis in PAML M8 model.

Table 2.4. Results of Fitmodel analysis for LW and UV opsin genes

			Parameters[c]					
Model[a]	Switching[b]	lnL	ω_0 (p_0)	ω_1 (p_1)	ω_2 (p_2)	Null	LRT	df
LW opsin								
M0 (one rate)	none	-6190.41	0.08 (1)	n/a	n/a	n/a		
M1a (neutral)	none	-6309.16	0 (0.71)	1 (0.29)	n/a	M0	-37.49	1
M2a (selection)	none	-6150.19	0.04 (0.91)	1 (0.09)	20 (0.001)	M1a	317.95*	2
M2aS1 (selection)	**equal**	**-6086.95**	**0.00 (0.92)**	**1 (0.07)**	**2.75 (0.01)**	**M2a**	**129.48***	**1**
M2aS2 (selection)	biased	-6086.15	0.00 (0.91)	1 (0.08)	2.04 (0.01)	M2aS1	1.59	2
M3 (discrete)	none	-6120.35	0.00 (0.65)	0.14 (0.28)	0.75 (0.07)	M1a	377.63*	3
M3S1 (discrete)	equal	-6084.41	0.00 (0.87)	0.41 (0.09)	1.65 (0.04)	M3	71.87*	1
M3S2 (discrete)	**biased**	**-6068.79**	**0.00[d] (0.10)**	**0.00[d] (0.88)**	**9.79 (0.01)**	**M3S1**	**31.25***	**2**
UV opsin								
M0 (one rate)	none	-6936.73	0.06 (1)	n/a	n/a	n/a		
M1a (neutral)	none	-7058.89	0 (0.73)	1 (0.27)	n/a	M0	-244.31	1
M2a (selection)	none	-6859.71	0.04 (0.94)	1 (0.06)	20 (0.004)	M1a	398.35*	2
M2aS1 (selection)	**equal**	**-6816.10**	**0.01 (0.88)**	**0.30 (0.06)**	**1[e] (0.06)**	**M2a**	**87.23***	**1**
M2aS2 (selection)	biased	-6824.05	0.01 (0.92)	1 (0.07)	18.04 (0.01)	M2aS1	-15.91	2
M3 (discrete)	none	-6824.91	0.00 (0.61)	0.10 (0.33)	0.54 (0.06)	M1a	467.96*	3
M3S1 (discrete)	equal	-6817.10	0.01 (0.89)	0.59 (0.11)	17.37 (0.0005)	M3	15.62*	1
M3S2 (discrete)	**biased**	**-6813.45**	**0.01 (0.89)**	**0.49 (0.11)**	**20 (0.0009)[f]**	**M3S1**	**7.29***	**2**

[a] Best models within each nested model set (selection, discrete) as given by LRTs shown in bold. Selection models are constrained so that $\omega_1 = 1$, while discrete models are constrained to have $\omega_0 < \omega_1 < \omega_2$.

[b] Switching scheme for each model. None = sites do not switch between ω classes, equal = sites have equal rates of switching between ω classes, biased = unequal rates of switching between ω classes.

[c] Estimated parameters for each ω class under the model. ω_x is the estimated omega for the 3 classes of sites (0, 1, 2). p_x is the estimated proportion of sites in each class.
[d] In this case, while the biased switching model was favored significantly over the equal switching model, ω_0 was estimated to be = ω_1, and the switching parameters was low (0, data not shown) due to the fact that these classes were virtually indistinguishable in evolutionary rate. The interpretation of this result is that most of the molecule is under constraint, while a small proportion of sites (0.01) is in the positively selected class (ω_2).
[e] ω_2 seems to violate the constraints of the selection model. Fitmodel estimates the ω values for each class with the constraint that $\omega_1 = 1$. However, if the estimated ω_2 is smaller than 1 at the end of analysis, then the omega values are re-ordered so that ω_2 is always has the largest ω value, in this case 1.
[f] The estimated proportion of sites is small compared to the alignment length (385 codons). Accordingly, we further examined mutations at the single site identified as in ω_2 by this model (see main text).
*significant at $p < 0.05$

Table 2.5. Partial correlations between change in selective constraint (ω) and change ecological and signaling traits

Δ ω[a]	Δ trait[b]	*pr*[c]	*p*	df
LW	Spectra	0.11	0.64	19
	Habitat	-0.23	0.31	
	Activity	0.54	0.01*	
UV	Spectra	0.23	0.29	21
	Habitat	-0.26	0.23	
	Activity	0.24	0.27	

[a] Change in ω values between branches obtained from PAML free-ratio branch models. Branches with ω > 900 (signifying a branch on which there were no synonymous substitutions) were removed prior to analysis.
[b] Change in traits determined using PIC values.
[c] Partial correlation coefficient after controlling for branch length

Fig. 2.1.

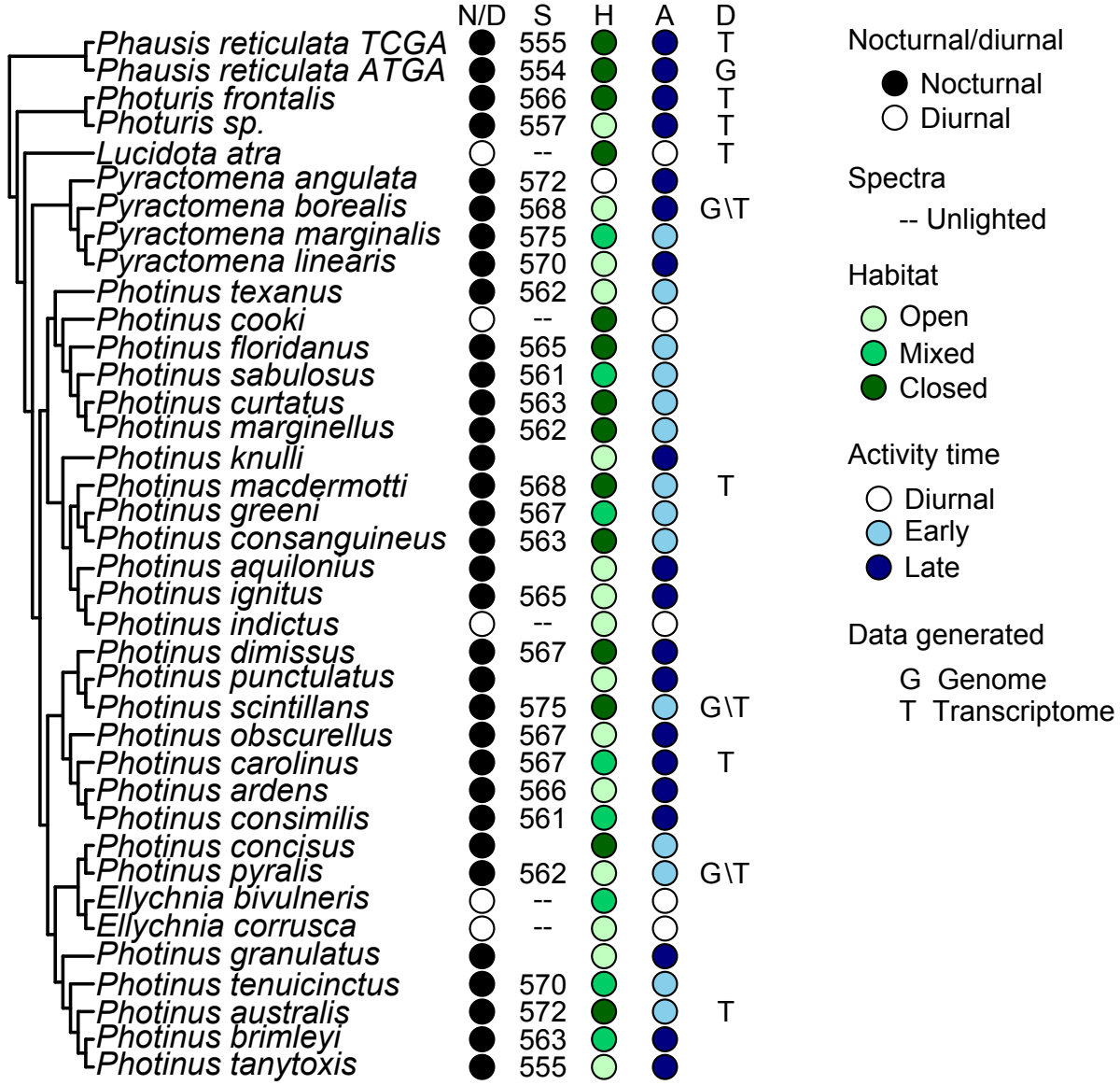


Figure 2.1. Cladogram showing signaling traits for 38 species used in this study

The topology of the phylogeny was obtained by adding nine additional taxa to the three-locus dataset described in Stanger-Hall and Lloyd (2015). Male signal and ecological traits are listed in columns to the right: (N/D) nocturnal/diurnal, (S) average male peak emission wavelength per species, (H) habitat in which signal is displayed, and (A) signaling activity start time. Values for traits are from personal observation of populations where specimens used in this study were captured and from the literature. (D) Data generated shows the species included in high-throughput sequencing for opsin identification. LW and UV opsins for all other taxa were obtained using Sanger sequencing.

Fig. 2.2.

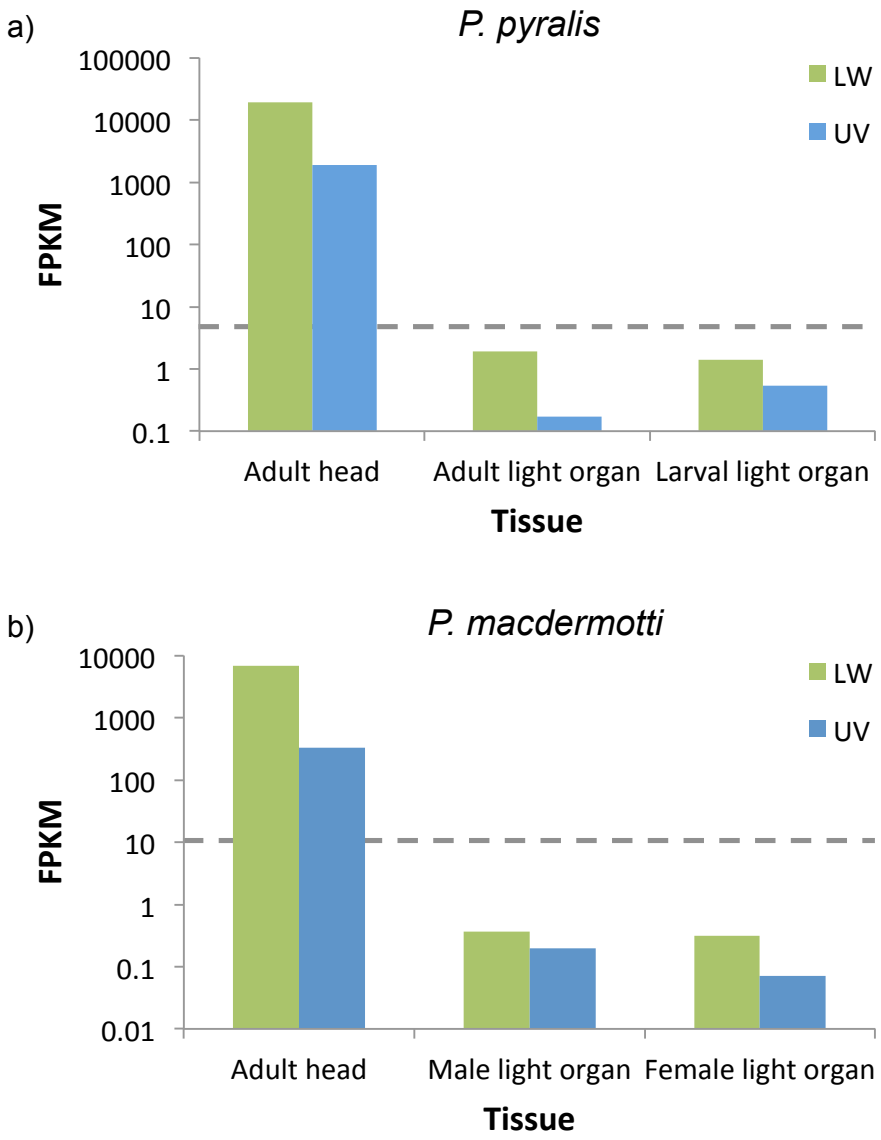


Figure 2.2. Opsins are highly expressed in head versus light organ tissue

Opsin expression levels (fragments per kilobase per million mapped reads, FPKM) in heads of (a) *Photinus pyralis* and (b) *Photinus macdermotti*. The gray dotted line shows median FPKM for all transcripts expressed in all tissues of each species (*Pn. Pyralis*: 1.29; *Pn. Macdermotti*: 0.91). Light organ tissues were sampled according to availability of larval/female specimens.

Fig. 2.3.

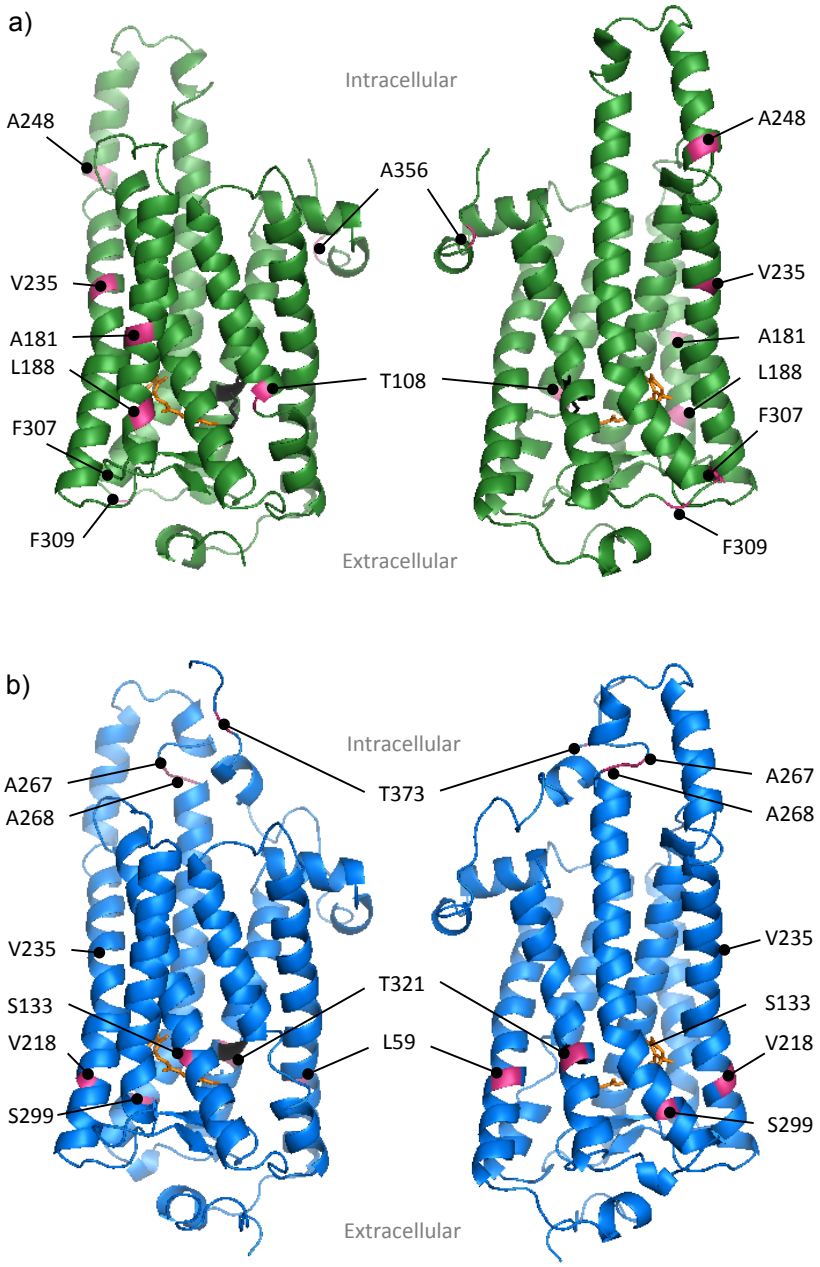


Figure 2.3. Homology models of *Pn. pyralis* LW opsin and UV opsin

Most of the positively selected sites identified by PAML and fitmodel do not interact with the chromophore. (a) LW opsin is shown in green and (b) UV opsin in blue. The orange sticks show the chromophore in the binding pocket. Black with a yellow star indicates the conserved lysine where the chromophore is bound to the opsin. The eight positively selected sites identified by PAML and fitmodel for each opsin are labeled and shown in magenta.

Fig. 2.4.

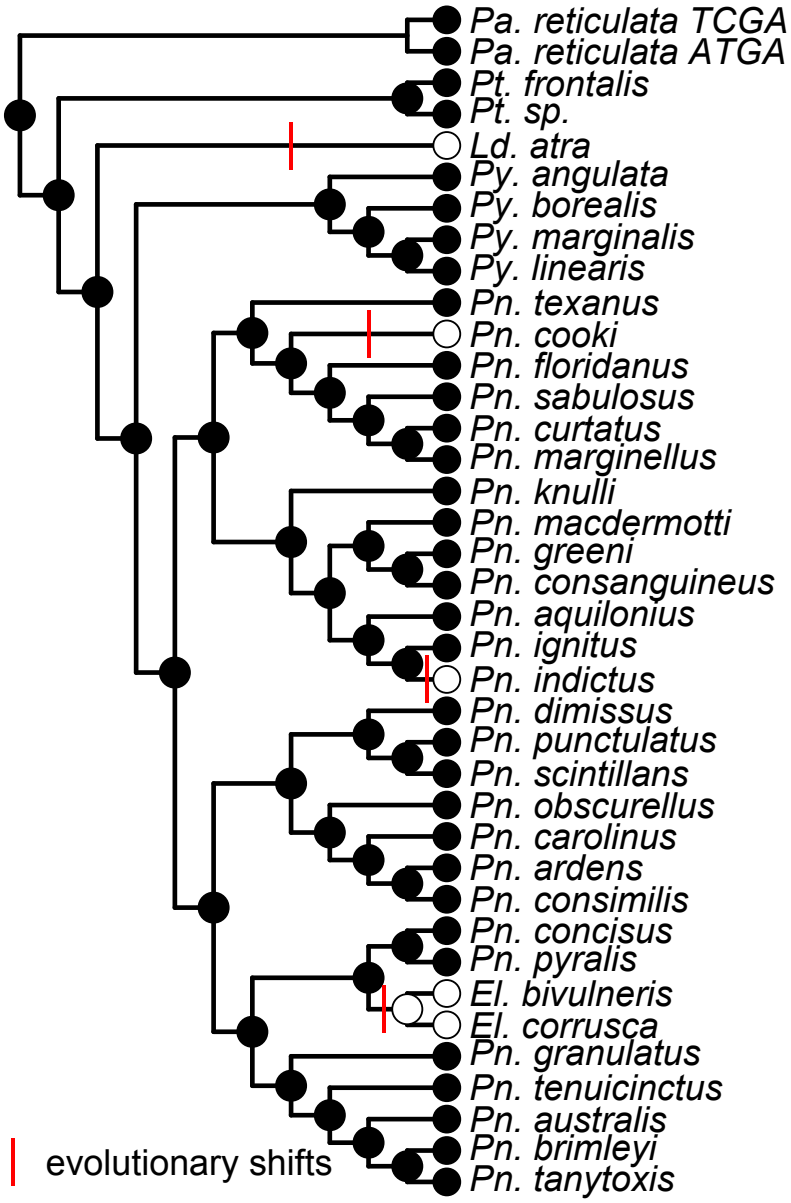


Figure 2.4. Four independent transitions from nocturnal to diurnal activity

Maximum parsimony reconstruction of nocturnal versus diurnal states. Filled/open circles indicate extant taxa and reconstructed ancestral taxa that are nocturnal/diurnal, respectively. *Pa*: *Phausis*, *Pt*: *Photuris*, *Ld*: *Lucidota*, *Py*: *Pyractomena*, *Pn*: *Photinus*, *El*: *Ellychnia*.

CHAPTER 3

GENOME SIZE EVOLUTION IN NORTH AMERICAN FIREFLIES[1]

[1] Sander, S.E., Korunes, K., Johnston, S., Hanrahan, S. and D.W. Hall. To be submitted to *Genome Biology and Evolution*.

Abstract

Eukaryotic genomes show tremendous variation across taxa. The lack of correlation between gene content and genome size, termed the "C-value paradox," has prompted numerous hypotheses to explain variation in genome size. Proximate explanations include ploidy variation and repetitive DNA, though the relative contributions of different repeat types (transposable elements, simple repeats) remains unexamined in many taxa. Ultimate explanations include selection on physiological correlates of genome size, such as cell size, which in turn influences body size, and metabolic rate, and a selection-drift barrier such that inefficient selection in species with small effective population sizes allows large genomes to accumulate. Few studies have examined both proximate and ultimate hypotheses within a single family of organisms. Here, we document 6-fold genome size variation in at least 20 species of North American fireflies (family Lampyridae) using flow cytometry. We then conduct low-coverage sequencing to identify common genomic repeats across species and combine these with measures of body size, proxies for metabolism (light production and female wing reduction), and proxies for effective population size (range size and nucleotide diversity) to test explanations for genome size. We find no evidence for selective explanations based on physiological correlates (body size, light production, female flight). In addition, repetitive content cannot account for the observed variation in genome size. However, we do find a negative correlation between range size and genome size, suggesting that effective population size and the selection-drift barrier may play a role. Fireflies, known for their light signals, offer an intriguing system to study genome size evolution, both across and within species.

Introduction

Eukaryotic genome sizes vary widely, from 2.19 Mb in the microsporidian fungus *Encephalitozoon romaleae* to 148,851.60 Mb in the angiosperm plant *Paris japonica* (Kullman *et al.* 2005; Bennett & Leitch 2012). Neither biological complexity nor the number of genes in the genome is predictive of genome size. For example, some salamander species have genomes that are 40 fold larger than the human genome (Gregory 2015). The lack of a relationship between genome size and complexity or gene content has long been termed the "C-value paradox" (Thomas 1971). Explanations for the C-value paradox include differences in ploidy level and/or the quantity of non-coding DNA, with the relative importance of each varying across taxa (reviewed in Gregory 2005). Further, the specific types of non-coding DNA that contribute to genome size variation also differs (Gaut & Ross-Ibarra 2008). While these factors can explain how genome size can vary, there is still a great deal that is unknown regarding their relative importance. In addition, the roles of natural selection and genetic drift in shaping genome-size variation are poorly understood.

Whole-genome duplication (WGD) changes ploidy and thus rapidly increases genome size. It has been observed in a variety of taxa (e.g. plants: Shi *et al*. 2010; McKain *et al.* 2012; Cannon *et al.* 2014; fish: Glasauer & Neuhauss 2014; and some insects: Jacobson *et al.* 2013). WGD occurs primarily through polyspermy (multiple sperm fertilizing the same egg) or through the production of unreduced gametes (Mable *et al.* 2011). Unreduced gamete production is more likely to occur in species without a dedicated germ line (i.e. where reproductive structures arise from somatic tissue) because there are more cell divisions between generations and thus more opportunities for an aberrant cell cycle in which cytokinesis does not occur. In addition, WGD is more likely in species lacking heteromorphic sex chromosomes (Orr 1990), because of the

resulting issues with relative dosage of autosomal versus sex-linked genes. For these reasons, WGD is common in plants and relatively rare (e.g. tetrapod vertebrates) or extremely rare (e.g. mammals) in many other taxa (reviewed in Mable 2004).

Other mechanisms that increase (or decrease) genome size have more modest effects than WGD. These include changes in the copy number of a single chromosome (aneuploidy), spontaneous insertions and deletions (indels), transposable element (TE) proliferation, and microsatellite sequence expansion and contraction (Petrov 2001). Spontaneous indels are usually mediated by recombination between homologous, but not orthologous, regions of DNA, implying that changes in genome size are more likely to occur when repetitive DNA is common.

The long-term fate of changes in genome size, which happen at the level of the individual, depends on the relative importance of natural selection and genetic drift. Genome size can affect organismal phenotypes on which selection acts. For example, there is a strong, positive correlation between cell size and genome size in most taxa, including bacteria, eukaryotic microbes and multicellular eukaryotes (Cavalier-Smith 1978; Shuter *et al.* 1983; Gregory 2001, 2002a; Beaulieu *et al.* 2008; Connolly *et al.* 2008). Cell size in turn can influence body size (Arendt 2007), which is one of the most ecologically important traits of an organism (Peters 1986). Thus, genome size variation is a likely target of selection through effects on body size. Further, metabolic and developmental correlates of genome size have also been described (reviewed in Gregory 2005). In birds, genome size is negatively correlated with both flight muscle and heart size, perhaps causally through effects on flight metabolism (Wright *et al.* 2014). Genome size also correlates with developmental rate and complexity in many taxa (e.g. Drosophila: Gregory & Johnston 2008, mosquitos: Ferrari & Rai 1989, ladybird beetles: Gregory *et al.* 2003), and it has been suggested that there is an upper genome-size limit of around 2 pg

(1,956 Mb) for insects that undergo metamorphosis (Gregory 2002b). Other physiological correlates with genome size have been examined though, in many cases, the link between genome size and the trait of interest remains unknown (reviewed in Gregory 2005). For example, it is not known what causes the positive correlation between genome size and song attractiveness in *Chorthippus biguttulus* grasshoppers (Schielzeth *et al.* 2014).

Genetic drift is also expected to play a significant role in the evolution of genome size, as it alters both the probability that a variant will initially increase in frequency, even when it is beneficial (Haldane's sieve, Haldane 1927), and the probability that a non-selected, or weakly selected variant will be lost or fixed. This implies that in small populations, selection will be inefficient unless it is strong enough to overcome the force of genetic drift. Previous work hypothesized that large genomes result when weak selection favoring reduced genome size is weaker than the strength of genetic drift, which results in a negative correlation between effective population size and genome size across broad taxonomic scales (Lynch & Conery 2003). This relationship has not been examined at lower taxonomic scales, such as within a family or genus.

There are relatively few animal systems in which genome-size variation has been measured, the proximate mechanisms underlying the variation identified, and the evolutionary hypotheses for the variation tested in a phylogenetic context. We chose to fill this gap by studying genome-size evolution in fireflies, which comprise the beetle family Lampyridae. Worldwide, there are over 2000 firefly species, which have a diverse and fascinating biology. They are best known for the nocturnal sexual displays of adults that utilize light production to signal to conspecifics. Light production uses a luciferase-luciferin reaction system, which has been extensively studied at the molecular level (reviewed in Fraga 2008), and is now widely used

in various gene expression reporter applications (e.g. Shapiro *et al.* 2005). The evolution of the lighted sexual displays have also been extensively studied, especially with respect to species recognition and mate choice (e.g. Lloyd 1966, 1969; Vencl & Carlson 1998; Lewis & Cratsley 2003, 2008; Lewis *et al.* 2004a; Ohba 2004; Cratsley & Lewis 2005; Tamura *et al.* 2005; Stanger-Hall & Lloyd 2015). In addition, fireflies have been used to address a variety of other questions, including the evolution of biological novelty (e.g. Day *et al.* 2009; Oba *et al*. 2010; Stansbury & Moczek 2014), aposematism (e.g. DeCock & Matthysen 1999; Lewis *et al*. 2012), and nuptial gifts (e.g. Cratsley 2004; Lewis *et al.* 2004b; South *et al.* 2011). Further, both morphological (Branham & Wenzel 2003) and molecular phylogenies (Stanger-Hall *et al.* 2007), including a three-gene phylogeny of North American *Photinus* fireflies (Stanger-Hall & Lloyd 2015), are available for comparative studies. Currently there is no published work on genome size variation in fireflies.

In this study, we use flow cytometry to determine the genome size for over 20 species of North American fireflies. The two beetle species with published genomes are 205 Mb (*Tribolium castaneum*: Juan & Petitpierre 1991) and 204 Mb (*Dendroctonus ponderosae*: Keeling *et al.* 2013) in size. In contrast, flow cytometry estimates from other sexually- reproducing beetle species range from 154-2650 Mb (Hanrahan & Johnston 2011). The closest sister taxon to fireflies that has been measured, *Phengodes fuscipes,* has a genome size of 2,233 Mb (Hanrahan & Johnston 2011) and so, we predict that fireflies will have genomes around 2,000 Mb in size. We then perform low coverage sequencing (to identify common repeats within genomes, e.g. Macas *et al.* 2007; Tenaillon *et al.* 2011) in 20 of the species, measure body size, geographic range and nucleotide diversity (as estimators of population size) in 21 species, and combine these data with a molecular phylogeny of the species to test whether: (1) there is a phylogenetic

component to genome-size variation such that closely-related taxa are more likely to have similar genome sizes; (2) repetitive DNA explains the variation in genome-size and, if so, what types of repetitive DNA are responsible; (3) genome-size is positively correlated with body size, as observed in other taxonomic groups; (4) genome size is larger in species with lower metabolic rate, measured indirectly as the presence/absence of adult lighted sexual displays and the presence/absence of reduced-winged females; and (5) range size and genetic diversity (as proxies for effective population size) are negatively correlated with genome size.

Materials and Methods

Specimen collection and identification for flow cytometry

Specimens were collected from natural populations and kept alive in 50 mL plastic conical tubes containing a piece of damp paper towel to retain moisture. Upon return to the lab, specimens were flash-frozen in liquid nitrogen and kept at -80 °C until individual heads could be harvested for flow cytometry. Specimens were identified to species in the field using flash pattern and morphology. This initial species identification was verified using the morphology of male genitalia (Green 1956, 1957; Fender 1966; Luk *et al.* 2011) and 376 bp of the mitochondrial *cytochrome oxidase I* (*COI*) locus, previously shown to be diagnostic to genus in fireflies (Stanger-Hall *et al.* 2007). Where species identification was ambiguous, an additional 895 bp of the *COI* locus were sequenced as well as two nuclear loci: 594 bp of *rudimentary* (*CAD*), and 420 bp of *wingless* (*WG*). These loci have been demonstrated to be phylogenetically informative for beetles (Wild & Maddison 2008) and for *Photinus* fireflies (Stanger-Hall & Lloyd 2015). For small specimens, whole bodies were required for flow cytometry; therefore, molecular data from proxy specimens were used to verify the field identification. Proxy

specimens were caught at the same location as flow cytometry specimens, typically on the same night. Bodies of large specimens and all proxies are retained in the KSH collection at the University of Georgia.

Flow cytometry

Genome size estimates were obtained from individual firefly heads using flow cytometry. Heads were prepared according to the best practice protocol described in (Hanrahan & Johnston 2011). Briefly, flow cytometry was performed on propidium iodide-stained nuclei isolated from head tissue using a Partec CyFlow with a solid-state laser emitting 532 nm. Two standards were used to ensure accurate measurement: *Drosophila virilis* (1C=328 Mb) and *Periplaneta americana* (1C=3338 Mb). Individual specimens were measured once on the flow cytometer using half of the head tissue. If the estimate showed a large deviation from other samples of the same species, the individual was re-measured using an independent preparation of the remaining head tissue. Where possible, five males and five females were measured per field-identified species (Table 3.1).

Statistical analyses of genome size

The variance in genome size estimates across taxonomic levels was investigated using standard least squares in JMP Pro 10 (SAS Institute Inc. 2012). The full model included the effects of genus (random), species nested within genus (random), and sex nested within species and genus (fixed). Subsequently, sex differences in genome size were explored using Student's t-tests for seven species that had estimates for at least two individuals of each sex (*Lucidota atra, Phausis sp. WAT, Photinus carolinus, Photinus curtatus, Photinus macdermotti, Photinus*

marginellus, and Photinus pyralis). The false discovery rate (FDR) for multiple comparisons was adjusted using the Benjamini-Hotchberg procedure (Benjamini & Hochberg 1995). Finally, species-level mean genome sizes for comparative analyses were obtained by averaging mean male and female values.

Morphological measurements

Body size measurements were obtained from photographs of ethanol-preserved KSH collection specimens on a 1 mm grid using ImageJ v.1.42 (Schneider *et al.* 2012, Appendix B, Table S1). Specimens used in size measurements did not overlap with those used in flow cytometry due to differences in storage requirements for downstream processes. All body size specimens were identified to species using morphology, flash behavior, and molecular methods (when necessary) as described above. The size measures examined were: pronotum length (straight line distance from base to apex), pronotum width (straight line distance between the two posterior points of the pronotum), pronotum area, elytron length (apex to humeral edge), and body length (top of head to apex of genital segment in the ventral aspect)(Figure 3.1). Dry body mass was not measured since it would compromise the utility of specimens for genetic work. Where possible, size measures were obtained from at least three males and three females per species.

All size measures were highly correlated (data not shown). Reduction of dimensionality using Principal Components Analysis (PCA) resulted in a single principal component with an eigenvalue > 1 that accounted for over 90% of the variance (depending on sex) and all size measurements had virtually identical loadings in the same direction along the axis. This was taken as an indication that using any one of the physical measurements would appropriately

represent a measure of size. Pronotum width was selected for use in the final analysis for several reasons: (1) body length, as measured, could be confounded by nutritional state/age (many firefly species do not eat as adults) and there was a relatively large variance in that dataset, (2) elytron length could be confounded by brachyptery in females of some species, and (3) previous work has shown the utility of pronotum width as a measure of size in fireflies (Vencl 2004). There were significant differences in pronotum width between the sexes (two-tailed t-test: $p = 0.017$) and subsequent analysis was performed on log-transformed species means for males and females separately. All statistical analyses were performed in JMP.

Phylogeny

Evolutionary relationships among species were reconstructed by extending the *Photinus* phylogeny of Stanger-Hall and Lloyd (2015) to include 10 taxa outside the genus. To do this, representative specimens from each species were sequenced at the three loci cited above, aligned with the Stanger-Hall and Lloyd dataset using MUSCLE (Edgar 2004) in Geneious R7 (Biomatters Ltd.), and manually reviewed. jModeltest2 (Darriba *et al.* 2012) was used to select an appropriate model of evolution for each locus (*WG*: K80+I+G, *CAD*: TIM3+I+G, *COI*: GTR+I+G). Ultrametric phylogenies were constructed in BEAST v.1.8 (Drummond *et al.* 2012) using an uncorrelated lognormal clock model to account for rate variation among linages. BEAST was run twice for 30 million generations each with 25% burn-in, until the estimated sample size for all parameters was over 200. Independent runs were assessed for convergence using Tracer and the majority-rule consensus tree produced in TreeAnnotater. The final tree was trimmed to include only those taxa used in this study.

Species biology and range

Data on signal mode (light/dark) and female wing reduction (full-size elytra present versus reduced elytra) were gathered from the literature (Green 1956; Green 1957) and field observations. Both signal mode and female wing reduction were coded as binary variables for tests of correlations with genome size. Range data was obtained for *Photinus* fireflies (Kathrin Stanger-Hall, personal communication; Stanger-Hall & Lloyd 2015) and quantified as the number of US states and Canadian provinces in which a species' presence has been recorded.

Nucleotide diversity

Sequences for multiple individuals of each species were extracted from the database of over 400 *COI* sequences used for species identification in the KSH collection. The 376 bp segments were aligned using MUSCLE in Geneious and imported into DNAsp v.5 (Librado & Rozas 2009) to calculate nucleotide diversity (π; Nei 1987), the average pairwise nucleotide differences per site, for each species. Positions with gaps in the alignment were excluded from analysis. The number of sequences available for analysis varied across species, however, there was no significant relationship between π and sample size ($p=0.68$). Statistical analyses were performed in JMP.

Independent contrasts

In order to determine if continuous traits (e.g. genome size, body size) could be considered independent of phylogeny, traits were assessed for phylogenetic signal, the tendency of species to resemble close relatives, using Blomberg's K (Blomberg *et al*. 2003) in the ape package v.3.0-11 (Paradis *et al.* 2004) in R 3.0.2 (R Core Team 2013). Genome size was log

transformed prior to analysis. A significant K, as determined by randomizing traits over tips of the phylogeny and calculating the distribution of K, is indicative of significant clustering of trait values among relatives. K = 1 indicates that the degree of clustering is as expected under the assumption of Brownian motion, while K < 1 and K > 1 indicate less or greater clumping than expected, respectively. If K was significant, then subsequent analyses of correlations accounted for relatedness between species using phylogenetic independent contrasts (PICs; Felsenstein 1985) in Mesquite v.2.75 (PDAP-PDTREE module v.1.15; Maddison & Maddison 2011; Midford *et al.* 2008). Branch lengths were assessed for statistical adequacy prior to using PICs (Maddison & Maddison 2011). No branch length transformations were necessary based on these diagnostic tests. Least-squares regressions of positivized contrasts computed through the origin were used to test for significant correlations. Pearson product-moment correlation coefficients and p-values are reported.

Ancestral states were estimated using the ape package in R.

454 sequencing

Low-coverage genomic sequencing was performed on 21 individuals, representing 20 species, to identify common repetitive elements that might account for variation in genome size (Appendix B, Table S2). In one species, *Pyropyga decipiens,* two individuals with different genome size types were sequenced. If sequencing is unbiased, then the proportion of repetitive elements in the sample should reflect their abundance in the genome (Macas *et al.* 2007; Swaminathan *et al.* 2007). Genomic DNA was isolated from thorax or whole body of single specimens using phenol-chloroform with RNAse digestion. Sequencing libraries for each specimen were uniquely barcoded and then all libraries pooled into two lanes of a 454 FLX

Titanium XLR70 (Roche Diagnostics Corporation). Library preparation, sample barcoding, sequencing, and demultiplexing were performed at the Georgia Genomics Facility (Athens, GA).

Sequences were assessed for quality using fastqc v.0.11.2 (Babraham Bioinformatics 2012) and subsequently trimmed for adapters and low-quality regions using the fastq-mcf program in ea-utils v.1.1.2 (parameters: -q 20 -p 10 –D1 5 -x 0.01 -w 20) (Aronesty 2011). Seqtk v.1.0 (https://github.com/lh3/seqtk) was used to trim 19 bases from the beginning of each read due to skewed base distributions and PCR duplicates were collapsed using the fastx toolkit v.0.0.13.2 (http://hannonlab.cshl.edu/fastx_toolkit). Mitochondrial sequences that would not contribute to flow cytometry estimates of genome size were identified and removed using BLASTn (evalue=1e-6; Altschul *et al.* 1990) of collapsed reads against a custom database of complete mitochondrial genomes from Elateroid beetles, including fireflies: *Luciola cruciata* (AB849456; Matsui & Amano, unpublished data), *Pyrocoelia rufa* (AF452048; Bae *et al.* 2004), *Rhagophthalmus ohbai* (AB267275; Li *et al.* 2007), *Rhagophthalmus lufengensis* (DQ888607; Li *et al.* 2007), *Pyrophorus divergens* (EF398270, Arnoldi *et al.* 2007), and *Chauliognathus opacus* (FJ613418; Sheffield *et al.* 2009). Prokaryotic contaminants were identified and removed using kraken v.0.10.5 (Wood & Salzberg 2014). Kraken uses a taxonomically-informed exact-match kmer analysis to identify microbial reads in high-throughput sequencing datasets. In this study, kmers from input reads were matched against a modified minikraken kmer library constructed from all RefSeq bacteria, archaea, plasmids, and virus sequences filtered for repetitive sequences using the BLAST+ dustmasker (Wood, personal communication). Reads that remained unclassified were retained in further analysis. Finally, all reads less than 80 bp were removed to eliminate short sequences that would decrease the efficiency of repetitive element identification and assembly.

Repetitive element identification and classification

Repetitive elements were identified using the RepeatExplorer Galaxy server with default parameters (Novák *et al.* 2013). RepeatExplorer identifies repetitive elements *de novo* by using a graph-based method to group reads into discrete clusters based on all-by-all blast similarity. It then annotates clusters using RepeatMasker (Smit *et al.* 2013-2015) using all or a subset of RepeatMasker databases and assembles contigs from the reads belonging to each cluster using CAP3 (parameters: -O '-p 80 -o 40; (Huang & Madan 1999; Novák *et al.* 2010). To be inclusive, we selected all RepeatMasker databases for annotation. Because the annotation step of RepeatExplorer is computationally intensive, only clusters consisting of more than 0.001% of the total reads were annotated. This cut-off was low enough to fully capture the distribution of repeats in all species (Appendix B, Figure S1), while remaining computationally tractable. Annotated clusters are termed "top" clusters. Remaining clusters are "bottom" clusters and were not annotated. Both top and bottom clusters were screened for contaminants by blasting assembled contigs against the NCBI nucleotide database (evalue: 1e-5) and excluding clusters with contigs that had high quality hits to mitochondrial, microbial, or human sequences (high quality = hits over 100 bp that were also over 50% of either the query or subject length). At least 60% identity was required to exclude mitochondrial and microbial contaminants, while at least 90% was required to exclude human sequences. Finally, top clusters were ordered by abundance for each species. The clusters that cumulatively accounted for at least 50% of the repetitiveness of top clusters within each species were manually curated using visual inspection of the RepeatExplorer output and contigs, tblastx (default parameters), and Tandem Repeats Finder (default parameters; Benson 1999). All clusters were manually curated for the smallest dataset, *Photinus pyralis*.

Top clusters were assigned to one of the following 10 repeat categories based on RepeatExplorer and manual annotations: 1. long terminal repeat (LTR), 2. long interspersed nuclear element (LINE), 3. DNA transposon (DNA), 4. rolling circle transposon (RC), 5. low complexity repeat, 6. simple repeat (short repeats of less than 20 bp), 7. tandem repeat (large repeats of more than 20 bp), 8. histone gene, 9. ribosomal gene, and 10. unknown repeat (no annotation) (Kapitonov & Jurka 2008; Wicker *et al.* 2007). In addition to these ten categories, an eleventh category comprised the sum of the bottom clusters ("low frequency" clusters). To examine patterns on a broader scale, another analysis was performed on a dataset created by assigning top cluster repeat categories to three broad groups: Class I TEs, or retrotransposons, (LTR, LINE), Class II DNA TEs (DNA, RC), and repeats (categories 5 - 9). The percentage of repetitive DNA present in each top cluster was tabulated for each species using R. To determine if there was a relationship between repeat composition and phylogenetic divergence across species, both the total number of top clusters shared and the phylogenetic distance between all pairs of taxa were calculated.

Repetitive element analysis

The distribution of abundance (percent of the genome sample) for each repeat category (or group) was assessed for normality using a Shapiro-Wilk test in R with Benjamini-Hochberg correction for multiple comparisons. In addition to the full analysis that included all species and all categories of top cluster repeats, analyses were performed on subsets of the data, removing an outlier taxon (*Photinus obscurellus*), an outlier cluster (CL1), or species with especially low coverage (less than 0.008x). Analyses of the reduced datasets were qualitatively identical and so only the results of the high coverage dataset are discussed. Linear regression was used to

examine the correlation between genome size and repetitiveness; including total repetitiveness the repetitiveness of top clusters, and each repeat category/group separately.

Genome size did not require transformation for normality. Among repeat categories, four did not require transformation (LTR, LINE, DNA, unknown), one (ribosomal elements) was log transformed, and the others (RC, low complexity, simple repeat, tandem repeat, and histone sequences) were not normal, regardless of the transformation utilized. Therefore, both parametric and nonparametric tests were used to explore the relationships between repeat categories and genome size. Analysis of variance (ANOVA) was performed (full model = all categories as random effects), assuming that ANOVA is robust to departures from normality. Model simplification was performed by calculating AICc values for each model and dropping terms in a stepwise fashion following the procedure of Crawley (2002). The best model(s) was determined by AICc weights and the statistical significance of retained parameters examined. For non-normally distributed traits, nonparametric regression with Spearman's rho was used. These analyses were performed in JMP. Variables were assessed for phylogenetic signal and PICs conducted as appropriate following the methods for independent contrasts described above.

Results

Genome size varies six-fold across North American firefly genera

Estimates obtained from a total of 151 specimens of 23 species across seven genera showed a six-fold variation in genome size across North American Lampyrids (range: 409 – 2572 Mb, Figure 3.2, Appendix B, Table S3). Approximately 72% of the genome size variation occurred at the genus level, 28% at the species level, and sex was a significant predictor of genome size ($p < 0.0001$, Appendix B, Text S1). Significant genome size differences between

the sexes were noted in three of the seven species for which we had measurements for at least two individuals of each sex: *Photinus curtatus* ($p = 0.007$, two-tailed), *Photinus macdermotti* ($p = 0.001$, two-tailed), and *Photinus marginellus* ($p = 0.004$, two-tailed; Appendix B, Text S2). The average difference in genome size between the sexes across all species ranged from 4 to 95 Mb. Fireflies generally have X0 sex determination (Dias *et al.* 2007) and, if the presence of the additional X in XX females accounts for the larger genome size than XO males, this result suggests that the size and content of the X chromosome varies among species.

Of the 30 total *Photuris* specimens measured, only the five individuals identified as *Photuris frontalis* were supported as a single species based on field-measured flash pattern and molecular sequences (data not shown). The remaining 25 specimens did not show concordance between species identity based on flash behavior and identity determined by molecular sequences, and genome size estimates did not show clustering by either flash pattern or DNA sequence (Appendix B, Figure S2). This is not surprising given that many *Photuris* have multiple flash patterns in their repertoire (Lloyd 1969) and the genus may represent a recent radiation (Stanger-Hall, personal communication). Genome size estimates within the unidentified *Photuris* specimens ranged from 2133 to 2572 Mb (Appendix B, Figure S2), with a standard deviation much larger than the maximum standard deviation of any other species in our dataset, suggesting that multiple species are present. Because of difficulty in distinguishing species, only *Photuris frontalis* was used in further comparative analyses. For repeat analysis using single specimens, the *Photuris* individual with the largest genome size estimate was also examined.

Cryptic lineages

While most within-species genome size variation could be attributed to sex, the almost two-fold difference within field-identified *Phausis reticulata* and *Pyropyga decipiens* specimens raised the possibility of cryptic species in these taxa. A molecular phylogeny revealed a cryptic *Phausis sp.* collected in Watkinsville, GA that is more closely related to the unlighted *Phausis inaccensa* than to other *Phausis reticulata* specimens, including specimens from a nearby site (21 km away) and voucher specimens from the Great Smokey Mountains National Park (241 km away)(Appendix B, Figure S3). In contrast, phylogenetic investigation of *Pyropyga decipiens* specimens suggested that these individuals comprise a single species with two genome size types, small (699 Mb) and large (1079 Mb)(Appendix B, Figure S4). Accordingly, the two *Phausis* were treated as separate species in comparative analyses. Because there were no observable morphological or genetic differences between the two *Pyropyga* genome size types, they were excluded from the comparative analyses of body size, proxies for metabolism, and proxies for effective population size. In the repeat analysis, the sequenced individuals of each genome size type were treated as separate lineages.

The evolutionary relatedness hypothesis

Blomberg's K values indicated that related species have genome sizes that are significantly more similar than expected by chance ($K = 0.63$, $p = 0.001$). Ancestral state reconstruction indicated that the most recent common ancestor of North American fireflies had a genome size of ~1200 Mb (Appendix B, Figure S5). There was a dramatic ~1 Gb expansion in *Photuris* lineages and large expansions and contractions of several hundred Mb (up to 2.6 fold)

within five of the six genera sampled. The exception was *Pyractomena*, in which the three species had similar genome sizes (789.57 - 768.04 Mb, a difference of less than 3%).

The physiological correlates hypotheses

We tested the prediction that large genome sizes are associated with large body sizes using correlations between PICs of pronotum width (a proxy for body size) and genome size for 21 species (excluding both *Pyropyga* and the unknown *Photuris*). Blomberg's K values revealed significant phylogenetic signal in both male and female pronotum width, though the signal was less than expected under the assumption of Brownian motion (log male pronotum width: K = 0.11, $p = 0.035$; log female pronotum width: K = 0.53, $p = 0.038$). Contrary to the prediction, neither male nor female body size was significantly correlated with genome size either before or after phylogenetic correction (PICs, male: r = -0.004, $p = 0.99$; female: r = -0.23, $p = 0.47$). We also tested two proxies for metabolism; signal mode (lighted versus unlighted fireflies) and female wing reduction. There was no significant relationship between genome size and signal mode when corrected for phylogenetic relatedness (16 lighted, 5 unlighted taxa; r = 0.06, p = 0.80). In addition, there was no significant relationship between genome size and female wing reduction (17 full-winged, 4 reduced-wing taxa; r = -0.06, p = 0.80).

The effective population size hypothesis

Blomberg's K analysis showed significant phylogenetic signal in π (K = 0.38, p = 0.023), but not range size (K = 0.38, p = 0.42). Without phylogenetic correction there was a significant negative relationship between range size and log genome size, but not π (Figure 3.3).

Neither relationship was significant when corrected for phylogeny (range size: $r = -0.48$, $p = 0.13$; π: $r = 0.11$, $p = 0.64$).

The repetitive element landscape in fireflies

We examined repetitive content using 454 whole genome shotgun sequences from 20 of the 23 species. We also included two *Pyropyga decipiens* specimens with differing genome size types- small (S) and large (L). Quality trimming and removal of microbial and mitochondrial contaminants resulted in a total of 1,094,914 reads used in the clustering analysis. Average depth of coverage (total number of nucleotides sequenced/genome size) ranged from 0.002 for *Photinus pyralis* to 0.109 for *Photinus indictus* (mean coverage across all species: 0.04). For the lowest coverage depth, genomic sequences present in at least 1000 copies in a genome are expected to show increased coverage in our dataset (2x) and thus be grouped into clusters by Repeat Explorer. For the highest sequencing depth, repeat sequences would only need to be present in at least 20 copies to be clustered. In total, 398,511 reads were grouped into 86,275 clusters, yielding a repetitiveness of 36% for the entire sample. Clusters were further screened for mitochondrial and microbial contaminants resulting in 85,863 clusters included in the final analysis. 1,548 clusters were identified as top clusters (representing at least 0.001% of reads within the sample) and annotated by RepeatMasker.

Total repetitive sequence (percentage of the sequences within a species clustered by Repeat Explorer) for each species ranged from 18 (*Photinus pyralis*) to 66 % (*Photinus obscurellus*)(Table 3.2). Surprisingly, in the high coverage dataset, repetitiveness ranged only 2-fold, whereas genome sizes ranged 5-fold. Total repetitive sequence was highly correlated with repetitiveness of top clusters (percentage of sequences within a species in top clusters) ($p <$

0.0001, R = 0.92), indicating that the repetitiveness of top clusters was an appropriate proxy for the total. In most species, bottom low frequency clusters made up the bulk of total repetitiveness (mean: 61% +/- 18%; Figure 3.4) and no single cluster was responsible for the majority of repetitiveness. The exception was in *Photuris sp*., in which one repeat cluster in the tandem repeat category represented 30% of the reads (Appendix B, Figure S1). Across species, top clusters in the unknown category were also abundant (mean: 17%). Class I TEs were, on average, more abundant than Class II (means: 8% and 5% respectively) and LINE elements were the most abundant TE category (Figure 3.4). The most abundant TE families (over 0.01% of the sample, averaged across species) for each TE category are shown in Table 3.3.

Among the top clusters, each was shared on average by 2.5 +/- 2.2 species. In contrast, a bottom cluster was shared on average by 1.2 +/- 0.4 species, indicating that bottom clusters are generally unique to a particular species (Appendix B, Text S3). A significant negative correlation between number of shared clusters and total branch length between species showed that more clusters were shared between closely related species (Spearman's rho = -0.34, $p < 0.0001$; Appendix B, Figure S6). Taken together, this suggests that many of the clusters identified in this analysis represent young repeats. The only cluster that was shared among all 20 species was a ribosomal (rRNA) gene sequence (Appendix B, Text S3).

Three outlier samples were identified based on total repetitiveness. *Photinus pyralis* had the lowest sequencing coverage and thus low frequency repeats would be substantially underestimated in this species. However, across species, sequencing effort (number of nucleotides sequenced) was not significantly correlated with total percent repetitive identified, but it was correlated with DNA, LINE, and LTR element abundance (DNA: p=0.03, R=0.22,

LINE: p=0.04, R-0.21, logLTR: p=0.03, R=0.23), suggesting that these TEs may be particularly underrepresented in *Photinus pyralis*.

The other two outlier species, *Photinus obscurellus* and *Photuris sp.* had greatly increased levels of unknown and tandem repeats, respectively, as compared to their sister taxa (Figure 3.5). We confirmed that the two species were not switched during library preparation by checking mitochondrial reads from the sample against mitochondrial sequences from known specimens. Several *Photinus obscurellus* clusters in the unknown category were examined during manual curation and none had high quality hits to known microbial contaminants suggesting that these clusters are in the *Photinus obscurellus* genome. Even though *Photuris sp*. had low coverage (0.004x; Table 3.2), it contained the most abundant cluster of any species, representing 29.7% of total sample repetitiveness. This cluster, unique to this species, was originally identified as an LTR; however, during manual curation, it was reassigned to the tandem repeat category since most of the RepeatMasker hits to LTR were short (less than 100 bp per read) and large tandem repeats were found across the assembled contigs.

The repetitive DNA hypothesis

Neither genome size, total repetitiveness, nor any of the broad repeat groups (Class I, Class II, repeats) showed significant phylogenetic signal in the high coverage dataset that included 17 species, after correcting for multiple comparisons (Appendix B, Table S4). Of the repeat categories, LTRs were the only category with significant signal ($K = 0.72$, $p = 0.001$), which is visible as phylogenetic clumping in Figure 3.5b. However, in the full dataset, DNA, and ribosomal sequences also had significant phylogenetic signal (Figure 3.5).

Genome size was not correlated with total percent repetitive (R -0.067, p = 0.97; Figure 3.6), nor any individual repeat group or category (Appendix B, Table S5). There was also no correlation between Class I and Class II element abundance (Appendix B, Figure S7).

Discussion

In this study, we documented genome size variation within and among at least 23 species of North American Lampyrids, and performed low coverage sequencing on 20 of them. We then used a molecular phylogeny, body size measurements, species biology, range sizes, and genetic diversity measures to comparatively test several proximate and ultimate hypotheses for genome size evolution.

Flow cytometry reveals a cryptic species and within species genome-size variation

Our genome size estimates led to the discovery of a cryptic species of *Phausis* from Watkinsville, Georgia that was confirmed using a molecular phylogeny. These specimens, which we term *Phausis sp.* WAT, grouped monophyletically with a congener, *Phausis inaccensa,* rather than with other members of *Phausis reticulata*. This pattern was consistent for both mitochondrial and nuclear markers indicating they are not hybrids, which was a possibility given that both *Phausis inaccensa* and *Phausis reticulata* males have been observed to attempt mating with heterospecific females in the field (L. Faust, personal communication). After closer examination, *Phausis sp.* WAT individuals can be distinguished morphologically from both *Phausis inaccensa* and *Phausis reticulata* (Appendix B, Text S4).

We also uncovered genome size variation across individuals of *Pyropyga decipiens,* with some having genomes of ~700 Mb and others of ~1080 Mb. Individuals were collected on the

same day on the same plants in the same field and available genetic data does not distinguish the two size types, so we are confident they represent individuals of the same species. However, we were not able to confirm this hypothesis because we were not expecting differences in genome size within a single population and our sample sizes across genome size types were low. Future collections of *Pyropyga decipiens* from this population and others are needed to validate this result and determine the extent of genome-size variation within this species. It will also be interesting to determine if the genome size variation is correlated with phenotypic variation and assortative mating.

Genome-size variation within a species opens up the intriguing questions of how and why this variation occurs. Some firefly species are known to have supernumerary B chromosomes (*Photinus pyralis, Pyractomena angulata*, *Aspisoma laterale*, reviewed in Dias *et al.* 2007). B chromosomes are selfish, unessential chromosomes, generally made of repetitive DNA, that are polymorphic across and within populations (Houben *et al.* 2014). The contribution of B chromosomes to genome size can be substantial. For example, B chromosomes in rice can contribute to a 155% increase in the DNA content of cells (Jones & Rees 1982). In fireflies, the extent of variation in size and number of B chromosomes across populations and species remains largely uninvestigated. In other beetle species, the number of B chromosomes can be quite variable (e.g. 0 to 2 in *Psylliodes dulcamarae*: Segarra & Petitpierre 1989; 3 to 9 in *Bubas bubalu*: Angus *et al.* 2007). However, while *Pg. decipiens* has not yet been karyotyped, we found no evidence for a concomitant increase in repetitive DNA between the two genome size types (S and L) to suggest that B chromosomes play a role.

Genome size variation within fireflies is substantial and not caused by WGD

The firefly species measured in this study ranged 6-fold in genome size (409-2572 Mb). In previous work, most Coleopteran species have 154-2650 Mb genome sizes (Hanrahan & Johnston 2011; Gregory 2015), though two species substantially exceed this range: a leaf beetle (*Chrysolina carnifex:* 3610 Mb; Petitpierre *et al.* 1993), and a weevil (*Aramigus tessellatus*: 3246 and 4909 Mb; Normark 1996). In these lineages, chromosomal counts correlate with genome size, and large genomes are associated with polyploidy and parthenogenesis. In contrast, firefly karyotypes are fairly constant and there are no known examples of parthenogenesis. Of the 26 species that have been karyotyped, 23 have nine autosomes and X0 sex determination (Smith 1953; Wasserman & Ehrman 1986; Dias *et al.* 2007). The three exceptions, which are Asian and South American taxa, have a reduced number of autosomes (seven or eight), likely due to fusion events. In one species this is accompanied by the development of a neoXY system (Dias *et al.* 2007). Instead of chromosome number, changes in genome size could be correlated with changes in chromosome size. Chromosome size remains to be investigated across firefly species.

The fold variation we observe in fireflies is on par with what has been observed in other beetle families: 11-fold (166-1936 Mb) in Chrysomelidae (N=64), 9-fold (185-1672 Mb) in Coccinellidae (N=31), and 5-fold (156-850 Mb) in Tenebrionidae (N=69)(Animal Genome Size Database; accessed Feb. 24, 2015). Thus fireflies, like other beetle families, show substantial variation in genome size.

Based on the empirical distribution of genome size across insects, it has been hypothesized that there is an upper limit of ~2 Gb for genome size in insects that undergo metamorphosis (Gregory 2002b). However, this study adds additional taxa that have genome

sizes that are inconsistent with this hypothesis. All of the *Photuris* specimens measured, which represent at least two species, exceeded the hypothesized 2 Gb threshold (Figure 3). Genome sizes over 2 Gb have also been observed in closely related bioluminescent taxa (Elateridae, Phengodes; Hanrahan & Johnston 2011).

No evidence for selection using physiological correlates

Variation in firefly biology sets up several specific expectations for correlates with genome size in this family. As observed at broad taxonomic levels, we expected a positive correlation between genome size and body size, due to the nucleotypic effects of cell size/volume (reviewed in Gregory 2001). In addition, previous work has shown negative correlations between genome size and proxies metabolic activity in birds (Wright *et al.* 2014). Many firefly species use nocturnal lighted displays as mating signals, with flying males signaling to females in the vegetation (Lloyd 1966). In contrast to lighted species, some fireflies are unlighted—they have lost light, are diurnal, and use long-distance pheromones to find mates (e.g. *Phosphaenus hemipterus:* DeCock & Matthysen 2005). If the energetic costs due to signal production or searching for mates are higher for lighted versus dark fireflies, then lighted species may have a higher metabolic rate and genome size would be expected to be smaller (Reinhold 1999). Firefly species also differ in female morphology, with wings reduced or absent in females of some species (brachyptery, neoteny), eliminating their ability to fly (reviewed in South *et al.* 2011). Species with females that do not fly may have lower resting metabolic rates and genome size would be expected to be larger. Similarly, flying insects have higher resting metabolic rates and are expected to have smaller genome sizes (Reinhold 1999).

We tested these hypotheses by determining whether genome size in fireflies was correlated with body size and two proxies for metabolism (light production and reduced wings in females). We found no evidence for a correlation between genome size and body size, light production or wing reduction in females. A positive correlation between body size and genome size has been noted in other insects, but studies of beetles have documented either negative or, like this study, no correlation (e.g. Tenebrionidae: Juan & Petitpierre 1991; Palmer *et al.* 2003; Coccinellidae: Gregory *et al.* 2003; Chrysomelidae: Petitpierre & Juan 1994). Metabolic rates do increase slightly (2-3%) during flashing behavior (Woods Jr. *et al.* 2007), but there is no significant difference in resting metabolic rate between lighted and unlighted species (N = 2 lighted, 2 unlighted). However, flight is expensive, representing 20-30% of the total energy budget of an individual adult firefly (Woods Jr. *et al.* 2007). It is thus perhaps surprising that there is no correlation between female wing reduction and genome size. One explanation is that our sample size was small, including only 4 species with flightless females. Flightless females have evolved several times independently in fireflies (Branham & Wenzel 2003) and so sample size could be increased in future studies. In addition, there are no data on resting metabolic differences between species with flighted versus flightless females, which would allow a more definitive prediction. Because males must still fly to search for females in both cases, resting metabolic rates may not significantly differ between these two types, in which case we would not expect genome-size differences.

Population size and genome size

We found a significant correlation between range size, as a proxy for effective population size, and genome size without correcting for phylogeny. Phylogenetic correction is likely

unnecessary for this dataset since range size data was limited to species in a single genus, *Photinus*, and no significant phylogenetic signal was observed. The pattern observed with range size suggests that effective population size does play a role in firefly genome size evolution, at least in this genus, with inefficient selection in species with small effective population sizes leading to the accumulation of large genomes (Lynch & Conery 2003). In the future, it will be interesting to see if this trend holds with the addition of other North American taxa, since the largest genomes we observed occurred outside of *Photinus*.

In contrast, we did not find a similar relationship between sequence diversity measures from the *COI* locus and genome size. One possible explanation for this lack of correlation is that the *COI* data did not have enough variation to generate an accurate picture of genetic diversity across species. While *COI* has been used with great success in determining the phylogenetic relationships across firefly species (e.g. Stanger-Hall *et al.* 2007, Stanger-Hall 2015), its utility in detecting diversity and population structure within species has been limited (e.g. Lee *et al.* 2003). We also had substantial variation in the number of samples across species (2-71), based on the availability of sequences in the database. In addition, as a mitochondrial gene, *COI* represents a single locus with a unique evolutionary history that may not reflect that of the organism as a whole (Ballard & Whitlock 2004). Future studies using multiple independent nuclear markers will provide a better estimate of effective population size across species.

Repetitive DNA

Using low-coverage 454 sequencing, we generated a comparative view of the repeat landscape across 20 firefly species. Repetitiveness across species ranged from 24 to 39% in high-quality samples (high coverage and few suspected contaminants). While the low-coverage

sequencing scheme used for this study underestimates total genome repetitiveness, these estimates are around the range of those found for the two beetle species with published genomes: *T. castaneum* (30%; Wang, et al. 2008) and *D. ponderosae* (17-23%; Keeling, et al. 2013).

Unlike in some plant species (e.g. Estep, et al. 2013), the repetitive landscape was generally not dominated by any single transposable element class or family, even in species with extremely large genome sizes. Instead, it was dominated by "bottom" low frequency repeats. Low frequency is somewhat of a misnomer since the coverage we obtained across species allowed us to detect repeats present in at least 1000 copies in the genome.

Surprisingly, we did not find a correlation between amount of repetitive DNA and genome size in fireflies when we examined either total repetive DNA, or the abundance of any category or group of repetitive DNA. This is especially curious given the large genome sizes in *Photuris* species. Since RepeatExplorer, and many other repeat-identification programs, are based on sequence similarity, old elements will not be detected. If there was a genome size expansion due to a transposable element early on in *Photuris* evolution, it is possible we would not be able to detect it because of mutational decay over time.

Numerous studies have shown that repetitive DNA contributes substantially to the genomes of many organisms (e.g. Black & Rai 1988; Kidwell 2002; Wang *et al*. 2008; Metcalfe *et al.* 2012; Estep *et al.* 2013). The lack of correlation between repetitive DNA and genome size in fireflies suggests that (a) our genome size estimates are incorrect or (b) our method of detecting repeats produced inaccurate measures, or (c) repeats were active long ago in fireflies and it is difficult to detect them because of divergence. However, for four species, the estimates of genome size obtained in this study correctly predicted the depth of coverage we obtained when doing genomic sequencing (Sander and Hall, *in press*), indicating that our size estimates

based on flow cytometry are accurate. Also, we also used a different method to detect repeats (Assisted Automated Assembler of Repeat Families; DeBarry *et al.* 2008) and obtained similar patterns in total repeat abundance across species. Instead, the repeats we detected in our low coverage sequencing may represent old divergent elements that are hard to detect. Future studies will require deeper sequencing depth and more aggressive annotation techniques to fully characterize the repeat landscape in fireflies.

Conclusion

In summary, our results provide no support for positive selection explanations for genome size evolution based on physiological correlates. Rather, the selection/drift barrier explains the relationship between genome size and range size. Intriguingly, the expectation of drift means that active repeats, even if mildly deleterious, will be able to proliferate in the genomes of species with small effective population sizes. However, the amount of repetitive DNA we detected cannot account for the observed 6-fold variation in genome size across the family. It will be interesting to see if this pattern holds in other beetle families with genome size variation.

Fireflies have been an important system for studies on the evolution of mating signal evolution and speciation, bioluminescence, nuptial gifts, and aposematism. The extensive genome-size variation present in the family suggest that fireflies will also be a valuable system for continued study of both the proximate and ultimate causes of genome size evolution within and across species.

Acknowledgements

The authors would like to thank Kathrin Stanger-Hall for the use of KSH specimens and Jim Lloyd and Lynn Faust for identification of many the specimens used in body size measurements. The authors would also like to thank the following people and organizations for collection assistance and permission: Allegheny National Forest, Ashley Brown, Megan Behringer, Tom Brightman (Longwood Gardens), the David Fisk family, the Entomological Society of Pennsylvania, Great Smoky Mountains National Park (permit GRSM-2011-SCI-0049), Illinois Department of Natural Resources (permit NH12.5615), Indiana University, Michael Marsh (Whitehall Experimental Forest, University of Georgia), Jerry McCollum (Wharton Conservation Center), David McNaughton (Fort Indiantown Gap), Jenna Pallansch, David Queller, Willem Roosenberg, State of Tennessee Department of Environment and Conservation (permit 2012-16), Tonya Saint John, Kevin Smith (Tyson Research Center), Joan Strassman, and Dorset Trapnell. This work was supported by a National Science Foundation Graduate Research Fellowship [SES]; a National Science Foundation Dissertation Improvement Grant [DEB-1311315 to DWH and SES]; and an award from the National Institute of General Medical Sciences of the National Institute of Health [award number T32GM007103 to SES].

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403-410.

Angus R, Wilson C, Mann D (2007) A chromosomal analysis of 15 species of Gymnopleurini, Scarabaeini and Coprini (Coleoptera: Scarabaeidae). *Tijdschrift voor Entomologie*, **150**, 201-211.

Arendt J (2007) Ecological correlates of body size in relation to cell size and cell number: patterns in flies, fish, fruits and foliage. *Biological Reviews*, **82**, 241-256.

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

Aronesty E (2011) ea-utils: "Command-line tools for processing biological sequencing data"; http://code.google.com/p/ea-utils

Bae JS, Kim I, Sohn HD, Jin BR (2004) The mitochondrial genome of the firefly, *Pyrocoelia rufa*: complete DNA sequence, genome organization, and phylogenetic analysis with other insects. *Molecular Phylogenetics and Evolution*, **32**, 978-985.

Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. *Molecular Ecology*, **13**, 729-744.

Beaulieu JM, Leitch IJ, Patel S, Pendharkar A, Knight CA (2008) Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytologist*, **179**, 975-986.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, **57**, 289-300.

Bennett MD, Leitch IJ (2012) Plant DNA C-values database (release 6.0, Dec. 2012)[cited 2015 February 27]. Available from: http://www.kew.org/cvalues/

Benson G (1999) Tandem Repeats Finder: a program to analyze DNA sequences. *Nucleic Acids Research*, **27**, 573-580.

Black WC, Rai KS (1988) Genome evolution in mosquitoes: intraspecific and interspecific variation in repetitive DNA amounts and organization. *Genetics Research*, **51**, 185-196.

Blomberg SP, Garland T, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, **57**, 717-745.

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

Cannon SB, McKain MR, Harkess A, Nelson MN, Dash S, Deyholos MK, *et al.* (2015) Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Molecular Biology and Evolution*, **32**, 193-210.

Cavalier-Smith T (1978) Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *Journal of Cell Science*, **34**, 247-278.

Connolly JA, Oliver MJ, Beaulieu JM, Knight CA, Tomanek L, Moline MA (2008) Correlated evolution of genome size and cell volume in diatoms (Bacillariophyceae). *Journal of Phycology*, **44**, 124-131.

Cratsley CK (2004) Flash signals, nuptial gifts and female preference in Photinus fireflies. *Integrative and Comparative Biology*, **44**, 238-241.

Cratsley CK, Lewis SM (2005) Seasonal variation in mate choice of *Photinus ignitus* fireflies. *Ethology*, **111**, 89-100.

Crawley MJ (2002) *Statistical Computing: An Introduction to Data Analysis using S-Plus*. Hoboken, NJ: John Wiley and Sons.

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, **9**, 772-772.

Day JC, Goodall TI, Bailey MJ (2009) The evolution of the adenylate-forming protein family in beetles: multiple luciferase gene paralogues in fireflies and glow-worms. *Molecular Phylogenetics and Evolution,* **50**, 93-101.

DeCock R, Matthysen E (1999) Aposematism and bioluminescence: experimental evidence from glow-worm larvae (Coleoptera: Lampyridae). *Evolutionary Ecology*, **13**, 619-639.

DeCock R, Matthysen E (2005) Sexual communication by pheromones in a firefly, *Phosphaenus hemipterus* (Coleoptera: Lampyridae). *Animal Behaviour*, **70**, 807-818.

DeBarry JD, Liu R, Bennetzen JL (2008) Discovery and assembly of repeat family pseudomolecules from sparse genomic sequence data using the Assisted Automated Assembler of Repeat Families (AAARF) algorithm. *BMC Bioinformatics*, **9**, 235.

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

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, **29**, 1969-1973.

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**.

Estep MC, DeBarry JD, Bennetzen JL (2013) The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. *Heredity*, **110**, 194-204.

Felsenstein J (1985) Phylogenies and the comparative method. *American Naturalist*, **125**, 1-15.

Fender KM (1966) The genus Phausis in America north of Mexico (Coleoptera-Lampyridae). *Northwest Science*, **40**, 83-95.

Ferrari JA, Rai KS (1989) Phenotypic correlates of genome size variation in *Aedes albopictus*. *Evolution*, **43**, 895-899.

Fraga H (2008) Firefly luminescence: a historical perspective and recent developments. *Photochemical & Photobiological Sciences*, **7**, 146-158.

Gaut BS, Ross-Ibarra J (2008) Selection on major components of angiosperm genomes. *Science*, **320**, 484-486.

Glasauer SK, Neuhauss SF (2014) Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular Genetics and Genomics*, **289**, 1045-1060.

Green J (1957) Revision of the nearctic species of Pyractomena (Coleoptera: Lampyridae). *Wasmann Journal of Biology*, **15**, 237-284.

Green JW (1956) Revision of the Nearctic species of Photinus (Lampyridae: Coleoptera). *Proceedings of the California Academy of Sciences*, **28**, 561-613.

Gregory TR (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biological Reviews*, **76**, 65-101.

Gregory TR (2002a) A bird's-eye view of the c-value enigma: genome size, cell size, and metabolic rate in the class aves. *Evolution*, **56**, 121-130.

Gregory TR (2002b) Genome size and developmental complexity. *Genetica*, **115**, 131-146.

Gregory TR (2005) *The Evolution of the Genome*. Burlington, MA: Elsevier Inc.

Gregory TR (2015) *Animal Genome Size Database*. http://www.genomesize.com.

Gregory TR, Johnston JS (2008) Genome size diversity in the family Drosophilidae. *Heredity*, **101**, 228-238.

Gregory TR, Nedved O, Adamowicz SJ (2003) C-value estimates for 31 species of ladybird beetles (Coleoptera: Coccinellidae). *Hereditas*, **139**, 121-127.

Haldane, J. B. S. 1927. A mathematical theory of natural and artificial selection, Part V: selection and mutation. *Mathematical Proceedings of the Cambridge Philosophical Society*, **23**, 838-844.

Hanrahan S, Johnston JS (2011) New genome size estimates of 134 species of arthropods. *Chromosome Research*, **19**, 809-823.

Houben A, Banaei-Moghaddam A, Klemme S, Timmis J (2014) Evolution and biology of supernumerary B chromosomes. *Cellular & Molecular Life Sciences*, **71**, 467-478.

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Research*, **9**, 868-877.

Jacobson AL, Johnston JS, Rotenberg D, Whitfield AE, Booth W, Vargo EL, *et al.* (2013) Genome size and ploidy of Thysanoptera. *Insect Molecular Biology*, **22**, 12-17.

Jones RN, Rees H (1982) *B Chromosomes*. Chicago, IL: Academic Press.

Juan C, Petitpierre E (1991) Evolution of genome size in darkling beetles (Tenebrionidae, Coleoptera). *Genome*, **34**, 169-173.

Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics*, **9**, 411-412.

Keeling CI, et al. 2013. Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. Genome Biology 14: R27-R27.

Keeling CI, Yuen MM, Liao NY, Docking TR, Chan SK, Taylor GA, *et al.* (2013) Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. *Genome Biology*, **14**, R27.

Kidwell MG (2002) Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, **115**, 49-63.

Kullman B, Tamm H, Kullman K (2005). *Fungal Genome Size Database.* http://www.zbi.ee/fungal-genomesize

Lee SC, Bae JS, Kim I, Suzuki H, Kim SR, Kim JG, *et al.* (2003) Mitochondrial DNA sequence-based population genetic structure of the firefly, *Pyrocoelia rufa* (Coleoptera: Lampyridae). *Biochemical Genetics*, **41**, 427-452.

Lewis SM, Cratsley CK (2003) Female preference for male courtship flashes in *Photinus ignitus* fireflies. *Behavioral Ecology*, **14**, 135-140.

Lewis SM, Cratsley CK (2008) Flash signal evolution, mate choice, and predation in fireflies. *Annual Review of Entomology*, **53**, 293-321.

Lewis SM, Cratsley CK, Demary K (2004a). Mate recognition and choice in Photinus fireflies. In. *Annales Zoologici Fennici*. Finnish Zoological and Botanical Publishing Board, pp. 809-821.

Lewis SM, Cratsley CK, Rooney JA (2004b). Nuptial gifts and sexual selection in Photinus fireflies. *Integrative and Comparative Biology*, **44**, 234-237.

Lewis SM, Faust L, DeCock R (2012) The dark side of the light show: predators of fireflies in the Great Smoky Mountains. *Psyche: A Journal of Entomology*.

Li X, Ogoh K, Ohba N, Liang X, Ohmiya Y (2007) Mitochondrial genomes of two luminous beetles, *Rhagophthalmus lufengensis* and *R. ohbai* (Arthropoda, Insecta, Coleoptera). *Gene*, **392**, 196-205.

Librado P, Rozas J (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451-1452.

Lloyd JE (1966) Studies on the flash communication system in Photinus fireflies. In. *Miscellaneous Publications*. Ann Arbor, Michigan: Museum of Zoology, University of Michigan.

Lloyd JE (1969) Flashes of Photuris fireflies: their value and use in recognizing species. *The Florida Entomologist*, **52**, 29-35.

Luk SPL, Marshall SA, Branham MA (2011) The fireflies of Ontario (Coleoptera: Lampyridae). *Canadian Journal of Athropod Identification*, **16**, 1-105.

Lynch M, Conery JS (2003) The origins of genome complexity. *Science*, **302**, 1401-1404.

Mable B (2004). 'Why polyploidy is rarer in animals than in plants': myths and mechanisms. *Biological Journal of the Linnean Society*, **82**, 453-466.

Mable BK, Alexandrou MA, Taylor MI (2011) Genome duplication in amphibians and fish: an extended synthesis. *Journal of Zoology*, **284**, 151-182.

Macas J, Neumann P, Navrátilová A (2007) Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics*, **8**, 427-427.

Maddison WP, Maddison DR (2011) Mesquite: a modular system for evolutionary analysis. Version 2.75.

McKain, MR, Wickett N, Zhang Y, Ayyampalayam S, McCombie WR, Chase MW, *et al.* (2012) Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). *American Journal of Botany*, **99**, 397-406.

Metcalfe CJ, Filée J, Germon I, Joss J, Casane D (2012) Evolution of the Australian lungfish (*Neoceratodus forsteri*) genome: a major role for CR1 and L2 LINE elements. *Molecular Biology and Evolution*, **29**, 3529-3539.

Midford PE, Garland T, Maddison WP (2008) PDAP Package of Mesquite. Version 1.15.

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press.

Normark BB (1996) Polyploidy of parthenogenetic *Aramigus tessellatus* (Say) (Coleoptera: Curculionidae). *The Coleopterists Bulletin*, **50**, 73-79.

Novák P, Neumann P, Macas J (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, **11**, 378.

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792-793.

Oba Y, Mori N, Yoshida M, Inouye S (2010) Identification and characterization of a luciferase isotype in the Japanese firefly, *Luciola cruciata*, involving in the dim glow of firefly eggs. *Biochemistry*, **49**, 10788-10795.

Ohba N (2004 Flash communication systems of Japanese fireflies. *Integrative and Comparative Biology*, **44**, 225-233.

Orr HA (1990) "Why polyploidy is rarer in animals than in plants" revisited. *American Naturalist*, **136**, 759-770.

Palmer M, Petitpierre E, Pons J (2003) Test of the correlation between body size and DNA content in Pimelia (Coleoptera: Tenebrionidae) from the Canary Islands. *European Journal of Entomology*, **100**, 123-129.

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**, 289-290.

Peters RH (1986) *The Ecological Implications of Body Size*. Cambridge University Press.

Petitpierre E, Juan C (1994) Genome size, chromosomes and egg-chorion ultrastructure in the evolution of Chrysomelinae. *Series Entomologica*, **50**, 213-225.

Petitpierre E, Segarra C, Juan C (1993) Genome size and chromosomal evolution in leaf beetles (Coleoptera, Chrysomelidae). *Hereditas*, **119**, 1-6.

Petrov DA (2001) Evolution of genome size: new approaches to an old problem. *TRENDS in Genetics*, **17**, 23-28.

R Development Core Team (2013) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org

Reinhold K (1999) Energetically costly behaviour and the evolution of resting metabolic rate in insects. *Functional Ecology*, **13**, 217-224.

Schielzeth H, Streitner C, Lampe U, Franzke A, Reinhold K (2014) Genome size variation affects song attractiveness in grasshoppers: evidence for sexual selection against large genomes. *Evolution*, **68**, 3629-3635.

Schneider CA, Rasband WS, Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, **9**, 671-675.

Segarra C, Petitpierre E (1989) Cytogenetic diversity of the European Psylliodes flea-beetles (Coleoptera, Chrysomelidae). *Hereditas*, **110**, 169-174.

Shapiro E, Lu C, Baneyx F (2005) A set of multicolored *Photinus pyralis* luciferase mutants for in vivo bioluminescence applications. *Protein Engineering Design and Selection*, **18**, 581-587.

Sheffield NC, Song H, Cameron SL, Whiting MF (2009) Nonstationary evolution and compositional heterogeneity in beetle mitochondrial phylogenomics. *Systematic Biology*, **58**, 381-394.

Shi T, Huang H, Barker MS (2010) Ancient genome duplications during the evolution of kiwifruit (Actinidia) and related Ericales. *Annals of Botany*, **106**, 497-504.

Shuter BJ, Thomas JE, Taylor WD, Zimmerman AM (1983) Phenotypic correlates of genomic DNA content in unicellular eukaryotes and other cells. *American Naturalist*, **122**, 26-44.

Smit AFA, Hubley R, Green P (2013-2015) *RepeatMasker Open-4.0*. http://www.repeatmasker.org

Smith SG (1953) Chromosome numbers of Coleoptera I. *Heredity*, **7**, 31-48.

South A, Stanger-Hall K, Jeng ML, Lewis SM (2011) Correlated evolution of female neoteny and flightlessness with male spermatophore production in fireflies (Coleoptera: Lampyridae). *Evolution*, **65**, 1099-1113.

Stanger-Hall KF, Lloyd JE (2015) Flash signal evolution in Photinus fireflies: Character displacement and signal exploitation in a visual communication system. *Evolution*, **69**, 666-682.

Stanger-Hall KF, Lloyd JE, Hillis DM (2007) Phylogeny of North American fireflies (Coleoptera: Lampyridae): implications for the evolution of light signals. *Molecular Phylogenetics and Evolution*, **45**, 33-49.

Stansbury MS, Moczek AP (2014) The function of Hox and appendage-patterning genes in the development of an evolutionary novelty, the Photuris firefly lantern. *Proceedings of the Royal Society of London B: Biological* Sciences, **281**, 20133333.

Swaminathan K, Varala K, Hudson ME (2007) Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. *BMC Genomics*, **8**, 132-132.

Tamura M, Yokoyama J, Ohba N, Kawata M (2005) Geographic differences in flash intervals and pre-mating isolation between populations of the Genji firefly, *Luciola cruciata*. *Ecological Entomology*, **30**, 241-245.

Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J (2011) Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biology and Evolutioni,* **3**, 219-229.

Thomas CA (1971) The genetic organization of chromosomes. *Annual Review of Genetics*, **5**, 237-256.

Vencl FV, Carlson AD (1998) Proximate mechanisms of sexual selection in the firefly *Photinus pyralis* (Coleoptera: Lampyridae). *Journal of Insect Behavior*, **11**, 191-207.

Vencl FV (2004) Allometry and proximate mechanisms of sexual selection in Photinus Fireflies and some other beetles. *Integrative and Comparative Biology,* 4, 242-249.

Wang S, Lorenzen MD, Beeman RW, Brown SJ (2008) Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome. *Genome Biology*, **9**, R61-R61.

Wasserman M, Ehrman L (1986) Firefly chromosomes, II (Lampyridae: Coleoptera). *The Florida Entomologist*, **69**, 755-757.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, *et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nature Reviews* Genetics, **8**, 973-982.

Wild AL, Maddison DR (2008) Evaluating nuclear protein-coding genes for phylogenetic utility in beetles. *Molecular Phylogenetics and Evolution*, **48**, 877-891.

Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, **15**, 1.

Woods Jr WA, Hendrickson H, Mason J, Lewis SM (2007) Energy and predation costs of firefly courtship signals. *The American Naturalist*, **170**, 702-708.

Wright NA, Gregory TR, Witt CC (2014) Metabolic 'engines' of flight drive genome size reduction in birds. *Proceedings of the Royal Society of London B: Biological Sciences*, **281**, 2013-2780.

Table 3.1. Collection dates and localities of specimens used in this study

Genus	*Species*	*N*[a]	*State(s) collected*	*Date(s)*
Photinus	*australis*	5	GA	June 2011
	brimleyi	2	TN	June 2012
	carolinus	7(5)	GA, PA	June 2011, June 2012
	cooki	(1)	TN	June 2012
	curtatus	5(2)	IL, OH	June 2012
	indictus	2	PA	June 2012
	macdermotti	10(6)	GA, PA	June 2011, April 2012, June 2012
	marginellus	6(5)	TN, PA	June 2012, July 2012
	obscurellus	3	PA	June 2012, July 2012
	pyralis	4(5)	GA, MO, MS, TN	June 2011, May 2012, June 2012
	sabulosus	3	OH	June 2012
	scintillans	6(1)	PA	June 2012
Ellychnia	*corrusca*	1	PA	June 2012
Pyropyga	*decipiens*	4(5)	PA	June 2012
Pyractomena	*angulata*	5	IN, MO	June 2012
	borealis	5	GA	March 2012
	marginalis	2	PA	June 2012
Lucidota	*atra*	5(5)	IL, OH, PA, TN	June 2012
	punctata	2	PA, TN	June 2012
Photuris	*frontalis*	5	GA	June 2011
	multiple sp.	17(8)[b]	IL, IN, MO, MS, PA, TN	June 2012
Phausis	*sp.*	2(5)	GA	March 2012
	reticulata	2	TN	June 2011

[a] Total number of males (females) per species

[b] A single unknown *Photuris* with the largest measured genome size was used in further analysis

Table 3.2. Repeat Explorer metrics

Species	Abv[a]	Genome Size (MB)[b]	% lost in QC	Reads[c]	Cov[d]	Exp[e]	Total Rep[f]	Top Clusters[g]		
								% Rep	N clusters	N curated
Photinus australis	Paust	1597.90	15.1	38348	0.016	16	38	10	191	17
Photinus brimleyi	Pbrim	1180.76	2.6	39510	0.021	21	31	10	190	21
Photinus carolinus	Pcaro	660.98	2.6	7876	0.007	7	28	8	75	2
Photinus cooki	Pcook	700.98	2.9	23217	0.021	21	28	8	156	8
Photinus indictus	Pindi	433.19	3.4	77037	0.109	109	36	15	275	26
Photinus macdermotti	Pmacd	508.99	2.3	32939	0.038	38	37	14	217	113
Photinus obscurellus	Pobsc	650.20	8.6	51506	0.055	55	66	57	373	64
Photinus pyralis	Ppyra	447.73	2.0	1554	0.002	2	18	3	15	15
Photinus sabulosus	Psabu	617.52	1.9	85682	0.091	91	30	8	205	17
Photinus scintillans	Pscin	1032.40	2.6	25880	0.017	17	34	13	386	52
Ellychnia corrusca	Ecorr	781.56	6.7	29161	0.024	24	24	7	99	5
Pyropyga decipiens S	PdeAS	700.92	2.1	64638	0.062	62	37	12	200	16
Pyropyga decipiens L	PdeBL	1079.61	2.4	40131	0.023	23	37	12	195	15
Pyractomena angulata	Pangu	789.57	2.9	48172	0.037	37	33	16	184	3
Pyractomena marginalis	Pmarg	768.04	25.2	50985	0.044	44	32	10	197	20
Lucidota atra	Latra	491.16	2.4	79663	0.103	103	37	15	201	14
Lucidota punctata	Lpunc	1300.06	1.9	149102	0.075	75	39	18	285	16
Photuris frontalis	Pfron	2154.37	1.7	72264	0.023	23	30	8	127	11
Photuris sp.	Pbigs	2460.56	3.9	19427	0.004	4	57	50	66	1
Phausis sp. WAT	PretW	1114.58	3.3	82203	0.046	46	35	16	127	9
Phausis reticulata	PretG	831.12	3.8	75619	0.058	58	31	10	143	9

[a] Four-letter species code
[b] Mean of average male and average female genome size estimates
[c] Total number of reads after qc and cleaning
[d] The total number of nucleotides sequenced/genome size
[e] Number of reads expected in sample if repeat has 1000 copies in the genome
[f] Total percent repetitive using all 85863 clusters
[g] Percent repetitive, number of clusters, and number of clusters manually curated in the top clusters (clusters that account for at least 0.001% of total reads)

Table 3.3. The most abundant (over 0.01% of the sample) families/subfamilies in each transposable element group (Class I and Class II) across species

Element	%*
Class I	
LTR	
Gypsy	0.717
Pao	0.031
Copia	0.012
DIRS	0.011
LINE	
Dong.R4	0.601
LOA	0.368
RTE.BovB	0.274
Penelope	0.228
I	0.133
Class II	
DNA	
Maverick	0.743
TcMar.Tc1	0.165
Crypton	0.127
TcMar.Mariner	0.114
RC	
Helitron	0.014

*Mean percent repetitive across species

Table 3.4. Chromosome counts from the literature in relationship to genome size estimates

Genus	Species	Chromosome #	Male Karyotype[a]	Genome size (Mb)[b]
Ellychnia	*corrusca*	19	9 + X	782
Photinus	*australis*	19	9 + X	1597
	macdermotti	19	8 + X	481
	pyralis	19 + S	9 + X	436
Pyractomena	*angulata*	19	9 + X+ S	789
	borealis	19	9 + X	775
Photuris	*pennsylvanica*	19	9 + X	2133-2572[c]
	congener	18 (sex?)		

[a] From data compiled in (Dias, et al. 2007)
[b] Species mean estimated from this study
[c] Range for *Photuris* specimens measured in this study

Fig. 3.1.

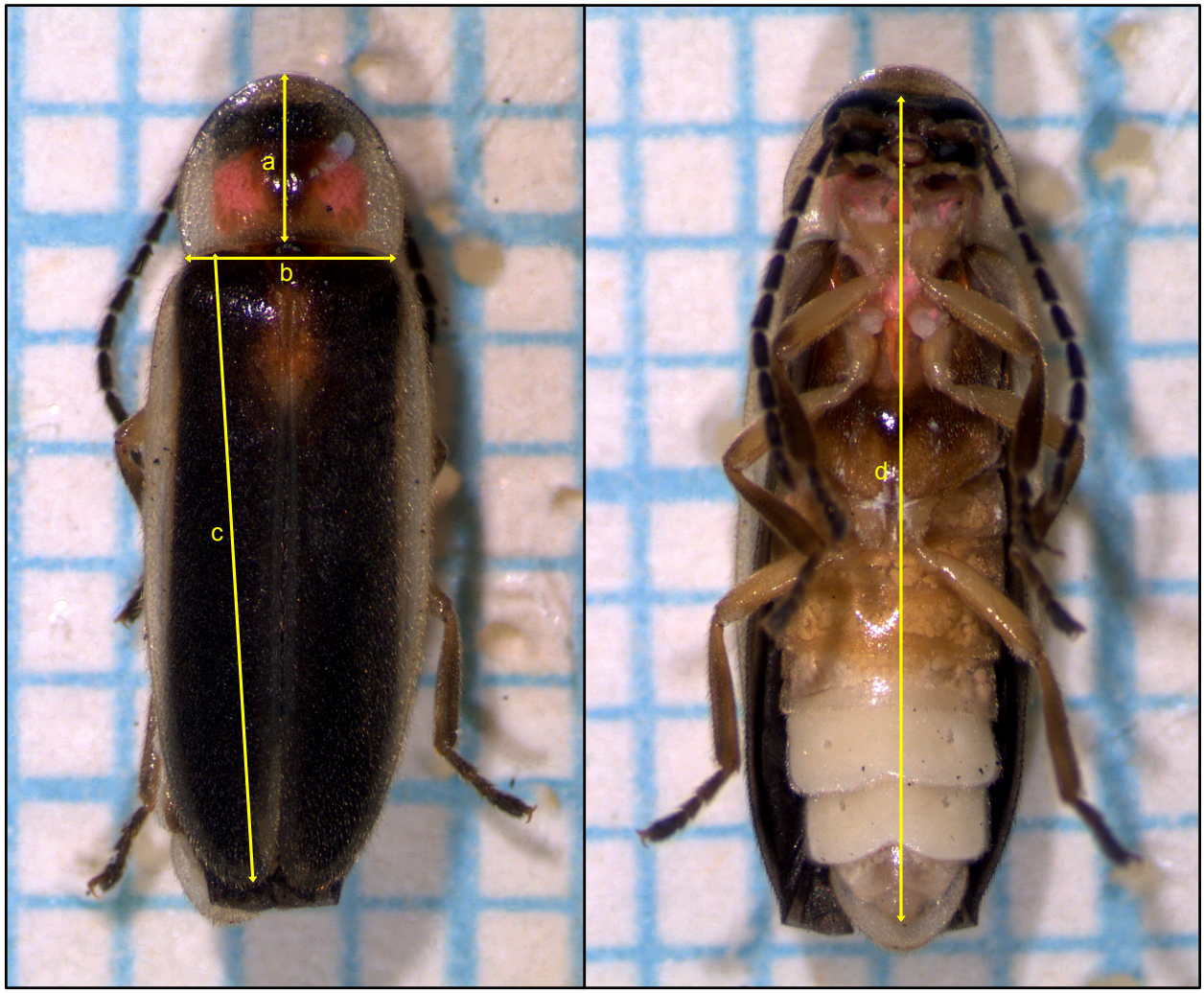


Figure 3.1. Body size measurements

Measurements were taken from photographs of ethanol-preserved specimens on a 1mm grid (pictured: KSH 9131, a male *Photinus curtatus*). All measures were highly correlated, thus only pronotum width was used in final analysis. Left: dorsal view. Right: ventral view. a) Pronotum length, b) pronotum width, c) elytron length, d) body length. The filled-in area of the pronotum was also measured.

Fig. 3.2.

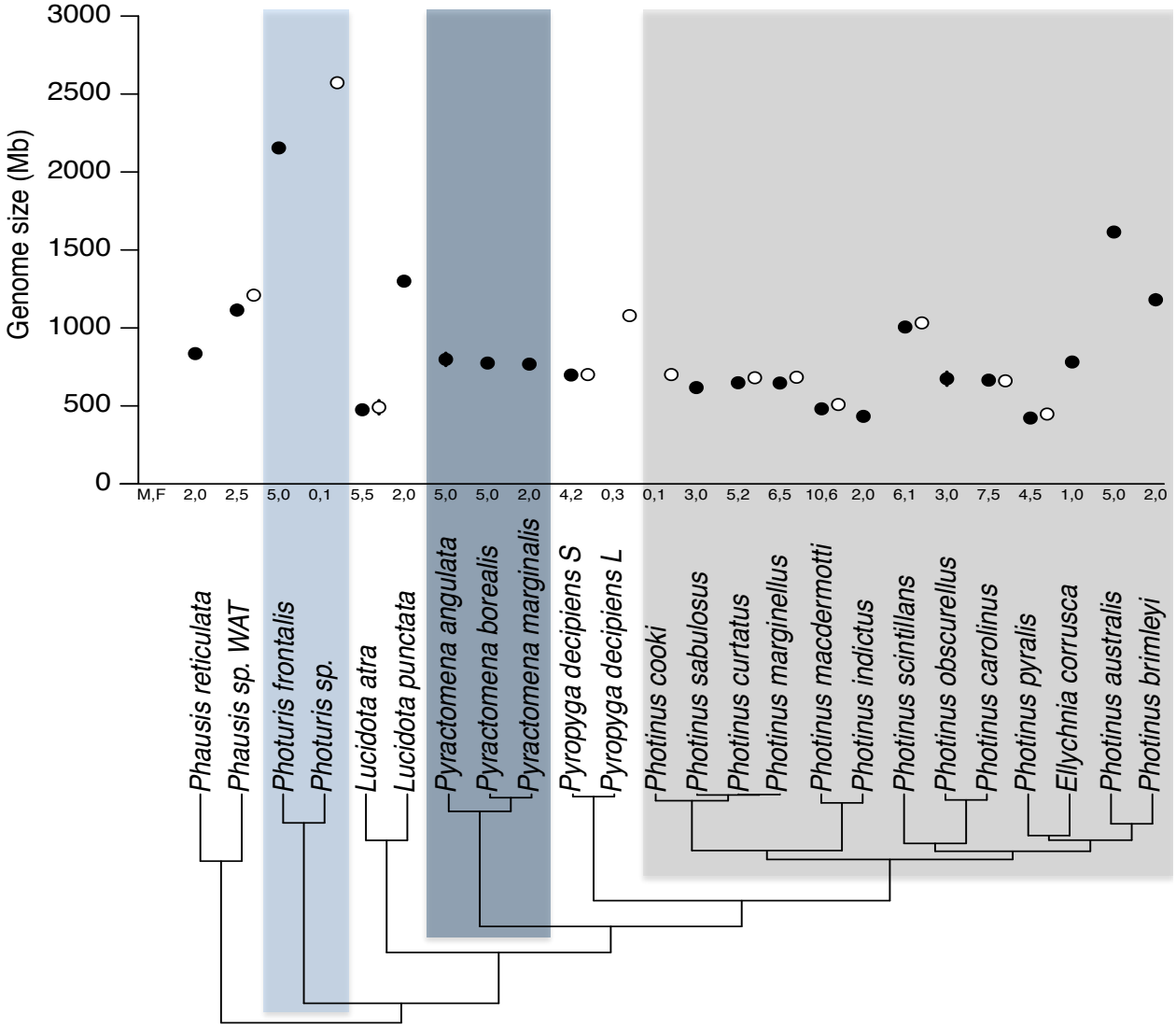


Figure 3.2. Genome size varies six-fold across 23 North American firefly species

Genome size ranges from 409 Mb (*Photinus pyralis*) to 2572 Mb (*Photuris sp.*) Bottom: An ultrametric molecular phylogeny generated from one mitochondrial and two nuclear loci with branch lengths proportional to relative time. Top: means and standard deviations for nuclear genome size (1C) estimates of males (filled circles) and females (empty circles) of each species. The values in the M, F row give the sample sizes for males and females, respectively. Where bars are not visible and multiple individuals were measured, the standard deviations are entirely covered by the mean circles. Across top and bottom, alternating shaded and empty boxes indicate divisions between genera.

Fig. 3.3.

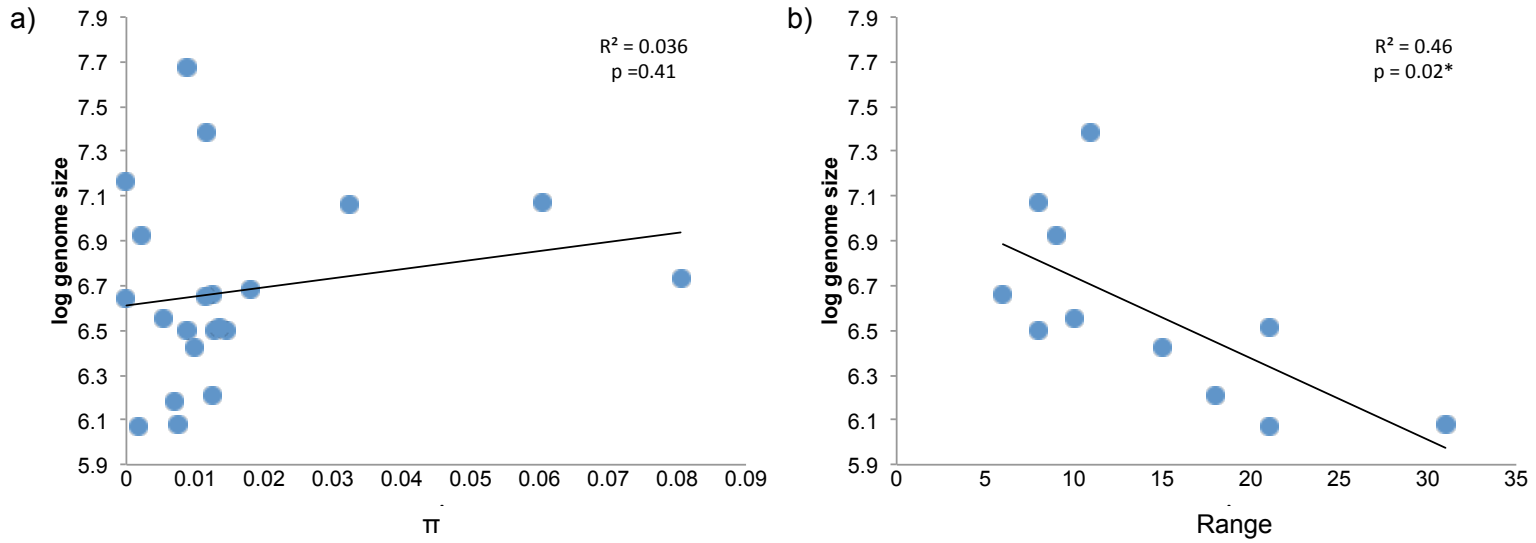


Figure 3.3. Relationship between genome size and proxies for effective population size

(a) Nucleotide diversity (π) shows a positive relationship that is not significant. (b) Range size (number of states with presence records) shows a significant, negative correlation. Values are not phylogenetically corrected.

Fig. 3.4.

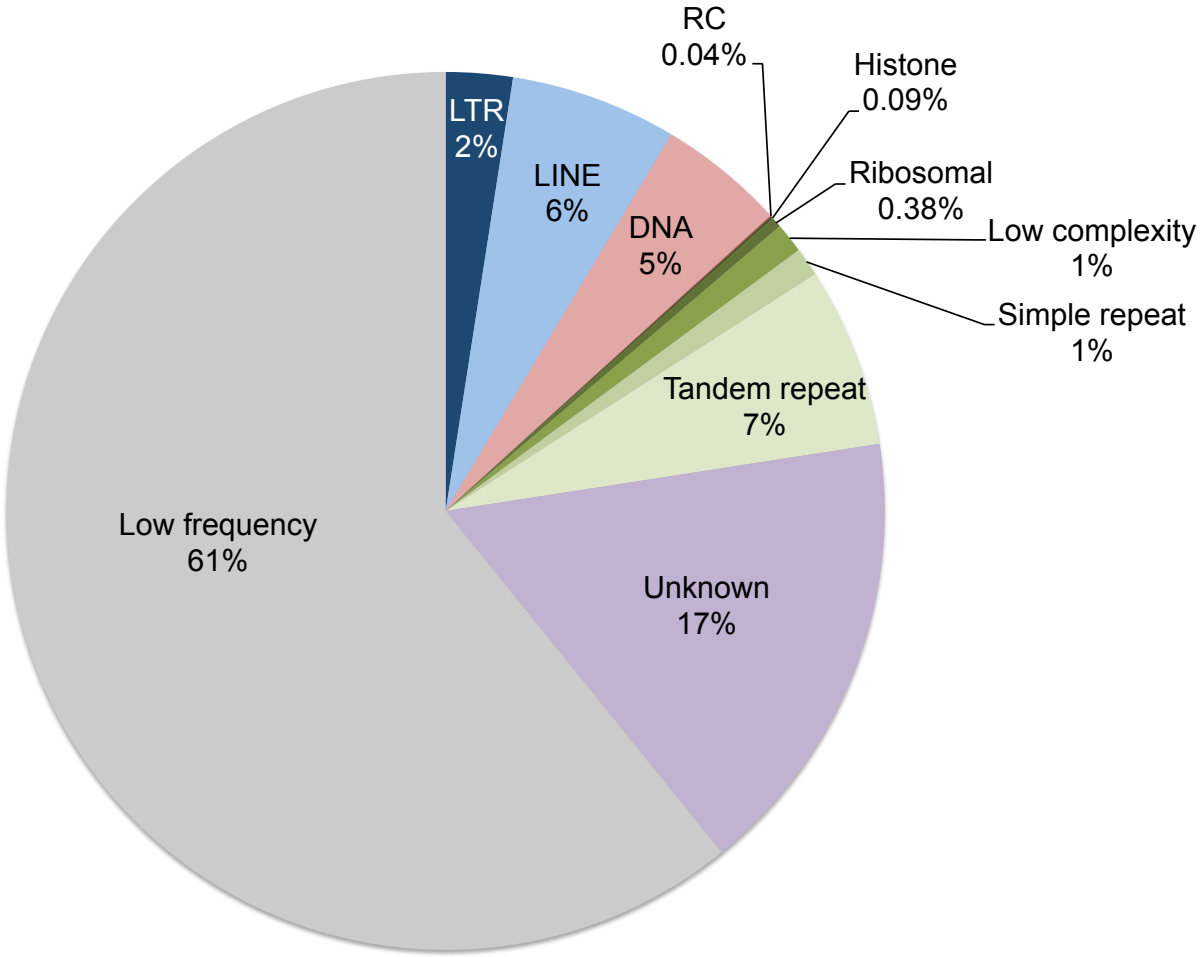


Figure 3.4. Average genomic composition of repetitive element categories

The mean percentage of genomic sample composition across species is given for each repetitive element category. Long terminal repeat (LTR), long interspersed nuclear element (LINE), DNA transposon (DNA), rolling circle (RC), low complexity, simple repeat (short repeats), tandem repeat (large tandem repeats ~100 bp or more), histone, ribosomal, unknown (no annotation), and low frequency (bottom clusters). Class I elements are shown in blues, Class II in reds, and other repeats in greens.

Figure 3.5. Repetitive element content across species

(a) Left: Ultrametric three-locus phylogeny with branch lengths in units of relative time. Middle: mean Genome size (Mb) per species (GS). Right: the percent repetitiveness of each sample, color-coded by repeat classification. Retrotransposons (Class I), DNA transposons (Class II), repeats, unknown top clusters and low frequency (bottom clusters). Total bar length is equal to total repetitiveness of the sample. *Photuris sp.* and *Photinus obscurellus* are clear outliers. *Photinus pyralis* is the least repetitive, likely because it had extremely low coverage in the dataset. (b-e) the contribution of 4 repeat orders: long terminal repeats (LTR), long interspersed nuclear elements (LINE), DNA transposons (DNA), and ribosomal repeats. In the full dataset LTR, DNA, and ribosomal repeats showed significant phylogenetic signal. In the high coverage dataset, only LTR had significant signal. Horizontal and vertical axes as in (a).

Fig. 3.5.

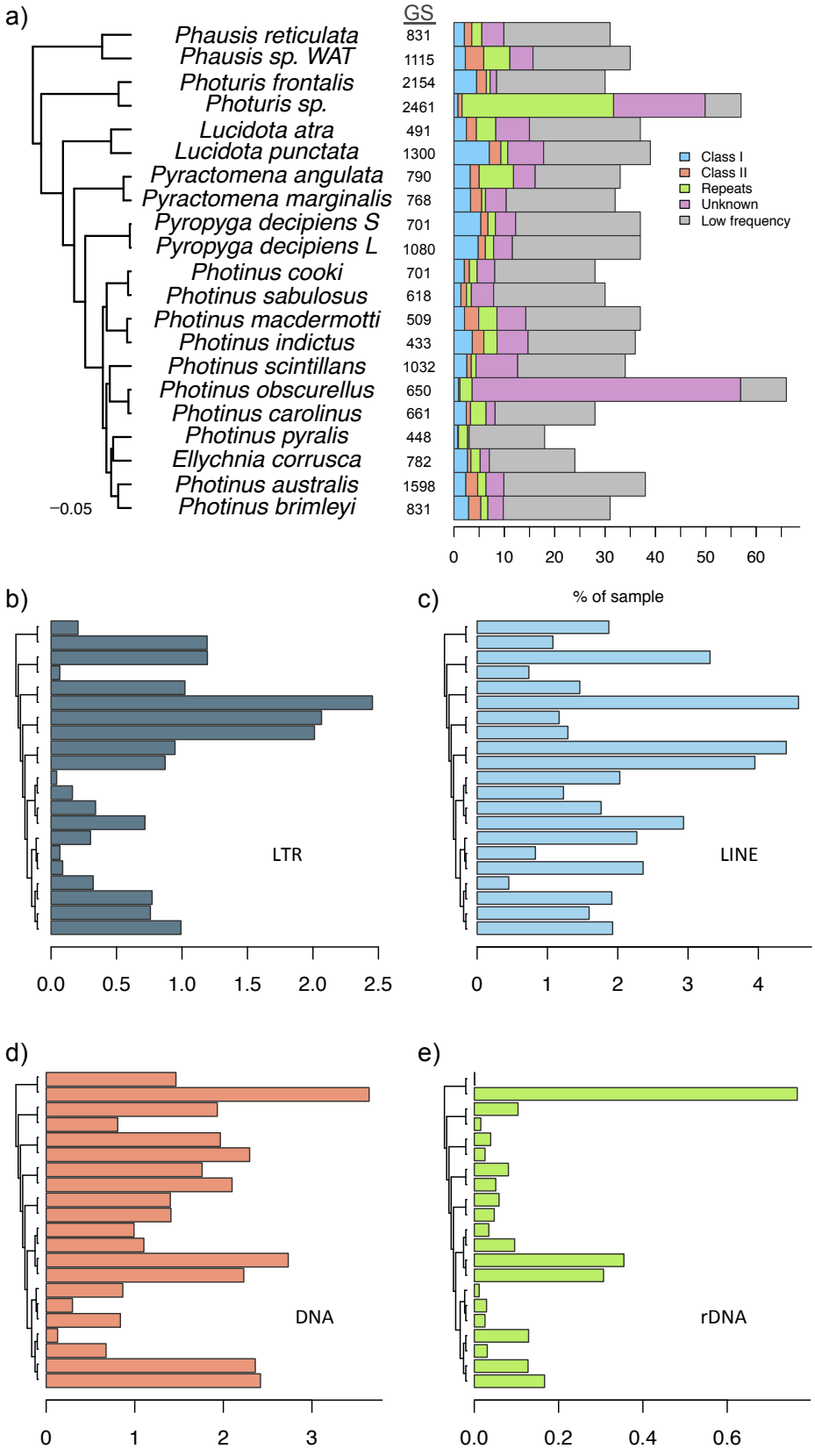

a)
GS
Phausis reticulata
Phausis sp. WAT
Photuris frontalis
Photuris sp.
Lucidota atra
Lucidota punctata
Pyractomena angulata
Pyractomena marginalis
Pyropyga decipiens S
Pyropyga decipiens L
Photinus cooki
Photinus sabulosus
Photinus macdermotti
Photinus indictus
Photinus scintillans
Photinus obscurellus
Photinus carolinus
Photinus pyralis
Ellychnia corrusca
Photinus australis
Photinus brimleyi
831
1115
2154
2461
491
1300
790
768
701
1080
701
618
509
433
1032
650
661
448
782
1598
831
−0.05
0
10
20
30
40
50
60
Class I
Class II
Repeats
Unknown
Low frequency
b)
LTR
0.0
0.5
1.0
1.5
2.0
2.5
c)
% of sample
LINE
0
1
2
3
4
d)
DNA
0
1
2
3
e)
rDNA
0.0
0.2
0.4
0.6

Fig. 3.6.

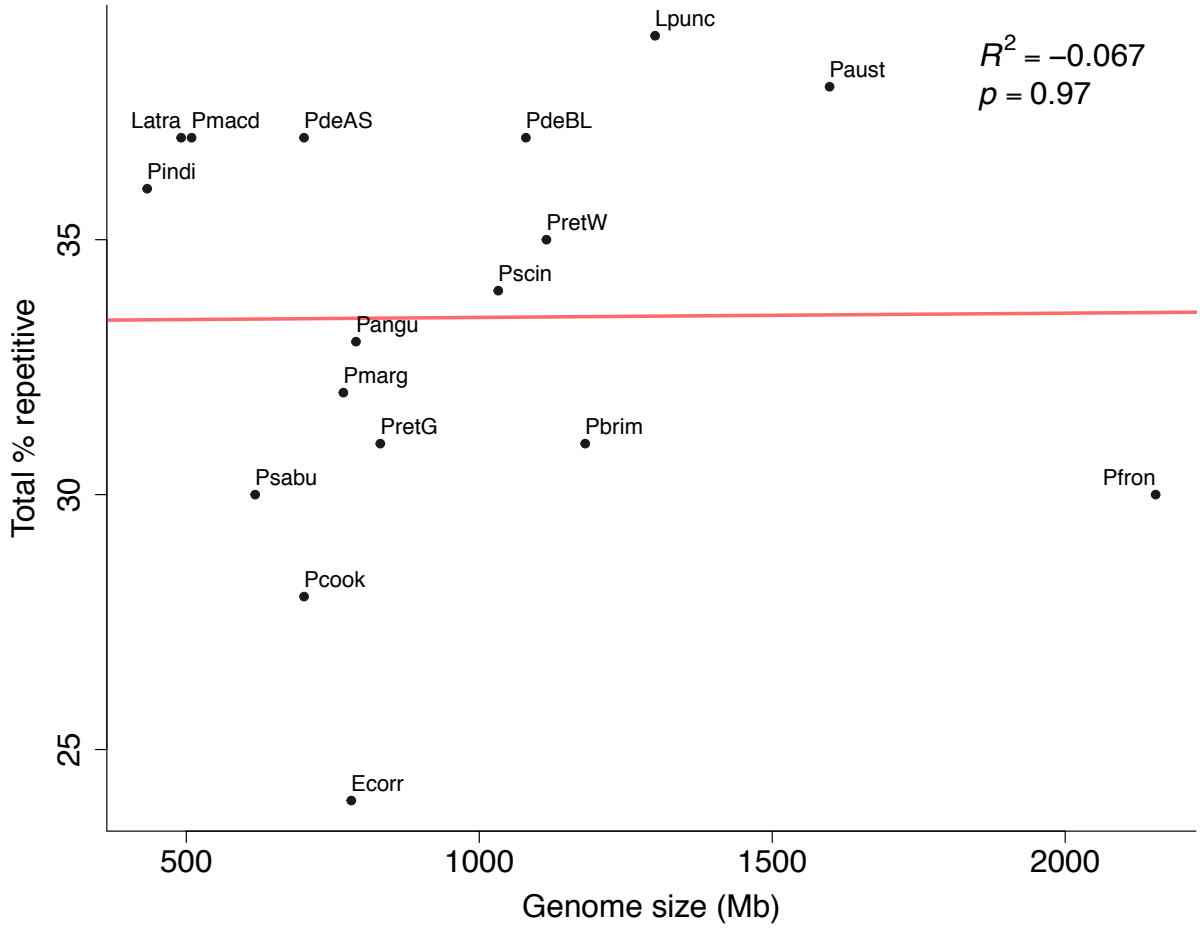


Figure 3.6. No correlation between total repetitiveness and genome size

Neither total repetitiveness nor genome size had significant phylogenetic signal, and so, were examined using linear regression.

CHAPTER 4

ANALYSIS OF MOLECULAR VARIATION ACROSS POPULATIONS OF A WIDESPREAD NORTH AMERICAN FIREFLY, *PHOTINUS PYRALIS*, REVEALS SELECTION ON LUCIFERASE BUT NOT OPSINS[1]

[1] Sander, S.E., Stanger-Hall, K., and D.W. Hall. To be submitted to *Molecular Ecology*.

Abstract

Genes underlying signal production and reception are expected to evolve to maximize signal detection in a specific environment. This expectation should hold both across and within species. Fireflies have long been a system in which to study signal production, reception, and signaling environment. Across species there are differences in flash signal color, visual sensitivity, and signaling environments. Differences in signal color have been hypothesized to be due to variation in the sequence of luciferase, the enzyme that catalyzes the light reaction. Differences in visual sensitivity have been hypothesized to be due to variation in the sequence of opsins, the protein component of the visual pigment. Here we ask if sequence differences in these genes also underlie variation in signal color and inferred visual sensitivity across populations of one widespread North American species, *Photinus pyralis*. We further test if selection has acted on these loci by examining their population-level differentiation relative to the distribution of differentiation derived from a genome-wide sample of single nucleotide polymorphisms. We find evidence for selection on luciferase, despite an absence of protein variation, but not on opsins. Our results challenge the paradigm of signal color being determined by sequence variation in luciferase in fireflies and suggest that either regulatory mutations or linked loci influence luciferase evolution.

Introduction

Communication is one of the most important capabilities of an organism (Hauser 1996) and has been the subject of intense study since Darwin (Darwin 1871). All communication systems involve the transmission of a signal that carries information from a sender to a receiver (Greenfield 2002). Natural selection is expected to favor signals that maximize detection when transmitted through a particular environment. As such, the evolution of any signaling system is affected ("driven") by the characteristics of both the species and the biotic and abiotic environment. The "sensory drive" framework posits that signal production and reception are expected to evolve to maximize signal detection in the specific environment in which signals are displayed (Endler 1992). The effects of natural and sexual selection driving the evolution of sensory systems in particular directions should be detectable in the genes underlying signal production and detection. Evidence for selection should be observable both across species as well as across populations that inhabit different environments within a species.

Fireflies, in the beetle family Lampyridae, are a long established system for the study of the evolution of mating signals due to their conspicuous and variable light displays (e.g. Lall *et al.* 1980, 1982; Seliger *et al.* 1982a, b; Lall 1993, 1994; Cronin *et al*. 2000; Lall & Worthy 2000; Sander & Hall, *in press*; Hall *et al.*, *in review*). Generally, in nocturnal species, flying males signal using species-specific flash patterns to stationary females in the vegetation. Flash pattern is known for use in species recognition and mate choice (Lloyd 1966, 1979; Lewis & Cratsley 2003, 2008; Lewis *et al*. 2004; Demary & Lewis 2007). Not only do the flash patterns vary in a species-specific manner, but there is also variation in signal emission color, ranging from green (554 nm) to yellow (579 nm) (Biggley *et al.* 1967; Lall *et al*. 1980; Seliger *et al.* 1982a, b; Eguchi *et al.* 1984; Cronin *et al.* 2000; Hall *et al.*, *in review*). Across species, emission color

correlates with the time of evening activity, with early-active (around sunset) species having yellower signals, perhaps to contrast with ambient environmental light (Lall *et al.* 1980, Hall *et al.*, *in review*). In addition, peak visual sensitivity matches peak emission color (Lall *et al.* 1980; Seliger *et al.* 1982a, b; Cronin *et al.* 2000).

In contrast to the relatively well-documented variation in signal color, visual sensitivity, and signaling environments across species, within-species variation of these traits remains largely uninvestigated. In a recent study, Hall and colleagues (*in review*) found significant within-species variation in emission color in three North American species. While these species did not vary in the onset of activity across populations, one species (*Photinus scintillans*) exhibited variation in emission color that was correlated with variation in the habitat (open field versus closed forest) in which signals were displayed.

One of the strengths of the firefly system is for studying the evolution of mating signals at the molecular level because the primary genes underlying both signal production (luciferase) and visual sensitivity (opsins) are known. To produce a signal, the enzyme luciferase interacts with its substrate, luciferin, in the presence of oxygen and ATP and releases a photon of light (e.g. Seliger & McElroy 1960, reviewed in Fraga 2008). The amino acid sequence of luciferase varies across firefly species and, because luciferase is widely used as a luminescent marker in molecular studies, specific substitutions within the enzyme have been identified that change the color of emitted light *in vitro* (Kajiyama & Nakano 1991; Conti *et al.* 1996; Branchini *et al.* 1998, 2001, 2003, 2007; Shapiro *et al*. 2005; Nakatsu *et al.* 2006). While fireflies have two luciferase paralogs, only one (LUC1) is expressed in the adult light organ and functions to produce flash signals (Day *et al.* 2009; Oba *et al.* 2010, 2013; Hall *et al.*, *in review*; unpublished data). A single gene used for signal production makes fireflies substantially less complicated in

comparison to other visual signaling systems that involve variation in color, pattern or both, and may involve numerous genes (e.g. butterfly wings: Brakefield & French 1999; Beldade & Brakefield 2002).

To detect a light signal, visual pigments in the photoreceptors of the eye absorb photons of light and then transduce the signal to the optic nerve. Visual pigments are composed of two parts: a signaling protein, opsin, and a light-sensing vitamin A-derived chromophore (Palczewski *et al.* 2000). Amino acid sequence variation in the opsin is known to affect visual sensitivity, especially substitutions at sites that interact with the chromophore (Yokoyama & Yokoyama 2000; Yokoyama 2008; Yokoyama, *et al.* 2008). Fireflies have only two opsins, one that detects long-wavelength light (LW opsin) and one that detects ultraviolet wavelengths (UV opsin)(Oba & Kainuma 2009, Martin *et al.* 2015; Sander and Hall, *in press*). Since there is no known UV component to the firefly signal, flash signals are likely detected solely by the LW opsin (Eguchi, *et al.* 1984), making this signal-reception system particularly uncomplicated in comparison to others. UV opsin may instead be used for navigation (Dacke *et al.* 2004) or in determining the onset of activity (Lall 1993, 1994).

We hypothesize that selection on coding variation in luciferase (the LUC1 paralog) and LW opsin underlies population differences in emission color, and inferred matching visual sensitivities. We test this hypothesis by determining whether within-species variation in the color of emitted light corresponds to genetic variation in luciferase and LW opsin in a widespread North American species, *Photinus pyralis*. We predict that there will be nonsynonymous substitutions in the coding sequences of luciferase (LUC1) and LW that are correlated with emission color, and that these substitutions will exhibit signatures of selection. Further, we predict that if there is evidence for selection, it will be driven by the light environment in which

signals are produced/received. If so, allele frequencies at selected loci are expected to correlate with differences in habitat. UV opsin is not expected to change across and thus serves as a control.

To test our hypotheses, we sequenced luciferase (LUC1), LW opsin and UV opsin across 12 *P. pyralis* populations exhibiting different peak emission colors and habitats. We tested for selection on these loci by examining their pattern of molecular variation relative to a set of genome-wide single nucleotide polymorphisms (SNPs) generated from double-digest restriction-site associated DNA sequencing (ddRADseq; Peterson *et al.* 2012). We then tested for correlations between spectra or habitat and genotypes at luciferase and opsin loci, while controlling for population structure determined using ddRADseq SNPs and mitochondrial sequences.

Materials and Methods

Study system

P. pyralis is a widespread and abundant firefly species that ranges from Arizona to New York (Lloyd 1966). Adult light displays can be seen over fields and in woods from June to October, depending on locality. *P. pyralis* adult light color is yellow (Seliger *et al.* 1964; Lall *et al.* 1980), and peak emission color ranges 10 nm (558 – 568 nm) across populations (Hall *et al.*, *in review*). There is no evidence for a correlation between peak emission color and habitat, but there is a positive correlation between absolute differences in peak emission color and geographic distance (Hall *et al.*, *in review*). *P. pyralis* luciferase was the first to be cloned (de Wet *et al.* 1985) and is widely used as a bioluminescent reporter (reviewed in: Fraga 2008;

Thorne *et al*. 2010). *In vitro*, the cloned *P. pyralis* luciferase emits at 560 nm (McElroy *et al.* 1969; White *et al.* 1971; Deluca & McElroy 1978).

Sampling and DNA extraction

P. pyralis individuals were collected from 42 natural populations (locations) during the summer in 2011-2013. For each population, the time of capture of the first specimen, temperature at the beginning of activity, and habitat type (open field, closed forest, or mixed) were recorded and emission spectra from at least five males measured (Hall *et al.*, *in review*). Specimens were identified to species using both flash pattern and morphology (Green 1956; Lloyd 1966), and preserved in 95% ethanol for later sequencing. We were confident in identification of *P. pyralis* individuals but, to be conservative, at least one specimen per population was confirmed molecularly by amplifying and sequencing 376 bp of *cytochrome oxidase I* (*COI*). This mitochondrial locus has been shown to be phylogenetically informative in *Photinus* fireflies (Stanger-Hall & Lloyd 2015). All specimens are retained in the permanent KSH collection at the University of Georgia.

Twelve populations were selected for sequencing of opsins and luciferase, and ddRADseq. These were chosen to capture variation in the mean male peak emission wavelength, habitat, and geographical distance (Figure 4.1; Table 4.1). The final sample included 16 individuals from each of six closed (forest) populations and six open (field) populations that captured most of the range of male *P. pyralis* peak emission (560 nm-569 nm; Hall *et al.*, *in review*; Appendix C, Table S1). Only male measurements and specimens were used since females are difficult to locate in the field. For all samples, genomic DNA was extracted from

thorax using a standard phenol chloroform isoamyl alcohol protocol with RNAse digestion (Sambrook *et al.* 1989).

LUC1 and LW and UV opsin sequencing and analysis

Adult-expressed luciferase (LUC1) and both opsins (LW and UV) were sequenced using Sanger sequencing. In order to amplify each gene in its entirety, PCR primers were designed using Primer3 (Rozen & Skeletsky 1998) from flanking sequences identified in *P. pyralis* transcriptome and genome sequences (Sander & Hall, *in press*). Both custom species-specific and previously published primers were then used to sequence each amplicon. All primer sequences and PCR cycling conditions are given in Appendix C, Table S2. Bidirectional sequencing was performed at the Georgia Genomics Facility (Athens, GA). The luciferase gene from one individual could not be amplified in its entirety and so this individual was excluded in luciferase analyses.

Sequences were assembled in Geneious R7 (Biomatters Ltd.) and manually inspected for errors and heterozygous sites. Full-length contigs were aligned using Muscle (Edgar 2004) in Geneious and annotated for exon-intron boundaries using coding sequences obtained from transcriptomes or downloaded from Genbank (LW and UV: Sander & Hall, *in press*; LUC1: M15077, de Wet *et al.* 1987). Singletons and gaps in introns due to indels were removed prior to downstream analysis. Alignments were phased in DNAsp v5 (Librado & Rozas 2009)(10,000 iterations, burnin of 10%) and haplotype diversity (Hd), pairwise nucleotide diversity (π), and Fst were calculated for both the whole molecule and the coding sequence.

COI sampling and analysis

To investigate population structure in *P. pyralis* we took advantage of the collection of over 400 mitochondrial *cytochrome oxidase I* (*COI*) partial sequences generated for fireflies by the Stanger-Hall lab over the past 8 years (primers: HCO-LCO; Stanger-Hall *et al.* 2007, Stanger-Hall & Lloyd 2015). In the lab these sequences are used for molecular verification of a specimen's species identity (Sander & Hall, *in press*, Hall *et al.*, *in review*). All *P. pyralis* sequences were extracted from the collection and combined with published *P. pyralis COI* sequences downloaded from Genbank for a total of 99 individuals collected from 45 localities across the Eastern United States during the summer from 1998-2013 (Appendix C, Table S3). A single *Photinus consisus*, three *Photinus carolinus* and one *Photuris quadrifulgens* served as outgroups (Stanger-Hall *et al.* 2007; Stanger-Hall & Lloyd 2015; Sander & Hall, *in press*).

Sequences were aligned using MUSCLE (Edgar 2004) in Geneious R7 (Biomatters Ltd.). The final alignment was trimmed to 555 bp of coding sequence with no gaps. jModeltest2 v.2.1.4 (Guindon & Gascuel 2003; Darriba *et al.* 2012) selected HKY + I + G as the best model of nucleotide substitution according to the AICc (corrected for small sample size; Posada 2008). Bayesian phylogenies were constructed in MrBayes v.3.2.1 (Ronquist *et al.* 2012) until the average standard deviation of split frequencies was below 0.01 (two independent runs, 5 million generations, ngammacat = 8, 25% burnin). Branches with support values below 50% were collapsed to yield the majority consensus tree.

ddRADseq library preparation, pooling, and sequencing

A subset of 15 individuals from each population was randomly chosen for RAD sequencing (Davey *et al*. 2011). From those 15, one was randomly chosen to generate an

additional, technical replicate for downstream optimization of RAD-locus assembly parameters (Mastretta-Yanes *et al.* 2015). In total, there were 16 libraries per population (15 individuals plus one technical replicate), resulting in 192 individual libraries across two 96-well plates. The 16 libraries from each population were equally divided among the plates and randomly assigned to wells to decrease sequencing bias due to library preparation. Of the replicate pairs, six were randomly assigned to wells within the same plate, while the other six were split between plates.

Library construction followed a 3RAD protocol (Travis Glenn, personal communication). 3RAD differs from ddRADseq (Peterson *et al*. 2012) by using three restriction enzymes, two to digest genomic DNA and one to cut adapter dimers. This increases the efficiency of sequencing shared loci across specimens because adapter-dimers are eliminated rather than sequenced, thus resulting in higher sequencing coverage depth of desired genomic DNA fragments. In this study, genomic DNA from each specimen was digested with claI and bamHI, and mspI was used as the adapter-dimer cutter. Following digestion, unique combinations of internal barcode adapters were ligated onto cut fragments. Internal barcodes ranged in length from 6-9 nucleotides ensuring appropriate library complexity for the Illumina sequencing platform. All samples on one plate were then pooled, divided into 3 aliquots, and each aliquot labeled with a unique combination of Illumina i5 and i7 adapters. This pooling scheme was intended to reduce bias in the libraries due to differential adapter amplication in the subsequent PCR amplification step. After amplification, aliquots were pooled by plate, resulting in two final libraries. Final libraries were then size selected using a Caliper LabChip XT (PerkinElmer) at the Savannah River Ecology Lab (Aiken, SC). The average size of fragments was 550bp +/- 12.5%. The two size-selected samples were pooled and run on 50% of 4 lanes of a PE75 NextSeq run at the Georgia

Genomics Facility (Athens, GA). All samples were included in all lanes to reduce the effect of lane on sequencing output.

ddRADseq Analysis

Samples were initially demultiplexed by outer Illumina adapters using bcl2fastq v2.16.0.10 (Illumina, Inc.). The reads for individual specimens were identified from the demultiplexed pools and cleaned using process_radtags in Stacks v1.29 (parameters: -q –r –renz_1 mspI –renz_2 bamHI –t 63, (Catchen *et al.* 2013). Non-eukaryote contaminants were then cleaned from each specimen using kraken (parameters: --paired –db minikraken_20141208; (Wood & Salzberg 2014). Attempts to avoid gut microbes by isolating DNA from thorax were generally successful (median: 0.02% reads identified as contaminants; range: 0.003-35%).

Cleaned paired-end reads were concatenated and run through the Stacks pipeline using default parameters. We identified 28/192 libraries (15%) that "failed", meaning that they had data for less than 50% of the loci that were shared across 80% of the samples. All of these libraries had fewer than 100,000 reads and were excluded from downstream analysis. The failed libraries included 3 of the technical replicate samples, leaving 9/12 samples with replicates for optimization. Replicates were used to find the optimal Stacks parameters according to the procedure of Mastretta-Yanes and colleagues (2015) (m = 3, M = 4, n = 3, *max_locus_stacks* = 3) with the default SNP calling model. In total, 154 samples were used in the final analysis, choosing the "best" sample from each replicate pair (i.e. the one with the most reads).

Stacks output was analyzed using the populations module in Stacks v1.30. SNP loci included in the final analysis were required to be present in all 154 individuals and have a minor allele frequency of at least 5%. Loci were annotated by comparing the consensus sequence for

each locus to all nucleotide sequences in Genbank (blastn, evalue: 1e-4, ID: 90%; (Altschul, et al. 1990). Each locus was tested for Hardy-Weinberg equilibrium prior to population structure analysis using pegas (Paradis 2010) in R with Bonferroni correction to account for multiple testing. Diversity measures were generated for the final dataset, again using the populations module in Stacks.

Fst outlier analysis

Loci with evidence for balancing and positive selection were identified using the FDIST method (Beaumont & Nichols 1996) as implemented in the LOSITAN Selection Workbench (Antao *et al.* 2008). This method estimates the background Fst distribution (i.e. for the likely neutral loci) and then identifies loci with Fst values falling in the tails of the distribution (i.e. more extreme than the 95% confidence interval). The method involves a two-step process. First data from all loci are utilized to estimate the mean Fst and then the mean is used to generate the expected neutral distribution by simulation. Second, loci falling outside of the 95% confidence interval are discarded as they are the most likely to be targets of selection and thus bias the estimate of Fst. Then the mean Fst from the remaining loci is re-estimated and the expected neutral distribution regenerated by simulation. This distribution is then used to identify loci that fall outside the Fst confidence interval. Each estimation step consisted of 50,000 simulations to approximate the mean neutral Fst distribution. Loci with Fst values larger or smaller than the Fst CI were considered to be under diversifying or balancing selection, respectively. All final RAD loci and all SNPs identified in luciferase, LW opsin, and UV opsin were included individually in this analysis. For each SNP, Fst was plotted against the heterozygosity among populations (He). Analysis of variance (ANOVA) was used to investigate differences in spectra and habitat among

the genotypes at positively selected signaling locus SNPs using JMP Pro 11 (SAS Institute Inc. 2013).

Population structure

Neighbor-joining dendrograms were constructed using Nei's distance (Nei 1978) between individuals and bootstrapped (1000 samples, 50% support cut-off) using the poppr v1.1.5 (Kamvar *et al.* 2014) and adegenet v1.4-2 (Jombart & Ahmed 2011) packages in R (R Core Team 2008).

We examined clustering using admixture models in STRUCTURE (Pritchard *et al.* 2000). All analyses were run for 1,000,000 iterations with a burn-in of 100,000. Analyses were repeated 20 times for each K value (1-13). Structure Harvester v0.6.94 (Earl & vonHoldt 2012) was used to identify the best number of clusters using using the Evanno method (ΔK; Evanno *et al*. 2005). STRUCTURE replicates were assessed in Clumpp v1.1.2 (Jakobsson & Rosenberg 2007) and visualized in Microsoft Excel for Mac v.14.4.4 (Microsoft Corporation 2011).

Testing for isolation by distance, by spectrum, and by habitat

We investigated hypotheses of genetic isolation by distance (IBD), by habitat (IBH), and/or by spectra (IBS) using Mantel and partial Mantel tests in Genodive (Meirmans & Van Tienderen 2004). Geographical distance was calculated as the Great Circle distance between each locality. Habitat values were coded as ordinal variables (0=open, 1=closed). Genetic distances for RAD loci, luciferase, LW opsin, and UV opsin were first tested for IBD, and, if significant, used in partial Mantel tests for IBH and IBS while controlling for geographic distance.

Results

mtDNA

COI sequencing of 99 individuals representing 45 localities resulted in 51 haplotypes. The resulting Bayesian phylogeny confirmed the monophyly of *P. pyralis* with high support (Figure 4.2). In addition, the topology showed that Western populations (Texas, Kansas, Arkansas and Mississippi) are basal to Eastern populations (remaining twelve states, Figure 4.2). Across the Eastern populations there was little phylogenetic resolution. The highest divergence occurred between specimens from Texas/Mississippi and New Jersey.

RAD loci

One half of a single run of an Illumina NextSeq High Output yielded 190,648,449 PE75 reads over all 192 libraries. After quality trimming, cleaning contaminants, and discarding unsuccessful libraries, there were 1,038,853 +/- 413,471 reads per specimen for 154 specimens. In the final data set there were 11-15 individuals per population. Running Stacks with optimized parameters resulted in 925 loci shared across all individuals, 716 of which were variable across populations. None of these loci could be ascribed to mitochondrial DNA or prokaryotic contaminants. All loci were in Hardy Weinberg equilibrium after (Bonferroni) correcting for multiple testing.

The variable SNPs developed from RAD loci allowed investigation of population structure on a finer scale than the *COI* data. Genetic distance-based neighbor-joining dendrograms showed that specimens from each population were generally monophyletic, though two pairs of populations showed mixing (Figure 4.3). The two populations from New Jersey (AMNJ and MANJ) were the closest pair (96 km) in our sampling scheme and showed

substantial inter-digitation in the neighbor-joining tree. The Athens, Georgia (ATGA) and Salisbury, North Carolina (SANC) populations showed less mixing and were the 5th-closest pair (336 km). The topology of the dendrogram showed that Texas populations are genetically distinct from the other populations, confirming the result from the COI tree. However, the topology suggested a mid-Western and an Eastern clade rather than an East-West split. The Eastern clade was less genetically diverse as evidenced by short branch lengths between populations ranging in latitude from Georgia to New Jersey. Interestingly, while the ATGA and SANC populations were intermixed, individuals collected from a valley in the Great Smokey Mountains National Park (HFTN), at an intermediate latitude between the two sites, formed a monophyletic group. Similarly, individuals from the St. Louis, Missouri (SLMO) population were surprisingly divergent from their nearest neighbors, Wynne, Arkansas (WYAR) and Byhalia, Mississippi (BYMS).

STRUCTURE results mirrored the phylogeny and thus provided further support for Texas, mid-Western, and Eastern clades. Analysis suggested K=3 as the most likely number of genetic clusters in the dataset (Figure 4.4). The three clusters corresponded to Texas, mid-West, and Eastern groups, with evidence of admixture between all three groups within mid-Western populations.

Variation at signaling loci

Adult luciferase (LUC1) had a single nonsynonymous mutation, V182I, in exon 3 that was heterozygous in four individuals from VATX. LW opsin had a single nonsynonymous mutation, A16V, in exon 1 that was heterozygous in four individuals, three from Dexter, Michigan (DEMI), and one from BYMS. UV opsin had no nonsynonymous substitutions. The

three loci differed in nucleotide diversity: 0.0010 in luciferase, 0.0016 in LW opsin and 0.0078 in UV opsin (Table 4.2). In sum, luciferase and LW opsin exhibited higher levels of nonsynonymous variation and lower levels of nucleotide diversity than UV opsin. However, nonsynonymous variation was very rare in both genes, found in less than 3% of individuals.

Fst outliers

LOSITAN (Antao *et al.* 2008) estimated the average Fst at neutral loci to be 0.38. Approximately 30% of loci were in the combined top and bottom tails of the distribution, with 9.6% in the top 5%. Three luciferase SNPs, each of which had high among-population levels of heterozygosity, were in the top 5% of Fst values indicating possible diversifying selection (site 723: He=0.52, Fst=0.80, p=0.003; site 1780: He=0.40, Fst=0.83, p=0.004; site 630: He=0.53, Fst=0.75, p=0.01)(Figure 4.6). The other 18 luciferase SNPs had lower levels of heterozygosity (≤ 0.11) and no evidence for extreme values of Fst. The single nonsynonymous mutation in luciferase was among these SNPs (site 667: He=0.013, Fst=0.04). For LW opsin, eight of the 24 SNPs were in the bottom 5% of the Fst distribution and were thus candidates for balancing selection. The remaining 16 SNPs were in the neutral Fst range, including the nonsynonymous substitution in LW (site 98: He=0.02, Fst=0.05). For UV opsin, five of the 41 SNPs were in the bottom 5% of the Fst distribution and were thus candidates for balancing selection. There was also one SNP in the top 5% of the Fst distribution (He=0.03, Fst=0.13, p=0.03). Examination of this SNP revealed that the high Fst value was due to the presence of private alleles in the Vanderpool, Texas (VATX) population. Given the restricted pattern of variation for this SNP, it was not possible to distinguish selection versus drift as an explanation for the high Fst.

No outlier RAD loci had significant BLAST hits. However, one RAD locus, 16212, was identified as part of *P. pyralis* luciferin regenerating enzyme (AB062786.1). Its Fst fell in the neutral range (Fst=0.20).

Selection at luciferase

The pattern of variation across populations for the three luciferase SNPs showing high Fst was examined in the context of phenotypic variation for peak emission color (spectra) and habitat (closed versus open). Since these traits are measured on different individuals (spectra), or at the population level (habitat), every individual in a population was given the same trait value. The value assigned was the population habitat value (open or closed) and the average spectra (measured in Hall *et al.*, *in review*). ANOVA was then used to determine whether there was a relationship between genotype and spectra or habitat. Genotypes at all three luciferase SNP outliers were significantly associated with both spectra and habitat (Appendix C, Figure S2). Sites 630 and 723 had similar patterns, with the heterozygote having a higher wavelength (i.e. more yellow) and a higher probability of being in an open habitat than either of the homozygotes, though the result for spectra was not significant for site 723. Site 1780, in the last intron of luciferase, showed a similar pattern for habitat, but the heterozygote was intermediate to the two heterozygotes for spectra.

Testing IBD, IBS, IBH

Mantel tests and partial Mantel tests were used to test hypotheses of genetic isolation by geographical distance, and isolation by habitat or spectra while controlling for geographical distance. Neither isolation by distance, habitat, or spectra were observed for the RAD dataset,

luciferase, LW opsin, or UV opsin, even when Eastern and mid-Western populations were tested separately to control for major geographic breaks.

Discussion

Here we undertook an investigation of variation in luciferase and opsin genes to examine the underlying genetic basis for differences in emission color and inferred differences in visual sensitivity across *P. pyralis* populations signaling in different habitats. We employed ddRADseq to generate a set of genome-wide SNPs against which to compare signaling locus SNPs in order to identify signatures of selection, and to investigate population structure within the species.

mtDNA and RAD loci suggest barriers to gene flow

P. pyralis is a widespread and abundant species in which both males and females can fly. Because it is common in the Eastern U.S. and found in a variety of habitats, including urban and disturbed areas, we expected to observe high levels of gene flow across the range, perhaps with limited gene flow between populations at large geographic distances or at certain geographic boundaries, such as the Appalachian mountains. The mtDNA *COI* sequence data suggested that this prediction might hold. Western populations were basal to Eastern populations and the greatest sequence divergence was between individuals collected at the extremes of the species range (Texas and New Jersey). In contrast, analysis using genome-wide SNPs suggested three genetic clusters: Texas, mid-Western, and Eastern populations. These clusters roughly correspond to putative geographical barriers to dispersal (i.e. the Appalachian Mountains and the Mississippi River Valley). These regions, particularly the Mississippi, are known geographic breaks for North American taxa (Soltis *et al.* 2006). The SNP data indicated Eastern populations

were genetically quite similar, suggesting a recent population expansion and/or high levels of ongoing gene flow.

It is not surprising that mtDNA and SNP results differed. Mitochondrial DNA has a long history in the study of population structure in animals (Avise 2000) and has many benefits including a high mutation rate and lack of recombination. Specifically the *COI* locus, used for barcoding, has been used in studies of firefly phylogenetics (Choi *et al.* 2003; Stanger-Hall *et al.* 2007; Osozawa *et al.* 2015; Stanger-Hall & Lloyd 2015) and population structure (Lee *et al.* 2003). However, *COI* was not able to resolve fine-scale population structure in *P. pyralis* in this study. Due to its lack of recombination, analysis of one mitochondrial DNA fragment captures the evolution of the entire mitochondrial genome. As such, the evolution of mitochondrial genes can be substantially affected by effects such as background selection (Ballard & Whitlock 2004), whereas the genome-wide SNPs are much more likely to be independent. Further, the mitochondrial genome migrates in females only. While they are able flyers, unmated females spent much of their time sitting in the vegetation waiting for a male (Buck 1937). Once inseminated, females then fly to find oviposition sites. It seems likely that males will fly substantially further in a night that a female, which would potentially result in different pattern of variation for mitochondrial versus nuclear loci.

Luciferase, but not opsins, show evidence of diversifying selection among populations

Contrary to our expectations, neither luciferase nor LW opsin showed substantial nonsynonymous changes in coding sequence. Previous site-directed mutagenesis of the single low-frequency nonsynonymous mutation site that we identified in luciferase showed that that substitutions at this site have increased thermostability without a shift in emission color (Law *et*

al. 2006). The single low frequency nonsynonymous site in LW opsin is not predicted to affect visual sensitivity based on its position relative to the chromophore binding pocket (Sander & Hall, *in press*).

Despite the lack of coding variation, luciferase did show a strong signature of selection based on the high Fst of three SNPs relative to the distribution generated from neutral genome-wide SNPs. High Fst for a SNP indicates that one variant is present at high frequency within some populations and is at low frequency in other populations. On its own, this finding suggests that there has either been selection on cis regulation of luciferase or there is another locus that is in strong linkage disequilibrium with luciferase and is the actual target of diversifying selection. The fact that we observe significant associations at the population level between SNP frequency and spectra for all three SNPs and between SNP frequency and habitat for two of the three SNPs strongly suggests that it is luciferase that is the target of selection. The 5' upstream region of *P. pyralis* luciferase has been examined in relation to related luciferase genes in other species, and a core promoter region has been identified (de Wet *et al.* 1987; Day 2005). In addition variation in 5' and 3' sequences flanking luciferase has been documented in the Asian firefly, *Luciola lateralis* (Cho *et al.* 1999). The presence of variation in such sequences remains to be investigated in *P. pyralis* and, if found to be present and exhibit a similarly strong signature of selection, the functional consequences of such variation determined. Interestingly, in *Lampyris noctiluca*, the European glowworm, an ancient transposon endonuclease domain is located 686 bp upstream of the start codon (Day 2005), though whether the insertion has functional significance is unknown.

Isolation by distance, spectra and habitat

There was no significant isolation by distance, spectra, or habitat in the genomic, luciferase, or opsin SNP data. Thus populations that were geographically closer, more similar in spectra or in identical habitats were not also more similar genetically. The lack of genetic isolation by spectra or habitat was not surprising given the fact that signal color is not thought to be important in mate recognition in fireflies (Ohba 2004) and there is no reason to think that habitat would play a role. In contrast, the lack of isolation by distance was somewhat surprising given that geographically closer populations are more similar in spectra than more distant populations (Hall et al. *in review*), which is consistent with higher gene flow on short geographic scales. However, our sample size (12 populations) is substantially smaller than that previous study and our populations were chosen so that they spanned the range of habitats and spectra, perhaps influencing the isolation by distance relationship. Clearly it would be useful to expand the study by genotyping the remainder of 42 populations.

Conclusion

Using RADseq markers, we were able to generate a null distribution for Fst and show evidence for balancing selection on three SNPs in luciferase across twelve *P. pyralis* populations. None of these SNPs cause a change in the amino acid sequence of the luciferase protein. Additional evidence that these sites are targets of selection is the significant differences in spectra and habitat usage between populations that differ in SNP genotypes at these sites. In contrast, LW opsin was under balancing selection. Future work will seek to substantially increase the number of populations analyzed and account for population structure while investigating differences in spectra and habitat.

Acknowledgments

This work was supported by the National Science Foundation (GRF to S.E.S and DDIG DEB-1311315 to D.W.H. and S.E.S.) and the NIGMS of the National Institute of Health (award number T32GM007103 to S.E.S.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to thank the following individuals for permission to collect and assistance with collections: Arkansas Department of Parks and Tourism, Gina Baucomb, Ashley Brown, Great Smoky Mountains National Park (permit GRSM-2011-SCI-0049 to KSH), Paul Kissel (Eisenhower State Park, TX), Michael Marsh (University of Georgia), Jenna Pallansch, David Queller, David Riskind (Texas Parks and Wildlife Department permit), Willem Roosenburg, Tonya Saint John, and Joan Strassman. The authors would also like to acknowledge Travis Glenn and Troy Kieran for help with ddRADseq study design and implementation.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403-410.

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method. *BMC Bioinformatics*, **9**, 323.

Avise JC (2000) *Phylogeography: the history and formation of species*. Harvard University Press.

Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. *Molecular Ecology*, **13**, 729-744.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London B: Biological Sciences*, **263**, 1619-1626.

Beldade P, Brakefield PM (2002) The genetics and evo–devo of butterfly wing patterns. *Nature Reviews Genetics*, **3**, 442-452.

Biggley WH, Lloyd JE, Seliger HH (1967) The spectral distribution of firefly light II. *The Journal of General Physiology,* **50**, 1681-1692.

Brakefield PM, French V (1999) Butterfly wings: the evolution of development of colour patterns. *BioEssays*, **21**, 391-401.

Branchini BR, Ablamsky DM, Murtiashaw MH, Uzasci L, Fraga H, Southworth TL (2007) Thermostable red and green light-producing firefly luciferase mutants for bioluminescent reporter applications. *Analytical biochemistry*, **361**, 253-262.

Branchini BR, Magyar RA, Murtiashaw MH, Anderson SM, Zimmer M (1998) Site-directed mutagenesis of histidine 245 in firefly luciferase: a proposed model of the active site. *Biochemistry*, **37**, 15311-15319.

Branchini BR, Magyar RA, Murtiashaw MH, Portier NC (2001) The role of active site residue arginine 218 in firefly luciferase bioluminescence. *Biochemistry*, **40**, 2410-2418.

Branchini BR, Southworth TL, Murtiashaw MH, Boije H, Fleet SE (2003) A mutagenesis study of the putative luciferin binding site residues of firefly luciferase. *Biochemistry*, **42**, 10429-10436.

Buck JB (1937) Studies on the firefly II the signal system and color vision in *Photinus pyralis*. *Physiological Zoology*, **10**, 412-419.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124-3140.

Cho KH, Lee JS, Choi YD, Boo KS (1999) Structural polymorphism of the luciferase gene in the firefly, *Luciola lateralis*. *Insect Molecular Biology,* **8**, 193-200.

Choi YS, Bae JS, Lee KS, Kim SR, Kim I, Kim JG, *et al.* (2003) Genomic structure of the luciferase gene and phylogenetic analysis in the Hotaria-group fireflies. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, **134**, 199-214.

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

Cronin TW, Jarvilehto M, Weckstrom M, Lall AB (2000) Tuning of photoreceptor spectral sensitivity in fireflies (Coleoptera: Lampyridae). *Journal of Comparative Physiology A*: 1-12.

Dacke M, Byrne MJ, Scholtz CH, Warrant EJ (2004) Lunar orientation in a beetle. *Proceedings of the Royal Society of London B: Biological Sciences*, **271**, 361-365.

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, **9**, 772-772.

Darwin, Charles (1871) *Sexual Selection and the Descent of Man*. Murray, London.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499-510.

Day JC (2005) Characterisation of the luciferase gene and the 5' upstream region in the European glow-worm *Lampyris noctiluca* (Coleoptera: Lampyridae). *European Journal of Entomology*, **102**, 787-791.

Day JC, Goodall TI, Bailey MJ (2009) The evolution of the adenylate-forming protein family in beetles: multiple luciferase gene paralogues in fireflies and glow-worms. *Molecular Phylogenetics and Evolution*, **50**, 93-101.

de Wet JR, Wood KV, DeLuca M, Helinski DR, Subramani S (1987) Firefly luciferase gene: structure and expression in mammalian cells. *Molecular and Cellular Biology*, **7**, 725-737.

de Wet JR, Wood KV, Helinski DR, DeLuca M (1985) Cloning of firefly luciferase cDNA and the expression of active luciferase in *Escherichia coli*. *Proceedings of the National Academy of Sciences U.S.A.*, **82**, 7870-7873.

Deluca M, McElroy WD (1978) Purification and properties of firefly luciferase. *Methods in Enzymology*, 3-15.

Demary KC, Lewis SM (2007) Male courtship attractiveness and paternity success in *Photinus greeni* fireflies. *Evolution*, **61**, 431-439.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359-361.

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**.

Eguchi E, Nemoto A, Meyer-Rochow VB, Ohba N (1984) A comparative study of spectral sensitivity curves in three diurnal and eight nocturnal species of Japanese fireflies. *Journal of Insect Physiology*, **30**, 607-612.

Endler JA (1992) Signals, signal conditions, and the direction of evolution. *American Naturalist*, **139**, S125-S153.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, **14**, 2611-2620.

Fraga H (2008) Firefly luminescence: a historical perspective and recent developments. *Photochemical & Photobiological Sciences*, **7**, 146-158.

Green JW (1956) Revision of the nearctic species of Photinus (Lampyridae: Coleoptera). *Proceedings of the California Academy of Sciences*, **28**, 561-613.

Greenfield MD (2002) *Signalers and Receivers: Mechanisms and Evolution of Arthropod Communication*. New York: Oxford University Press.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**, 696-704.

Hauser, Marc D (1996) *The Evolution of Communication*. MIT press.

Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics*, **132**, 583-589.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801-1806.

Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, **27**, 3070-3071.

Kajiyama N, Nakano E (1991) Isolation and characterization of mutants of firefly luciferase which produce different colors of light. *Protein Engineering*, **4**, 691-693.

Kamvar ZN, Tabima JF, Grünwald NJ (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, **2**, e281.

Lall A, Lord E, Trouth CO (1982) Vision in the firefly *Photuris lucicrescens* (Coleoptera: Lampyridae): spectral sensitivity and selective adaptation in the compound eye. *Journal of Comparative Physiology*, **147**, 195-200.

Lall AB (1993) Action spectra for the initiation of bioluminescent flashing activity in males of twilight-active firefly *Photinus scintillans* (Coleoptera: Lampyridae). *Journal of Insect Physiology*, **39**, 123-127.

Lall AB (1994) Spectral cues for the regulation of bioluminescent flashing activity in the males of twilight-active firefly *Photinus scintillans* (Coleoptera: Lampyridae) in nature. *Journal of Insect Physiology*, **40**, 359-363.

Lall AB, Seliger HH, Biggley WH, Lloyd JE (1980) Ecology of colors of firefly bioluminescence. *Science*, **210**, 560-562.

Lall AB, Worthy KM (2000) Action spectra of the female's response in the firefly *Photinus pyralis* (Coleoptera: Lampyridae): evidence for an achromatic detection of the bioluminescent optical signal. *Journal of Insect Physiology*, **46**, 965-968.

Law GHE, Gandelman Olga A, Tisi Laurence C, Lowe Christopher R, Murray James AH (2006) Mutagenesis of solvent-exposed amino acids in *Photinus pyralis* luciferase improves thermostability and pH-tolerance. *Biochemical Journal*, **397**, 305-312.

Lee SC, Bae, JS, Kim I, Suzuki H, Kim SR, Kim JG, *et al.* (2003) Mitochondrial DNA sequence-based population genetic structure of the firefly, *Pyrocoelia rufa* (Coleoptera: Lampyridae). *Biochemical Genetics*, **41**, 427-452.

Lewis SM, Cratsley CK (2003) Female preference for male courtship flashes in *Photinus ignitus* fireflies. *Behavioral Ecology*, **14**, 135-140.

Lewis SM, Cratsley CK (2008) Flash signal evolution, mate choice, and predation in fireflies. *Annual Reviews of Entomology*, **53**, 293-321.

Lewis SM, Cratsley CK, Demary K (2004) Mate recognition and choice in Photinus fireflies. In. *Annales Zoologici Fennici*. Finnish Zoological and Botanical Publishing Board, pp 809-821.

Librado P, Rozas J (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451-1452.

Lloyd JE (1966) Studies on the flash communication system in Photinus fireflies. In. *Miscellaneous Publications*. Ann Arbor, Michigan: Museum of Zoology, University of Michigan.

Lloyd JE (1979) Sexual selection in luminescent beetles. In. *Sexual Selection and Reproductive Competition in Insect*s. Edited by Blum MS, Blum NA. New York: Academic Press, pp. 293-342.

Martin GJ, Lord NP, Branham MA, Bybee SM (2015) Review of the firefly visual system (Coleoptera: Lampyridae) and evolution of the opsin genes underlying color vision. *Organisms Diversity & Evolution*, 1-14.

Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson BC (2015) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, **15**, 28-41.

McElroy WD, Seliger HH, White EH (1969) Mechanism of bioluminescence, chemi-luminescence and enzyme function in the oxidation of firefly luciferin. *Photochemistry and Photobiology*, **10**, 153-170.

Meirmans PG, Van Tienderen PH (2004) GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, **4**, 792-794.

Nakatsu T, Ichiyama S, Hiratake J, Saldanha A, Kobashi N, Sakata K, *et al.* (2006) Structural basis for the spectral difference in luciferase bioluminescence. *Nature*, **440**, 372-376.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89**, 583-590.

Oba Y, Furuhashi M, Bessho M, Sagawa S, Ikeya H, Inouye S (2013) Bioluminescence of a firefly pupa: involvement of a luciferase isotype in the dim glow of pupae and eggs in the Japanese firefly, *Luciola lateralis*. *Photochemical & Photobiological Sciences*, **12**, 854-863.

Oba Y, Kainuma T (2009) Diel changes in the expression of long-wavelength sensitive and ultraviolet-sensitive opsin genes in the Japanese firefly, *Luciola cruciata*. *Gene*, **436**, 66-70.

Oba Y, Mori N, Yoshida M, Inouye S (2010) Identification and characterization of a luciferase isotype in the Japanese firefly, *Luciola cruciata*, involving in the dim glow of firefly eggs. *Biochemistry*, **49**, 10788-10795.

Ohba N (2004) Flash communication systems of Japanese fireflies. *Integrative and Comparative Biology*, **44**, 225-233.

Osozawa S, Oba Y, Kwon H-Y, Wakabayashi J (2015) Vicariance of Pyrocoelia fireflies (Coleoptera: Lampyridae) in the Ryukyu islands, Japan. *Biological Journal of the Linnean Society*, Early view.

Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, *et al.* (2000) Crystal structure of rhodopsin: a G protein-coupled receptor. *Science*, **289**, 739-745.

Paradis E (2010) pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, **26**, 419-420.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Posada D (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253-1256.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetic,* **155**, 945-959.

R Development Core Team (2013) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, *et al.* (2012) MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, **61**, 539-542.

Rozen S, Skeletsky HJ (1998) Primer3. Code available at http://www.genome.wi.mit.edu/genome_software/other/primer3.html

Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning*. New York: Cold Spring Harbor Laboratory Press.

Seliger HH, Buck JB, Fastie WG, McElroy WD (1964) The spectral distribution of firefly light. *Journal of General Physiology*, **48**, 95-104.

Seliger HH, Lall AB, Lloyd JE, Biggley WH (1982a) The colors of firefly bioluminescence—I. Optimization model. *Photochemistry and Photobiology*, **36**, 673-680.

Seliger HH, Lall AB, Lloyd JE, Biggley WH (1982b) The colors of firefly bioluminescence—II. Experimental evidence for the optimization model. *Photochemistry and Photobiology*, **36**, 681-688.

Seliger HH, McElroy WD (1960) Spectral emission and quantum yield of firefly bioluminescence. *Archives of Biochemistry and Biophysics*, **88**, 136-141.

Shapiro E, Lu C, Baneyx F (2005) A set of multicolored *Photinus pyralis* luciferase mutants for in vivo bioluminescence applications. *Protein Engineering Design and Selection*, **18**, 581-587.

Soltis DE, Morris AB, McLachlan JS, Manos PS, Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, **15**, 4261-4293.

Stanger-Hall KF, Lloyd JE (2015) Flash signal evolution in Photinus fireflies: Character displacement and signal exploitation in a visual communication system. *Evolution*, **69**, 666-682.

Stanger-Hall KF, Lloyd JE, Hillis DM (2007) Phylogeny of North American fireflies (Coleoptera: Lampyridae): implications for the evolution of light signals. *Molecular Phylogenetics and Evolution*, **45**, 33-49.

Thorne N, Inglese J, Auld DS (2010) Illuminating insights into firefly luciferase and other bioluminescent reporters used in chemical biology. *Chemistry & Biology*, **17**, 646-657.

White EH, Rapaport E, Seliger HH, Hopkins TA (1971) The chemi-and bioluminescence of firefly luciferin: an efficient chemical production of electronically excited states. *Bioorganic Chemistry*, **1**, 92-122.

Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15, R46.

Yokoyama S (2008) Evolution of dim-light and color vision pigments. *Annual Review of Genomics and Human Genetics*, **9**, 259-282.

Yokoyama S, Yang H, Starmer WT (2008) Molecular basis of spectral tuning in the red- and green-sensitive (m/lws) pigments in vertebrates. *Genetics*, **179**, 2037-2043.

Yokoyama S, Yokoyama R (2000) Comparative Molecular Biology of Visual Pigments. In. *Molecular Mechanisms in Visual Transduction.* Edited by Stavenga DG, DeGrip WJ, Pugh Jr. EN. Amsterdam: Elsevier, pp. 257-296.

Table 4.1. Populations selected for sequencing

The twelve populations selected for investigation span the range of habitats and wavelengths measured for *Photinus pyralis* males (Hall et al. 2015). The four-letter population code (Code), locality, latitude (Lat), longitude (Lon), mean male peak emission wavelength (Spectra), standard error of male peak emission wavelength (SE), number of males measured for spectra (N), and habitat type (O: open, C: closed) are given for each population.

Code	Locality	Lat	Lon	N	Spectra	SE	Habitat
AMNJ	Amwell, NJ	40.24	-74.52	5	566.89	0.76	C
AMOH	Amesville, OH	39.41	-82.00	4	560.73	0.55	O
ATGA	Athens, GA	33.89	-83.37	2	560.70	0.48	O
BYMS	Byhalia, MS	34.88	-89.69	6	565.83	0.46	O
DEMI	Dexter, MI	42.38	-83.92	5	561.60	1.03	C
DETX	Denison, TX	33.82	-96.61	6	564.12	1.34	C
HFTN	Hickory Flats Branch, TN	35.68	-83.55	4	561.32	0.70	C
MANJ	Mahwah, NJ	41.08	-74.19	6	559.59	0.10	C
SANC	Salisbury, NC	35.63	-80.35	6	564.04	0.82	C
SLMO	St. Louis, MO	38.64	-90.29	8	562.86	0.50	O
VATX	Vanderpool, TX	29.81	-99.57	8	562.00	0.55	O
WYAR	Wynne, AR	35.17	-90.71	5	565.92	1.51	O

Table 4.2. Genetic diversity and gene flow estimates for luciferase (LUC1), and LW and UV opsin.

Measures are given for the alignment of entire sequences (ALL) and just coding sequences (CDS). The number of individuals sequenced (N), number of variable sites (VS), number of synonymous (Syn) and nonsynonymous (NSyn) mutations, the position of nonsynonymous mutations in the nucleotide alignment (Site), haplotype diversity (Hd +/- (SD)), nucleotide diversity (π +/- (SD)), and Fst (Hudson *et al.* 1992) are given.

Locus	Length (bp)	N	VS	Syn	NSyn	Site	Hd	π	Fst
LUC1: ALL	1966	190	21	13	1	667	0.69 (0.021)	0.00099 (0.00004)	0.61
LUC1: CDS	1650	190	14	13	1	544	0.63 (0.018)	0.00080 (0.00003)	0.65
LW: ALL	1459	191	24	18	1	98	0.79 (0.019)	0.00161 (0.00006)	0.20
LW: CDS	1134	191	19	18	1	47	0.78 (0.019)	0.00200 (0.00007)	0.20
UV: ALL	1440	191	41	26	0	N/A	0.91 (0.007)	0.00382 (0.00010)	0.31
UV: CDS	1152	191	24	26	0	N/A	0.90 (0.007)	0.00351 (0.00009)	0.34

Fig. 4.1.

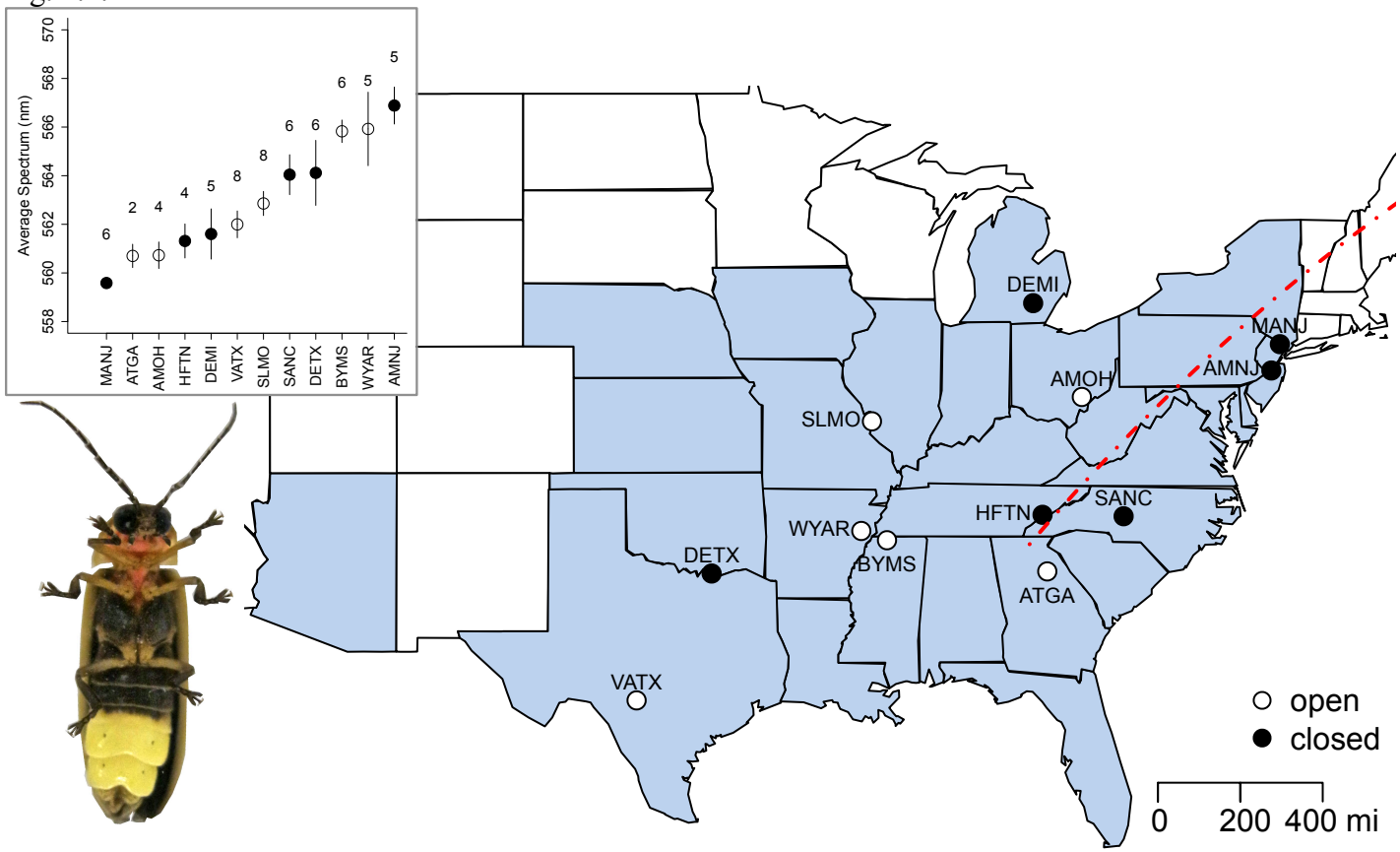


Figure 4.1. Distribution and sampling sites of *Photinus pyralis*

Shading indicates states with at least one record of the species presence (Lloyd 1966). Michigan was added in Hall and colleagues (*in review*). Populations are indicated by their 4-letter location code and colored by habitat type. The red dotted line indicates the approximate position of the Appalachian Mountains. Inset (top left) is the distribution of male mean peak emission spectrum (+/- 1 SE) ordered by increasing wavelength (from yellow to more orange). Sample sizes for emission measurements are shown above. Lower left shows the ventral aspect of an adult male.

Fig. 4.2.

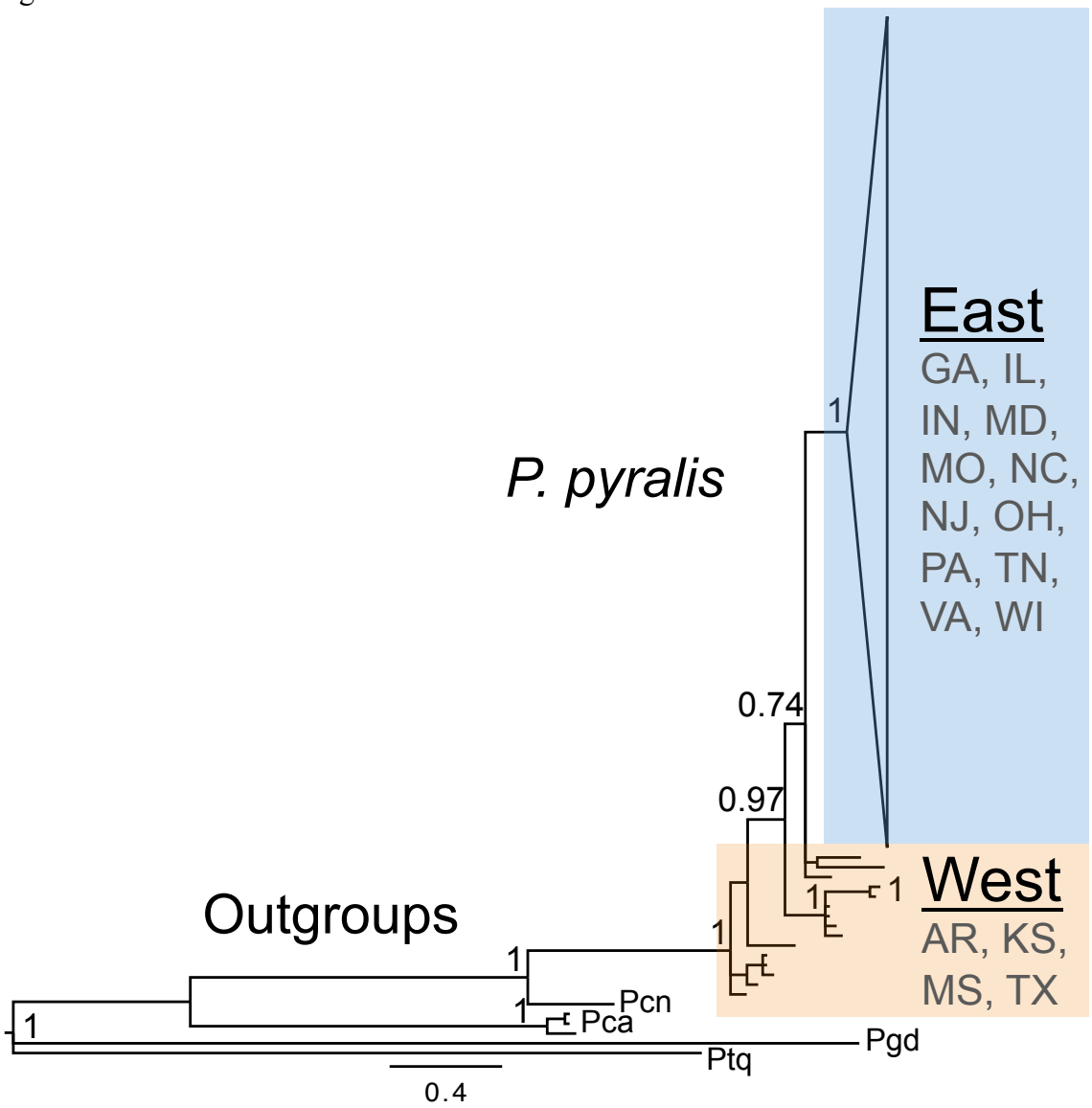


Figure 4.2. *Photinus pyralis* from Eastern states are derived from Western states

This phylogeny, created from 555 bp of *COI* sequence, confirms the monophyly of *P. pyralis* and shows high support for Western populations being basal to Eastern populations. Abbreviations beneath regions are the states in which the specimens were collected. The cartoon depicts a large polytomy at the base of all Eastern specimens. Numbers next to nodes indicate Bayesian support values. Values over 0.7 are shown. Pcn: *Photinus concisus*, Pca: *Photinus carolinus*, Pgd: *Pyropyga decipiens*, Ptq: *Photuris quadrifulgens*. The full phylogeny is given in Appendix C, Figure S1.

Fig. 4.3.

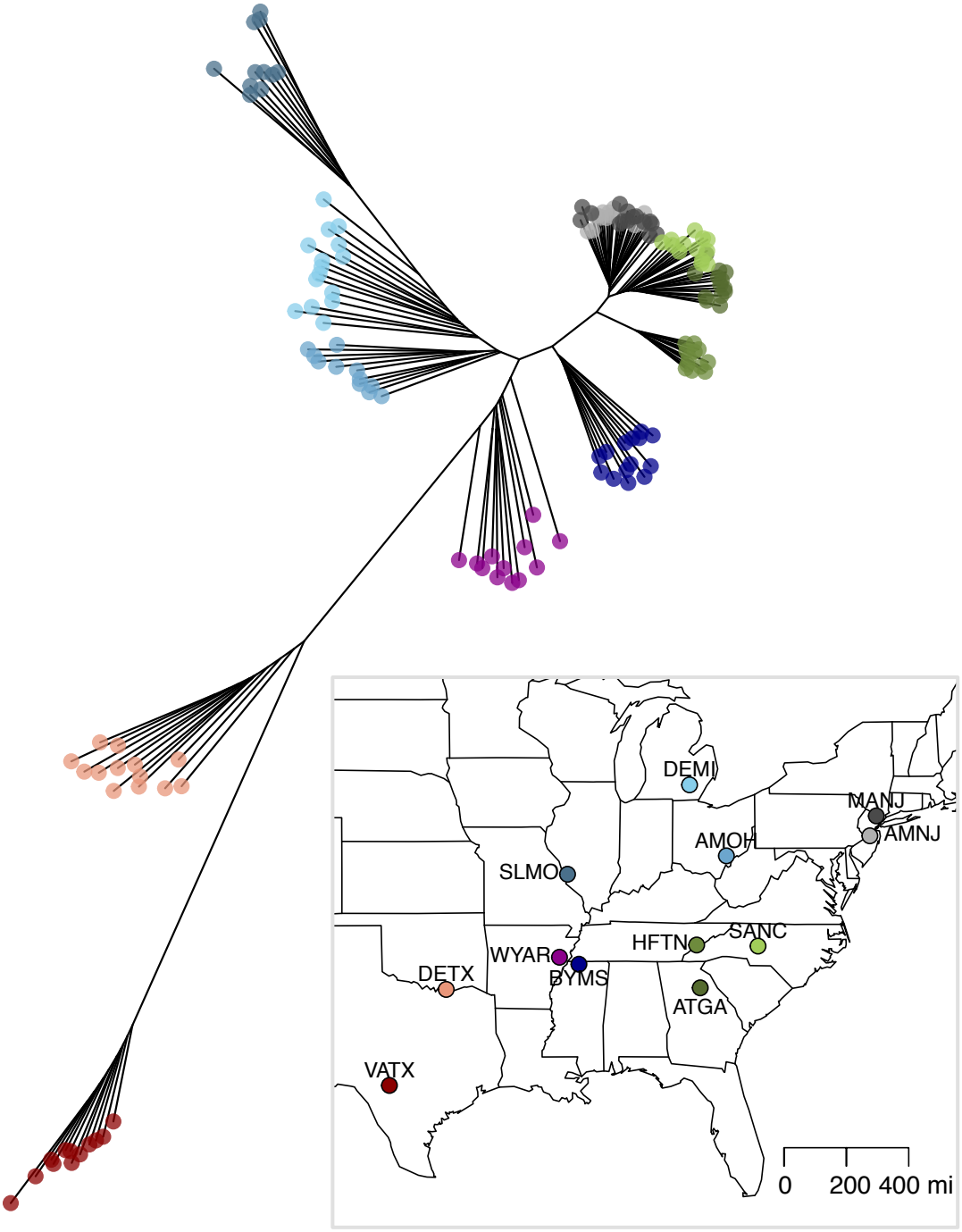


Figure 4.3. Neighbor-joining dendrogram of 154 *Photinus pyralis* from 12 populations

Unrooted dendrogram constructed from genetic distances between individuals (Nei 1978) based on 716 variable SNPs. Individual specimens are labeled using colored circles that correspond to their population of origin (inset map). The four-letter location codes correspond to the populations given in Table 4.1.

Fig. 4.4.

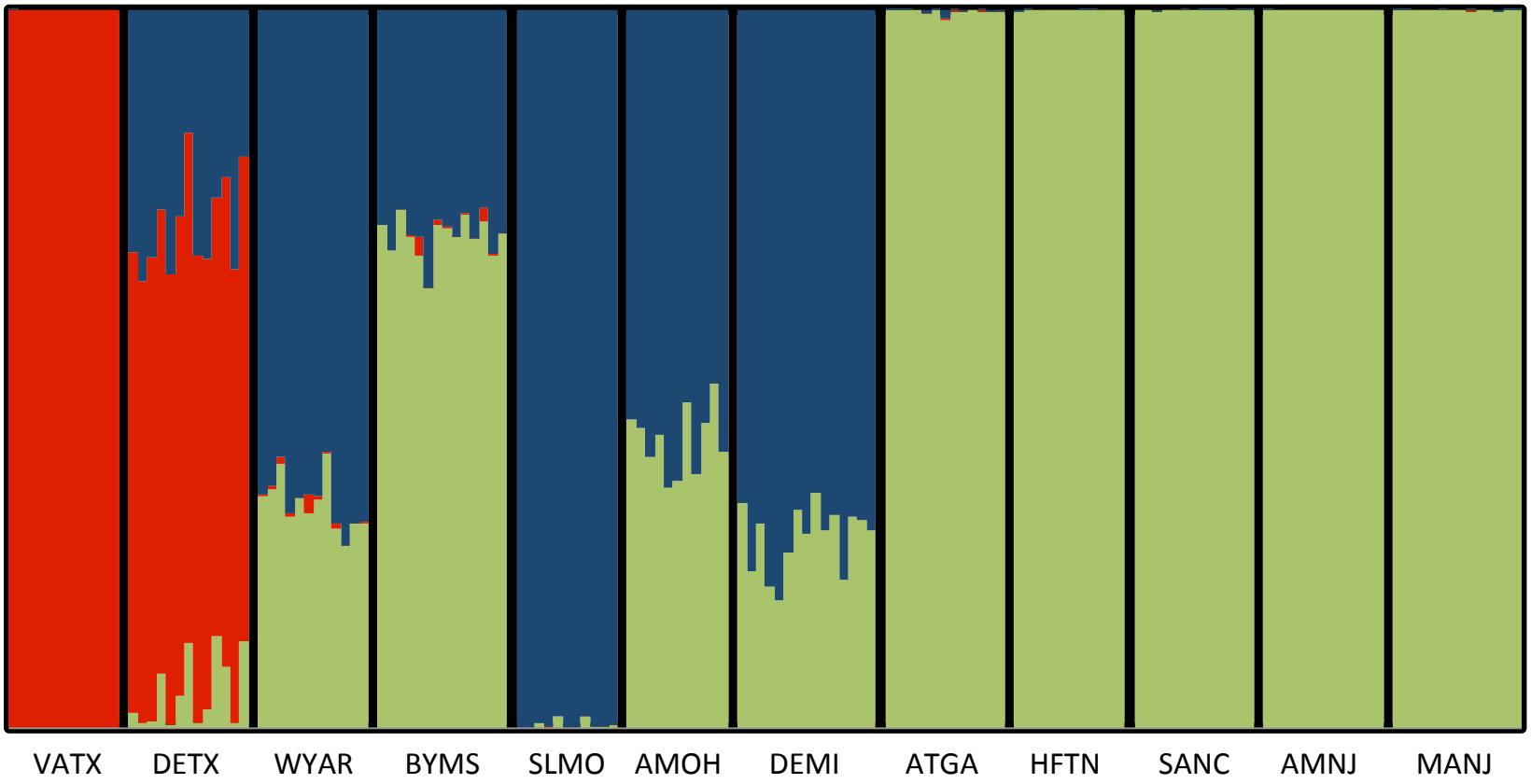


Figure 4.4. STRUCTURE results for K=3

Posterior probabilities of membership in each of three clusters across 154 individuals. Columns show stacked probabilities of membership for each individual. Dark lines separate individuals by population of origin, indicated by the locality code below. Populations are ordered left to right by Great Circle distance from VATX. These three clusters roughly correspond to three regions: Texas (red), mid-Western (blue), and Eastern (green), with evidence for admixture between adjacent regions.

Fig. 4.5.

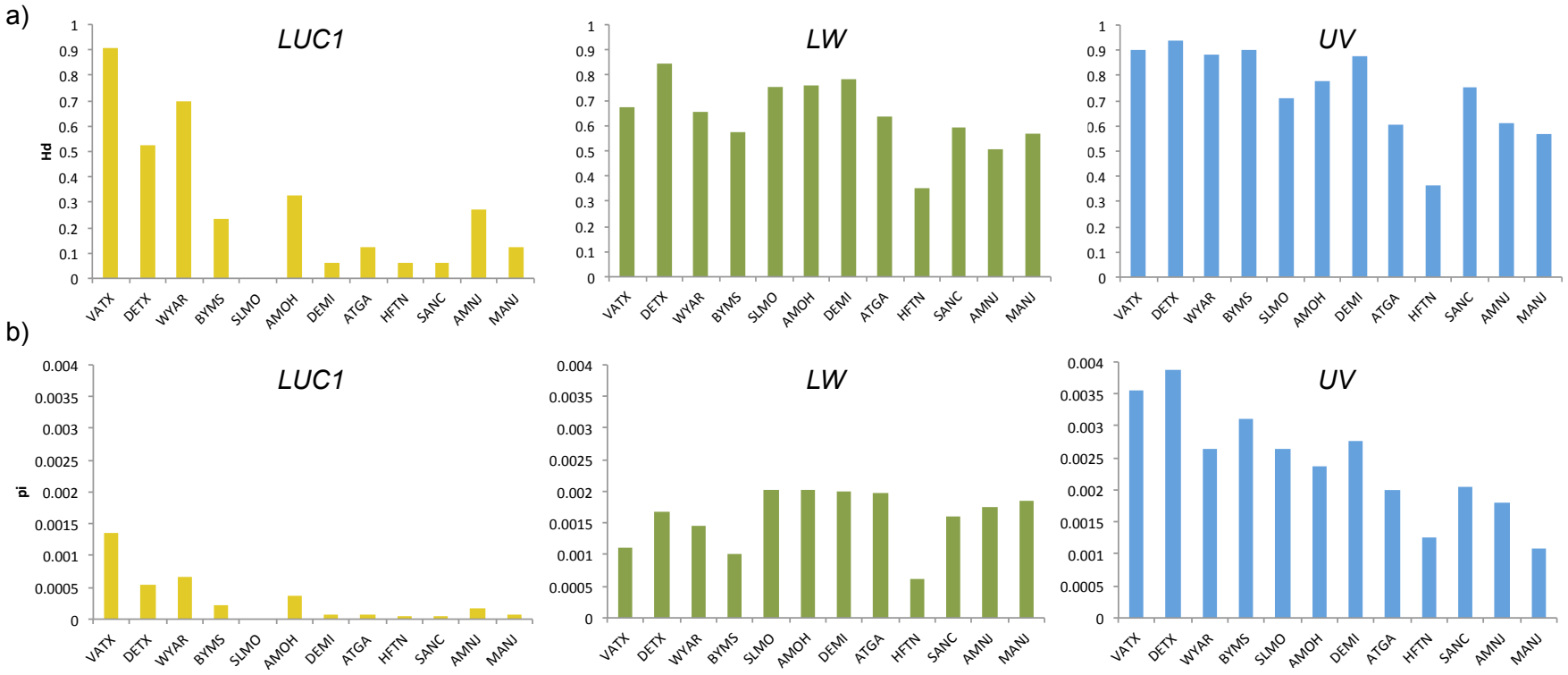


Figure 4.5. Diversity statistics across populations at three loci involved in signal production/reception

Luciferase has much lower diversity than LW or UV opsin. (a) Haplotype diversity (Hd), (b) nucleotide diversity (π), adult luciferase (gold), LW opsin (green), UV opsin (blue).

Fig. 4.6.

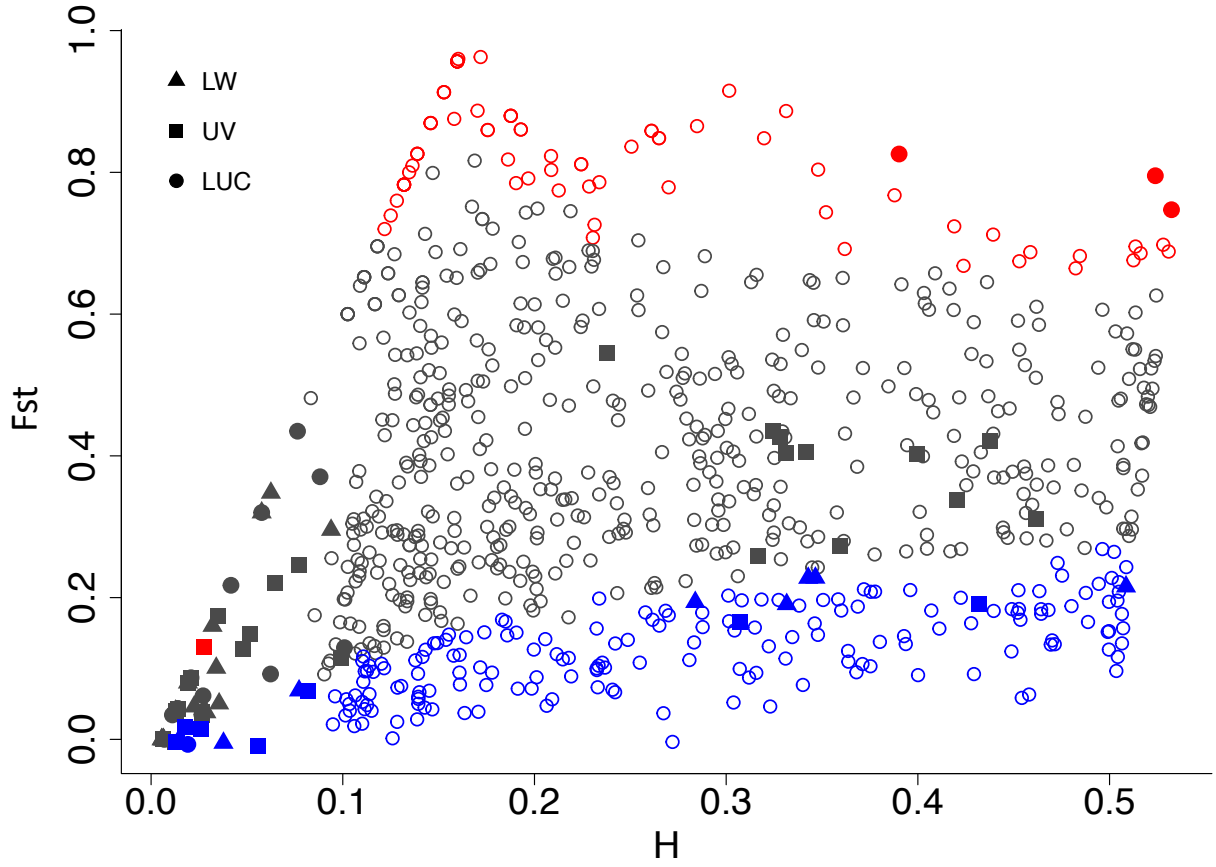


Figure 4.6. Fst outlier plot

Fst versus expected heterozygosity (H) for 716 RAD loci (open circles) plus SNPs (filled symbols) from luciferase (LUC), LW opsin, and UV opsin. Simulations in LOSITAN (Antao *et al.* 2008) estimated the distribution of Fst to identify outlier loci. Loci in red (top 5% of Fst values) and blue (bottom 5% of Fst values) are candidates for diversifying and balancing selection respectively. Loci in gray are candidates for neutral evolution. The three luciferase SNPs with high heterozygosity (top right) are also in the top 5% of Fst values.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

The diversity of animal signals is a long-standing enigma in evolutionary biology (Darwin 1871). In order for a signal to be successful, it must pass through the environment to be detected by a receiver (Greenfield 2002), leading to the expectation that signals, receptors, and the environments in which signals are displayed are evolutionarily linked (Endler 1992). Similarly, the genes underlying signal production and reception are expected to evolve in concert. However, studying how signals and receptors (and their genetic basis) evolve with respect to the environment is often difficult because both signal production and reception can be complex physical or behavioral traits underlain by the effects of multiple genes.

Fireflies have long been used as a system to study the evolution of signals, receptors, and environments (e.g. Biggley, et al. 1967; Lloyd 1966). While fireflies are primarily known for their species-specific flash patterns, they vary across species in signal mode (lighted flashes versus unlighted pheromone signals) and, both across and within nocturnal species, in light signal color. Nocturnal species also vary in the time and place they are active, leading to differences in the ambient light environments in which signals are displayed. Finally, signal production and visual reception are thought to have a simple genetic basis, with signal color governed by the enzyme luciferase and visual reception by opsin proteins in the eye.

In this work, I capitalized on these advantages to conduct two studies on the variation in signal reception and/or production genes with respect to variation in signal color and light environment both across and within species. Looking across species, I performed transcriptome

and genome sequencing to discover opsins and found only two opsins, one sensitive to long wavelengths and one to ultraviolet wavelengths, in the 10 species examined. This was surprising, given that fireflies photoreceptors are sensitive to blue wavelengths as well. Examining sequence variation in these opsins and an additional 28 species led to the discovery of specific amino acid changes that are correlated with changes in signal mode. Several of these sites have been shown to be important in other insect lineages.

Across 12 populations within a single species, I found no amino acid changes in either luciferase or long wavelength opsin that could account for the observed differences in signal color, and inferred differences in visual sensitivity (based on work showing exact matches between visual sensitivity and light color; e.g. Lall, et al. 1980) and light environment (based on vegetation cover; Endler 1993). Unexpectedly, I found signatures of selection on luciferase, despite the lack of coding variation, suggesting that the selection is on luciferase regulation or a locus in linkage disequilibrium with luciferase. Luciferase allele frequencies correlated with differences in signal color and vegetation cover, further pointing to selection on luciferase rather than a linked locus. This represents the first evidence of selection on luciferase across firefly populations.

In developing the genomic resources to conduct the first two projects, I discovered substantial variation in genome size across and within North American firefly species. I investigated proximate and ultimate explanations for this variation and found that recent expansions in repetitive DNA could not explain the variation in genome size I observed. A negative relationship between genome size and range size suggests that the selection-drift barrier plays a role in genome size evolution in fireflies, where selection is less efficient in species with

small effective population sizes, leading to the accumulation of large genome sizes (Lynch and Conery 2003).

The findings from these studies represent some of the first fruits of bringing the study of fireflies signals into the genomic era. Now that candidate sites involved in visual sensitivity have been identified in opsins, functional studies can explore the effects of the identified substitutions. Recent development of efficient insect opsin expression systems will aid in this effort (Frentiu, et al. 2015). Genomic sequencing, assembly, and annotation of *Photinus pyralis* would help to investigate regulatory domains and genes near luciferase. This sequencing is already underway. Genome re-sequencing of the populations in this study and investigation of the regions near luciferase that show signatures of selection will help develop hypotheses as to genes or putative regulatory regions that may be involved. While this represents a substantial amount of work, it will dramatically expand the genetic resources for fireflies, with implications for biomedical research. Since firefly luciferase is widely used as a bioluminescent reporter, there is interest in particular interest in investigating variation in color (Branchini, et al. 2007; de Wet, et al. 1987; de Wet, et al. 1985; Kajiyama and Nakano 1991) and the genetic basis of luciferin (its substrate) synthesis (Day, et al. 2004).

In addition, this work has opened a new avenue of investigation in fireflies- variation in genome size and content within and across species. Future studies using deeper sequencing depth will be able to develop a more complete picture of the repeat landscape in fireflies. Validation of the presence and location of abundance repeats using fluorescent in situ hybridization (FISH) may illuminate the contribution of B chromosomes to repeat content and genome size variation within species. From this work, it is clear that fireflies will continue to be a fascinating system in which to investigate signal evolution, as well as many other fascinating biological questions.

References

Biggley WH, Lloyd JE, Seliger HH (1967) The spectral distribution of firefly light II. *The Journal of General Physiology,* **50**, 1681-1692.

Branchini BR, Ablamsky DM, Murtiashaw MH, Uzasci L, Fraga H, Southworth TL (2007) Thermostable red and green light-producing firefly luciferase mutants for bioluminescent reporter applications. *Analytical biochemistry*, **361**, 253-262.

Darwin, Charles (1871) *Sexual Selection and the Descent of Man*. Murray, London.

Day JC, Tisi LC, Bailey MJ (2004) Evolution of beetle bioluminescence: the origin of beetle luciferin. *Luminescence*, **19**, 8-20.

de Wet JR, Wood KV, DeLuca M, Helinski DR, Subramani S (1987) Firefly luciferase gene: structure and expression in mammalian cells. *Molecular and Cellular Biology*, **7**, 725-737.

de Wet JR, Wood KV, Helinski DR, DeLuca M (1985) Cloning of firefly luciferase cDNA and the expression of active luciferase in *Escherichia coli*. *Proceedings of the National Academy of Sciences U.S.A.*, **82**, 7870-7873.

Endler JA (1993) The color of light in forests and its implications. *Ecological Monographs*, **63**, 2-27.

Endler JA (1992) Signals, signal conditions, and the direction of evolution. *The American Naturalist*,**139**, S125-S153.

Frentiu FD, Yuan F, Savage WK, Bernard GD, Mullen SP, Briscoe AD (2015) Opsin clines in butterflies suggest novel roles for insect photopigments. *Molecular Biology and Evolution*, **32**, 368-379.

Greenfield MD (2002) *Signalers and Receivers: Mechanisms and Evolution of Arthropod Communication*. New York: Oxford University Press.

Kajiyama N, Nakano E (1991) Isolation and characterization of mutants of firefly luciferase which produce different colors of light. *Protein Engineering*, **4**, 691-693.

Lall AB, Seliger HH, Biggley WH, Lloyd JE (1980) Ecology of colors of firefly bioluminescence. *Science*, **210**, 560-562.

Lloyd JE (1966) Studies on the flash communication system in Photinus fireflies. In. *Miscellaneous Publications*. Ann Arbor, Michigan: Museum of Zoology, University of Michigan.

Lynch M, Conery JS (2003) The origins of genome complexity. *Science*, **302**, 1401-1404.

APPENDIX A

Supporting information for Sander and Hall, *in press*
Molecular Ecology

Table S1. Collection and Genbank information for specimens and loci used in this study.
The KSH column gives the unique identification number of each specimen in the Stanger-Hall collection at the University of Georgia. Unless indicated by superscripts, opsin sequences were acquired using PCR of genomic DNA. Where more than one specimen number is listed per species, the consensus opsin sequence was used in the final analysis. Collection localities for specimens listed as (1) are given in Stanger-Hall and Lloyd (2015).

			Accession Number	
Species	KSH	State	LW-opsin	UV-opsin
Ellychnia bivulneris	465	(1)		
Ellychnia corrusca	611, 9734	(1), PA		
Lucidota atra	8874[mh]	GA		
*Phausis reticulata ATGA**	8700[g]	GA		
*Phausis reticulata TCGA**	8729-31[mh], 8733[mh], 8734[mh], 8737[mh]	GA		
Photinus aquilonius	1383	(1)		
Photinus ardens	1763, 1764	(1)		
Photinus australis	8109[mh], 8850	GA		
Photinus brimleyi	9013, 9014, 9015	TN		
Photinus carolinus	8325[mh], 9522, 10098	PA, GA		
Photinus concisus	226	(1)		
Photinus consanguineus	7A	(1)		
Photinus consimilis	1054, 1055	(1)		
Photinus cooki	1689, 9026	(1), TN		
Photinus curtatus	9173	IL		
Photinus dimissus	750, 11059	(1), TX		
Photinus floridanus	1016	(1)		
Photinus granulatus	1780	KS		
Photinus greeni	Pg23	(1)		
Photinus ignitus	417	(1)		
Photinus indictus	1569	(1)		
Photinus knulli	3059	(1)		
Photinus macdermotti	8702[mh], 8705[ml], 8706[fl], 8713	GA		
Photinus marginellus	9855, 10258	PA		

Photinus obscurellus	10076	PA
Photinus punctulatus	13A	(1)
Photinus pyralis	8175[mh], 8176 [ml], 8369 [ll], 8819 [g], 9311	GA, TN, OH
Photinus sabulosus	11139	AL
Photinus scintillans	8001[g], 9865[mh], 10230	PA
Photinus tanytoxis	Pt1, Pt2	(1)
Photinus tenuicinctus	1441	(1)
Photinus texanus	11026	TX
Photuris frontalis	8048[mh]	GA
Photuris sp.[a]	8869, 8879[mh], 8870	GA
Pyractomena angulata	9060, 9224, 9225	MO, IN
Pyractomena borealis	8630[g], 8637[mh], 10861	GA
Pyractomena linearis	10104	PA
Pyractomena marginalis	9348	OH

[mh] transcriptome: adult male head, [ml] transcriptome: adult male light organ, [fl] transcriptome: adult female light organ, [ll] transcriptome: larval light organ, [g] whole-genome sequencing

* ATGA: Athens, GA; TCGA: Tate City, GA

[a] Our inability to identify the *Photuris* species is not unexpected given that members of this genus are often morphologically indistinguishable, exhibit variable flash patterns in the field, and molecular markers do not always resolve species. *Pt. frontalis* is an exception—individuals are easily distinguishable in the field due to their synchronous flash pattern and form a monophyletic group using COI sequence data (data not shown). In order to sample thoroughly across genera, we included the maximum number of *Photuris* we could confidently assign to different species, including *Pt. frontalis* specimens and a single unidentified *Photuris* specimen for which we had light emission data.

Table S2. Forward (F) and reverse (R) primer sequences and their empirically determined annealing temperatures (T_A).
Primers were specifically designed for this study and are shown in pairs (F and R). Most primers were capable of amplifying most or all species from one or more genera (panel A). Species-specific primers had to be designed for UV opsin for three species (panel B). The five genera are *Photinus* (*Pn*), *Pyractomena* (*Py*), *Lucidota* (*Ld*), *Photuris* (*Pt*) and *Phausis* (*Pa*).

A. Genus-specific primers

Primer	Sequence 5'-3'	T_A (ºC)	Amplification*				
LW opsin			*Pn*	*Py*	*Ld*	*Pt*	*Pa*
LWT2_-97F	VGSGRGTTGCAGTTTAGRGC	64-57	x	x	x		
LWT2_568R	WAYGCAATCATSGTCATDGTCC						
LWT2_329F	TGGMAAYGGMATGGTWATCTAC	60-53	x	x	x	x	x
LWT2_1004R	HGGWGTCCAMGCCAWRAACC						
LWT2_868F	MRGYHATGAGAGAACAAGC	50	x	x	x		
LWT2_1231R	RSMYGSYTTTTCWTCTGAYG						
LWP_807F	ATGTTCCGGAAGGAAACWTGGC	60-53	x	x			
LWP_1503R	TTCTTCTGATGYCSCRGTGGCTGC						
LWT1_-8F	ATMTYAYAATGTCRGTGTTG	53				x	x
LWT1_1072R	GACCAKATTGTYGCWAWRGG						
LWT1_660F	TCRNTMGGSTGGACTRTYGC	56				x	x
LWT1_1395R	TTGCAAGGRTTTATTSTG						
LWphauIA_F	CCTCTTTTAYTCCTGGTTTG	59-52					x
LWphauIF_R	TGTTCTCTCATTGCCTTTTC						
LWphauC_F	CCTTTCTTGGTGGTTGTG	60-53					x
LWphau_R	CCAGCATTCCATATATTTGAC						

UV opsin							
UVP_-113F	ATGYAATTAATTAAAGGTCG	65-58	x		x		
UVP_2352R	GTGCGAGTTARTTTYCCTTC						
UV_-10F	GRARACATGMTNNTGCATAACGC	65-58	x	x			
UV_5211R	GWRCAHKRNGTNARRAAHCC						
UVa_-14F	GTTTGAATGAARACATGCTNS	59-52	x	x			
UVP_2352R	Sequence given above						
UV_593F	HATGATGHTRAARAYRCC	59-52	x	x	x		x
UV_1134R	BARRAARCANACCATWATDGC						
UVP_997F	SANGTHATGCAYCAYGARMARGC	65-58	x				
UVP_1435R	THMAWYNHYTACGCYGTYGC						
UV_999F	RGTHATGCAYCAYGARAARGC	64-57	x	x			
UV_1425R	GTHGSDGCYKGYTCRKTHG						
UphauBA_F	AATGTGAGTGAGATCAAATG	58-51		x	x		x
UphauBD_R	CGAACAGACAAAGAGAACAG						
UphauEA_F	AAATTAACTCGCACAAAAGC	60-53					x
UphauEE_R	CATCAAAACGTAAACTGCTTC						
UVptrmidA_F	TGTAAATTTGGCATTTTGTG	60-53				x	
UVptrmidF_R	TAGAAATAAATTATCAAAGACATTG G						
UptrIA_F	AGATACATGGAATAAACCTCAG					x	
UptrIE_R	TCCCACCATTTTCTCTAAAC						

UptrEB_F	GTTATTTGGTGCTACACGTTC	60-53		x	x	x	x
UptrEE_R	CTTAGGAACTGGGTGTTGAG						
UV_181F	ATYCACATWCCCGWMCATTGG	65-58	x				
UVP_2352R	Sequence given above						
UVpyrmidD_F	TTTTCATTGGTTACAGATCG	60-53		x			
UVpyrmidE_R	TTTTCATGGTGCATAACTTG						
UpyEA_F	GAAAACTAACTCGCACAAAAG	60-53		x			
UpyEC_R	AGTTCGTTGTCCATTTATCG						
UVphmidH_F	ACCGCCTCCAATATGTTC	60-53					
UvphmidF_R	ACTCCCTCTTCCTTAATTGC						
Uphiz_F	TTKTAGGCMAAAAARATG	56	x				
Uphiz_R	ATGTATKAATGAAAACAATTS						
Uphic_F	GTTATGCATCACGAAAAGG	55-48	x				
Uphif_R	TACTTTGGCAATTTTGATTTC						

* x = successful amplification

B. Species-specific primers

Primer	Sequence	T_A (ºC)	Species
UV opsin			
UVP_-113F	Sequence given above		
UiE2A_R	AAATGTAGAGCAAACCAAGG	57-50	*Pn. indictus*
UmacEA_F	CGTTGTCTTATTCATTTGCTC		
UmacEE_R	TCGCCTCCTTGATTTAGAC	60-53	*Pn. macdermotti, Pn. greeni, Pn. consanguineus*
UaE2A_F	GAATGAAAACATGCTCGTG		
UaE2_R	ATGTACACCAAGGCAAGAAG	60-53	*Ld. atra*
UaE3B_F	ATTCACATTCCCGTACATTG		
UaE3E_R	GACAAAGAACCCGTGAAAC	61-54	*Ld. atra*
UaE4C_F	GCGAAGTAACCAATCTATGC		
UaE4E_R	GTTGTGCCACCCTGTATATC	60-53	*Ld. atra*

PCR cycling parameters: Initial denaturation at 94°C for 3 minutes. Then 35 3-step cycles of 94°C for 45 seconds, the appropriate annealing temperature (see above) for 1 minute, and 72°C for 1-2 minutes. There was a final extension at 72°C for 3 minutes. Where a range of annealing temperatures is noted, primers were run in a touchdown protocol with the following modification: 7 cycles of annealing temperature starting at high end of range listed above, stepping down 1°C each cycle, then 28 cycles of annealing at the low temperature (total=35 cycles).

Sanger sequencing: PCR products were bidirectionally sequenced using a BigDye® Terminator v3.1 Cycle Sequencing Kit (Life Technologies) on an Applied Biosystems 3730xl at the Georgia Genomics Facility (Athens, GA).

Table S3. Genbank information for species used to extend the phylogeny of Stanger-Hall and Lloyd (2015).
WG: wingless, CAD: rudimentary, COI: cytochrome oxidase.

		Accession Number		
Species	KSH	WG	CAD	COI
Phausis reticulata ATGA	8700			
*Phausis reticulata GSNP**	8410			
Photuris frontalis	8180			
Photuris sp.	9032			
Pyractomena angulata	9060			
Pyractomena borealis	10861			
Pyractomena linearis	10104			
Pyractomena marginalis	9348			

*GSNP: Great Smoky Mountains National Park. *Pa. reticulata* from the Great Smoky Mountains form a monophyletic clade with *Pa. reticulata* from Tate City, GA, the specimen from which we were able to amplify opsins (data not shown).

Table S4. Blomberg's K for ecological and signal traits.

Blomberg's K is a measure of phylogenetic signal (tendency of species to resemble each other) in the observed data relative to that of traits that evolved under a Brownian motion model (Blomberg et al. 2003). K = 1 indicates that the observed data exhibit the amount of signal expected under Brownian motion, K < 1 indicates less signal than expected and K > 1 indicates more resemblance (phylogenetic "clumping") than expected. Presence of statistically significant phylogenetic signal is calculated using permutation tests that randomly shuffle tip taxa to generate a distribution of K values. For the two traits with significant signal, spectra and activity time, K values are less than 1, indicating that, while sister taxa resemble each other, the resemblance is less than expected under a Brownian motion model of trait evolution.

Trait	K	P
Diurnal/nocturnal	0.08	0.16
Spectra	0.14	0.002*
Habitat	0.06	0.06
Activity time	0.12	0.01*

*=significant at $p < 0.05$.

Table S5. Amino acid sites that interact with the chromophore as determined by homology modeling.

Amino acid sequence at these sites is generally conserved across species within opsin type. Site numbers are given in reference to the full-length *Pn. pyralis* LW opsin and UV opsin amino acid sequences. -- = a binding site in one opsin without an equivalent in the other.

LW-opsin sites	Amino acid(s)	UV-opsin sites	Amino acid(s)
105	M	102	M
--	--	105	K
109	M	--	--
132	Y	128	Y, F
133	G	129	G, A
136	G	132	G
137	S	133	S, A
140	G	136	G
--	--	137	I
144	I	--	--
197	Y	193	F
200	E	196	E
205	V, A	201	S
--	--	202	C
207	G	203	S, T
208	T	204	F
210	N, Y	--	--
224	Y	220	L
--	--	221	F
228	V	224	S
229	Y	225	Y
291	W	288	F
295	W	292	W
298	Y	295	Y
299	L	296	A
302	N	299	A, S
303	Y	--	--
320	S	319	A

Total: 24 sites each

Table S6. Comparisons of substitutions in firefly and other organisms

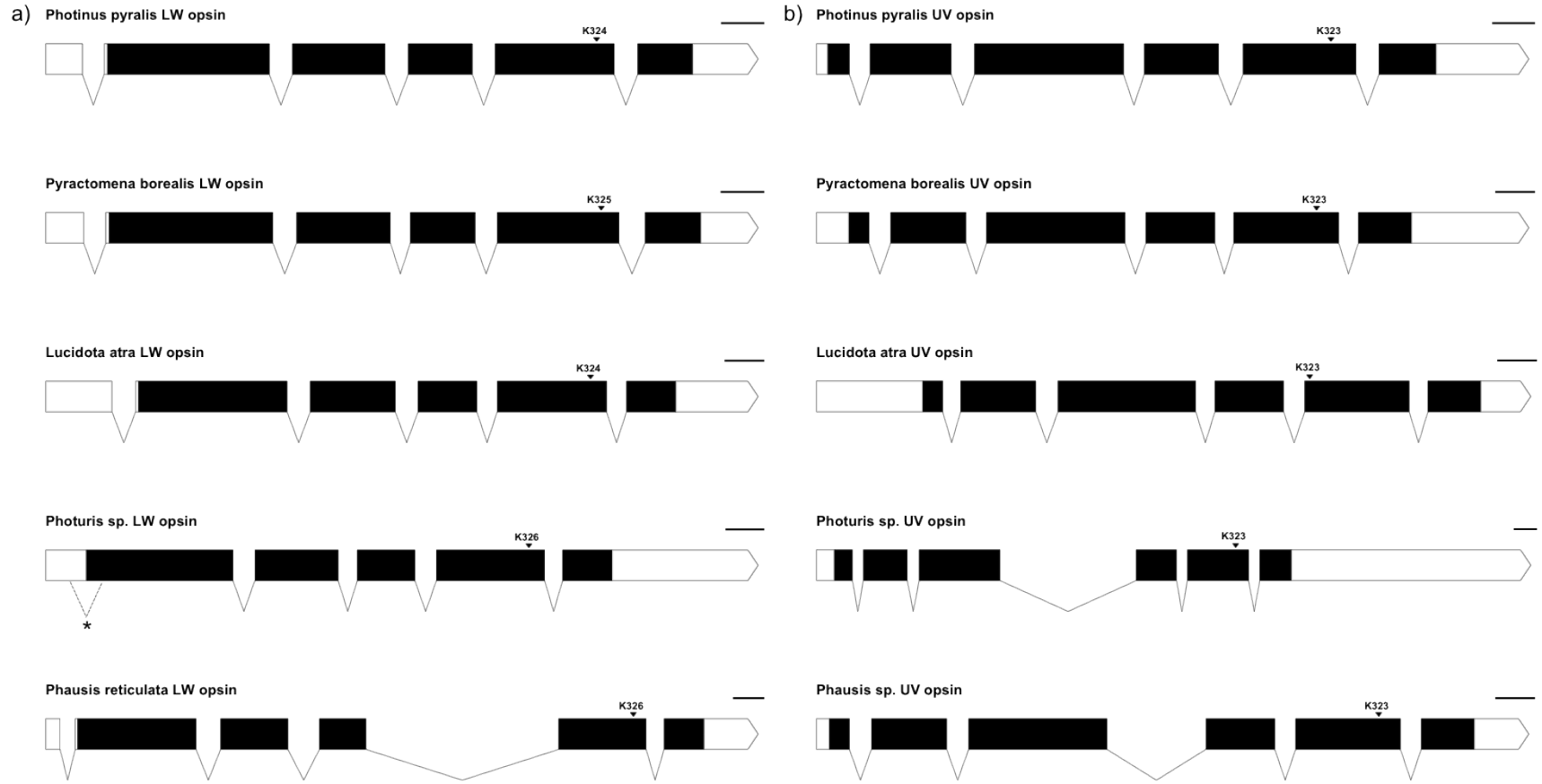
Individual substitutions are examined. Amino acid position (Site) given in reference to the full *Pn. pyralis* sequence. FS: substitution in firefly (before/after), FSP: polarity of amino acids involved in firefly substitution (P: polar, NP: nonpolar), Lineage: branch along which the substitution occurred (MRCA: most recent common ancestor), M: signal mode (D: diurnal, N: nocturnal), ΔSpectra: inferred change in male peak emission- either red shifted or blue shifted (nm), ΔHabitat: inferred change in habitat (0 = open, 1 = mixed, 2 = closed), ΔActivity: inferred change in activity (0 = early, 1 = late), SO: substitution in (other taxon), SOP: polarity of amino acids involved in other substitution (P: polar, NP: nonpolar), ΔAbs: change in absorbance either inferred or measured where applicable. References given in main text Table 3. Inferred changes in spectra, habitat, and activity taken from ancestral state reconstructions using phylogenetic independent contrasts (see Materials and Methods).

Site	FS	FSP	Lineage	M	ΔSpectra	ΔHabitat	ΔActivity	SO	SOP	ΔAbs
LW										
108	C/T	P/P	*Ld. atra*	D	N/A	N/A	N/A			
			MRCA Photinus	N	None	Closed (0.09)	Early (0.18)			
	T/M	P/NP	MRCA *Pn. dimissus, Pn. carolinus*	N	Red (2)	Closed (0.12)	Late (0.2)	V63M Heliconius	NP/NP	Red (20)
			Pn. australis	N	Red (8)	Closed (0.94)	Early (0.63)			
			Pn. brimleyi	N	Red (4)	Closed (0.5)	None			
	-	-	-	-	-	-	-	F112V Liminitis	NP/NP	No data
181	M/L	NP/NP	*Pa reticulata*	N	Blue (7)	Closed (0.8)	Late (0.12)			
			Pt. frontalis	N	Red (6)	Closed (1)	None			
	M/T	NP/P	*Ld. atra*	D	N/A	N/A	N/A			
	M or L/L	NP/NP	MRCA Pyractomena	N	None	Open (0.13)	(0.29)			
	M or L/A	NP/NP	MRCA Photinus	N	None	Closed (0.09)	Early (0.18)	V136A Heliconius	NP/NP	Red (20)
	A/L	NP/NP	MRCA *Pn. dimissus, Pn. carolinus* clades	N	Red (2)	Closed (0.12)	Late (0.2)	V136L Papilio	NP/NP	Red (2)
	A/V	NP/NP	*Pn. concisus*	N	No data	Closed (1)	None			
			El. bivulneris	D	N/A	N/A	N/A			
	A/G	NP/NP	*El. corrusca*	D	N/A	N/A	N/A			
	-	-	-	-	-	-	-	V136M Vanessa	NP/NP	Blue (20)
188	V or I/L	NP/NP	MRCA Pyractomena	N	None	Open (0.13)	Late (0.29)	V143L Papilio	NP/NP	Red (45)
	V/L	NP/NP	MRCA *Pn. obscurellus* clade	N	Blue (3)	Open (0.58)	Late (0.19)			
	V/L	NP/NP	*Pn. tenuicinctus*	N	Red (3)	Open (0.06)	Early (0.44)			
	V/I	NP/NP	MRCA *Pn. macdermotti* clade	N	Red (1)	Open (0.26)	Late (0.35)	V143I Heliconius	NP/NP	Red (20)

	V/I	NP/NP	*Pn. granulatus*	N	No data	Open (0.75)	Late (.4)			
	V/I	NP/NP	*Pn. brimleyi*	N	Red (4)	Closed (0.5)	None			
356	A/S	NP/P	*Pa reticulata*	N	Blue (7)	Closed (0.8)	Late (0.12)			
			Pt. frontalis	N	Red (6)	Closed (1)	None			
			MRCA Ellychnia	D	N/A	N/A	N/A			
UV										
			Pn. australis	N	Red (8)	Closed (0.94)	Early (0.63)			
	A/G	NP/NP	MRCA Pyractomena	N	None	Open (0.13)	Late (0.29)			
59	V/L	NP/NP	MRCA Pyractomena	N	None	Open (0.13)	Late (0.29)			
			Pn. indictus	D	N/A	N/A	N/A			
			MRCA *Pn. obscurellus* clade	N	Blue (3)	Open (0.58)	Late (0.19)			
			MRCA *Pn. concisus, Pn. pyralis*	N	No data	Closed (0.17)	Early (0.41)			
			El. bivulneris	D	N/A	N/A	N/A			
	V/M	NP/NP	MRCA *Pn. punctulatus, Pn. scintillans*	N	No data	Open (0.41)	Early (0.2)			
			Pn. granulatus	N	No data	Open (0.25)	Late (0.4)			
	M/L	NP/NP	*Pn. punctulatus*	N	No data	Open (1)	Late (0.5)			
133	S/A	P/NP	*El. bivulneris*	D	N/A	N/A	N/A	S116A *Pieris*	P/NP	Blue (13)

Figure S1. Opsin gene structure is conserved across species.

Typical gene structure for each genus: (a) LW opsin and (b) UV opsin. The conserved lysine residue at the chromophore binding site for each opsin is noted. Interestingly, LW opsins show a conserved 5' UTR intron upstream of exon. Scale bar: 100 bp, light bar = UTR, dark bar = translated.

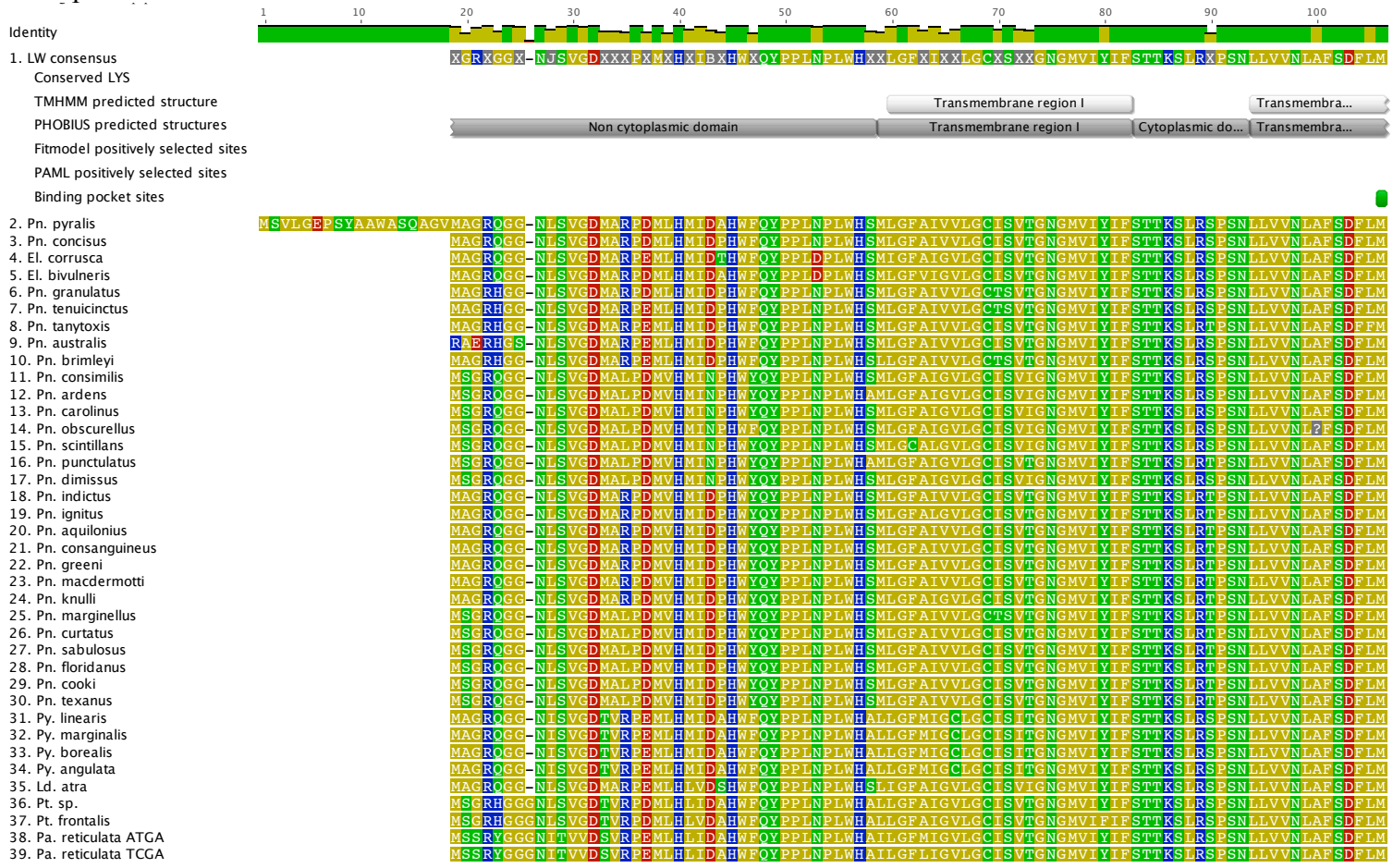


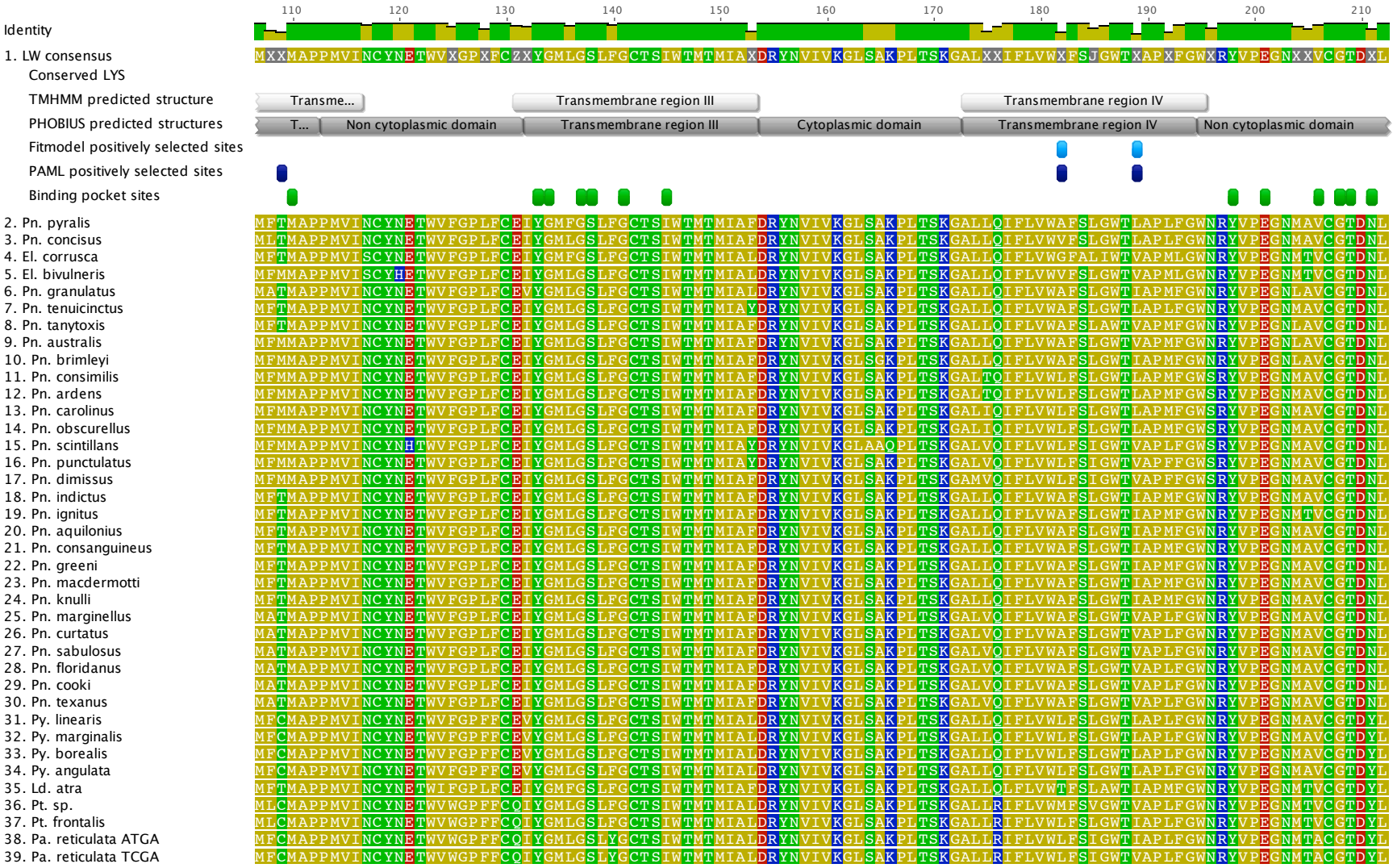
*Given the conservation of the 5' UTR intron across the rest of the species' LW-opsins, *Photuris* LW was expected to have a 5' UTR intron. However, our degenerate primers were not able to amplify across the start site, nor was genomic sequencing performed on a *Photuris* specimen. Therefore, a confirmed 5' UTR intron is not shown, but is likely present given the conservation across all other species.

Figure S2. LW and UV amino acid alignments used in PAML and fitmodel analysis.

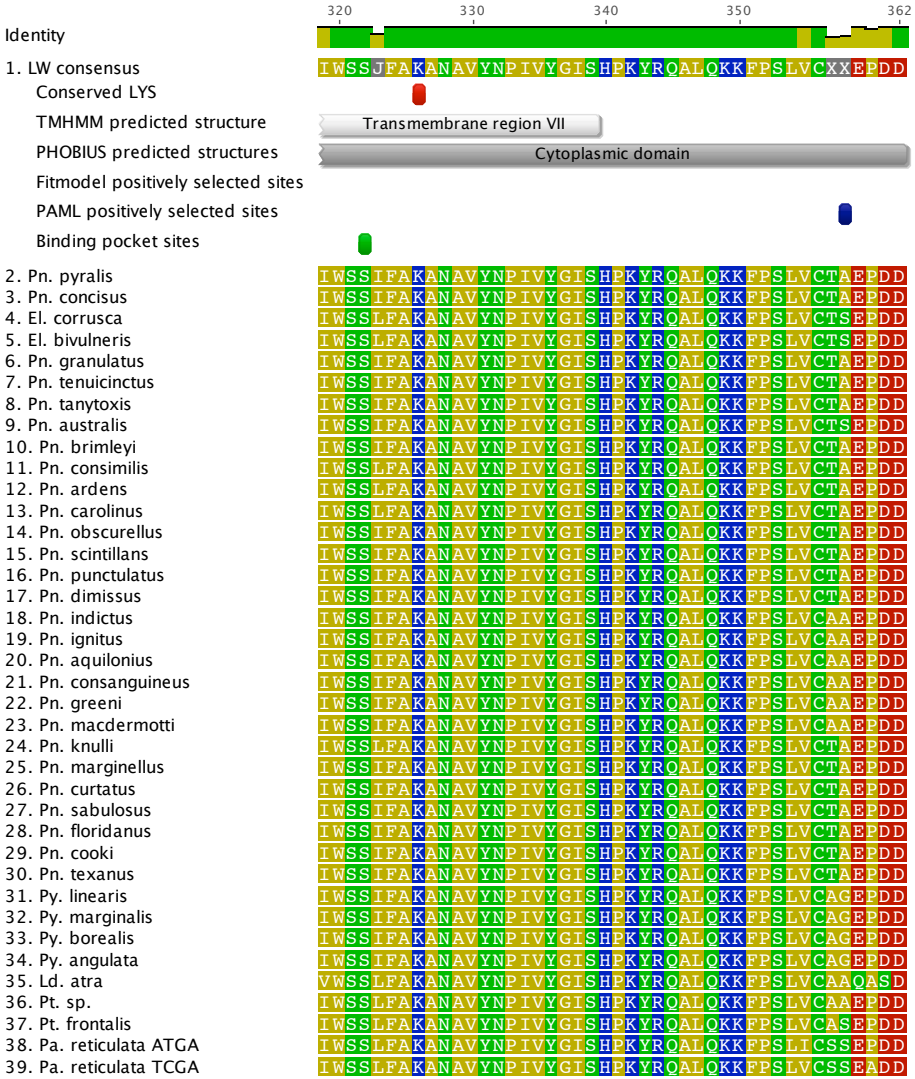
Opsin amino acid sequence is highly conserved within (a) LW and (b) UV opsin. Due to primer constraints, the LW opsin alignment used in analysis begins 18 amino acids after the start site and ends 18 amino acids upstream of the stop codon. Below, the LW sequence from *Pn. pyralis* is shown from the start codon for illustration. The UV alignment begins at the start site and ends 6 amino acids upstream of the stop codon. Amino acid sites are shaded by polarity. In addition to alignment, structural motifs and sites that are identified as being under selection are indicated.

a) LW opsin




110
120
130
140
150
160
170
180
190
200
210
Identity
1. LW consensus
Conserved LYS
TMHMM predicted structure
Transme...
Transmembrane region III
Transmembrane region IV
PHOBIUS predicted structures
T...
Non cytoplasmic domain
Transmembrane region III
Cytoplasmic domain
Transmembrane region IV
Non cytoplasmic domain
Fitmodel positively selected sites
PAML positively selected sites
Binding pocket sites
2. Pn. pyralis
3. Pn. concisus
4. El. corrusca
5. El. bivulneris
6. Pn. granulatus
7. Pn. tenuicinctus
8. Pn. tanytoxis
9. Pn. australis
10. Pn. brimleyi
11. Pn. consimilis
12. Pn. ardens
13. Pn. carolinus
14. Pn. obscurellus
15. Pn. scintillans
16. Pn. punctulatus
17. Pn. dimissus
18. Pn. indictus
19. Pn. ignitus
20. Pn. aquilonius
21. Pn. consanguineus
22. Pn. greeni
23. Pn. macdermotti
24. Pn. knulli
25. Pn. marginellus
26. Pn. curtatus
27. Pn. sabulosus
28. Pn. floridanus
29. Pn. cooki
30. Pn. texanus
31. Py. linearis
32. Py. marginalis
33. Py. borealis
34. Py. angulata
35. Ld. atra
36. Pt. sp.
37. Pt. frontalis
38. Pa. reticulata ATGA
39. Pa. reticulata TCGA


220
230
240
250
260
270
280
290
300
310
Identity
1. LW consensus
Conserved LYS
TMHMM predicted structure
PHOBIUS predicted structures
Fitmodel positively selected sites
PAML positively selected sites
Binding pocket sites
Transmembrane region V
Transmembrane region VI
Non cy...
Transmembrane region V
Cytoplasmic domain
2. Pn. pyralis
3. Pn. concisus
4. El. corrusca
5. El. bivulneris
6. Pn. granulatus
7. Pn. tenuicinctus
8. Pn. tanytoxis
9. Pn. australis
10. Pn. brimleyi
11. Pn. consimilis
12. Pn. ardens
13. Pn. carolinus
14. Pn. obscurellus
15. Pn. scintillans
16. Pn. punctulatus
17. Pn. dimissus
18. Pn. indictus
19. Pn. ignitus
20. Pn. aquilonius
21. Pn. consanguineus
22. Pn. greeni
23. Pn. macdermotti
24. Pn. knulli
25. Pn. marginellus
26. Pn. curtatus
27. Pn. sabulosus
28. Pn. floridanus
29. Pn. cooki
30. Pn. texanus
31. Py. linearis
32. Py. marginalis
33. Py. borealis
34. Py. angulata
35. Ld. atra
36. Pt. sp.
37. Pt. frontalis
38. Pa. reticulata ATGA
39. Pa. reticulata TCGA

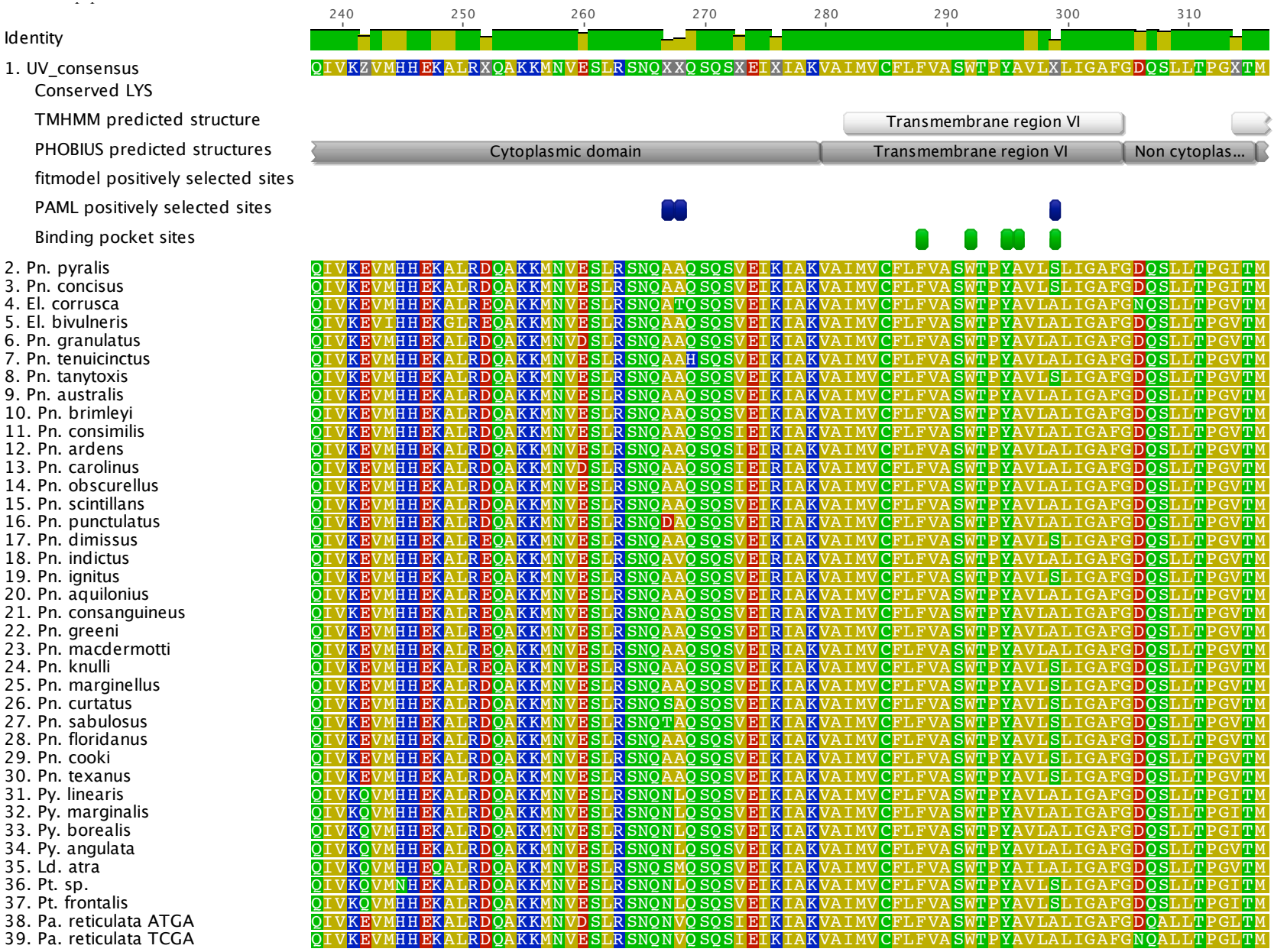

320
330
340
350
362
Identity
1. LW consensus
Conserved LYS
TMHMM predicted structure
Transmembrane region VII
PHOBIUS predicted structures
Cytoplasmic domain
Fitmodel positively selected sites
PAML positively selected sites
Binding pocket sites
2. Pn. pyralis
3. Pn. concisus
4. El. corrusca
5. El. bivulneris
6. Pn. granulatus
7. Pn. tenuicinctus
8. Pn. tanytoxis
9. Pn. australis
10. Pn. brimleyi
11. Pn. consimilis
12. Pn. ardens
13. Pn. carolinus
14. Pn. obscurellus
15. Pn. scintillans
16. Pn. punctulatus
17. Pn. dimissus
18. Pn. indictus
19. Pn. ignitus
20. Pn. aquilonius
21. Pn. consanguineus
22. Pn. greeni
23. Pn. macdermotti
24. Pn. knulli
25. Pn. marginellus
26. Pn. curtatus
27. Pn. sabulosus
28. Pn. floridanus
29. Pn. cooki
30. Pn. texanus
31. Py. linearis
32. Py. marginalis
33. Py. borealis
34. Py. angulata
35. Ld. atra
36. Pt. sp.
37. Pt. frontalis
38. Pa. reticulata ATGA
39. Pa. reticulata TCGA

b) UV opsin


1
10
20
30
40
50
60
70
Identity
1. UV_consensus
Conserved LYS
TMHMM predicted structure
PHOBIUS predicted structures
fitmodel positively selected sites
PAML positively selected sites
Binding pocket sites
Transmembrane region I
Non cytoplasmic domain
Transmembrane region I
2. Pn. pyralis
3. Pn. concisus
4. El. corrusca
5. El. bivulneris
6. Pn. granulatus
7. Pn. tenuicinctus
8. Pn. tanytoxis
9. Pn. australis
10. Pn. brimleyi
11. Pn. consimilis
12. Pn. ardens
13. Pn. carolinus
14. Pn. obscurellus
15. Pn. scintillans
16. Pn. punctulatus
17. Pn. dimissus
18. Pn. indictus
19. Pn. ignitus
20. Pn. aquilonius
21. Pn. consanguineus
22. Pn. greeni
23. Pn. macdermotti
24. Pn. knulli
25. Pn. marginellus
26. Pn. curtatus
27. Pn. sabulosus
28. Pn. floridanus
29. Pn. cooki
30. Pn. texanus
31. Py. linearis
32. Py. marginalis
33. Py. borealis
34. Py. angulata
35. Ld. atra
36. Pt. sp.
37. Pt. frontalis
38. Pa. reticulata ATGA
39. Pa. reticulata TCGA


80
90
100
110
120
130
140
150
Identity
1. UV_consensus
Conserved LYS
TMHMM predicted structure
Transmembrane region II
Transmembrane region III
PHOBIUS predicted structures
Cytoplasmic d...
Transmembrane region II
Non cytoplas...
Transmembrane region III
Cytoplasmi...
fitmodel positively selected sites
PAML positively selected sites
Binding pocket sites
2. Pn. pyralis
3. Pn. concisus
4. El. corrusca
5. El. bivulneris
6. Pn. granulatus
7. Pn. tenuicinctus
8. Pn. tanytoxis
9. Pn. australis
10. Pn. brimleyi
11. Pn. consimilis
12. Pn. ardens
13. Pn. carolinus
14. Pn. obscurellus
15. Pn. scintillans
16. Pn. punctulatus
17. Pn. dimissus
18. Pn. indictus
19. Pn. ignitus
20. Pn. aquilonius
21. Pn. consanguineus
22. Pn. greeni
23. Pn. macdermotti
24. Pn. knulli
25. Pn. marginellus
26. Pn. curtatus
27. Pn. sabulosus
28. Pn. floridanus
29. Pn. cooki
30. Pn. texanus
31. Py. linearis
32. Py. marginalis
33. Py. borealis
34. Py. angulata
35. Ld. atra
36. Pt. sp.
37. Pt. frontalis
38. Pa. reticulata ATGA
39. Pa. reticulata TCGA


160
170
180
190
200
210
220
230
Identity
1. UV_consensus
Conserved LYS
TMHMM predicted structure
Transmembrane region IV
Transmembrane region V
PHOBIUS predicted structures
Cyto...
Transmembrane region IV
Non cytoplasmic domain
Transmembrane region V
fitmodel positively selected sites
PAML positively selected sites
Binding pocket sites
2. Pn. pyralis
3. Pn. concisus
4. El. corrusca
5. El. bivulneris
6. Pn. granulatus
7. Pn. tenuicinctus
8. Pn. tanytoxis
9. Pn. australis
10. Pn. brimleyi
11. Pn. consimilis
12. Pn. ardens
13. Pn. carolinus
14. Pn. obscurellus
15. Pn. scintillans
16. Pn. punctulatus
17. Pn. dimissus
18. Pn. indictus
19. Pn. ignitus
20. Pn. aquilonius
21. Pn. consanguineus
22. Pn. greeni
23. Pn. macdermotti
24. Pn. knulli
25. Pn. marginellus
26. Pn. curtatus
27. Pn. sabulosus
28. Pn. floridanus
29. Pn. cooki
30. Pn. texanus
31. Py. linearis
32. Py. marginalis
33. Py. borealis
34. Py. angulata
35. Ld. atra
36. Pt. sp.
37. Pt. frontalis
38. Pa. reticulata ATGA
39. Pa. reticulata TCGA


240
250
260
270
280
290
300
310
Identity
1. UV_consensus
Conserved LYS
TMHMM predicted structure
PHOBIUS predicted structures
fitmodel positively selected sites
PAML positively selected sites
Binding pocket sites
Transmembrane region VI
Cytoplasmic domain
Transmembrane region VI
Non cytoplas...
2. Pn. pyralis
3. Pn. concisus
4. El. corrusca
5. El. bivulneris
6. Pn. granulatus
7. Pn. tenuicinctus
8. Pn. tanytoxis
9. Pn. australis
10. Pn. brimleyi
11. Pn. consimilis
12. Pn. ardens
13. Pn. carolinus
14. Pn. obscurellus
15. Pn. scintillans
16. Pn. punctulatus
17. Pn. dimissus
18. Pn. indictus
19. Pn. ignitus
20. Pn. aquilonius
21. Pn. consanguineus
22. Pn. greeni
23. Pn. macdermotti
24. Pn. knulli
25. Pn. marginellus
26. Pn. curtatus
27. Pn. sabulosus
28. Pn. floridanus
29. Pn. cooki
30. Pn. texanus
31. Py. linearis
32. Py. marginalis
33. Py. borealis
34. Py. angulata
35. Ld. atra
36. Pt. sp.
37. Pt. frontalis
38. Pa. reticulata ATGA
39. Pa. reticulata TCGA

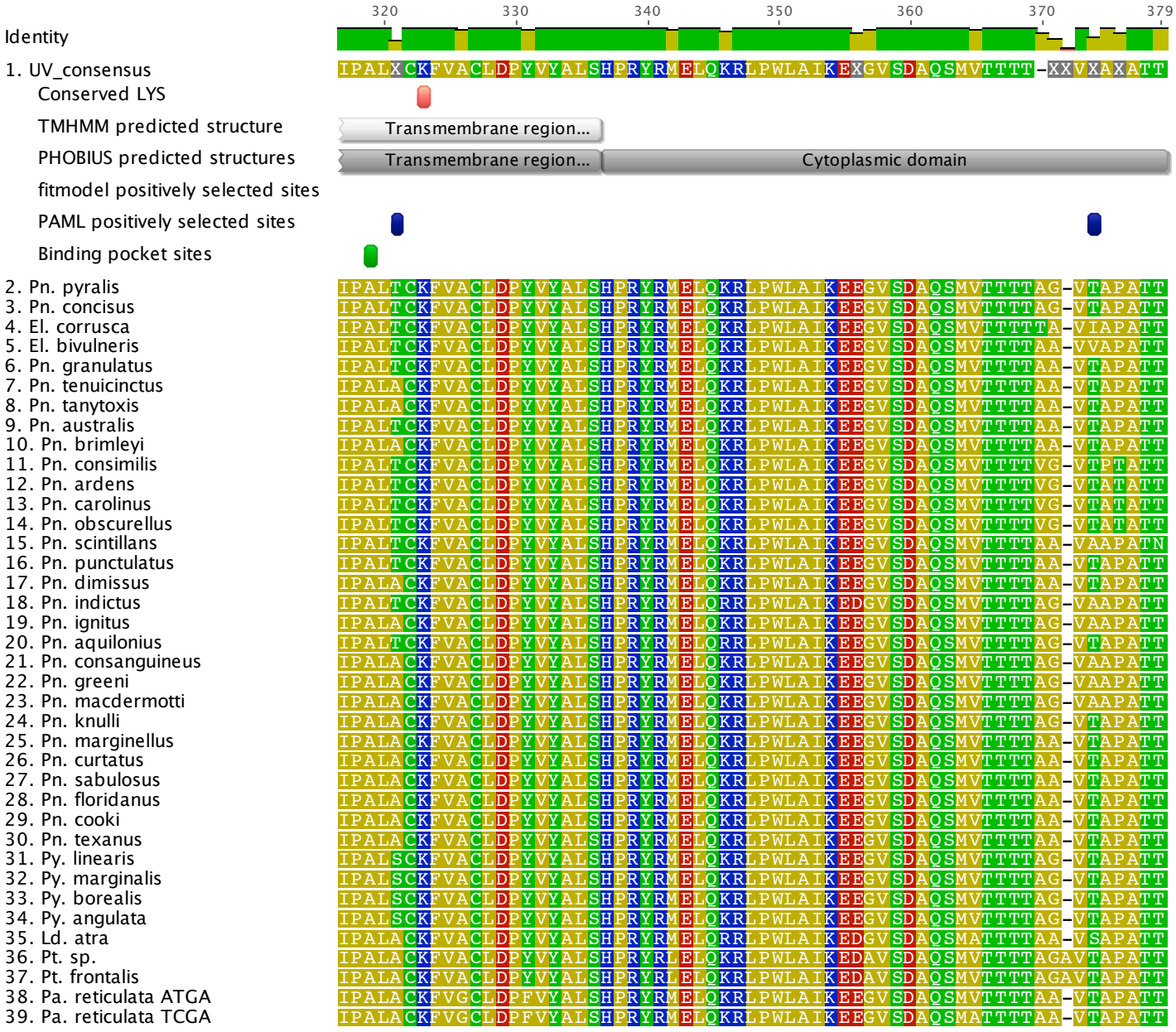

320
330
340
350
360
370
379
Identity
1. UV_consensus
Conserved LYS
TMHMM predicted structure
Transmembrane region...
PHOBIUS predicted structures
Transmembrane region...
Cytoplasmic domain
fitmodel positively selected sites
PAML positively selected sites
Binding pocket sites
2. Pn. pyralis
3. Pn. concisus
4. El. corrusca
5. El. bivulneris
6. Pn. granulatus
7. Pn. tenuicinctus
8. Pn. tanytoxis
9. Pn. australis
10. Pn. brimleyi
11. Pn. consimilis
12. Pn. ardens
13. Pn. carolinus
14. Pn. obscurellus
15. Pn. scintillans
16. Pn. punctulatus
17. Pn. dimissus
18. Pn. indictus
19. Pn. ignitus
20. Pn. aquilonius
21. Pn. consanguineus
22. Pn. greeni
23. Pn. macdermotti
24. Pn. knulli
25. Pn. marginellus
26. Pn. curtatus
27. Pn. sabulosus
28. Pn. floridanus
29. Pn. cooki
30. Pn. texanus
31. Py. linearis
32. Py. marginalis
33. Py. borealis
34. Py. angulata
35. Ld. atra
36. Pt. sp.
37. Pt. frontalis
38. Pa. reticulata ATGA
39. Pa. reticulata TCGA

Figure S3. Distribution of expression for *Pn. pyralis* transcripts in head tissue.

The left panel shows a histogram of Trinity-assembled *Pn. pyralis* transcripts binned by their log expression (fragments per kilobase of transcript per million mapped reads, FPKM). The right shows an outlier boxplot of all transcripts (box: quartiles, whiskers: 1.5*interquartile range, red bracket: shortest half of the data (densest)). Orange indicates transcripts that were in the top 95% of expression as ranked by FPKM values. LW opsin is ranked 3rd highest expressed in head (99.99 percentile) and is shown in green. UV opsin is ranked 25th highest expression in head (99.93 percentile) and is shown in blue. The results for the other 10 species were similar, with both opsins in the top 5% of expressed transcripts in head tissue and LW opsin expressed at a higher level than UV opsin.

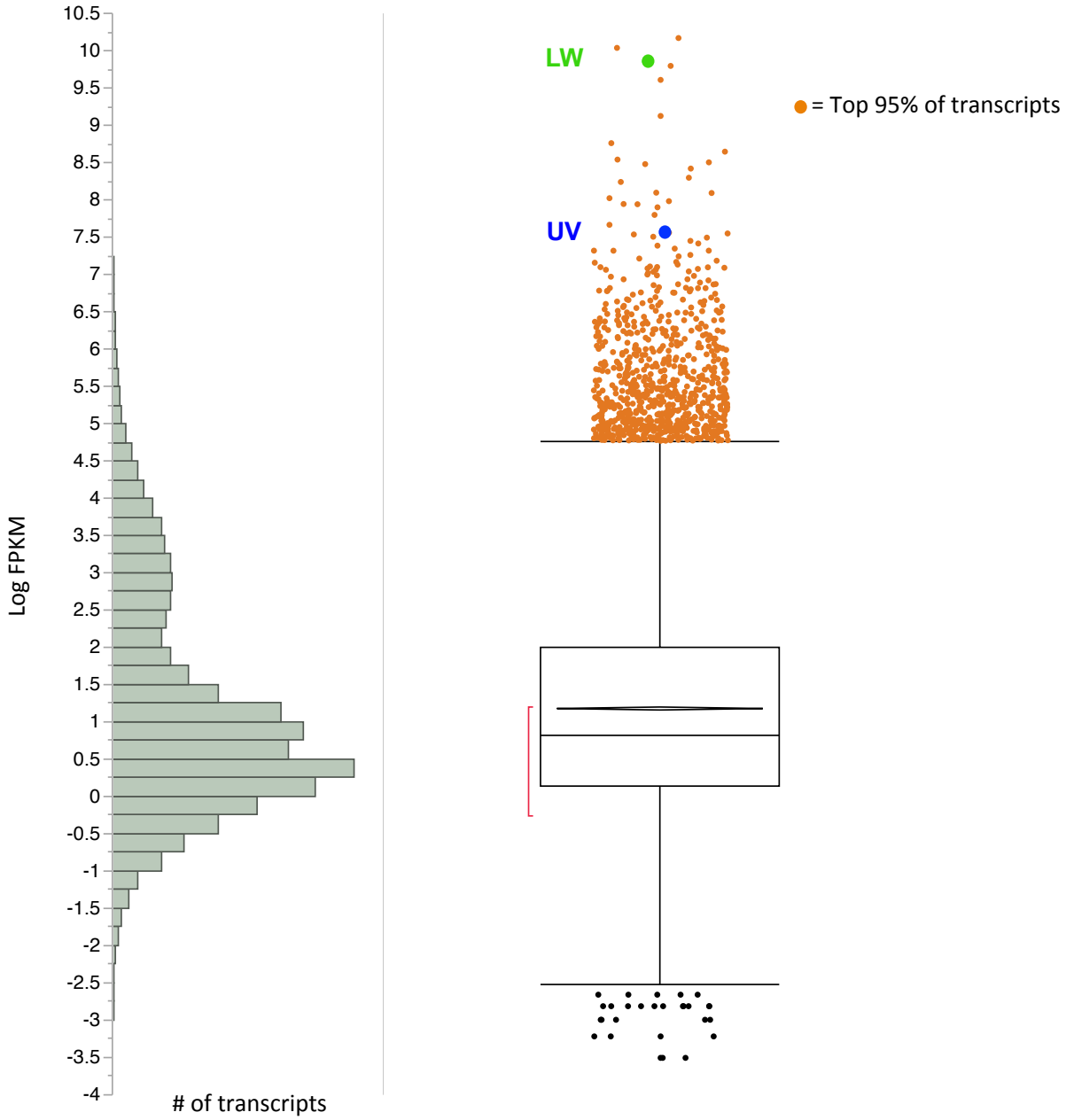


Figure S4. Insect opsin phylogeny supports homology-inferred firefly opsin function

Firefly LW and UV opsins lie in clades with conserved LW and UV function, respectively. Briefly, opsin translated amino acid sequences were downloaded from Genbank: *Pieris rapae* UV (AB208673.1), B (AB208675.1), V (AB208674.1), LW1 (AB177984.1), LW2 (AB188567.1); *Heliconius melpomene* UV1 (GU324678.1), UV2 (AY918896.1), B (AY918897.1), LW (EU480690.1); *Manduca sexta* Manop1 (L78080.1), Manop2 (L78081.1), Manop3 (AD001674.1); *Drosophila melanogaster* UV3 (AAA28854.1), UV4 (NP_476701.1), UV5B (AAC47426.1), UV7 (NP_524035), LMS6 (CAB06821.1), LMS1 (NP_524407.1), LMS2 (AAA28734.1); *Thermonectus marmoratus* UV1 (EU921226.1), UV2 (EU921227.1), LW (EU921225.1); *Tribolium castaneum* Uv (ABW06837.1), LW (ABA00706.1); *Anopheles gambiae* UV5 (XP_001688790), UV7 (XP_308329), LMS (CAA76825.1); *Apis mellifera* UV (NP_001011605.1), B (NP_001011606.1), LW1 (NP_001011639.2), LW2 (NP_001071293.1); *Megoura viciae* UV (AAG17120.1), LWS (AAG17119.1). These sequences were aligned using Muscle with default parameters in Geneious, and used to construct a neighbor-joining tree. The *T. pacificus* opsin used in homology modeling (X70498.1) and *Bos taurus* LW (NP_776991.1) and SW (NP_776992) opsins served as outgroups. The final alignment was 317 amino acids in length. Numbers at nodes show bootstrap support. UV: ultraviolet-sensitive; B: blue-sensitive; LW: long wavelength-sensitive; OG: outgroup.

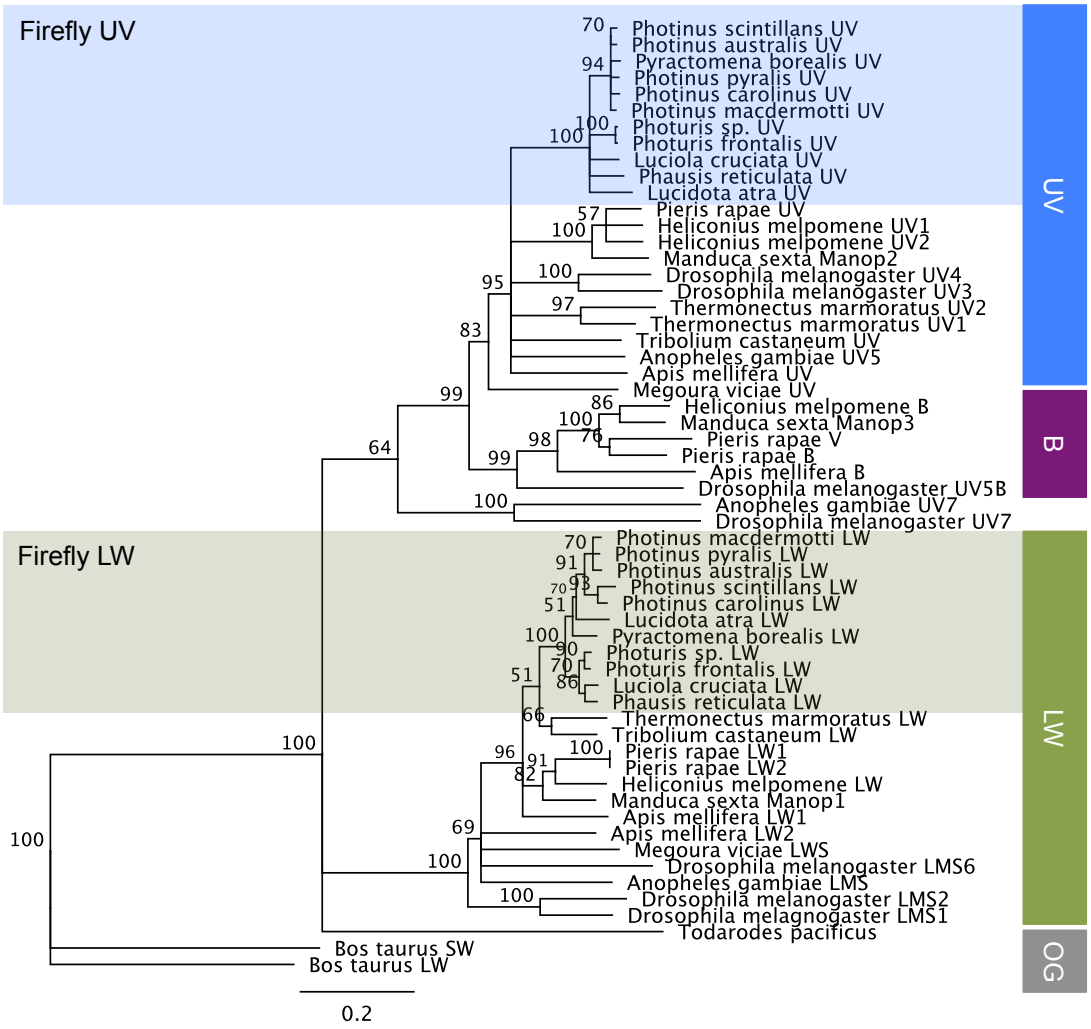
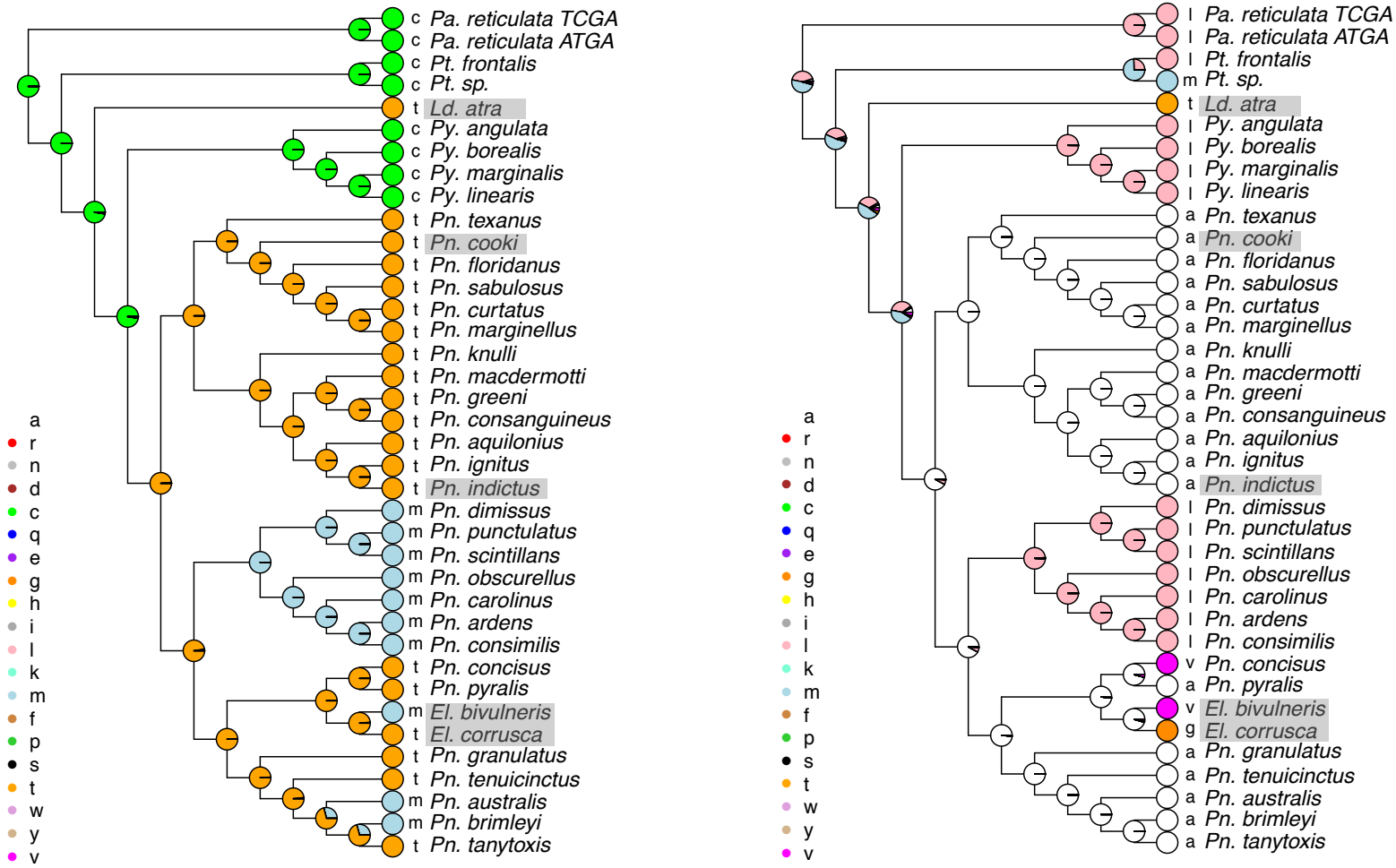
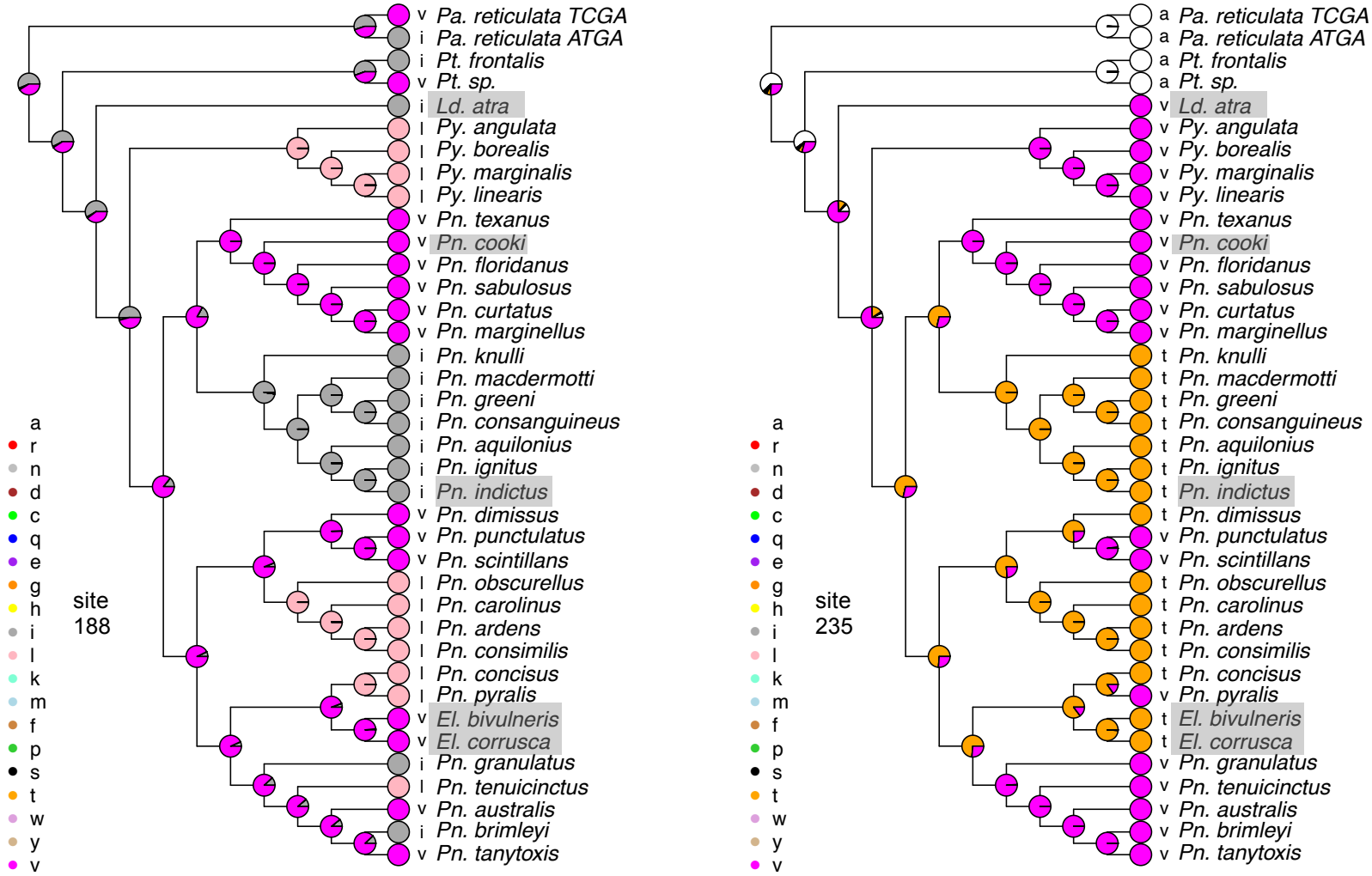
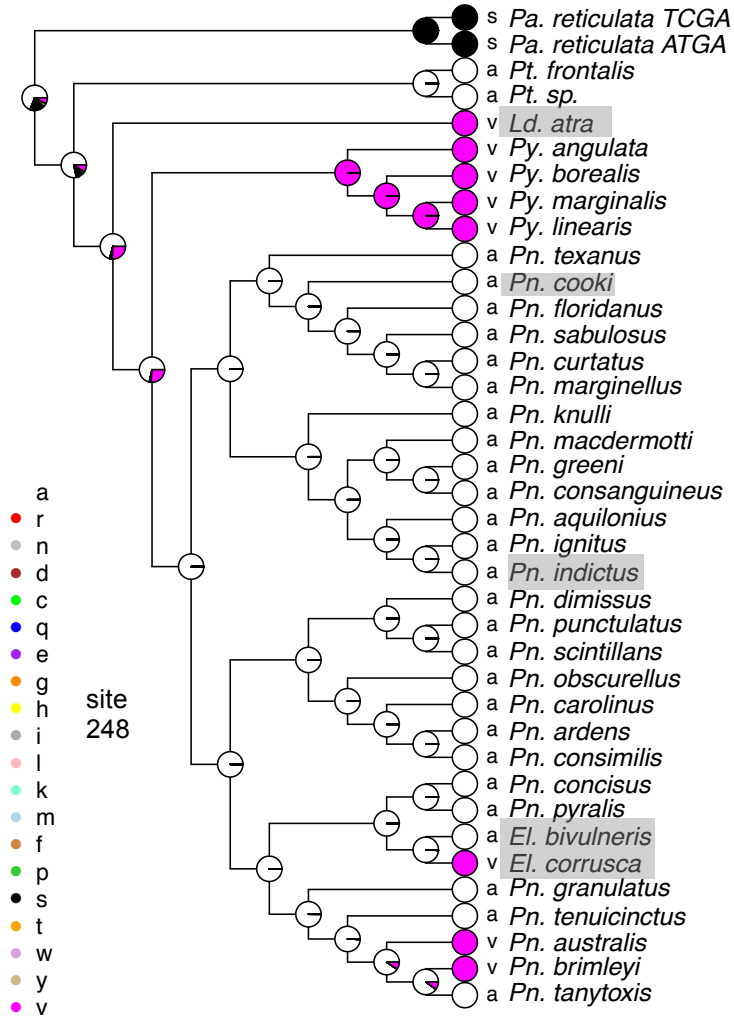


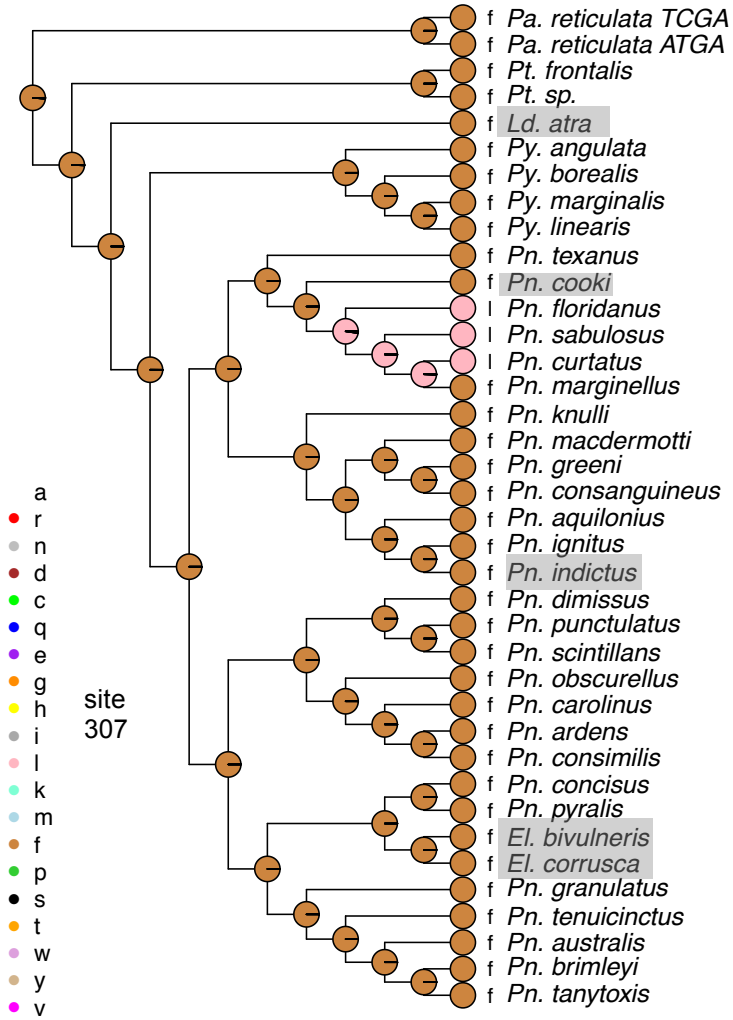
Figure S5. Ancestral reconstruction of eight positively selected amino acid sites in LW opsin.

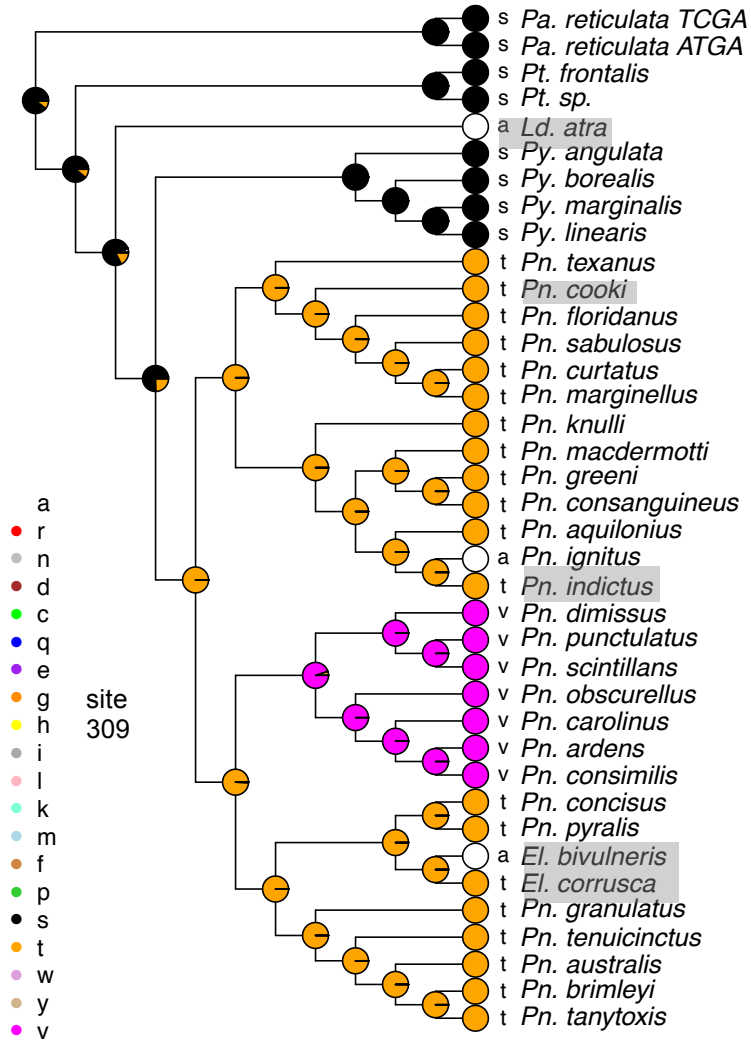
Reconstructions of the eight positively selected amino acid positions, using *Pn. pyralis* as reference, identified by PAML and fitmodel analyses. Prottest3 selected a LG model of amino acid substitution for use in reconstructions. None of these sites are in the chromophore binding pocket according to the homology model. *Pa*: *Phausis*, *Pt*: *Photuris*, *Ld*: *Lucidota*, *Py*: *Pyractomena*, *Pn*: *Photinus*, *El*: *Ellychnia*. Legend is the color of each amino acid, using single letter code. Gray boxes show diurnal taxa.




v Pa. reticulata TCGA
i Pa. reticulata ATGA
i Pt. frontalis
v Pt. sp.
i Ld. atra
l Py. angulata
l Py. borealis
l Py. marginalis
l Py. linearis
v Pn. texanus
v Pn. cooki
v Pn. floridanus
v Pn. sabulosus
v Pn. curtatus
v Pn. marginellus
i Pn. knulli
i Pn. macdermotti
i Pn. greeni
i Pn. consanguineus
i Pn. aquilonius
i Pn. ignitus
i Pn. indictus
v Pn. dimissus
v Pn. punctulatus
v Pn. scintillans
l Pn. obscurellus
l Pn. carolinus
l Pn. ardens
l Pn. consimilis
l Pn. concisus
l Pn. pyralis
v El. bivulneris
v El. corrusca
i Pn. granulatus
l Pn. tenuicinctus
v Pn. australis
i Pn. brimleyi
v Pn. tanytoxis
a
r
n
d
c
q
e
g
h
i
l
k
m
f
p
s
t
w
y
v
site
188
a Pa. reticulata TCGA
a Pa. reticulata ATGA
a Pt. frontalis
a Pt. sp.
v Ld. atra
v Py. angulata
v Py. borealis
v Py. marginalis
v Py. linearis
v Pn. texanus
v Pn. cooki
v Pn. floridanus
v Pn. sabulosus
v Pn. curtatus
v Pn. marginellus
t Pn. knulli
t Pn. macdermotti
t Pn. greeni
t Pn. consanguineus
t Pn. aquilonius
t Pn. ignitus
t Pn. indictus
t Pn. dimissus
v Pn. punctulatus
v Pn. scintillans
t Pn. obscurellus
t Pn. carolinus
t Pn. ardens
t Pn. consimilis
t Pn. concisus
v Pn. pyralis
t El. bivulneris
t El. corrusca
v Pn. granulatus
v Pn. tenuicinctus
v Pn. australis
v Pn. brimleyi
v Pn. tanytoxis
a
r
n
d
c
q
e
g
h
i
l
k
m
f
p
s
t
w
y
v
site
235


s Pa. reticulata TCGA
s Pa. reticulata ATGA
a Pt. frontalis
a Pt. sp.
v Ld. atra
v Py. angulata
v Py. borealis
v Py. marginalis
v Py. linearis
a Pn. texanus
a Pn. cooki
a Pn. floridanus
a Pn. sabulosus
a Pn. curtatus
a Pn. marginellus
a Pn. knulli
a Pn. macdermotti
a Pn. greeni
a Pn. consanguineus
a Pn. aquilonius
a Pn. ignitus
a Pn. indictus
a Pn. dimissus
a Pn. punctulatus
a Pn. scintillans
a Pn. obscurellus
a Pn. carolinus
a Pn. ardens
a Pn. consimilis
a Pn. concisus
a Pn. pyralis
a El. bivulneris
v El. corrusca
a Pn. granulatus
a Pn. tenuicinctus
v Pn. australis
v Pn. brimleyi
a Pn. tanytoxis
a
r
n
d
c
q
e
g
h
i
l
k
m
f
p
s
t
w
y
v
site
248


f Pa. reticulata TCGA
f Pa. reticulata ATGA
f Pt. frontalis
f Pt. sp.
f Ld. atra
f Py. angulata
f Py. borealis
f Py. marginalis
f Py. linearis
f Pn. texanus
f Pn. cooki
l Pn. floridanus
l Pn. sabulosus
l Pn. curtatus
f Pn. marginellus
f Pn. knulli
f Pn. macdermotti
f Pn. greeni
f Pn. consanguineus
f Pn. aquilonius
f Pn. ignitus
f Pn. indictus
f Pn. dimissus
f Pn. punctulatus
f Pn. scintillans
f Pn. obscurellus
f Pn. carolinus
f Pn. ardens
f Pn. consimilis
f Pn. concisus
f Pn. pyralis
f El. bivulneris
f El. corrusca
f Pn. granulatus
f Pn. tenuicinctus
f Pn. australis
f Pn. brimleyi
f Pn. tanytoxis
a
r
n
d
c
q
e
g
h
i
l
k
m
f
p
s
t
w
y
v
site
307


s Pa. reticulata TCGA
s Pa. reticulata ATGA
s Pt. frontalis
s Pt. sp.
a Ld. atra
s Py. angulata
s Py. borealis
s Py. marginalis
s Py. linearis
t Pn. texanus
t Pn. cooki
t Pn. floridanus
t Pn. sabulosus
t Pn. curtatus
t Pn. marginellus
t Pn. knulli
t Pn. macdermotti
t Pn. greeni
t Pn. consanguineus
t Pn. aquilonius
a Pn. ignitus
t Pn. indictus
v Pn. dimissus
v Pn. punctulatus
v Pn. scintillans
v Pn. obscurellus
v Pn. carolinus
v Pn. ardens
v Pn. consimilis
t Pn. concisus
t Pn. pyralis
a El. bivulneris
t El. corrusca
t Pn. granulatus
t Pn. tenuicinctus
t Pn. australis
t Pn. brimleyi
t Pn. tanytoxis
a
r
n
d
c
q
e
g
h
i
l
k
m
f
p
s
t
w
y
v
site
309

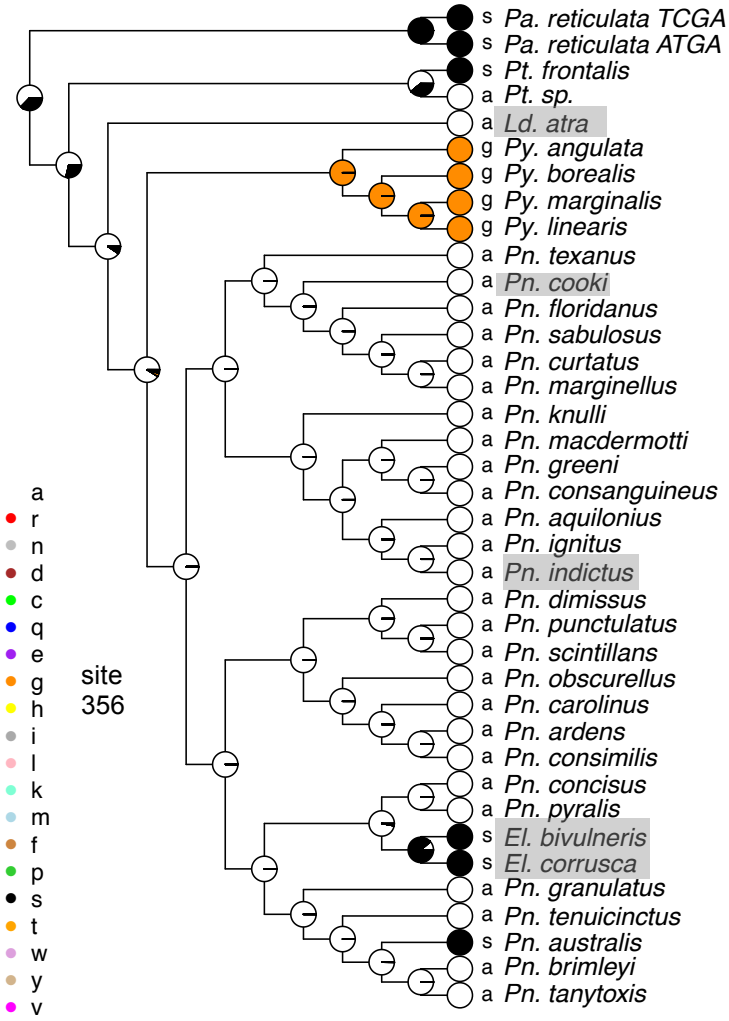

s Pa. reticulata TCGA
s Pa. reticulata ATGA
s Pt. frontalis
a Pt. sp.
a Ld. atra
g Py. angulata
g Py. borealis
g Py. marginalis
g Py. linearis
a Pn. texanus
a Pn. cooki
a Pn. floridanus
a Pn. sabulosus
a Pn. curtatus
a Pn. marginellus
a Pn. knulli
a Pn. macdermotti
a Pn. greeni
a Pn. consanguineus
a Pn. aquilonius
a Pn. ignitus
a Pn. indictus
a Pn. dimissus
a Pn. punctulatus
a Pn. scintillans
a Pn. obscurellus
a Pn. carolinus
a Pn. ardens
a Pn. consimilis
a Pn. concisus
a Pn. pyralis
s El. bivulneris
s El. corrusca
a Pn. granulatus
a Pn. tenuicinctus
s Pn. australis
a Pn. brimleyi
a Pn. tanytoxis
a
r
n
d
c
q
e
g
h
i
l
k
m
f
p
s
t
w
y
v
site
356

Figure S6. Ancestral reconstruction of the two variable binding pocket sites in LW opsin.

Reconstructions of the two variable binding pocket sites amino acid positions, using *Pn. pyralis* as reference, identified by the homology model. An LG model of amino acid evolution was used as in Figure S5. Both sites show evidence of a single amino acid change. *Pa: Phausis, Pt: Photuris, Ld: Lucidota, Py: Pyractomena, Pn: Photinus, El: Ellychnia*. Legend is the color of each amino acid, using single letter code. Gray boxes show diurnal taxa.

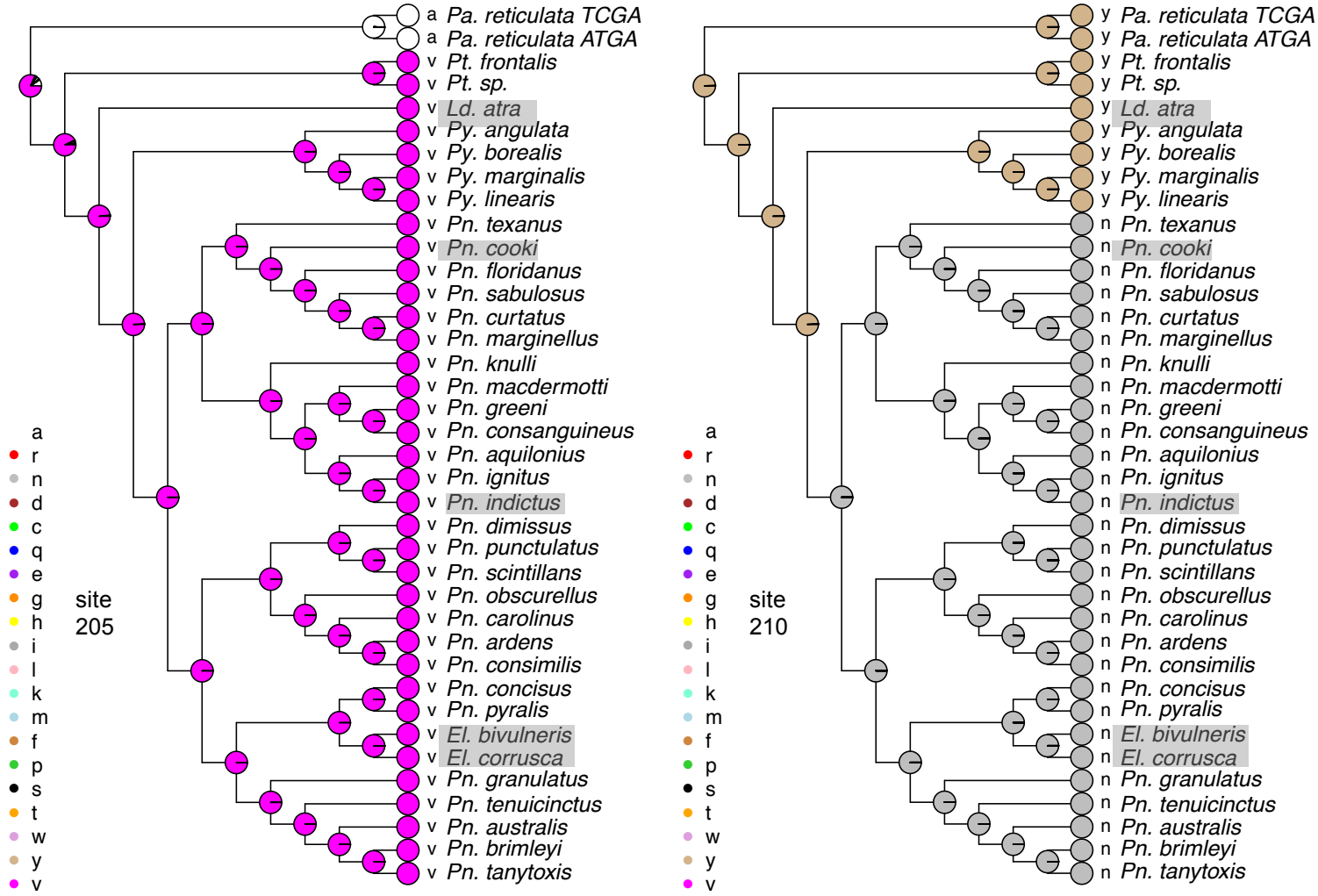


Figure S7. Ancestral reconstruction of two sites in LW opsin with parallel/convergent changes in diurnal taxa.

The two sites below show changes only in diurnal lineages. Sites are labeled according to the *Pn. pyralis* full-length amino acid sequence. An LG model of amino acid evolution was used as in Figure S5. *Pa: Phausis, Pt: Photuris, Ld: Lucidota, Py: Pyractomena, Pn: Photinus, El: Ellychnia*. Legend is the color of each amino acid, using single letter code. Gray boxes show diurnal taxa.

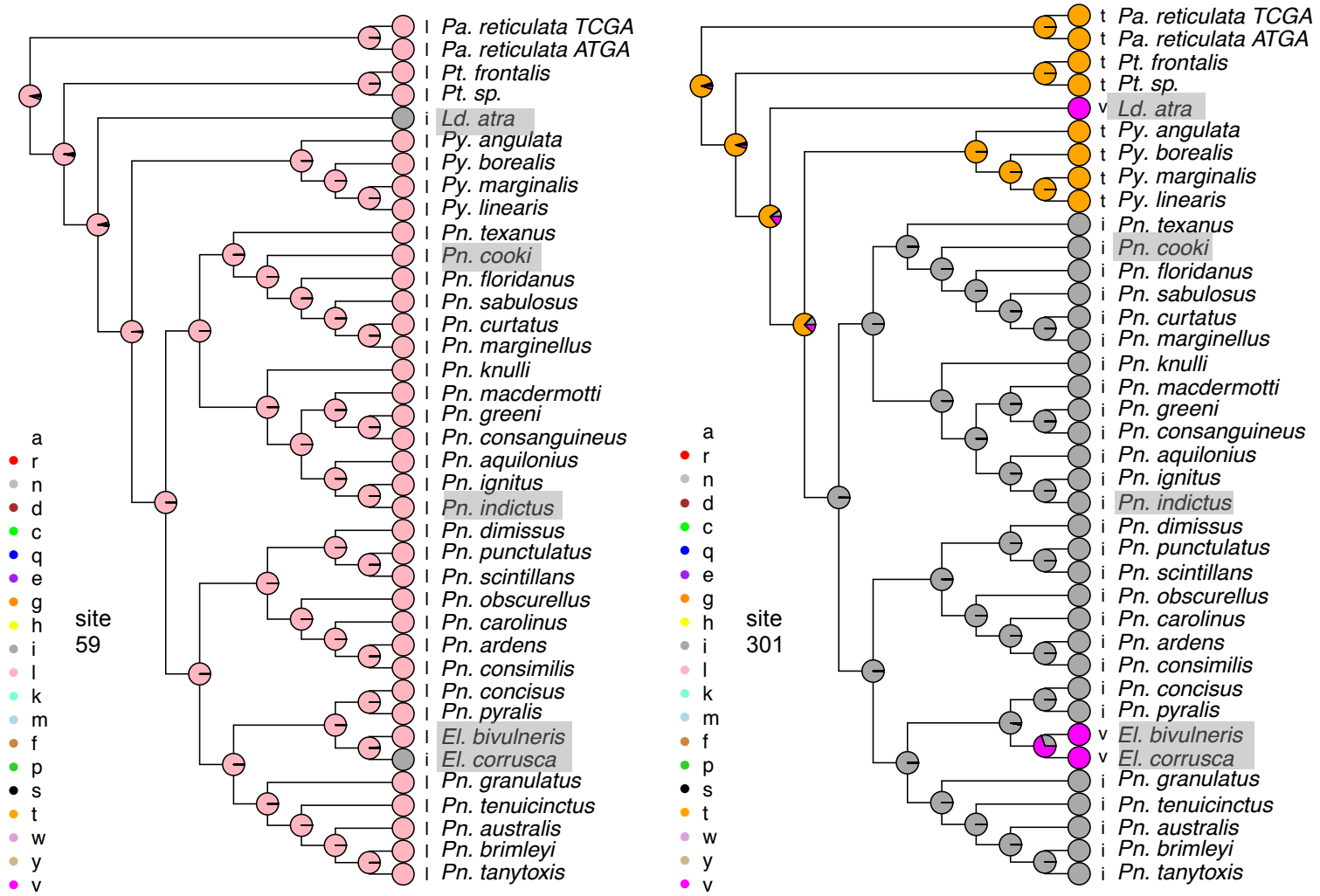
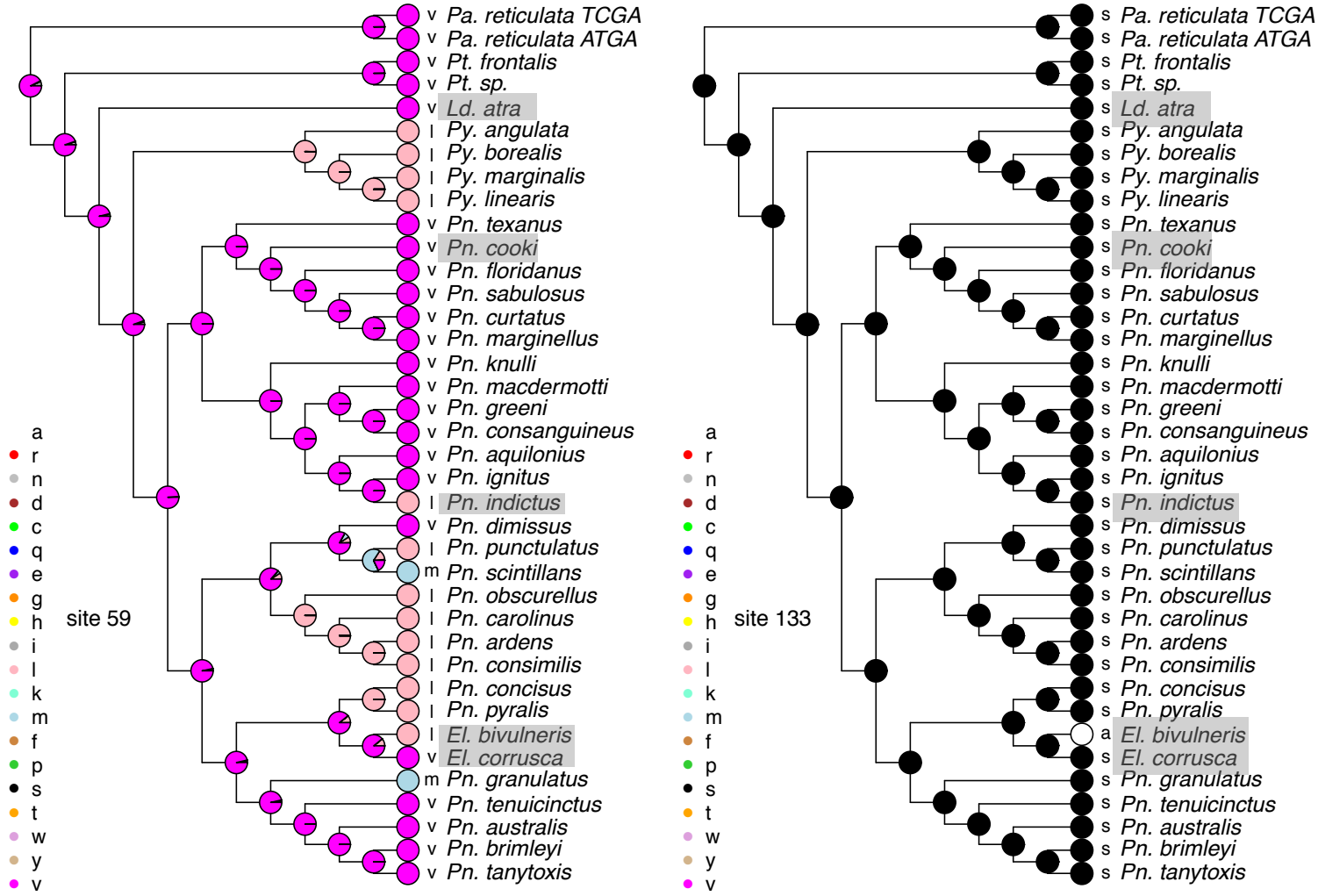
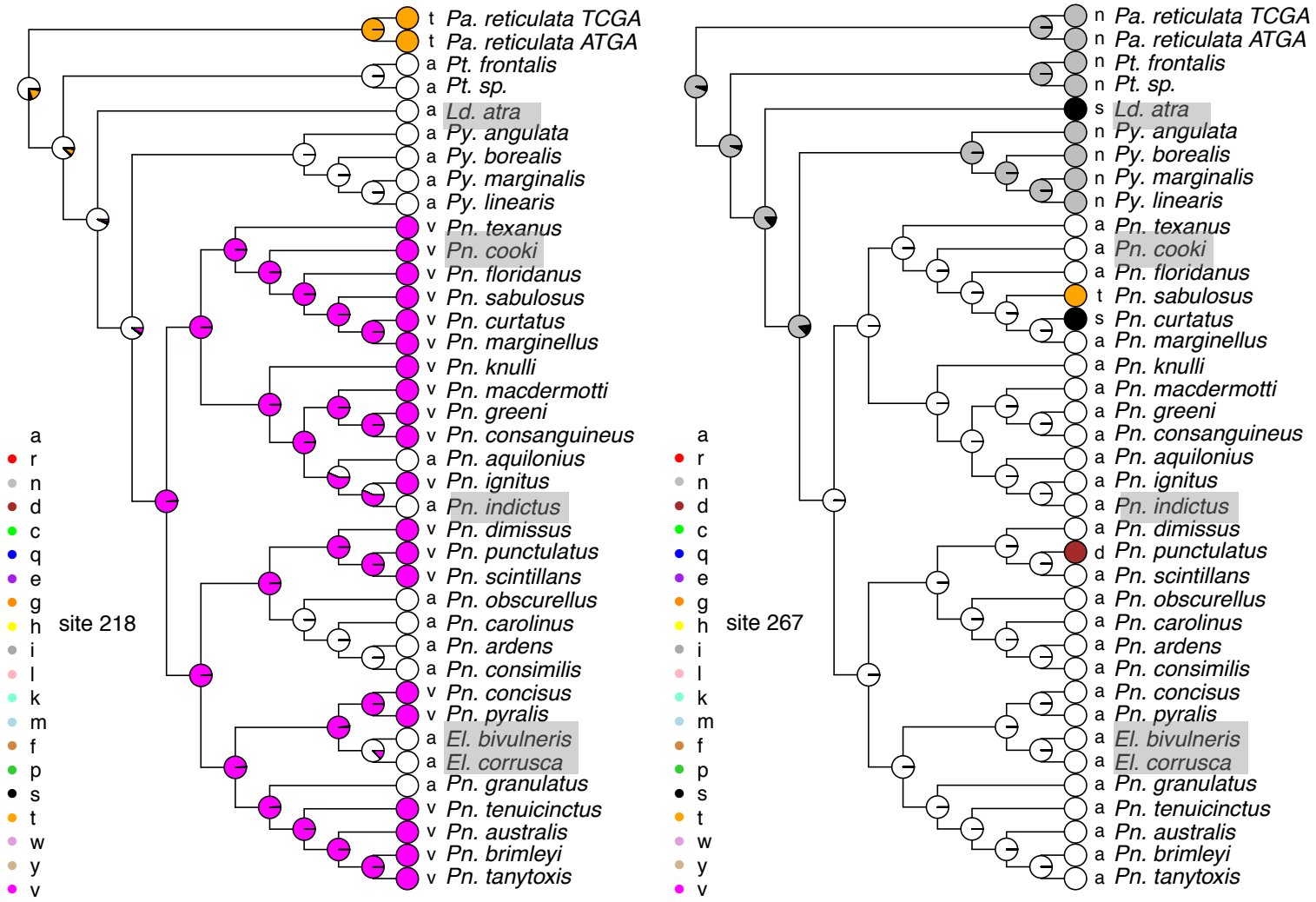
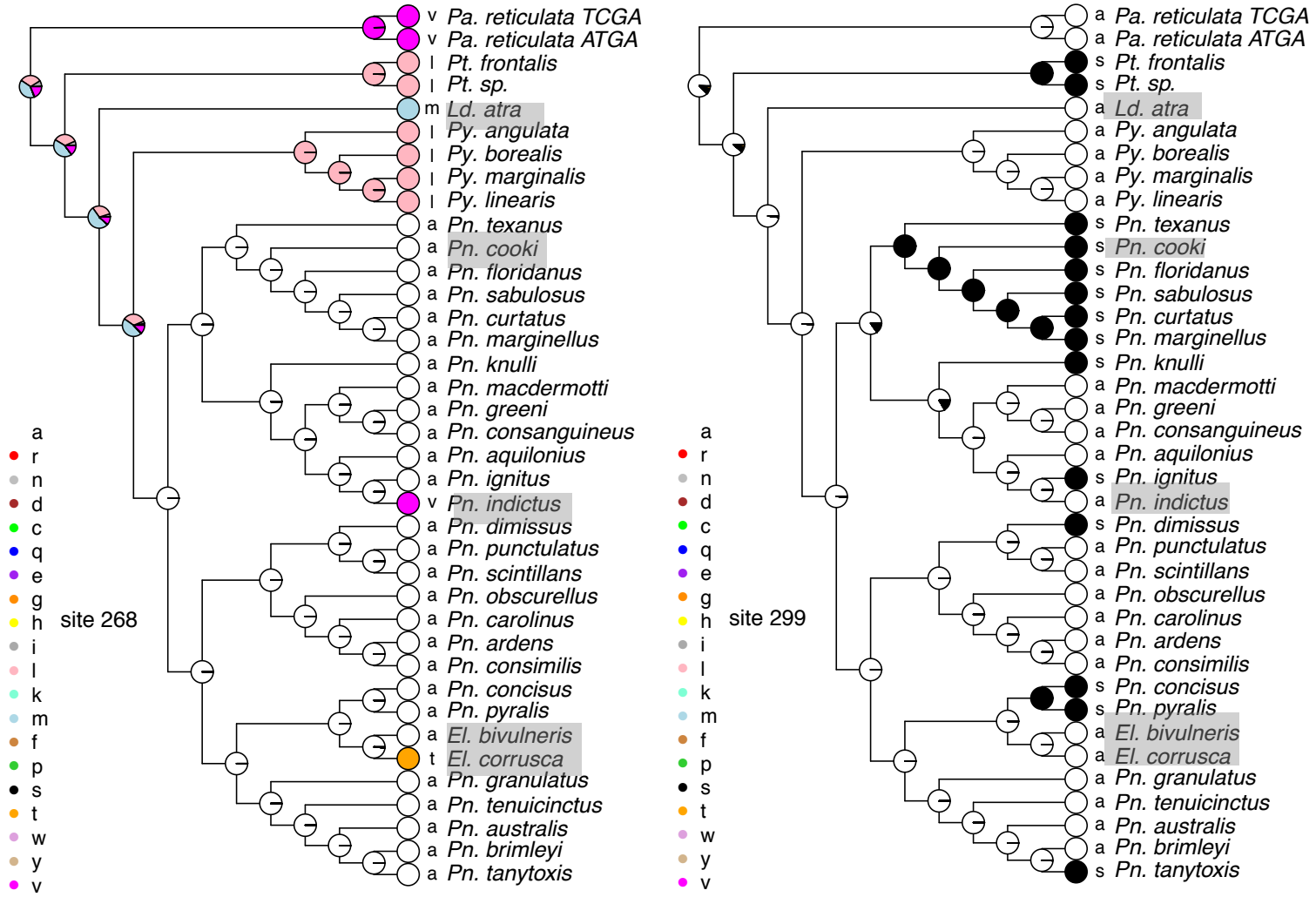


Figure S8. Ancestral reconstruction of eight positively selected amino acid sites in UV opsin.

Reconstructions of the eight positively selected amino acid positions, using *Pn. pyralis* as reference, identified by PAML and fitmodel analyses. Prottest3 selected a JTT model of amino acid substitution for use in reconstructions. Sites 133 and 299 are in the binding pocket. *Pa: Phausis, Pt: Photuris, Ld: Lucidota, Py: Pyractomena, Pn: Photinus, El: Ellychnia*. Legend is the color of each amino acid, using single letter code. Gray boxes show diurnal taxa.




t Pa. reticulata TCGA
t Pa. reticulata ATGA
a Pt. frontalis
a Pt. sp.
a Ld. atra
a Py. angulata
a Py. borealis
a Py. marginalis
a Py. linearis
v Pn. texanus
v Pn. cooki
v Pn. floridanus
v Pn. sabulosus
v Pn. curtatus
v Pn. marginellus
v Pn. knulli
v Pn. macdermotti
v Pn. greeni
v Pn. consanguineus
a Pn. aquilonius
v Pn. ignitus
a Pn. indictus
v Pn. dimissus
v Pn. punctulatus
v Pn. scintillans
a Pn. obscurellus
a Pn. carolinus
a Pn. ardens
a Pn. consimilis
v Pn. concisus
v Pn. pyralis
a El. bivulneris
a El. corrusca
a Pn. granulatus
v Pn. tenuicinctus
v Pn. australis
v Pn. brimleyi
v Pn. tanytoxis
site 218
a r n d c q e g h i l k m f p s t w y v
n Pa. reticulata TCGA
n Pa. reticulata ATGA
n Pt. frontalis
n Pt. sp.
s Ld. atra
n Py. angulata
n Py. borealis
n Py. marginalis
n Py. linearis
a Pn. texanus
a Pn. cooki
a Pn. floridanus
t Pn. sabulosus
s Pn. curtatus
a Pn. marginellus
a Pn. knulli
a Pn. macdermotti
a Pn. greeni
a Pn. consanguineus
a Pn. aquilonius
a Pn. ignitus
a Pn. indictus
a Pn. dimissus
d Pn. punctulatus
a Pn. scintillans
a Pn. obscurellus
a Pn. carolinus
a Pn. ardens
a Pn. consimilis
a Pn. concisus
a Pn. pyralis
a El. bivulneris
a El. corrusca
a Pn. granulatus
a Pn. tenuicinctus
a Pn. australis
a Pn. brimleyi
a Pn. tanytoxis
site 267
a r n d c q e g h i l k m f p s t w y v


v Pa. reticulata TCGA
v Pa. reticulata ATGA
l Pt. frontalis
l Pt. sp.
m Ld. atra
l Py. angulata
l Py. borealis
l Py. marginalis
l Py. linearis
a Pn. texanus
a Pn. cooki
a Pn. floridanus
a Pn. sabulosus
a Pn. curtatus
a Pn. marginellus
a Pn. knulli
a Pn. macdermotti
a Pn. greeni
a Pn. consanguineus
a Pn. aquilonius
a Pn. ignitus
v Pn. indictus
a Pn. dimissus
a Pn. punctulatus
a Pn. scintillans
a Pn. obscurellus
a Pn. carolinus
a Pn. ardens
a Pn. consimilis
a Pn. concisus
a Pn. pyralis
a El. bivulneris
t El. corrusca
a Pn. granulatus
a Pn. tenuicinctus
a Pn. australis
a Pn. brimleyi
a Pn. tanytoxis
a r n d c q e g h i l k m f p s t w y v
site 268
a Pa. reticulata TCGA
a Pa. reticulata ATGA
s Pt. frontalis
s Pt. sp.
a Ld. atra
a Py. angulata
a Py. borealis
a Py. marginalis
a Py. linearis
s Pn. texanus
s Pn. cooki
s Pn. floridanus
s Pn. sabulosus
s Pn. curtatus
s Pn. marginellus
s Pn. knulli
a Pn. macdermotti
a Pn. greeni
a Pn. consanguineus
a Pn. aquilonius
s Pn. ignitus
a Pn. indictus
s Pn. dimissus
a Pn. punctulatus
a Pn. scintillans
a Pn. obscurellus
a Pn. carolinus
a Pn. ardens
a Pn. consimilis
s Pn. concisus
s Pn. pyralis
a El. bivulneris
a El. corrusca
a Pn. granulatus
a Pn. tenuicinctus
a Pn. australis
a Pn. brimleyi
s Pn. tanytoxis
a r n d c q e g h i l k m f p s t w y v
site 299

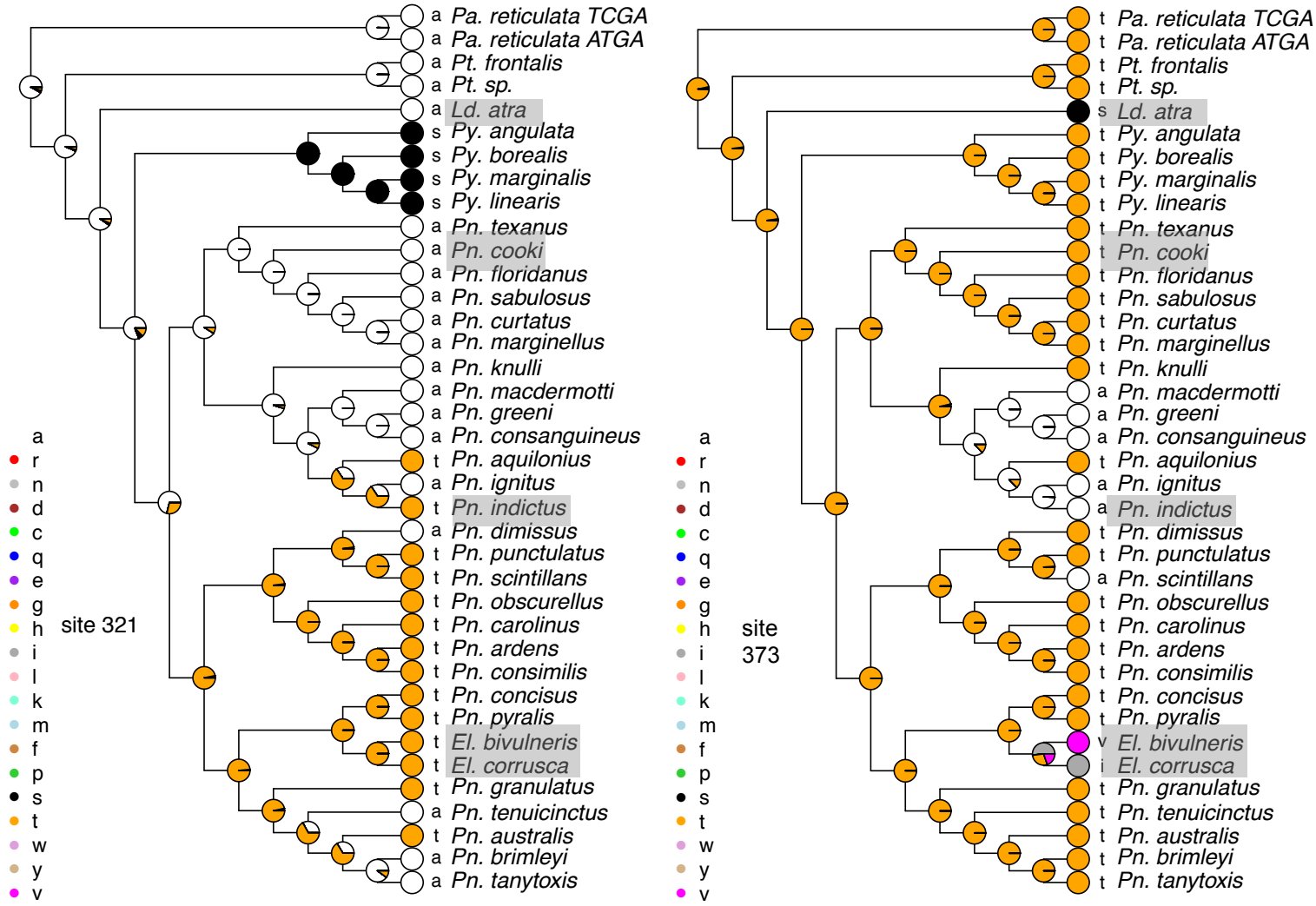
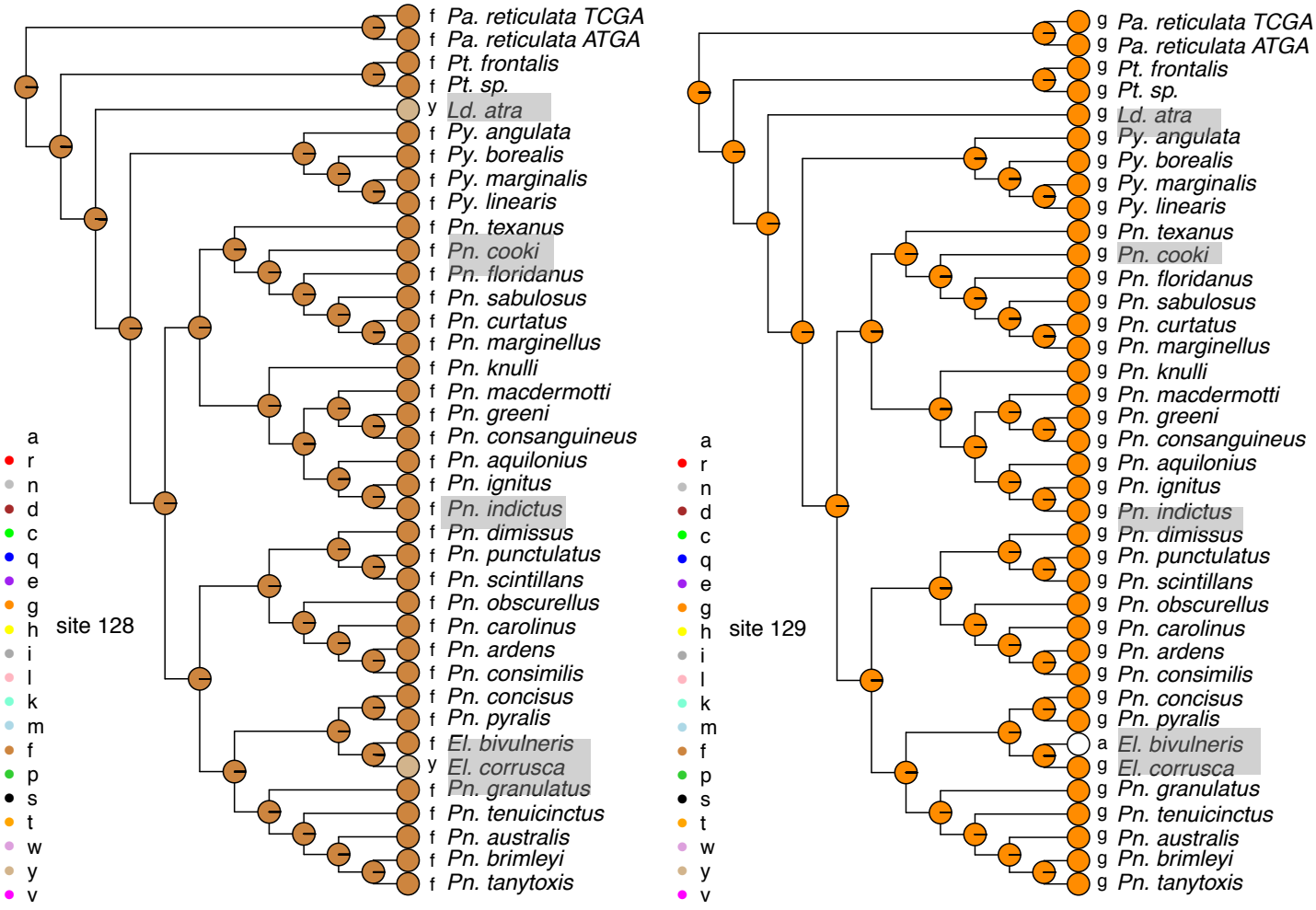

a Pa. reticulata TCGA
a Pa. reticulata ATGA
a Pt. frontalis
a Pt. sp.
a Ld. atra
s Py. angulata
s Py. borealis
s Py. marginalis
s Py. linearis
a Pn. texanus
a Pn. cooki
a Pn. floridanus
a Pn. sabulosus
a Pn. curtatus
a Pn. marginellus
a Pn. knulli
a Pn. macdermotti
a Pn. greeni
a Pn. consanguineus
t Pn. aquilonius
a Pn. ignitus
t Pn. indictus
a Pn. dimissus
t Pn. punctulatus
t Pn. scintillans
t Pn. obscurellus
t Pn. carolinus
t Pn. ardens
t Pn. consimilis
t Pn. concisus
t Pn. pyralis
t El. bivulneris
t El. corrusca
t Pn. granulatus
a Pn. tenuicinctus
t Pn. australis
a Pn. brimleyi
a Pn. tanytoxis
a r n d c q e g h i l k m f p s t w y v
site 321
t Pa. reticulata TCGA
t Pa. reticulata ATGA
t Pt. frontalis
t Pt. sp.
s Ld. atra
t Py. angulata
t Py. borealis
t Py. marginalis
t Py. linearis
t Pn. texanus
t Pn. cooki
t Pn. floridanus
t Pn. sabulosus
t Pn. curtatus
t Pn. marginellus
t Pn. knulli
a Pn. macdermotti
a Pn. greeni
a Pn. consanguineus
t Pn. aquilonius
a Pn. ignitus
a Pn. indictus
t Pn. dimissus
t Pn. punctulatus
a Pn. scintillans
t Pn. obscurellus
t Pn. carolinus
t Pn. ardens
t Pn. consimilis
t Pn. concisus
t Pn. pyralis
v El. bivulneris
i El. corrusca
t Pn. granulatus
t Pn. tenuicinctus
t Pn. australis
t Pn. brimleyi
t Pn. tanytoxis
a r n d c q e g h i l k m f p s t w y v
site 373

Figure S9. Reconstruction of three of the five variable binding pocket amino acid sites in UV-opsin

Reconstructions of three of the five variable binding pocket sites in UV-opsin, with position corresponding to *Pn. pyralis*. A JTT model of amino acid evolution was used as in Figure S8. Variable binding pocket sites 133 and 299 were also identified as under positive selection; their reconstructions are included in Figure S8. *Pa: Phausis, Pt: Photuris, Ld: Lucidota, Py: Pyractomena, Pn: Photinus, El: Ellychnia*. Legend is the color of each amino acid, using single letter code. Gray boxes show diurnal taxa.



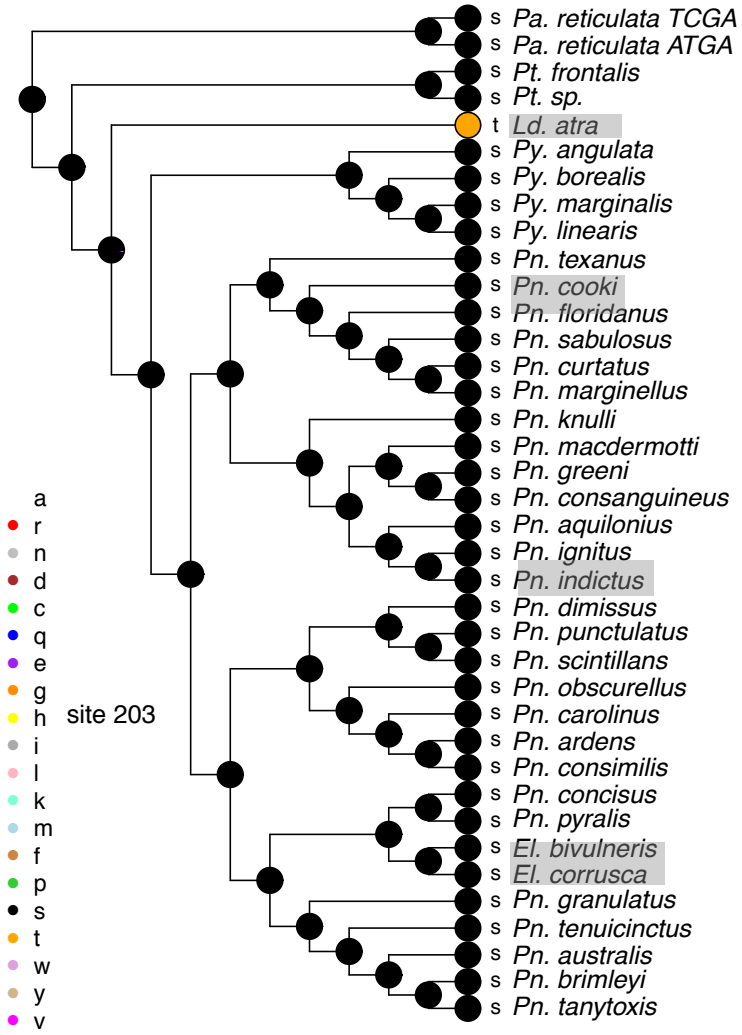

s Pa. reticulata TCGA
s Pa. reticulata ATGA
s Pt. frontalis
s Pt. sp.
t Ld. atra
s Py. angulata
s Py. borealis
s Py. marginalis
s Py. linearis
s Pn. texanus
s Pn. cooki
s Pn. floridanus
s Pn. sabulosus
s Pn. curtatus
s Pn. marginellus
s Pn. knulli
s Pn. macdermotti
s Pn. greeni
s Pn. consanguineus
s Pn. aquilonius
s Pn. ignitus
s Pn. indictus
s Pn. dimissus
s Pn. punctulatus
s Pn. scintillans
s Pn. obscurellus
s Pn. carolinus
s Pn. ardens
s Pn. consimilis
s Pn. concisus
s Pn. pyralis
s El. bivulneris
s El. corrusca
s Pn. granulatus
s Pn. tenuicinctus
s Pn. australis
s Pn. brimleyi
s Pn. tanytoxis
site 203
a
r
n
d
c
q
e
g
h
i
l
k
m
f
p
s
t
w
y
v

Figure S10. Ancestral reconstruction of two sites in UV opsin with parallel changes in diurnal taxa.

The two sites below show changes exclusive to diurnal lineages. Sites are labeled according to the *Pn. pyralis* full-length amino acid sequence. A JTT model of amino acid evolution was used as in Figure S8. *Pa: Phausis, Pt: Photuris, Ld: Lucidota, Py: Pyractomena, Pn: Photinus, El: Ellychnia*. Legend is the color of each amino acid, using single letter code. Gray boxes show diurnal taxa.

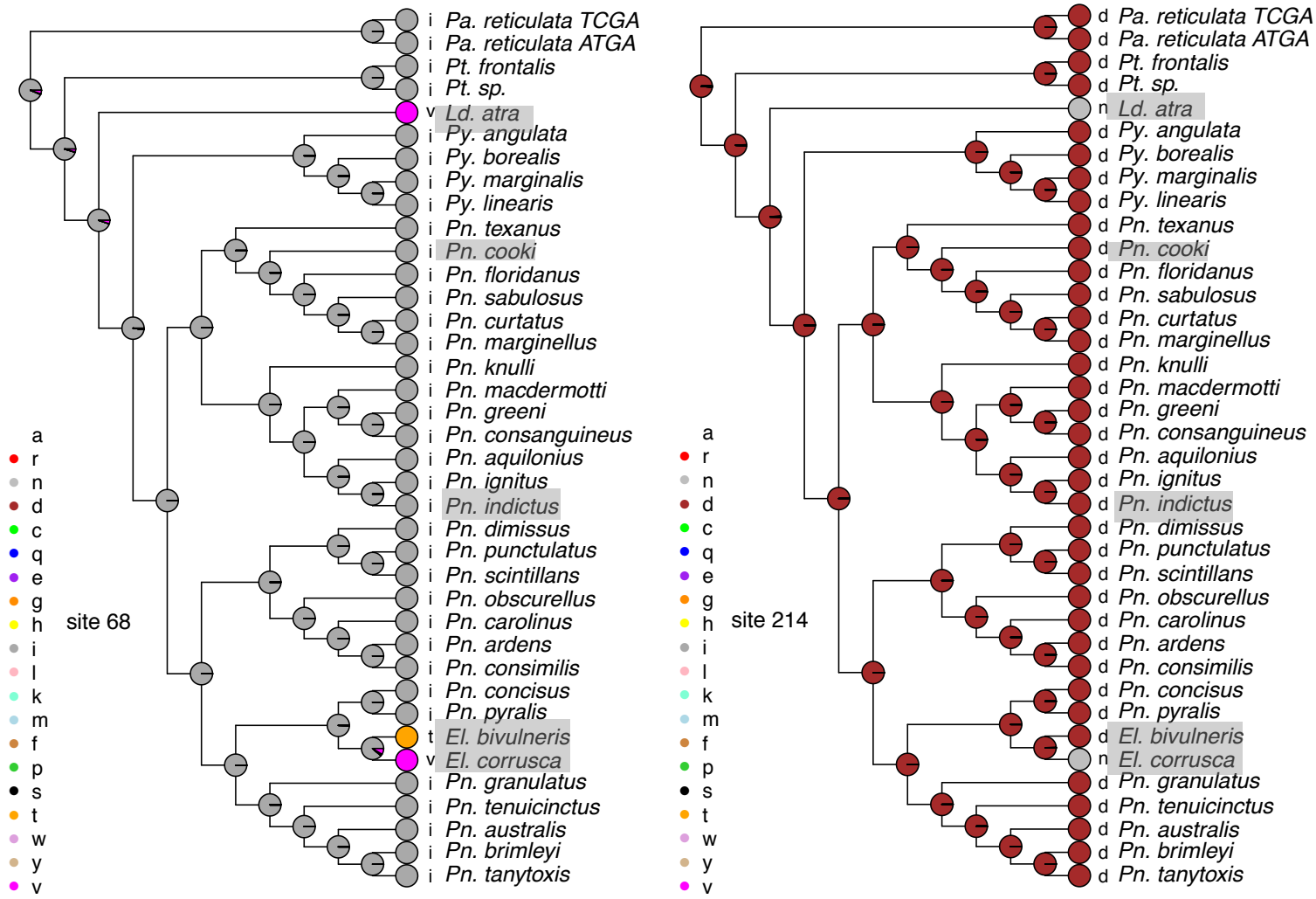
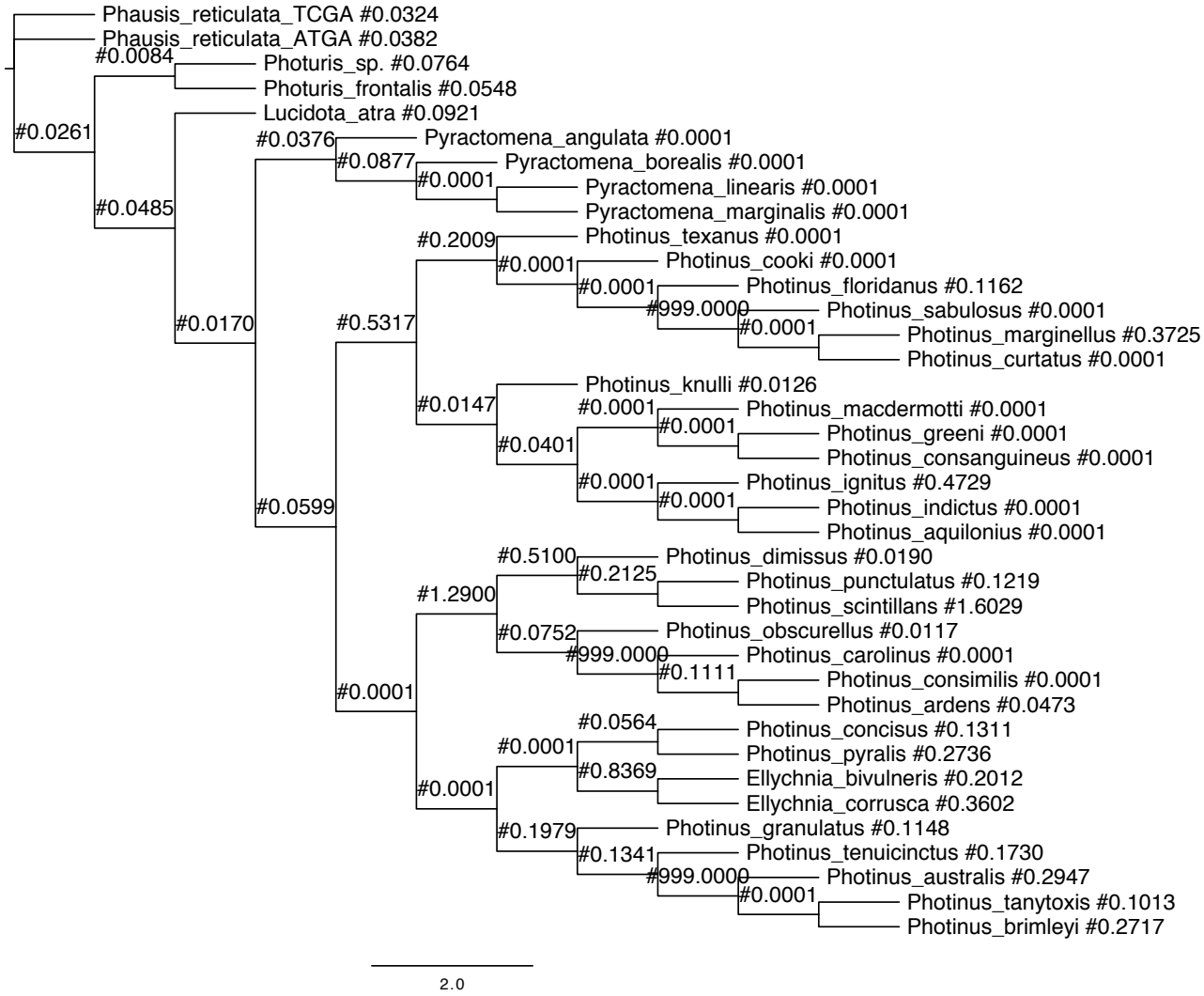
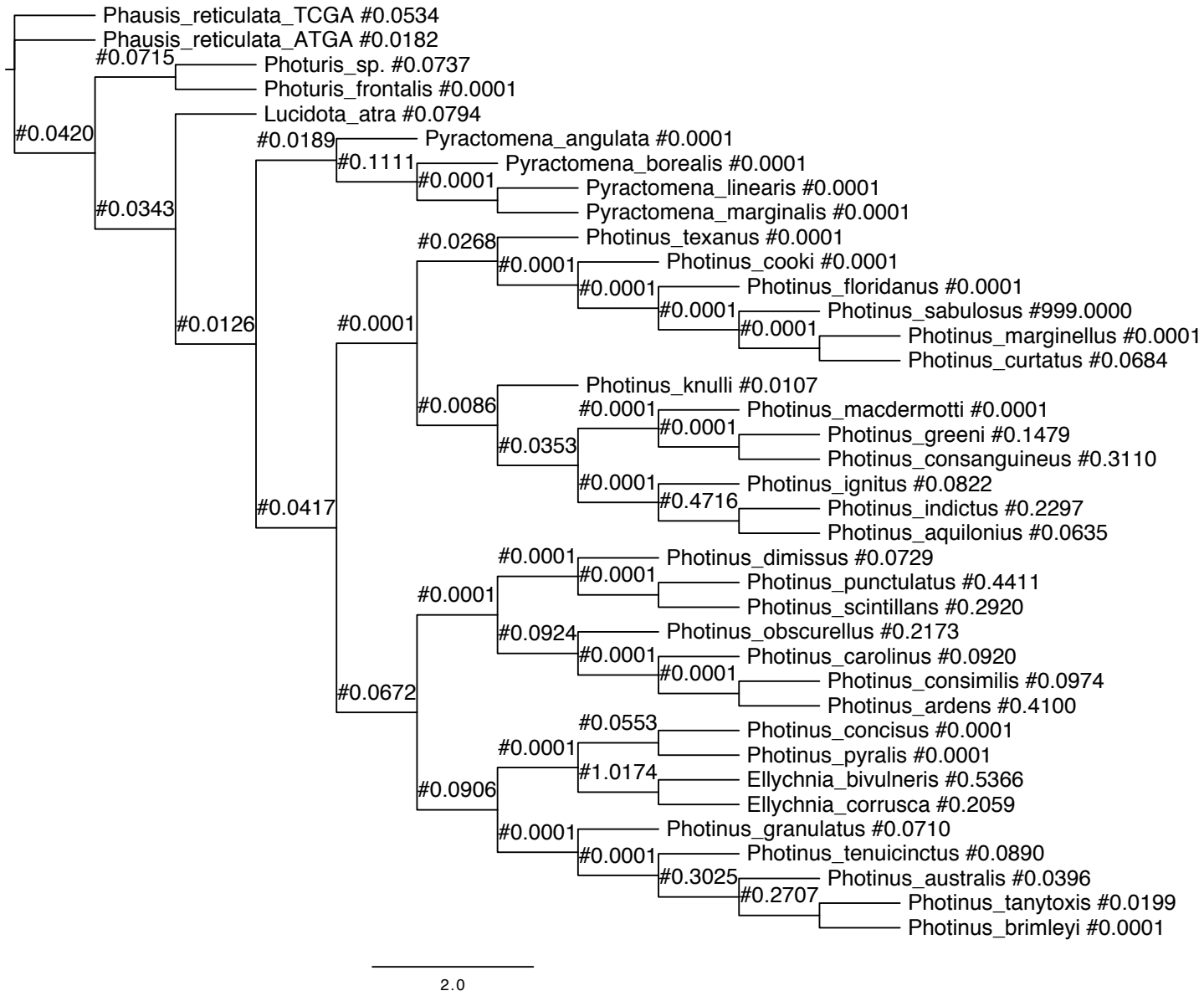


Figure S11. Free-ratio ω values on phylogeny for use in statistical tests of associations between selective constraint and ecological and signaling traits. (a) LW-opsin and (b) UV-opsin. ω values for each branch are also shown.

a) LW-opsin



b) UV-opsin


Phausis_reticulata_TCGA #0.0534
Phausis_reticulata_ATGA #0.0182
#0.0715
Photuris_sp. #0.0737
Photuris_frontalis #0.0001
Lucidota_atra #0.0794
#0.0420
#0.0189
Pyractomena_angulata #0.0001
#0.1111
Pyractomena_borealis #0.0001
#0.0001
Pyractomena_linearis #0.0001
Pyractomena_marginalis #0.0001
#0.0343
#0.0268
Photinus_texanus #0.0001
#0.0001
Photinus_cooki #0.0001
#0.0001
Photinus_floridanus #0.0001
#0.0001
Photinus_sabulosus #999.0000
#0.0001
Photinus_marginellus #0.0001
Photinus_curtatus #0.0684
#0.0126
#0.0001
Photinus_knulli #0.0107
#0.0086
#0.0001
Photinus_macdermotti #0.0001
#0.0001
Photinus_greeni #0.1479
#0.0353
Photinus_consanguineus #0.3110
#0.0001
Photinus_ignitus #0.0822
#0.0417
#0.4716
Photinus_indictus #0.2297
Photinus_aquilonius #0.0635
#0.0001
Photinus_dimissus #0.0729
#0.0001
Photinus_punctulatus #0.4411
#0.0001
Photinus_scintillans #0.2920
#0.0924
Photinus_obscurellus #0.2173
#0.0001
Photinus_carolinus #0.0920
#0.0001
Photinus_consimilis #0.0974
#0.0672
Photinus_ardens #0.4100
#0.0553
Photinus_concisus #0.0001
#0.0001
Photinus_pyralis #0.0001
#1.0174
Ellychnia_bivulneris #0.5366
#0.0906
Ellychnia_corrusca #0.2059
#0.0001
Photinus_granulatus #0.0710
#0.0001
Photinus_tenuicinctus #0.0890
#0.3025
Photinus_australis #0.0396
#0.2707
Photinus_tanytoxis #0.0199
Photinus_brimleyi #0.0001
2.0

Text S1**.** Measuring nocturnal species' light emissions in the field.

In addition to emission spectra, activity time, and habitat data gathered from the literature, we used data from specimens opportunistically collected across the Eastern United States from April through July of 2012 and 2013. At each field site (population) visited, we attempted to collect emission spectra on a minimum of five individuals of each species. In addition, we recorded temperature, elevation, longitude, latitude, time of initiation of activity, and classified the habitat as closed, open, or mixed. We focused on adult males, because they are easier to find in large numbers than females. In our final trait analysis, we used the average male peak emission wavelength, and the activity time and habitat classifications, for the specific population from which the specimen used for opsin sequencing was derived. These data are given in the table following the text.

Measuring spectra

Emission spectra for each individual were recorded soon after capture in the field using an Ocean Optics Jaz Modular Sensing Suite, and processed with SpectraSuite software v. 12.2 (Ocean Optics 2008). Individuals were held so that the unobstructed light organ was in line with the end of a fiber optic cable. During measurement, each flash was transmitted along the cable to the spectrophotometer, which output the emission spectrum to a laptop (Dell Inspiron Mini10v) as the intensity of signal versus wavelength. The SpectraSuite software trigger function was used to capture single flashes in real time once the signal at 550 nm, integrated over 100 ms time intervals, reached an intensity of 2500 counts. The intensity threshold was increased in environments with more ambient light to avoid capturing background spectra. The duration of integration was determined empirically as that which captured sufficient signal to produce a detectable emission while not saturating the sensor. We attempted to record the emission spectra of at least five flashes for each individual.

Because of measurement noise across the spectrum we could not simply use the highest value of the emission spectrum as the appropriate peak emission. Instead, a cubic polynomial was fit to the emission spectrum in the region of maximum intensity to estimate the peak emission wavelength, using a custom program in Mathematica v. 8 (Wolfram Research, 2010). The wavelength at the maximum of the cubic fit in the region of the emission peak was identified as the peak wavelength. When the program was written, several different polynomials were evaluated, and cubic was found to perform best when checked by visual inspection. Peak wavelengths from every spectrum were averaged to determine the value for each individual. We found no effect of the intensity of the spectrum, which was affected primarily by the distance between the light organ and the detector, on the peak wavelength of light emissions.

To assess the reliability of the field spectrophotometer measurements, we routinely checked its precision by measuring spectra for a 545 nm LED at each field site at the beginning and end of each night. In the lab, we examined the effect of running time (i.e. since being turned on) and optic cable aperture size. In no case did we find that readings were affected by the tested variables. Since we were planning to hold some of the smaller specimens cupped in our hand to record light emissions, we tested the effect of various methods of holding a specimen on the emission spectra. These measurements established that reflections from the skin (by cupping a specimen or holding the light organ too close to the skin) sometimes caused a small red shift, on the order of a few nm, in the readings. As a result, we took great care that all measurements in our data set were taken with an unobstructed light organ (by holding the head and thorax of

specimens), or by measuring them through a clear conical collection tube. We confirmed with a 545 nm LED that no shifting in wavelength occurred between measurements within and outside of the collection tube.

Habitat classification

Habitat was independently characterized by two observers based on the density of vegetation at each field site: fields and areas with minimal canopy cover were designated as open, wooded areas and areas with thick vegetation were designated closed, and areas where activity localized in both open and closed habitats, or on the edge of these two regions, were designated as mixed. Populations were assigned to habitat based on the location in which most individuals were active.

Activity time classification

Each species collected or observed at each field site was assigned to an activity time category based on its onset of activity relative to sunset. The time of sunset was determined from the latitude, longitude, and date of collection using the National Oceanic and Atmospheric Administration (NOAA) Solar Calculator (http://www.esrl.noaa.gov/gmd/grad/solcalc/). Species were divided into early or late categories following Lall and coworkers (1980): early species were defined as those that became active up to thirty minutes after sunset, while late species commenced flashing activity after this time. Populations were assigned to activity categories based on when the first individuals of a species were caught at a location.

Additional details about the methods and determination of species averages can be found in Hall et al. (*in review*).

			Spectra[c]			
Species[a]	Location[b]	N	Mean	SD	Habitat[d]	Activity[e]
Pa. reticulata	Athens, GA	5	554.11	0.81	2	1
Pa. reticulata	Tate City, GA	8	554.54	0.70	2	1
Pn. australis	Athens, GA	9	571.72	3.96	2	0
Pn. brimleyi	Jackson Co., TN	2	562.80	1.15	1	1
Pn. carolinus	ANF, PA	16	566.67	1.26	1	1
Pn. consimilis	FITG, PA	1	561.37	0.16	1	1
Pn. curtatus	Charleston, IL	6	562.66	1.29	2	0
Pn. dimissus	Denison, TX	1	566.71	0.67	2	1
Pn. greeni	Amherst, MA	5	567.29	1.40	1	0
Pn. macdermotti	Athens, GA	6	567.61	2.78	2	0
Pn. marginellus	FITG, PA; Kutztown, PA	9	561.81	0.51	2	0
Pn. obscurellus	ANF, PA	1	567.47	0.30	0	1
Pn. pyralis	Athens, GA; Athens, OH	8	561.88	1.67	0	0
Pn. sabulosus	Andalusia, AL	5	560.64	0.76	1	0
Pn. scintillans	FITG, PA	7	574.85	2.89	2	0
Pn. texanus	Vanderpool, TX	3	562.26	0.95	0	0
Pt. frontalis	Athens, GA	5	565.51	0.57	2	1
Pt. sp.	Tate City, GA	1	556.80	0.33	0	1

Py. angulata	ANF, PA; Wynne, AR	2	572.25	0.25	0.5	1
Py. borealis	Athens, GA	3	568.25	0.47	0	1
Py. linearis	ANF, PA	4	569.92	4.58	0	1
Py. marginalis	Amesville, OH	1	575.29	0.42	1	0

[a] Genus abbreviations: *Pa: Phausis, Pt: Photuris, Ld: Lucidota, Py: Pyractomena, Pn: Photinus, El: Ellychnia*.

[b] Location of population(s) sampled for emission spectra. ANF: Allegheny National Forest, FITG: Fort Indiantown Gap.

[c] Trait data are given as the value for the specific population where the specimen used for confirmatory Sanger sequencing of opsins was caught. N: number of specimens measured; Mean: average male peak emission wavelength (nm); SD: standard deviation across individuals within the population. Where specimens from multiple locations were used to get opsin consensus sequences, this is the total number of specimens measured and the mean/SD across those specific populations. Where N = 1, SD represents the within-individual standard deviation.

[d] Habitat category coded as an ordinal variable. 0: open, 1: mixed, 2: closed.

[e] Activity time category coded as an ordinal variable. 0: early, 1: late.

Text S2. Homology modeling

Model construction

The best opsin template structure was determined using the SwissModel online workspace (Arnold et al. 2006; http://swissmodel.expasy.org), followed by manual review of electron density maps of the suggested templates to determine which had the best resolution, particularly in the chromophore binding pocket. *Pn. pyralis* sequences for LW and UV opsin were separately used as the target sequence in the template search. The final template selected for both opsins was squid opsin (*Todarodes pacificus*), 2z73A, with a resolution of 2.5 A (Murakami and Kouyama 2008). Homology models for LW and UV opsins were then constructed in SwissModel. The resulting models had low QMEAN scores (LW: -7.04, UV: -7.83; Benkert et al. 2011), indicating that models were not optimal, most likely due to the low sequence identity between the target and template (LW: 36%, UV: 30%). Firefly opsins have long tails at either end that are divergent from squid opsin and were not successfully aligned. QMEAN scores improved marginally when these areas were removed. A loop-out in the LW-opsin model was fixed by altering Y206 to F in order to preserve the helical structure at this position. Other differences between the homology models and secondary structure predictions from PsiPred (Jones 1999) could not be fixed given the divergence between the template (squid) and target (firefly) sequences.

Binding pocket site identification

Sites able to interact with the chromophore in the chromophore binding pocket were determined by visual inspection of Van der Waals forces around the chromophore in the homology models for LW and UV-opsin. If the forces from the chromophore and the opsin amino acid site touched, as calculated in PyMol, they were considered binding pocket sites. If the Van der Waals forces did not touch, but were oriented toward the binding pocket in such a way that no other forces were interfering between the opsin and the chromophore, they were considered "possible interactors" and included in the list of candidates.

Electron density maps of the template structure were particularly well resolved in the binding pocket area, including several conserved amino acid sequence motifs, leading to high confidence in the identification of individual sites with the potential to interact with the chromophore in LW opsin. Several binding pocket sites had large changes in structure between the LW and UV opsin, consistent with the idea that the chromophore is held differently in the UV opsin binding pocket and supporting the putative functional difference between these distinct opsin types.

Text S3**.** Investigating robustness of phylogeny.

Methods for phylogeny construction following Stanger-Hall and Lloyd (2015)

Species tree construction methodology followed the procedures given in Stanger-Hall and Lloyd (2015) and extended these procedures to explore the robustness of the tree to methods of construction. Briefly, the final alignment was tested for appropriate models of sequence evolution using jModeltest2 v1.4 (Guindon and Gascuel 2003; Darriba et al. 2012) and competing models were evaluated using the Akaike Information Criterion, corrected for small sample size (AICc). The best model of substitution for the final phylogeny partitioned the alignment by locus (*WG*: K80+I+G, *CAD*: TIM3+I+G, *COI*: GTR+I+G). PartitionFinder v.1.1.1 (Lanfear et al. 2012) was also used to identify the best partitioning scheme of full alignments using AICc. Trees constructed using different substitution models and partitioning schemes were compared using lnL values from maximum likelihood analyses and harmonic means in Bayesian analyses.

Phylogenies were constructed using MrBayes 3.2 (Ronquist et al. 2012), Garli 2.0 (Zwickl 2006), and the lognormal clock model in BEAST v. 1.8 (Drummond et al. 2012). MrBayes was run until the average standard deviation of split frequencies between two simultaneous runs was less than 0.01, for at least five million generations with four chains, a heating parameter of 0.02, and a 25% burn-in. Garli was run until there was no significant improvement in topology over 20,000 generations and the total improvement in tree score was less than 0.05 over 500 generations. BEAST was run for 30 million generations, until the estimated sample sizes (ESS) for all parameters were over 200. Two independent runs were performed and assessed for convergence using Tracer v1.4.

The opsin phylogeny

In addition to the neighbor-joining phylogeny shown in the main text, we constructed maximum likelihood and Bayesian phylogenies with beetle opsin nucleotide sequences to better resolve the evolution of opsins across beetles, and, since many insect molecular phylogeny studies include opsins in their construction, ascertain the utility of opsins as a phylogenetic marker in fireflies. To do this several large alignments were constructed that included other beetle opsins: *Tribolium castaneum* (LW: gi|242117920, UV: gi|91092881), *Thermonectus marmoratus* (LW: EU921225, UV I: EU921226, UV II: EU921227), the known firefly opsins from *Luciola cruciata* (LW: AB300328, UV: AB300329), and two outgroups: the squid rhodopsin used in homology modeling (*Todarodes pacificus*: X70498.1) and *Bos taurus* opsins (LW: NM_174566, SW: NM_174567). Alignments were generated by aligning translated coding sequences using ClustalW (Larkin et al. 2007) with default parameters in Geneious and manually reviewed. Nucleotides were then forced onto the amino acid alignment using RevTrans (Wernersson and Pedersen 2003) and reviewed a second time. In total 3 alignments were examined: (1) beetle opsins + outgroups, (2) only LW opsins, and (3) only UV opsins.

Phylogenetic analysis of the beetle opsin alignments with divergent opsin outgroups confirmed the ancient split of UV and LW opsins observed in the neighbor-joining phylogeny. However, within-firefly species relationships were not well resolved. Garli maximum likelihood analysis resulted in topologies with low bootstrap support and many polytomies, similar to the neighbor-joining tree presented in Figure S4, whereas Bayesian trees had higher support, but the topology differed at key branches from the tree of Stanger-Hall and Lloyd (2015), using two nuclear markers and one mitochondrial marker. Incongruence between the species tree and opsin

tree is expected if there has been convergent evolution of opsins in divergent lineages that have similar ecological or signaling traits. It is also possible that LW and UV opsin singly could not yield a fully resolved species tree because of the lack of variable sites in these extremely conserved genes or due to incomplete lineage sorting, especially in recently diverged taxa.

The species phylogeny and taxon sampling

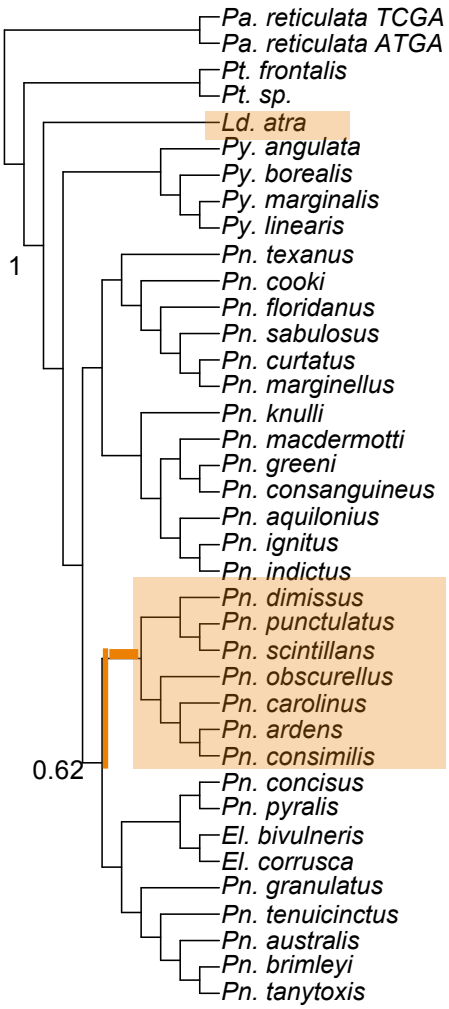
Because the opsin tree was not resolved enough for analysis, we performed our subsequent analysis using two species tree topologies obtained from extending the three-locus dataset described in Stanger-Hall and Lloyd (2015)(see main text for more details). In order to examine the robustness of the phylogeny to taxon sampling, two alignments were generated, one with 14 additional North American taxa and one with those 14 plus 18 international specimens. The resulting sequences were aligned and used to create phylogenies as described in the main text. In all analyses the species topology was constrained to have *Luciola cerata* as outgroup. The placement of *Lu. cerata* as sister to North American firefly species is supported by the topology of the firefly worldwide phylogeny (Stanger-Hall et al. 2007) based on ribosomal and mitochondrial genes. Tree construction method, either Bayesian or maximum likelihood, had no effect on topology.

Two slightly different topologies resulted from the two alignments, as shown on the following page (numbers show Bayesian support values for conflicting nodes). Subsequent analysis was performed on both topologies. Results from the 76-taxon topology are presented in the text. Topology did not substantially affect the identification of positively selected sites in PAML. However, one additional site in LW, site 248, was identified using the 94-taxon phylogeny and was included in further analysis (Table 3, main text).

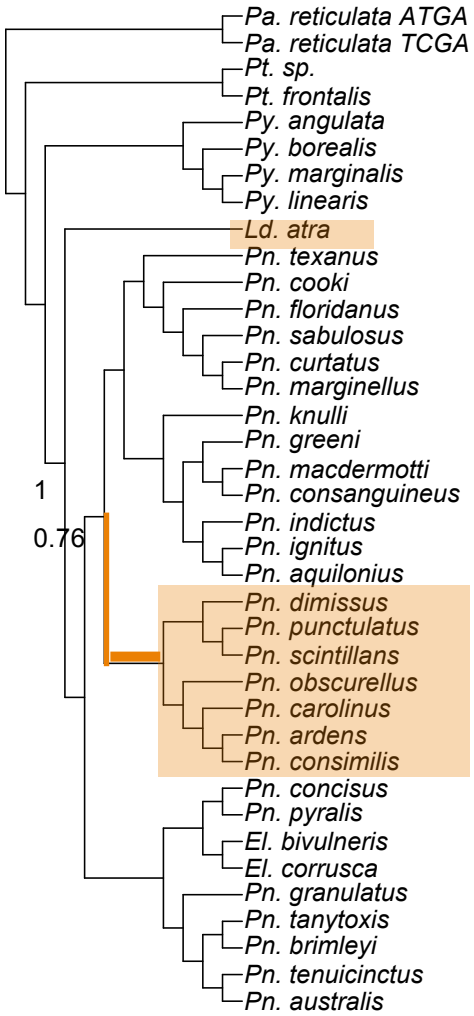
The branch lengths of the phylogeny (assuming no clock) with the additional basal lineages suggested that there might be substitution rate heterogeneity among lineages that might affect tree inference, especially branch length. Therefore, we also constructed two ultrametric trees using the two taxon sets using a lognormal clock model in BEAST v1.8 (Drummond et al. 2012). For each tree, BEAST was run for 30 million generations, until the estimated sample sizes (ESS) for all parameters were over 200. Two independent runs were performed and assessed for convergence using Tracer v1.4 (Rambaut and Drummond 2007). The ucld.stdev parameter of both runs differed significantly from 0, indicating that our data was subject to rate heterogeneity.

Analyses performed on these ultrametric trees resulted in some differences in PAML and fitmodel results (data available upon request). A consensus set of positively selected sites was constructed from the PAML and fitmodel analyses across all four taxon (76 vs 94) by method (Garli vs BEAST) topologies (see main text). Results and figures presented in the main text are based on the BEAST 76-taxon topology.

All final trees were trimmed to include one individual per species that was included in our dataset of opsin sequences and ecological and signaling traits.



Tree 1: Lucidota basal (+14 taxa) Ellychnia/pyralis derived



Tree2: Pyractomena basal (+32 taxa) Ellychnia/pyralis basal

Text S4. Details on specific amino acid substitutions in relation to signal mode

LW opsin

Reconstructions of the LW opsin amino acid sequences showed convergent or parallel changes in diurnal taxa for two of the eight positively selected sites identified by both PAML and fitmodel: A248V in *Lucidota atra* and *Ellychnia corrusca*, a substitution that does not change the polarity at that site, and S\T309A in *Ld. atra* and *El. bivulneris*, a polar to nonpolar change (Figure S5). These changes occurred in no more than two diurnal taxa (rather than all diurnal lineages) and were also observed in at least one nocturnal lineage. In contrast, reconstruction of amino acids at the two variable binding pocket sites in LW did not show any pattern with nocturnal to diurnal activity shifts (Figure S6). Rather, the two variable sites, 205 and 210, both experienced substitutions of amino acids of similar chemical property along a single branch. V205A (both nonpolar) changed on the branch to the genus *Phausis*; Y210N (both polar) changed on the branch to the MRCA of *Photinus*. Reconstruction of all variable sites showed two sites outside the binding pocket with parallel or convergent changes exclusive to diurnal lineages (sites 59 and 301; Figure S7). These sites, especially those with changes in polarity, are candidates to investigate in future functional studies.

UV opsin

Of the eight UV opsin sites with $\omega > 1$, there were parallel changes at two sites on diurnal lineages: V59L in *Pn. indictus* and *El. bivulneris*, and V218A in *Pn. indictus* and the MRCA of *Ellychnia* (Figure S8). Again, these changes were not limited to diurnal lineages and also occurred in a nocturnal lineage. It is interesting to note that the only amino acid changes at site 268 in *Photinus* occurred in the diurnal lineages *Pn. indictus* and *El. corrusca*, though these were not parallel changes and only the substitution in *El. corrusca* involved a change in polarity. Of the five variable binding pocket sites, four had amino acid changes that were associated with diurnal lineages: two independent F128Y (both nonpolar) on the branches to *Ld. atra* and *El. corrusca*, G129A (both nonpolar) in *El. bivulneris*, S133A (polar to nonpolar) in *El. bivulneris*, S203T (both polar) in *Ld. atra*. (Figure S9). These changes were exclusive to diurnal lineages. The sixth variable binding pocket site had six independent changes, A299S (nonpolar to polar), in nocturnal lineages only (Figure S8). Two other variable sites had evidence for parallel amino acid changes exclusive to diurnal taxa: sites 68 and 214 (Figure S10).

Text S5. Testing predictions of LW spectral absorbance.

Cronin proposed a model for tuning that included an opsin-plus-chromophore tuned to 540 nm in *Pt. versicolor* and *Pn. pyralis*, and 560 nm in *Pn. scintillans*, with screening pigments that allowed further tuning of spectral sensitivity to each species-specific signal color (Cronin et al. 2000). Given this model, it is expected that *Pn. pyralis* and *Pt. versicolor* LW opsins should be more similar to one another at spectral tuning sites than *Pn. pyralis* and *Pn. scintillans*. This pattern is observed when examining the polarity of amino acids at 2/8 of the positively selected sites, but none of the variable binding pocket sites, identified in this study. At site 108 *Pn. scintillans* has a nonpolar methionine while *Pn. pyralis* and the two *Photuris* species have polar cysteine and threonine residues, suggesting that there may be functional consequences for the methionine substitution. At site 309, *Pn. scintillans* has a nonpolar valine while the other three species have polar threonine and serine residues. This pattern is purely speculative and needs further functional study.

References

Arnold K, Bordoli L, Kopp J, Schwede (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics,* **22**, 195-201.

Benkert P, Biasini M, Schwede T (2011) Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, **27**, 343-350.

Biggley WH, Lloyd JE, Seliger HH (1967) The spectral distribution of firefly light II. *The Journal of General Physiology*, **50**, 1681-1692.

Blomberg SP, Garland T, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, **57**, 717-745.

Cronin TW, Jarvilehto M, Weckstrom M, Lall AB (2000) Tuning of photoreceptor spectral sensitivity in fireflies (Coleoptera: Lampyridae). *Journal of Comparative Physiology A*:1-12.

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, **9**, 772-772.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with beauti and the BEAST 1.7. *Molecular Biology and Evolution*, **29**, 1969-1973.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**, 696-704.

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, **292**, 195-202.

Lall AB, Seliger HH, Biggley WH, Lloyd JE (1980) Ecology of colors of firefly bioluminescence. *Science*, **210**, 560-562.

Lanfear R, Calcott B, Ho SYW, Guindon S (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, **29**, 1695-1701.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947-2948.

Murakami M, Kouyama T (2008) Crystal structure of squid rhodopsin. *Nature*, **453**, 363-367.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, **61**, 539-42.

Seliger HH, Buck JB, Fastie WG, McElroy WD (1964) The spectral distribution of firefly light. *Journal of General Physiology*, **48**, 95-104.

Stanger-Hall KF, Lloyd JE, Hillis DM (2007) Phylogeny of North American fireflies (Coleoptera: Lampyridae): implications for the evolution of light signals. *Molecular Phylogenetetics and Evolution*, **45**, 33-49.

Wernersson R, Pedersen AG (2003) RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Research*, **31**, 3537-3539.

Zwickl, DJ (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. Dissertation, The University of Texas at Austin.

APPENDIX B

Supplementary material for Sander *et al.*

To be submitted to *Genome Biology and Evolution*

Table S1. Morphological measures dataset

Unique identification number in the Stanger-Hall collection (KSH), state in which collected (Loc), pronotum length (ProL), pronotum width (ProW), pronotum area (ProA), elytron length (ElyL), and body length (BodyL) measurements for each specimen used in t morphological analysis. Genus abbreviations: Ellychnia (El), Lucidota (Ld), Phausis (Pa), Photinus (Pn), Photuris (Pt), Pyractomena (Py), Pyropyga (Pg). N/A indicates a measurement that could not be taken due to specimen damage.

KSH	Genus	Species	Loc	Sex	ProL	ProW	ProA	ElyL	BodyL
609	*El*	*corrusca*	MI	F	3.005	4.28	11.261	9.903	10.521
2095	*El*	*corrusca*	VT	F	2.572	3.666	7.485	8.528	8.555
1138	*El*	*corrusca*	MA	F	1.951	4.744	6.925	10.881	11.788
706	*El*	*corrusca*	WI	M	2.178	3.225	5.777	7.059	7.413
2094	*El*	*corrusca*	VT	M	2.132	2.991	4.878	6.75	N/A
2090	*El*	*corrusca*	VT	M	2.179	3.341	5.533	7.919	N/A
11502	*El*	*corrusca*	MA	M	2.264	3.643	6.671	7.54	8.298
11503	*El*	*corrusca*	MA	M	2.011	3.144	5.379	7.48	8.273
11227	*Ld*	*atra*	PA	F	2.96	4.92	10.981	11.03	12.88
9800	*Ld*	*atra*	PA	F	2.792	4.71	9.696	10.886	9.771
8357	*Ld*	*atra*	TN	F	2.418	3.744	7.202	9.146	9.235
8550	*Ld*	*atra*	NY	M	1.837	3.138	4.728	8.044	9.794
11399	*Ld*	*atra*	MA	M	2.263	3.269	5.743	8.7	8.77
9325	*Ld*	*atra*	OH	M	2.661	3.786	7.86	8.758	10.845
9011	*Ld*	*punctata*	TN	M	1.137	1.59	1.403	5.516	5.39
8439	*Ld*	*punctata*	TN	M	1.125	1.757	1.496	5.417	5.104
8355	*Ld*	*punctata*	TN	M	1.436	1.774	1.766	5.352	5.736
8408	*Pa*	*reticulata*	TN	M	1.69	2.141	3.185	7.122	N/A
8409	*Pa*	*reticulata*	TN	M	1.385	1.791	2.258	5.799	6.289
8430	*Pa*	*reticulata*	TN	M	1.499	1.954	2.544	5.734	N/A
8033	*Pa*	*reticulata*	GA	M	1.278	1.51	1.848	5.517	6.107
8718	*Pa*	*reticulata*	GA	M	1.175	1.387	1.619	4.948	5.447
8663	*Pa*	*sp. WAT*	GA	M	1.266	1.564	1.773	5.395	5.774
8667-2	*Pa*	*sp. WAT*	GA	M	0.834	1.179	0.873	4.333	4.085
8667-4	*Pa*	*sp. WAT*	GA	M	0.769	1.172	0.828	3.957	3.747
1742	*Pn*	*australis*	TN	F	1.487	2.178	2.718	5.404	6.635
1740	*Pn*	*australis*	TN	F	1.548	2.449	3.048	5.779	7.941
8052	*Pn*	*australis*	GA	F	1.835	2.648	3.743	7.545	9.007
2010	*Pn*	*australis*	MS	M	1.932	2.725	4.579	6.733	8.333
1029	*Pn*	*australis*	MS	M	1.999	2.765	4.671	7.597	8.385
588	*Pn*	*australis*	MO	M	1.775	2.354	3.367	6.093	8.849
490	*Pn*	*brimleyi*	AR	M	2.142	2.944	5.243	7.63	9.643
546	*Pn*	*brimleyi*	AR	M	1.801	2.312	3.424	5.663	10.466
660	*Pn*	*brimleyi*	AR	M	2.335	2.856	5.432	7.547	9.524

9705	*Pn*	*carolinus*	PA	F	2.594	4.178	8.376	10.003	13.914
11341	*Pn*	*carolinus*	PA	F	2.008	3.629	7.335	10.001	13.972
9712	*Pn*	*carolinus*	PA	F	2.365	4.115	8.337	10.093	11.847
8824	*Pn*	*carolinus*	GA	M	2.103	3.94	6.019	11.589	11.345
9521	*Pn*	*carolinus*	PA	M	1.744	2.929	3.946	10.977	11.446
8218	*Pn*	*carolinus*	GA	M	2.384	3.776	7.724	10.482	13.617
1696	*Pn*	*cooki*	TN	F	1.405	1.782	2.206	5.112	5.862
1689	*Pn*	*cooki*	TN	F	1.481	1.989	2.433	5.403	7.192
1690	*Pn*	*cooki*	TN	F	1.541	1.878	2.374	5.235	6.572
1697	*Pn*	*cooki*	TN	M	1.198	1.504	1.633	5.17	N/A
1660	*Pn*	*cooki*	KY	M	1.315	1.93	2.196	5.538	7.248
1688	*Pn*	*cooki*	TN	M	1.409	1.707	2.158	3.819	N/A
1490	*Pn*	*curtatus*	NE	F	1.386	1.73	2.03	5.173	6.619
1529	*Pn*	*curtatus*	MN	F	1.513	2.027	2.499	5.873	7.496
1486	*Pn*	*curtatus*	NE	F	1.645	1.974	2.693	5.234	6.319
1499	*Pn*	*curtatus*	MN	M	1.692	2.093	3.007	6.554	7.397
1579	*Pn*	*curtatus*	IN	M	1.344	1.767	1.986	5.203	N/A
1581	*Pn*	*curtatus*	IN	M	1.73	2.655	3.97	6.552	N/A
9131	*Pn*	*curtatus*	IL	M	1.476	1.913	2.537	5.691	7.202
9172	*Pn*	*curtatus*	IL	M	1.542	1.978	2.81	5.794	6.866
1511	*Pn*	*curtatus*	MN	M	1.36	1.879	2.32	5.91	6.225
1780	*Pn*	*granulatus*	KS	M	1.575	1.78	2.576	5.44	7.212
1461	*Pn*	*granulatus*	KS	M	1.895	2.113	3.608	5.226	N/A
66	*Pn*	*granulatus*	KS	M	1.306	1.71	1.894	4.464	N/A
625	*Pn*	*indictus*	MI	F	1.687	2.335	3.438	5.384	8.202
1569	*Pn*	*indictus*	IN	F	1.808	2.588	4.031	5.431	7.361
624	*Pn*	*indictus*	MI	M	1.772	2.25	3.81	5.081	N/A
9685	*Pn*	*indictus*	PA	M	1.631	2.237	2.908	5.595	N/A
9687	*Pn*	*indictus*	PA	M	1.411	2.035	2.635	5.315	N/A
991	*Pn*	*macdermotti*	CT	F	2.516	3.353	7.272	9.145	12.436
8609	*Pn*	*macdermotti*	MA	F	2.062	2.993	5.288	7.909	9.617
8608	*Pn*	*macdermotti*	MA	F	2.047	2.961	5.089	8.127	10.286
707	*Pn*	*macdermotti*	WI	F	1.894	2.727	4.192	6.906	8.723
918	*Pn*	*macdermotti*	CT	M	1.663	2.365	3.125	6.778	7.523
1042	*Pn*	*macdermotti*	TN	M	1.751	2.577	3.527	7.429	8.689
8304	*Pn*	*macdermotti*	GA	M	1.699	2.535	3.497	7.489	8.487
2046	*Pn*	*macdermotti*	MO	M	2.236	2.863	5.309	6.601	N/A
267	*Pn*	*macdermotti*	NC	M	2.356	2.976	5.878	7.558	9.307
1041	*Pn*	*macdermotti*	TN	M	1.801	2.496	3.375	6.414	N/A
694	*Pn*	*marginellus*	IL	F	1.634	2.078	2.804	6.627	N/A
946	*Pn*	*marginellus*	MI	F	1.278	2.467	2.865	6.175	8.139
9850	*Pn*	*marginellus*	PA	F	1.503	1.863	2.264	5.819	7.352
10251	*Pn*	*marginellus*	PA	F	1.611	2.159	3.044	6.196	6.417

959	*Pn*	*marginellus*	NY	M	1.508	2.121	2.842	6.214	7.476
8362	*Pn*	*marginellus*	TN	M	1.606	2.044	2.808	6.215	7.743
9849	*Pn*	*marginellus*	PA	M	1.593	1.929	2.687	6.278	6.681
1063	*Pn*	*obscurellus*	VT	M	1.695	2.531	3.602	8.026	8.043
2089	*Pn*	*obscurellus*	VT	M	1.537	2.407	2.829	6.669	8.47
1060	*Pn*	*obscurellus*	VT	M	1.743	2.709	3.686	7.933	8.78
930	*Pn*	*pyralis*	PA	F	2.417	3.657	7.691	9.865	14.734
830	*Pn*	*pyralis*	WV	F	3.124	3.963	10.567	10.841	N/A
9059	*Pn*	*pyralis*	MO	F	2.126	2.894	5.563	8.538	11.517
9318	*Pn*	*pyralis*	OH	F	2.335	3.206	6.674	9.302	13.763
693	*Pn*	*pyralis*	IL	M	3.017	3.695	9.814	11.571	13.401
872	*Pn*	*pyralis*	MD	M	2.174	2.602	4.895	9.691	12.674
795	*Pn*	*pyralis*	WV	M	2.116	2.439	4.559	8.447	11.144
10176	*Pn*	*pyralis*	PA	M	3.309	3.881	10.573	12.391	14.974
8492	*Pn*	*sabulosus*	VA	M	1.565	1.805	2.524	5.974	7.333
8508	*Pn*	*sabulosus*	VA	M	1.338	1.652	2.142	5.808	6.805
8578	*Pn*	*sabulosus*	MD	M	1.404	1.734	2.294	5.934	6.667
11405	*Pn*	*scintillans*	NJ	F	1.404	2.055	2.505	1.894	6.676
11209	*Pn*	*scintillans*	PA	F	1.433	1.993	2.823	2.271	7.562
10242	*Pn*	*scintillans*	PA	F	1.222	1.458	1.859	2.128	5.655
1582	*Pn*	*scintillans*	IN	M	1.612	2.305	3.133	6.614	7.849
1575	*Pn*	*scintillans*	IN	M	1.4	1.968	2.075	5.401	N/A
851	*Pn*	*scintillans*	PA	M	1.453	2.574	3.25	7.183	8.297
11440	*Pn*	*scintillans*	MD	M	1.527	2.216	2.901	6.166	8.064
8049	*Pt*	*frontalis*	GA	M	2.412	3.209	6.493	11.315	11.306
8186	*Pt*	*frontalis*	GA	M	2.657	3.422	8.169	10.131	11.941
8032	*Pt*	*frontalis*	GA	M	2.53	3.407	7.308	9.709	11.491
9033	*Pt*	*sp.*	MS	M	2.598	2.952	6.221	9.22	10.16
9034	*Pt*	*sp.*	MS	M	2.36	2.792	5.616	9.936	10.658
8614	*Py*	*angulata*	MA	M	2.826	4.32	9.399	9.504	11.344
9072	*Py*	*angulata*	MO	M	2.492	4.214	7.932	9.355	8.908
9076	*Py*	*angulata*	MO	M	2.465	3.57	7.131	8.565	8.879
10863	*Py*	*borealis*	GA	F	3.434	4.47	13.341	12.13	13.787
8091	*Py*	*borealis*	GA	M	4.121	4.08	9.41	9.711	12.091
10861	*Py*	*borealis*	GA	M	3.976	4.167	13.55	12.114	15.566
8801	*Py*	*borealis*	GA	M	3.594	4.28	13.224	11.872	12.25
10105	*Py*	*marginalis*	PA	M	1.95	2.196	3.695	6.968	N/A
698	*Pg*	*decipiens*	WI	F	1.289	1.831	2.177	4.374	5.801
1501	*Pg*	*decipiens*	MN	F	1.207	2.261	2.183	5.325	6.583
1552	*Pg*	*decipiens*	MI	F	1.222	1.858	1.947	4.17	6.152
1555	*Pg*	*decipiens*	MI	M	0.965	1.373	1.181	3.78	4.101
1553	*Pg*	*decipiens*	MI	M	1.201	1.808	1.832	4.149	5.753
1130	*Pg*	*decipiens*	MD	M	0.989	1.437	1.188	4.472	4.45

Table S2. Specimens used in 454 sequencing

Specimen[a]	Species[b]	Sex	Locality
9734	*El. corrusca*	M	Allegheny National Forest, PA
9324	*Ld. atra*	F	Athens, OH
9010	*Ld. punctata*	M	Fall Creek Falls State Park, TN
8410	*Pa. reticulata*	M	Great Smoky Mountains National Park, TN
8136	*Pa. sp. WAT*	M	Watkinsville, GA
9623	*Pg. decipiens L*	F	Allegheny National Forest, PA
9622	*Pg. decipiens S*	F	Allegheny National Forest, PA
8234	*Pn. australis*	M	Watkinsville, GA
9015	*Pn. brimleyi*	M	Cummins Mill State Park, TN
9692	*Pn. carolinus*	F	Allegheny National Forest, PA
9026	*Pn. cooki*	F	Mongomery Bell State Park, TN
9686	*Pn. indictus*	M	Allegheny National Forest, PA
9592	*Pn. macdermotti*	F	Allegheny National Forest, PA
10076	*Pn. obscurellus*	M	Allegheny National Forest, PA
9020	*Pn. pyralis*	F	Cummins Mill State Park, TN
9363	*Pn. sabulosus*	M	Amesville, OH
10217	*Pn. scintillans*	F	Longwood Gardens; Kennett Square, PA
8178	*Pt. frontalis*	M	Whitehall Forest; Athens, GA (University of Georgia)
9032	*Pt. sp.*	F	Byhalia, MS
9223	*Py. angulata*	M	Kent Farm; Bloomington, IN (Indiana University)
10084	*Py. marginalis*	M	Allegheny National Forest, PA

[a] Refers to the unique identifying number in the Stanger-Hall collection at the University of Georgia
[b] Genus abbreviations: Ellychnia (El), Lucidota (Ld), Phausis (Pa), Pyropyga (Pg) Photinus (Pn), Photuris (Pt), Pyractomena (Py)

Table S3. Genome size estimates for specimens used in this study

Specimen[a]	Genus	Species	Sex	Genome size (Mb)
9734	*Ellychnia*	*corrusca*	M	781.56
9324	*Lucidota*	*atra*	F	512.31
9609	*Lucidota*	*atra*	F	463.47
9610	*Lucidota*	*atra*	F	557.55
9646	*Lucidota*	*atra*	F	457.01
9766	*Lucidota*	*atra*	F	465.45
9007	*Lucidota*	*atra*	M	463.88
9008	*Lucidota*	*atra*	M	478.21
9009	*Lucidota*	*atra*	M	514.96
9108	*Lucidota*	*atra*	M	461.59
9219	*Lucidota*	*atra*	M	457.24
9010	*Lucidota*	*punctata*	M	1294.29
9681	*Lucidota*	*punctata*	M	1305.82
8392	*Phausis*	*reticulata*	M	835.50
8432	*Phausis*	*reticulata*	M	835.62
8652	*Phausis*	*sp. WAT*	F	1212.05
8653	*Phausis*	*sp. WAT*	F	1194.4
8654	*Phausis*	*sp. WAT*	F	1260.45
8655	*Phausis*	*sp. WAT*	F	1173.44
8656	*Phausis*	*sp. WAT*	F	1210.31
8650	*Phausis*	*sp. WAT*	M	1090.98
8651	*Phausis*	*sp. WAT*	M	1138.18
8137	*Photinus*	*australis*	M	1622.34
8138	*Photinus*	*australis*	M	1616.9
8140	*Photinus*	*australis*	M	1628.59
8142	*Photinus*	*australis*	M	1606.40
8234	*Photinus*	*australis*	M	1601.54
9014	*Photinus*	*brimleyi*	M	1157.25
9015	*Photinus*	*brimleyi*	M	1204.26
9692	*Photinus*	*carolinus*	F	701.99
9693	*Photinus*	*carolinus*	F	627.07
9694	*Photinus*	*carolinus*	F	657.01
9710	*Photinus*	*carolinus*	F	669.04
9711	*Photinus*	*carolinus*	F	649.77
8331	*Photinus*	*carolinus*	M	664.26
8332	*Photinus*	*carolinus*	M	685.05
8334	*Photinus*	*carolinus*	M	682.26
8335	*Photinus*	*carolinus*	M	660.66
8336	*Photinus*	*carolinus*	M	645.95
8340	*Photinus*	*carolinus*	M	676.03

8341	*Photinus*	*carolinus*	M	642.99
9026	*Photinus*	*cooki*	F	700.98
9135	*Photinus*	*curtatus*	F	684.18
9461	*Photinus*	*curtatus*	F	677.00
9132	*Photinus*	*curtatus*	M	635.88
9134	*Photinus*	*curtatus*	M	652.33
9136	*Photinus*	*curtatus*	M	647.18
9138	*Photinus*	*curtatus*	M	652.33
9159	*Photinus*	*curtatus*	M	655.47
9684	*Photinus*	*indictus*	M	435.86
9686	*Photinus*	*indictus*	M	430.51
9552	*Photinus*	*macdermotti*	F	511.48
9592	*Photinus*	*macdermotti*	F	531.43
9696	*Photinus*	*macdermotti*	F	503.42
9697	*Photinus*	*macdermotti*	F	497.69
9716	*Photinus*	*macdermotti*	F	504.72
9731	*Photinus*	*macdermotti*	F	505.19
8235	*Photinus*	*macdermotti*	M	481.23
8328	*Photinus*	*macdermotti*	M	475.14
8709	*Photinus*	*macdermotti*	M	472.88
8710	*Photinus*	*macdermotti*	M	485.05
8712	*Photinus*	*macdermotti*	M	506.97
9570	*Photinus*	*macdermotti*	M	471.12
9571	*Photinus*	*macdermotti*	M	487.56
9574	*Photinus*	*macdermotti*	M	481.56
9575	*Photinus*	*macdermotti*	M	483.02
9645	*Photinus*	*macdermotti*	M	468.89
10075	*Photinus*	*marginellus*	F	709.52
10316	*Photinus*	*marginellus*	F	684.91
10849	*Photinus*	*marginellus*	F	679.06
9606	*Photinus*	*marginellus*	F	669.35
9607	*Photinus*	*marginellus*	F	675.55
10082	*Photinus*	*marginellus*	M	634.99
10259	*Photinus*	*marginellus*	M	649.24
10846	*Photinus*	*marginellus*	M	644.51
10847	*Photinus*	*marginellus*	M	647.07
10848	*Photinus*	*marginellus*	M	650.07
10850	*Photinus*	*marginellus*	M	655.14
10076	*Photinus*	*obscurellus*	M	641.09
10103	*Photinus*	*obscurellus*	M	659.30
10252	*Photinus*	*obscurellus*	M	722.63
9020	*Photinus*	*pyralis*	F	441.95
9021	*Photinus*	*pyralis*	F	470.10

9025	*Photinus*	*pyralis*	F	433.71
9028	*Photinus*	*pyralis*	F	440.52
9077	*Photinus*	*pyralis*	F	452.38
8371	*Photinus*	*pyralis*	M	408.73
8871	*Photinus*	*pyralis*	M	421.76
8872	*Photinus*	*pyralis*	M	411.82
8880	*Photinus*	*pyralis*	M	447.03
9361	*Photinus*	*sabulosus*	M	611.54
9362	*Photinus*	*sabulosus*	M	618.98
9363	*Photinus*	*sabulosus*	M	622.03
10173	*Photinus*	*scintillans*	F	1032.40
10222	*Photinus*	*scintillans*	M	1003.87
10223	*Photinus*	*scintillans*	M	989.24
10224	*Photinus*	*scintillans*	M	1000.55
9879	*Photinus*	*scintillans*	M	1014.70
9884	*Photinus*	*scintillans*	M	1017.58
9889	*Photinus*	*scintillans*	M	1010.13
8180	*Photuris*	*frontalis*	M	2182.92
8182	*Photuris*	*frontalis*	M	2169.13
8183	*Photuris*	*frontalis*	M	2159.09
8184	*Photuris*	*frontalis*	M	2148.06
8185	*Photuris*	*frontalis*	M	2112.67
9030	*Photuris*		F	2486.35
9031	*Photuris*		F	2489.62
9032	*Photuris*		F	2572.44
9057	*Photuris*		F	2338.91
9640	*Photuris*		F	2543.14
9641	*Photuris*		F	2427.12
9642	*Photuris*		F	2366.32
9643	*Photuris*		F	2485.21
9017	*Photuris*		M	2455.33
9018	*Photuris*		M	2290.44
9055	*Photuris*		M	2301.44
9064	*Photuris*		M	2142.73
9153	*Photuris*		M	2386.88
9166	*Photuris*		M	2221.54
9179	*Photuris*		M	2281.16
9260	*Photuris*		M	2221.38
9262	*Photuris*		M	2188.70
9263	*Photuris*		M	2413.88
9267	*Photuris*		M	2161.6
9269	*Photuris*		M	2219.67
9717	*Photuris*		M	2223.11

9718	*Photuris*		M	2426.73
9719	*Photuris*		M	2157.83
9720	*Photuris*		M	2133.18
9721	*Photuris*		M	2181.62
9060	*Pyractomena*	*angulata*	M	774.07
9222	*Pyractomena*	*angulata*	M	771.98
9223	*Pyractomena*	*angulata*	M	763.72
9224	*Pyractomena*	*angulata*	M	851.48
9225	*Pyractomena*	*angulata*	M	830.91
8630	*Pyractomena*	*borealis*	M	761.25
8631	*Pyractomena*	*borealis*	M	766.75
8632	*Pyractomena*	*borealis*	M	801.71
8633	*Pyractomena*	*borealis*	M	760.65
8634	*Pyractomena*	*borealis*	M	783.76
10084	*Pyractomena*	*marginalis*	M	781.32
10104	*Pyractomena*	*marginalis*	M	754.75
9621	*Pyropyga*	*decipiens S*	F	704.31
9622	*Pyropyga*	*decipiens S*	F	697.53
9615	*Pyropyga*	*decipiens S*	M	688.31
9617	*Pyropyga*	*decipiens S*	M	704.17
9618	*Pyropyga*	*decipiens S*	M	704.01
9619	*Pyropyga*	*decipiens S*	M	695.97
9623	*Pyropyga*	*decipiens L*	F	1072.29
9624	*Pyropyga*	*decipiens L*	F	1086.92
9625	*Pyropyga*	*decipiens L*	F	1078.08

[a] Refers to the unique identifying number in the Stanger-Hall collection at the University of Georgia

Table S4. Testing for phylogenetic signal in repeats in the high coverage dataset

Blomberg's K values and their significance are given. * indicates significance after Benjamini-Hochberg correction for multiple comparisons.

Analysis 1:

Trait	K	p	Comparison	B-H correction
LTR	0.72	0.001*	12	0.004
Low complexity	0.54	0.012	11	0.005
DNA	0.44	0.012	10	0.005
Total percent repetitive	0.39	0.013	9	0.006
Unknown	0.50	0.013	8	0.006
Log ribosomal	0.36	0.025	7	0.007
Histone	0.43	0.055	6	0.008
LINE	0.34	0.056	5	0.010
Genome size	0.40	0.069	4	0.013
Simple repeat	0.32	0.2	3	0.017
RC	0.27	0.244	2	0.025
Tandem repeat	0.16	0.47	1	0.050

Analysis 2:

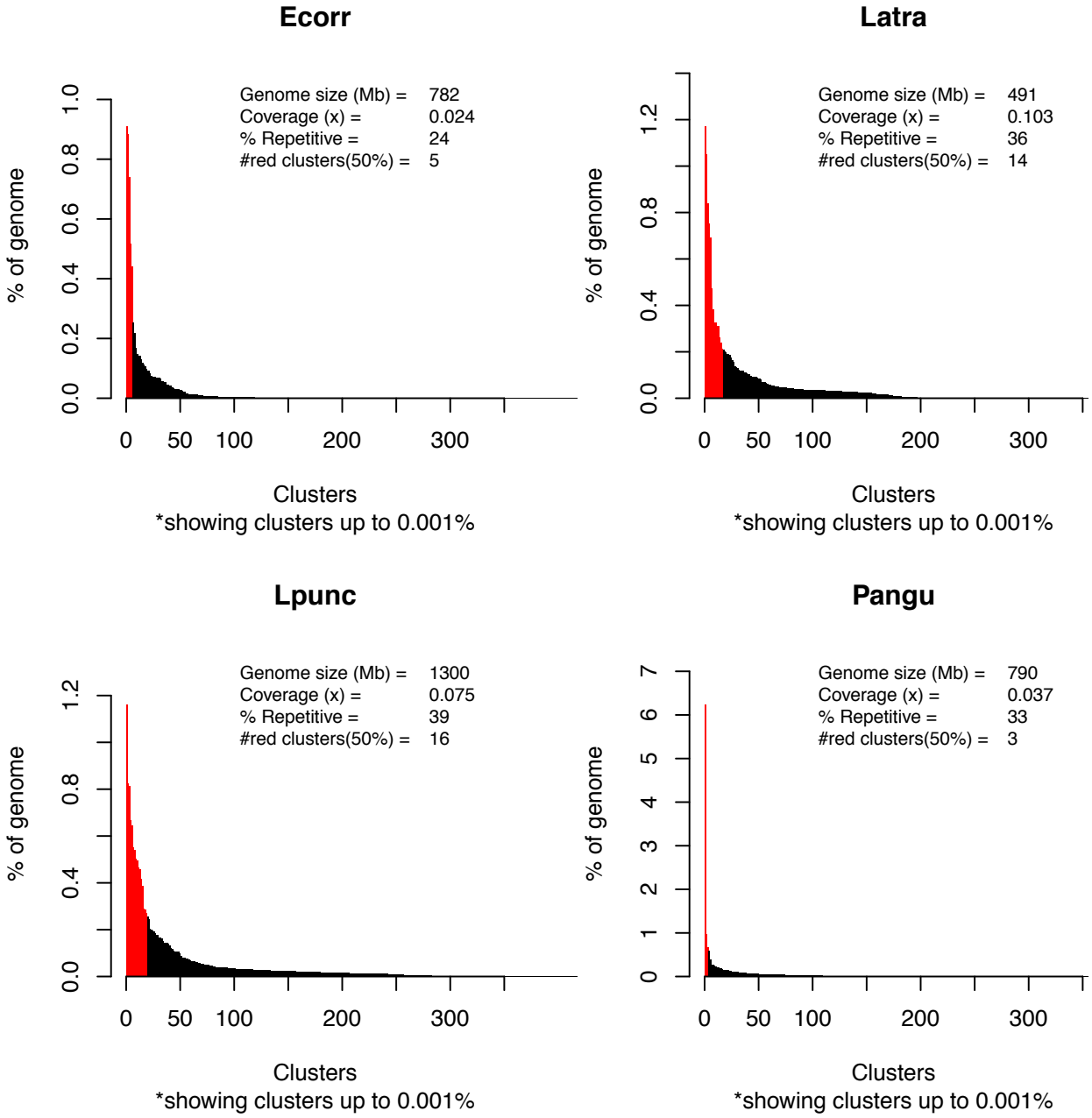
Final traits to use	K	p	Comparison	B-H correction
Total percent repetitive	0.39	0.013	6	0.008
Unknown	0.50	0.013	5	0.010
Class II	0.43	0.014	4	0.013
Class I	0.36	0.052	3	0.017
Genome size	0.40	0.069	2	0.025
Log repeats	0.17	0.295	1	0.050

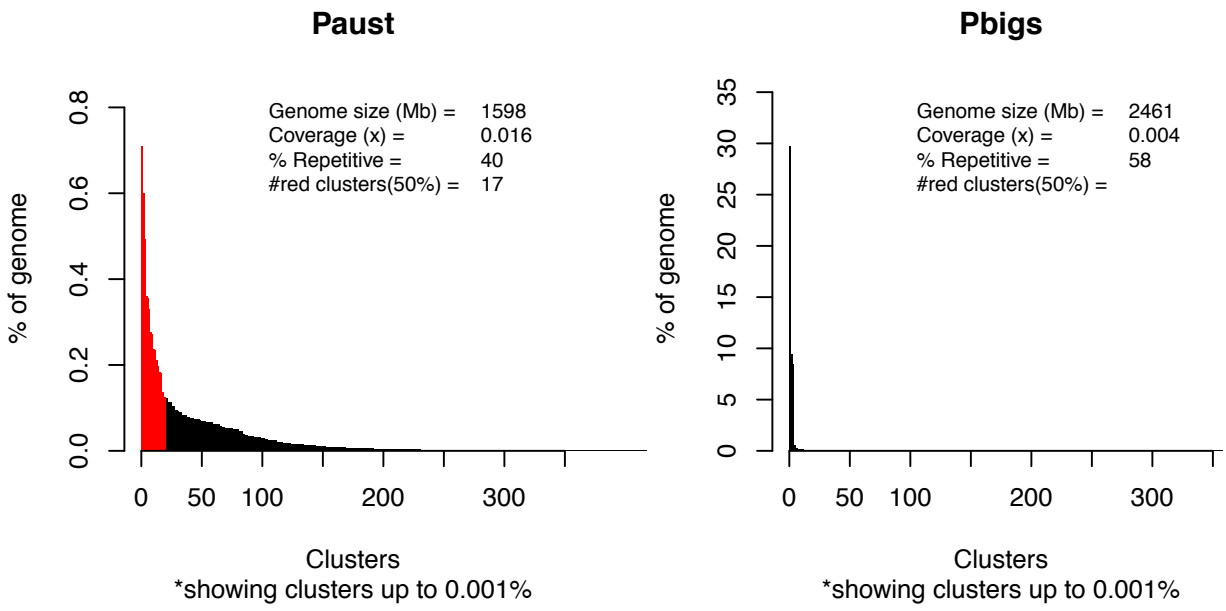
Table S5. Model selection and ANOVA

Model: GS =	Error SS	K	AICc	delta-AICc	AICc Weight	Sig parameters
Analysis 1, no phylogenetic correction						
DNA + H + LINE + LC + LTR + RC + logR + SR + TR +U	897351	12	335.10	74.44	2.71E-17	none
DNA + H + LC + LTR + RC + logR + SR + TR +U	897422	11	307.90	47.24	2.18E-11	none
DNA + H + LC + LTR + RC + SR + TR +U	897971	10	289.78	29.12	1.88E-07	RC (0.045)
DNA + H + LC + LTR + RC + SR + TR	1148312	9	281.01	20.34	1.51E-05	none
DNA + H + LC + LTR + RC + SR	1503169	8	275.87	15.21	0.00020	none
DNA + H + LTR + RC + SR	1522044	7	268.53	7.86	0.0077	none
DNA + H + LTR + RC	1942037	6	266.63	5.96	0.020	none
DNA + H + RC	1956878	5	261.81	1.15	0.22	none
DNA + H	2361771	4	260.89	0.22	0.35	none
H	2861844	3	260.66	0.00	0.40	none
Analysis 2, no phylogenetic correction						
Class I + Class II + R+ U + LF	1828843	7	271.65	12.55	0.0013	none
Class I + Class II + R+ U	1832661	6	265.64	6.55	0.025	none
Class II + R + U	2055417	5	262.65	3.55	0.11	none
Class II + U	2455227	4	261.55	2.45	0.20	none
unknown	2609644	3	259.10	0.00	0.67	none

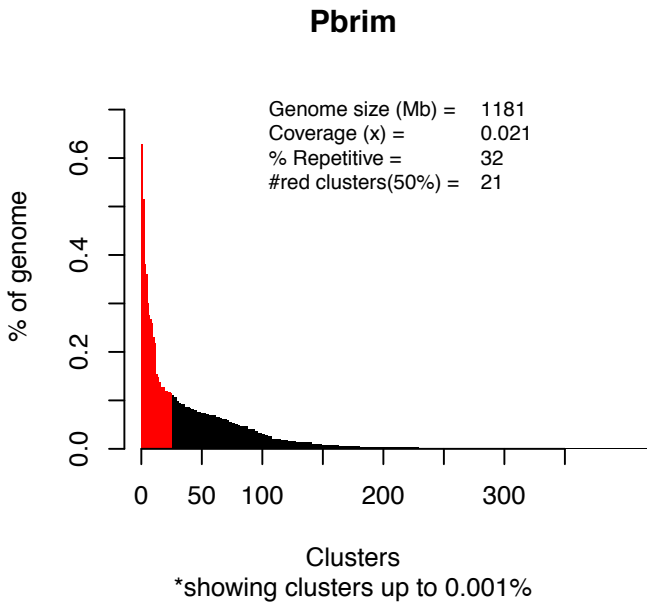
Figure S1. Distribution of top cluster abundance in each species prior to contaminant screening and manual curation

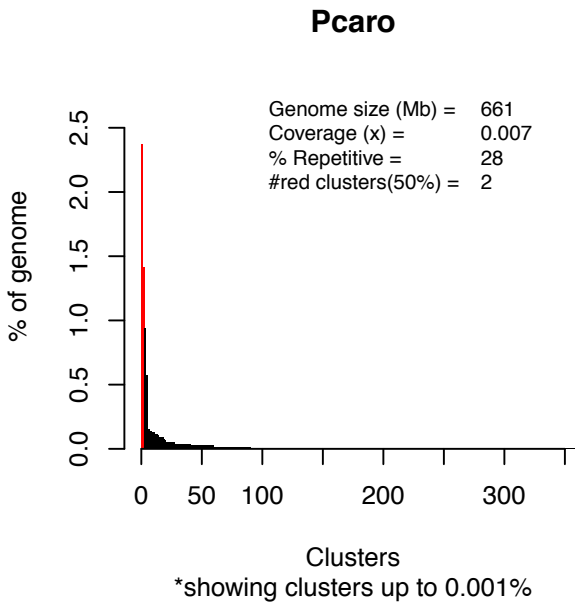
Species abbreviations are as in the main text (Table 2). Top clusters are a subset of all clusters that account for at least 0.001% of reads (20 reads) from the 454 dataset. Clusters are ordered by abundance in total sample (% of genome) and those that cumulatively sum to 50% of the repetitiveness of top clusters are shown in red. Average genome size, coverage, total percent repetitive of the sample (% repetitive), and the number of clusters in red are shown for each species. Where no red clusters are shown, the first cluster accounts for over 50% of the repetitiveness of top clusters.

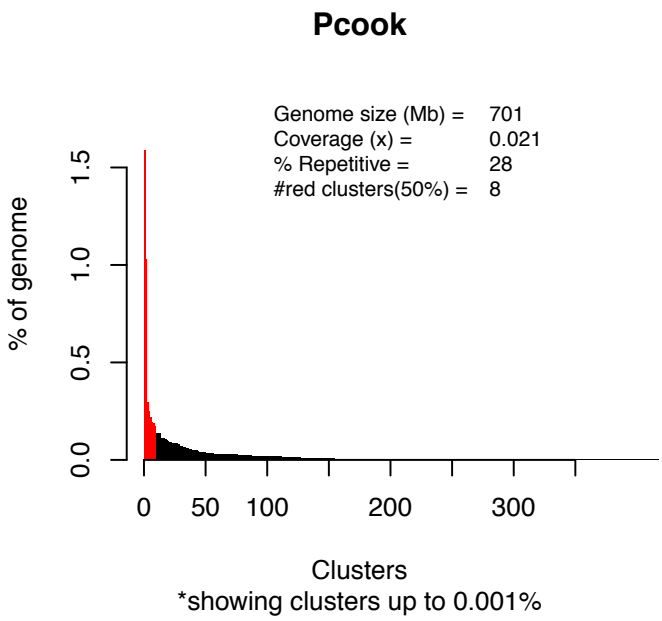


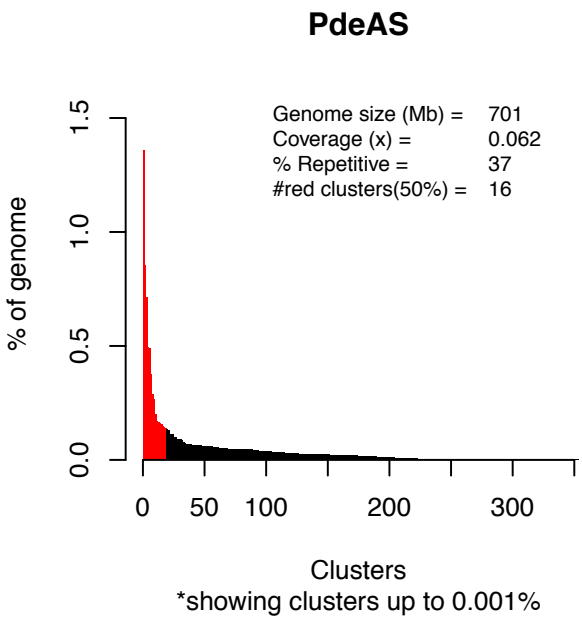

Paust
Genome size (Mb) = 1598
Coverage (x) = 0.016
% Repetitive = 40
#red clusters(50%) = 17
% of genome
0.0
0.2
0.4
0.6
0.8
0
50
100
200
300
Clusters
*showing clusters up to 0.001%

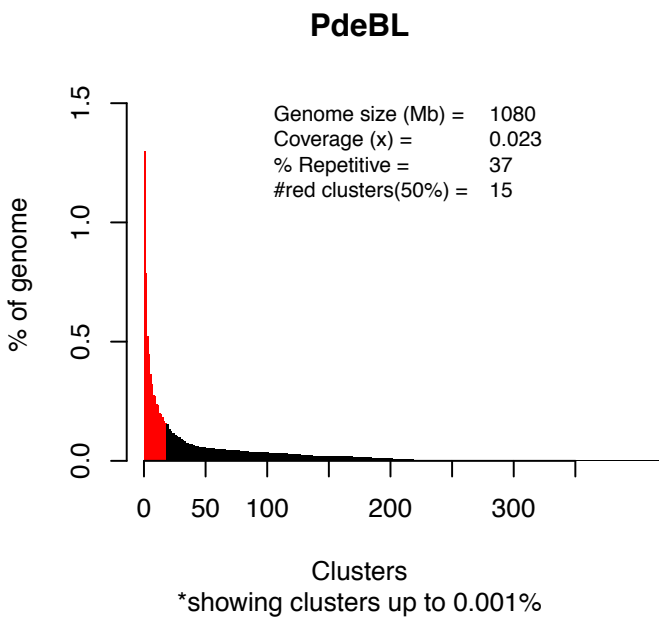
Pbigs
Genome size (Mb) = 2461
Coverage (x) = 0.004
% Repetitive = 58
#red clusters(50%) =
% of genome
0
5
10
15
20
25
30
35
0
50
100
200
300
Clusters
*showing clusters up to 0.001%

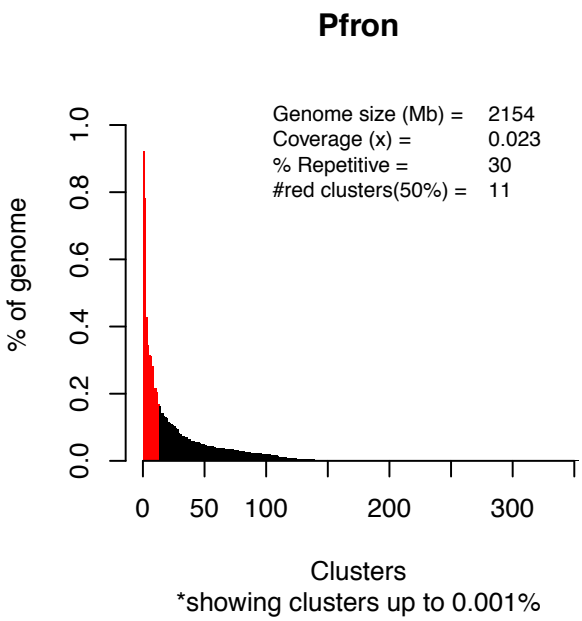

Pbrim
Genome size (Mb) = 1181
Coverage (x) = 0.021
% Repetitive = 32
#red clusters(50%) = 21
% of genome
0.0
0.2
0.4
0.6
0
50
100
200
300
Clusters
*showing clusters up to 0.001%

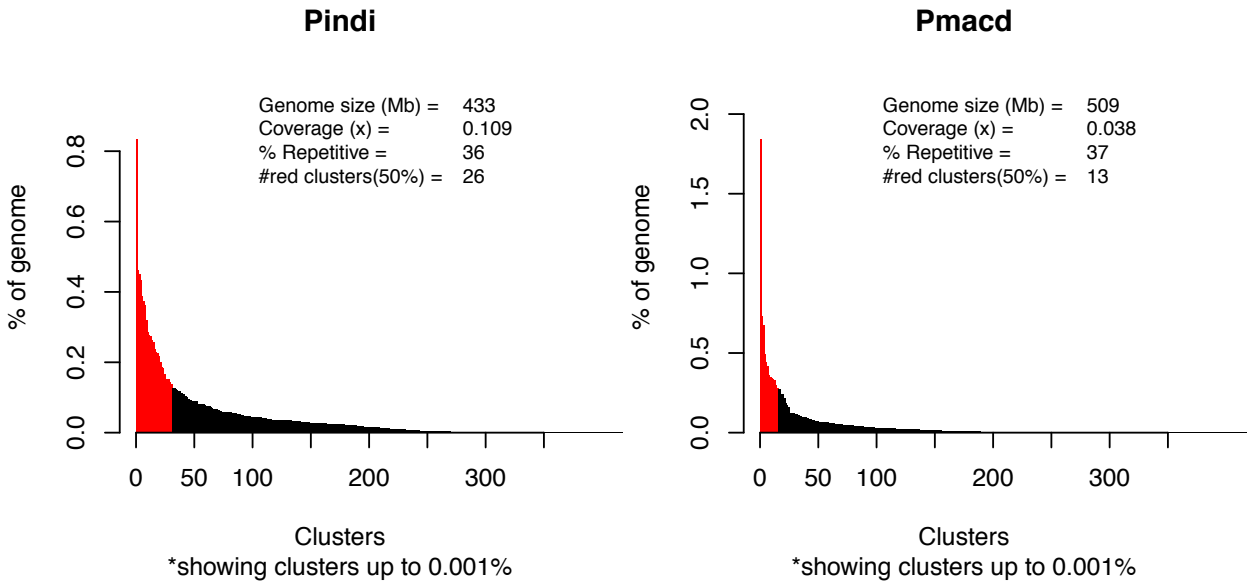

Pcaro
Genome size (Mb) = 661
Coverage (x) = 0.007
% Repetitive = 28
#red clusters(50%) = 2
% of genome
0.0
0.5
1.0
1.5
2.0
2.5
0
50
100
200
300
Clusters
*showing clusters up to 0.001%


Pcook
Genome size (Mb) = 701
Coverage (x) = 0.021
% Repetitive = 28
#red clusters(50%) = 8
% of genome
0.0
0.5
1.0
1.5
0
50
100
200
300
Clusters
*showing clusters up to 0.001%

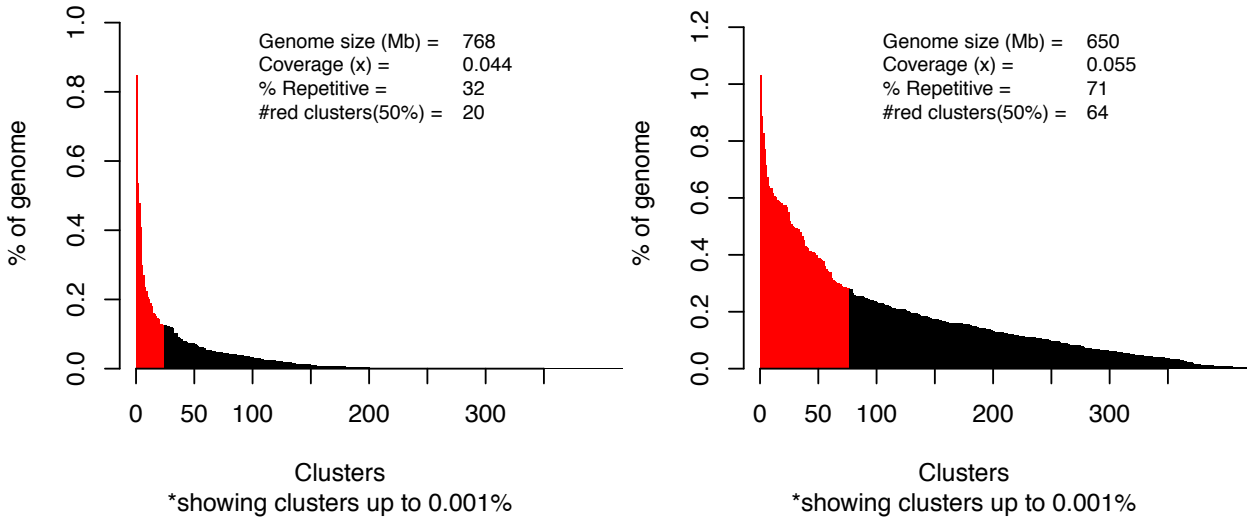

PdeAS
Genome size (Mb) = 701
Coverage (x) = 0.062
% Repetitive = 37
#red clusters(50%) = 16
% of genome
0.0
0.5
1.0
1.5
0
50
100
200
300
Clusters
*showing clusters up to 0.001%


PdeBL
Genome size (Mb) = 1080
Coverage (x) = 0.023
% Repetitive = 37
#red clusters(50%) = 15
% of genome
0.0
0.5
1.0
1.5
0
50
100
200
300
Clusters
*showing clusters up to 0.001%

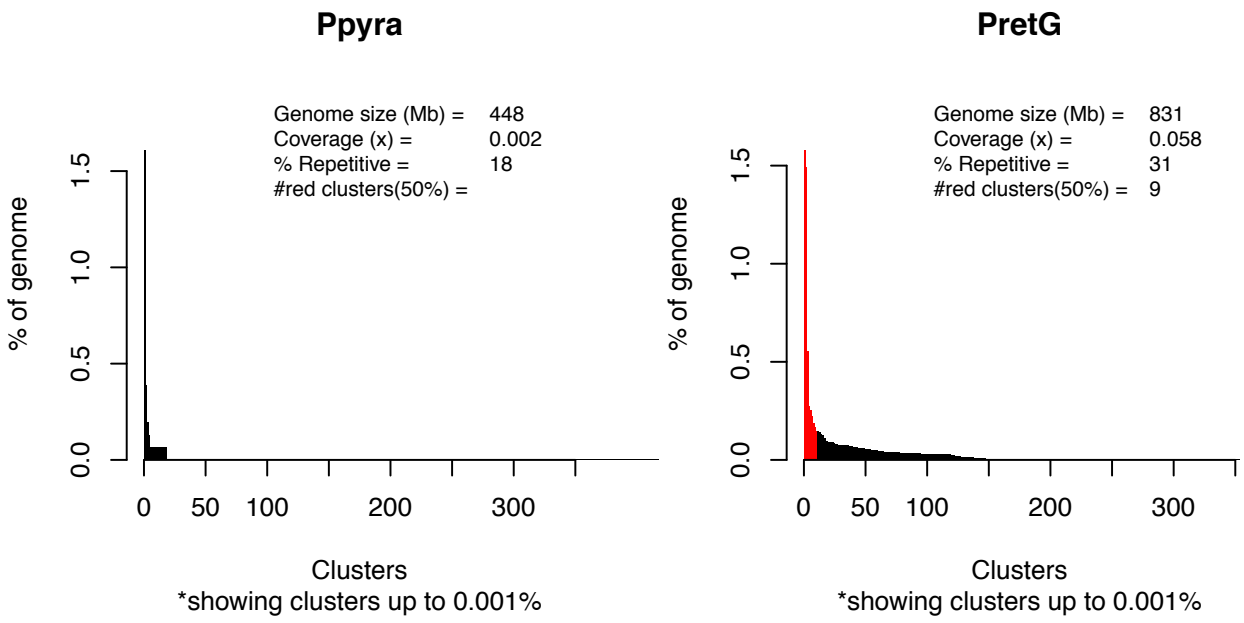

Pfron
Genome size (Mb) = 2154
Coverage (x) = 0.023
% Repetitive = 30
#red clusters(50%) = 11
% of genome
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
200
300
Clusters
*showing clusters up to 0.001%


Pindi
Genome size (Mb) = 433
Coverage (x) = 0.109
% Repetitive = 36
#red clusters(50%) = 26
% of genome
0.0
0.2
0.4
0.6
0.8
0
50
100
200
300
Clusters
*showing clusters up to 0.001%

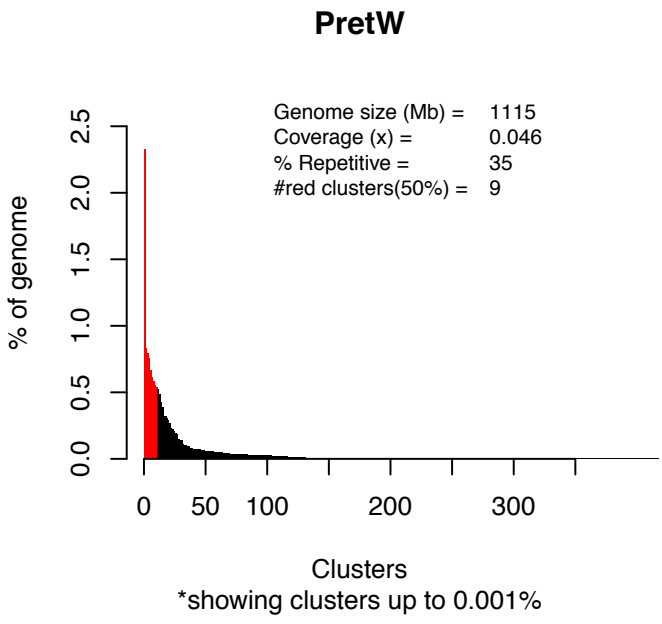
Pmacd
Genome size (Mb) = 509
Coverage (x) = 0.038
% Repetitive = 37
#red clusters(50%) = 13
% of genome
0.0
0.5
1.0
1.5
2.0
0
50
100
200
300
Clusters
*showing clusters up to 0.001%

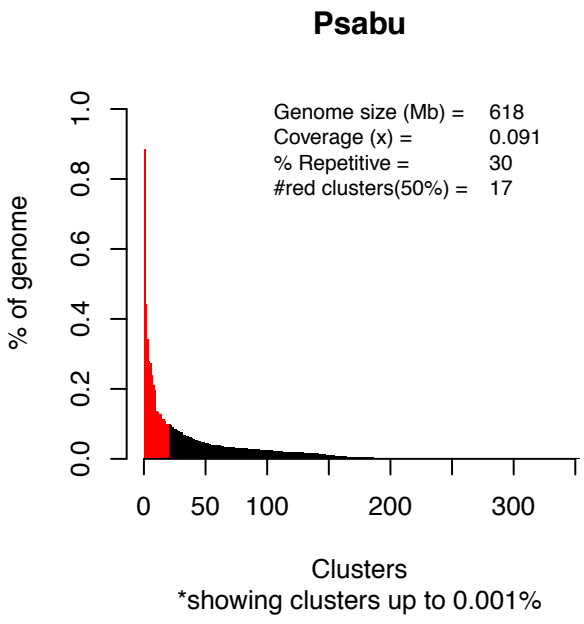

Pmarg
Genome size (Mb) = 768
Coverage (x) = 0.044
% Repetitive = 32
#red clusters(50%) = 20
% of genome
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
200
300
Clusters
*showing clusters up to 0.001%

Pobsc
Genome size (Mb) = 650
Coverage (x) = 0.055
% Repetitive = 71
#red clusters(50%) = 64
% of genome
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
50
100
200
300
Clusters
*showing clusters up to 0.001%


Ppyra
Genome size (Mb) = 448
Coverage (x) = 0.002
% Repetitive = 18
#red clusters(50%) =
% of genome
0.0
0.5
1.0
1.5
0
50
100
200
300
Clusters
*showing clusters up to 0.001%

PretG
Genome size (Mb) = 831
Coverage (x) = 0.058
% Repetitive = 31
#red clusters(50%) = 9
% of genome
0.0
0.5
1.0
1.5
0
50
100
200
300
Clusters
*showing clusters up to 0.001%


PretW
Genome size (Mb) = 1115
Coverage (x) = 0.046
% Repetitive = 35
#red clusters(50%) = 9
% of genome
0.0
0.5
1.0
1.5
2.0
2.5
0
50
100
200
300
Clusters
*showing clusters up to 0.001%


Psabu
Genome size (Mb) = 618
Coverage (x) = 0.091
% Repetitive = 30
#red clusters(50%) = 17
% of genome
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
200
300
Clusters
*showing clusters up to 0.001%

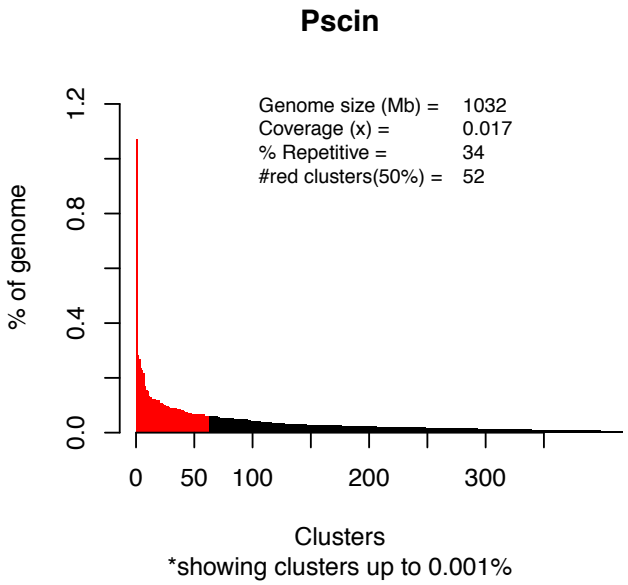

Pscin
Genome size (Mb) = 1032
Coverage (x) = 0.017
% Repetitive = 34
#red clusters(50%) = 52
% of genome
0.0
0.4
0.8
1.2
0
50
100
200
300
Clusters
*showing clusters up to 0.001%

Figure S2. Photuris genome sizes do not cluster by flash behavior or locality

Individual Photuris specimen genome size as estimated by flow cytometry as shown, ordered by locality, and colored by the type of flash behavior observed in the field. Filled circles indicate males, while oen circles indicate females. Locality abbreviations: Allegheny National Forest, PA (ANPA); Athens, GA (ATGA); Bloomington, IN (BLIN); Byhalia, MS (BYMS); Charleston, IL (CHIL); Eureka, MO (EUMO); Jackson County, TN (JCTN); St. Louis, MO (SLMO).

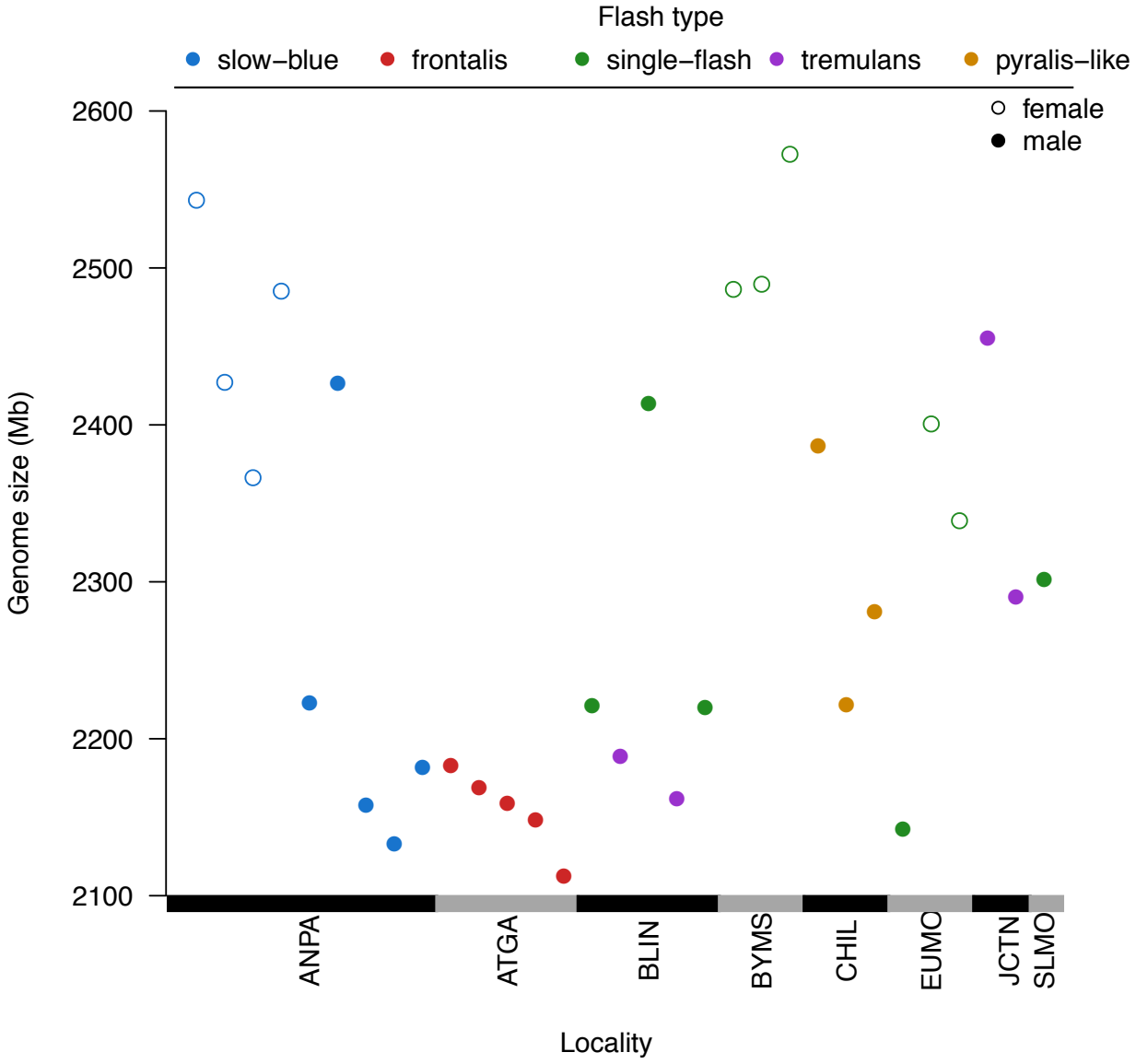


Figure S3. A three-gene phylogeny shows that *Phausis* specimens are two distinct species

This Bayesian phylogeny, based on mitochondrial *COI*, and nuclear *rudimentary* and *wingless* loci (Stanger-Hall and Lloyd 2015) shows that lighted Phausis from Watkinsville, GA (WAGA), cluster with unlighted *Phausis inaccensa* from Tennessee and are reciprocally monophyletic with lighted *Phausis reticulata* from both Athens, GA (ATGA) and Great Smoky Mountains National Park (GSMNP). The majority rule consensus phylogeny is shown. Nodes with support less than 50% were collapsed into polytomies. Otherwise, node values are not shown because all had 100% Bayesian support. Scale is in units of substitutions per site. Numbers in front of species names are the unique KSH numbers given to specimens in the Stanger-Hall collection at the University of Georgia. Numbers following species names represent the year of collection.

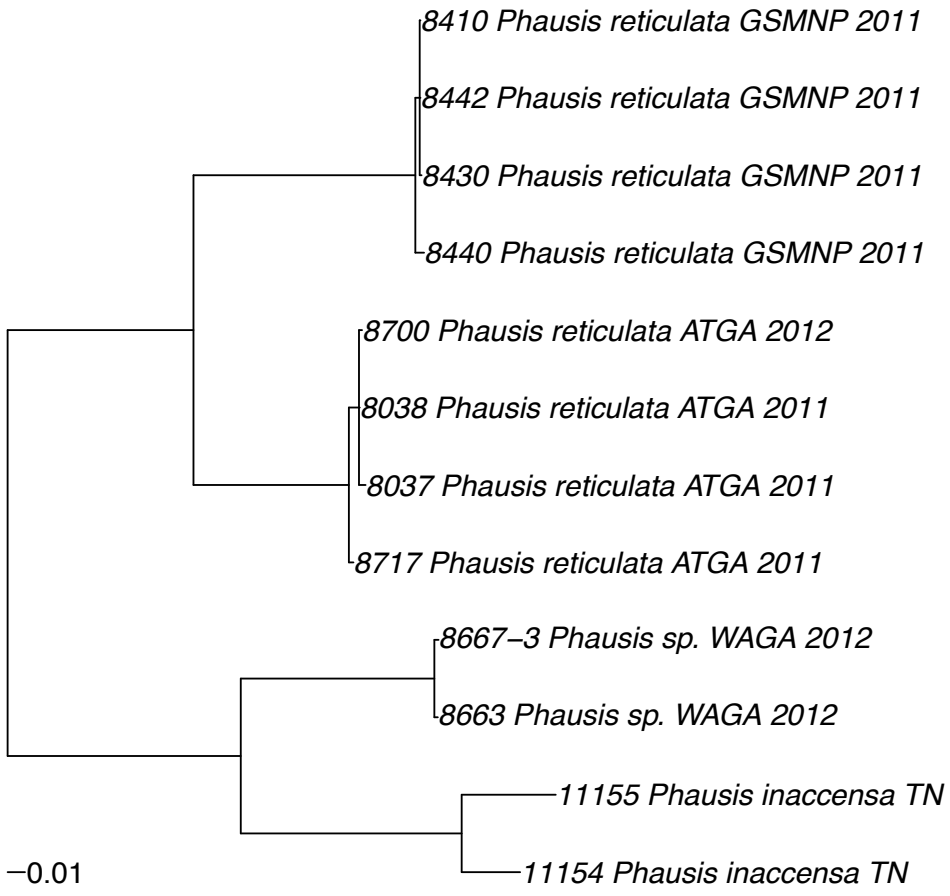


Figure S4. The three-gene phylogeny shows that *Pyropyga* specimens are a single species, despite genome size differences

This Bayesian phylogeny, based on mitochondrial *COI*, and nuclear *rudimentary* and *wingless* sequences (Stanger-Hall and Lloyd 2015), suggests that *Pyropyga* specimens make up a single species that is polymorphic for genome size. While the two "small" specimens (genome size ~ 700 Mb) group together with high support, they are not reciprocally monophyletic to the specimens with "large" genome sizes (~1100 Mb). Pyractomena specimens are shown as an outgroup and to give an idea of the scale of sequence variation between genera and species. Node values give Bayesian support values for the majority-rule consensus phylogeny. Nodes with support less than 50% were collapsed into polytomies. Scale is in substitutions per site. Numbers in front of species names are the unique KSH numbers given to specimens in the Stanger-Hall collection at the University of Georgia. The *Pyractomena borealis* sequences were downloaded from Genbank (COI: KP121568.1, wingless: KP121506.1, rudimentary: KP121630.1; Stanger-Hall and Lloyd 2015).

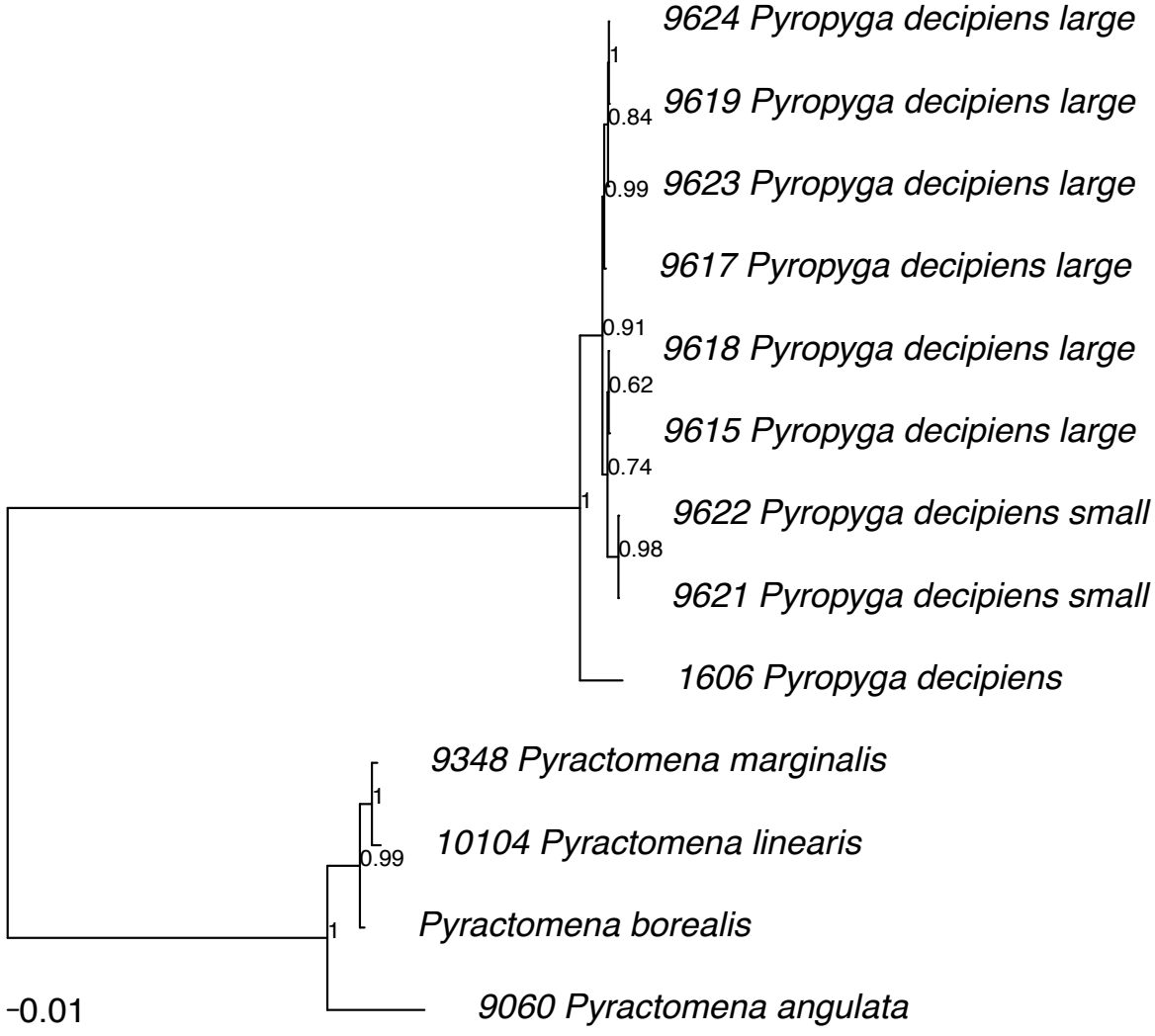


Figure S5. Ancestral state reconstruction of firefly genome size

Reconstruction of ancestral genome size (Mb) using phylogenetic independent contrasts (red to blue indicates large to small genome size, respectively). The phylogeny is the ultrametic BEAST phylogeny used in all other analyses with branch lengths called to units of relative time. Numbers at terminal nodes show extant taxon genome sizes. While expansions are noted in Photuris and the branches leading Photinus australis/brimleyi, there is generally a pattern of loss along the phylogeny. The shrinkage hypothesis is concordant with the large genome size estimate for *Phengodes fuscipes* (2,233 Mb; Hanrahan and Johnston 2011), the closest relative to fireflies with an estimated genome size.

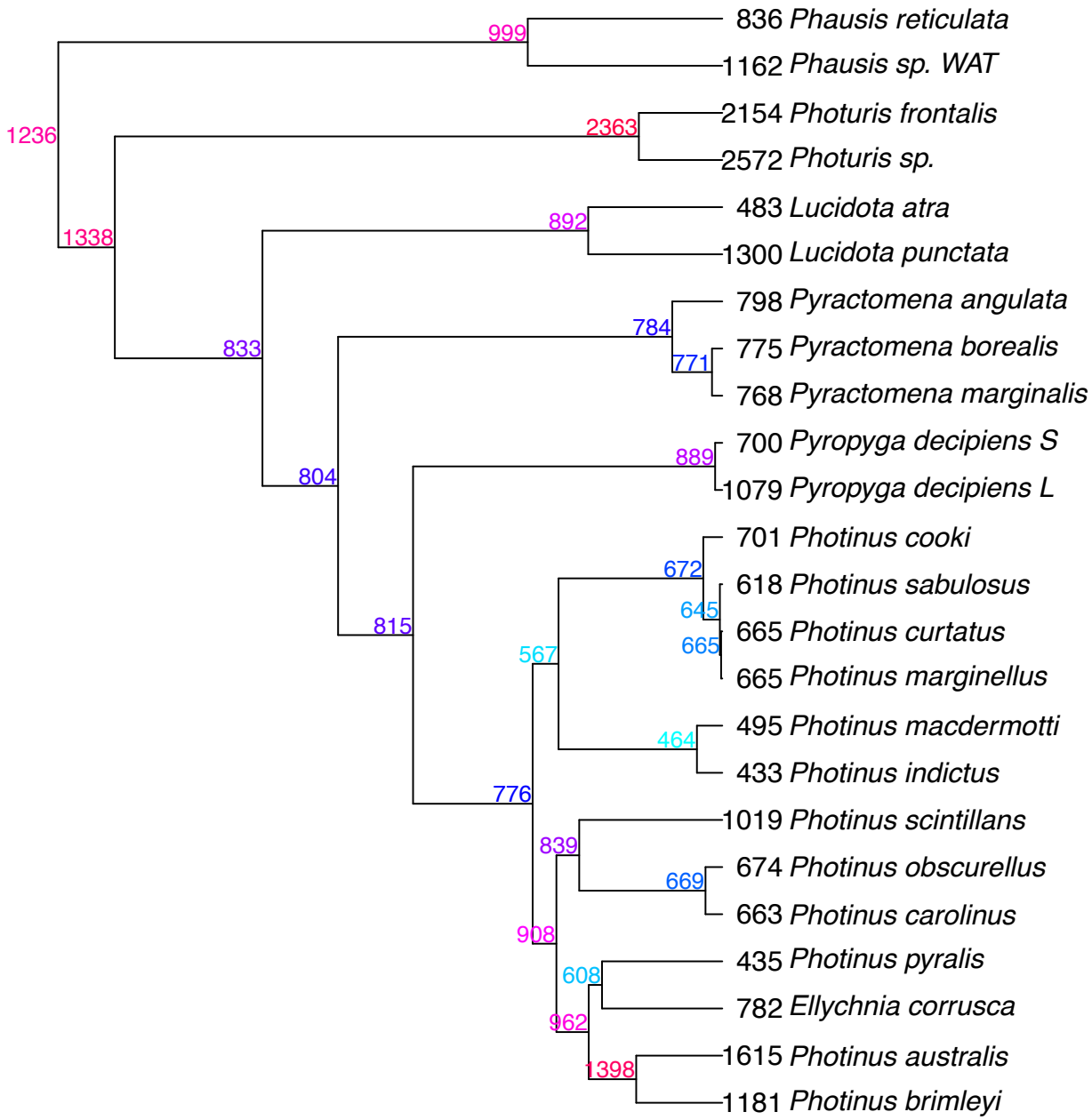


Figure S6. The relationship between shared clusters and divergence

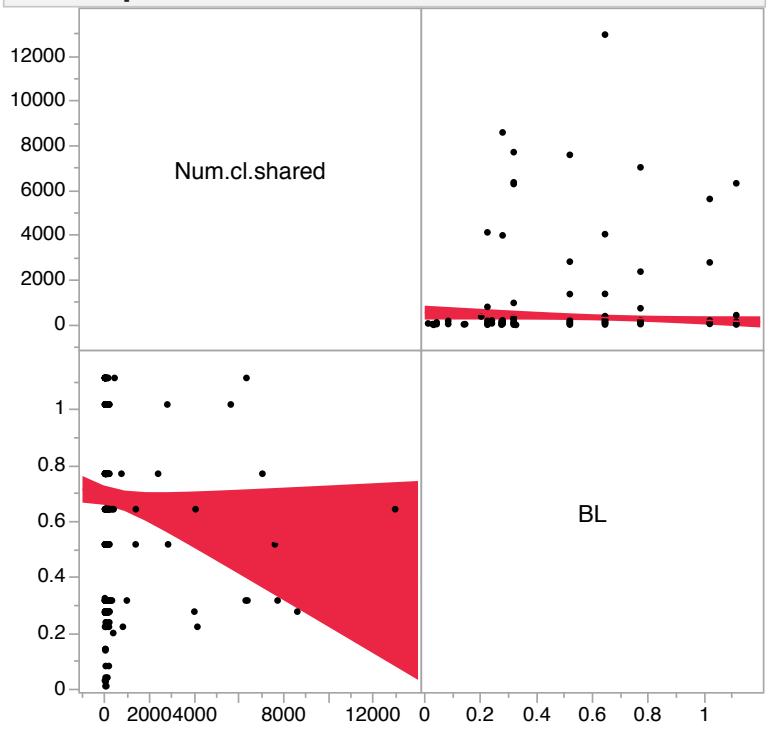
More closely related species share more clusters. The number of shared clusters and the total phylogenetic distance between taxa were calculated for all species pairs.

Multivariate

Correlations

	Num.cl.shared	BL
Num.cl.shared	1.0000	-0.0866
BL	-0.0866	1.0000

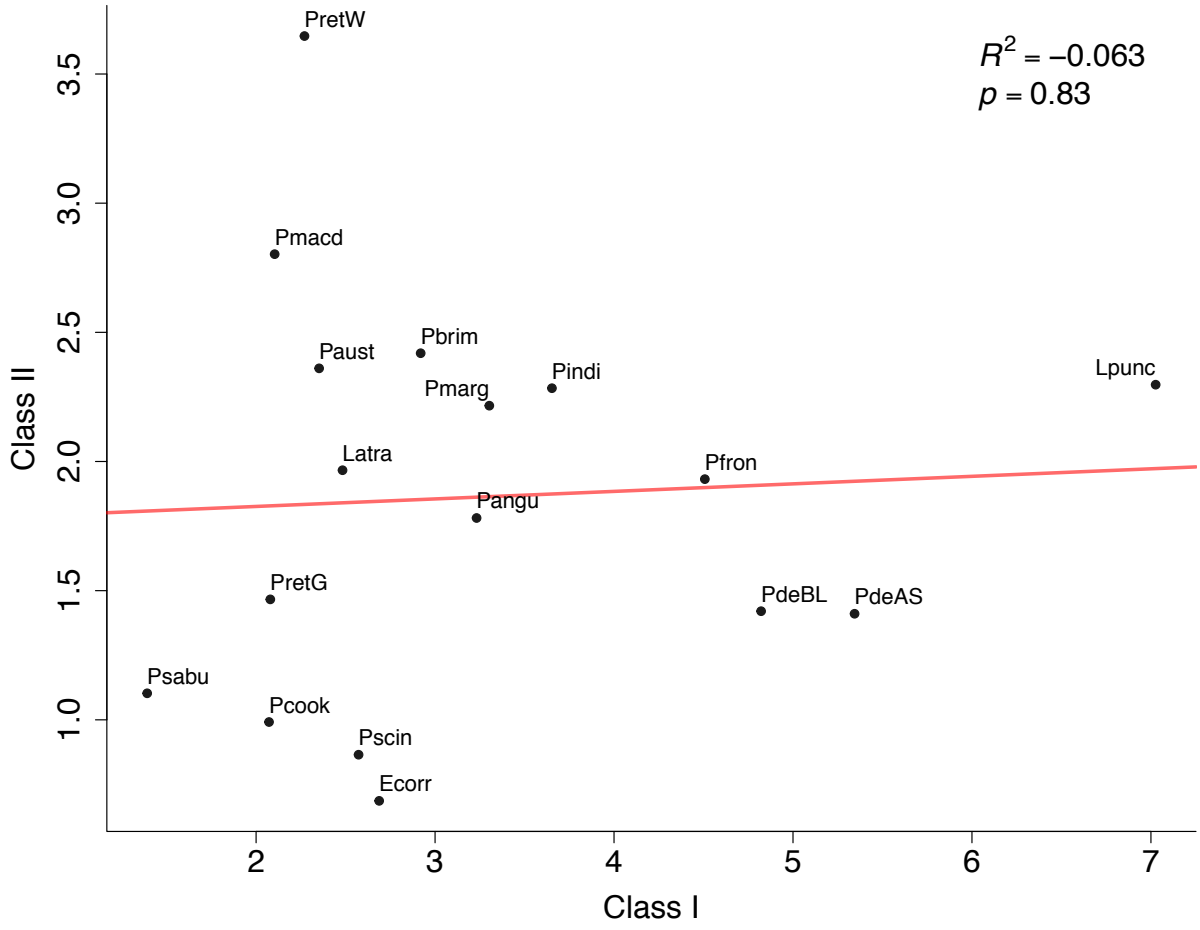
Scatterplot Matrix



Nonparametric: Spearman's ?

Variable	by Variable	Spearman ?	Prob>|?|	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8
BL	Num.cl.shared	-0.3405	<.0001*	

Figure S7. No significant relationship between Class I and Class II element abundance



Text S1. Partitioning the variance in genome size across taxonomic levels

Standard least squares methods were used in JMP in JMP Pro 10 (SAS Institute Inc. 2012). Full model: Genome size = Genus (random) + Species[Genus] (random) + Sex[Species, Genus]. Genus accounted for 72% of the variance (mostly due to Photuris), while species accounted for 28%. There was a significant effect of sex.

Summary of Fit

RSquare	0.998175
RSquare Adj	0.998034
Root Mean Square Error	21.05002
Mean of Response	810.6488
Observations (or Sum Wgts)	127

REML Variance Component Estimates

Random Effect	Var Ratio	Var Component	Std Error	95% Lower	95% Upper	Pct of Total
Genus	629.19485	278798.44	192632.71	-98754.72	656351.61	71.693
Species[Genus]	247.43384	109638.8	37710.818	35726.959	183550.65	28.193
Residual		443.10351	64.633217	339.41821	603.05665	0.114
Total		388880.35	192649.8	178454.03	1404953.7	100.000

-2 LogLikelihood = 1238.1430032

Note: Total is the sum of the positive variance components.

Total including negative estimates = 388880.35

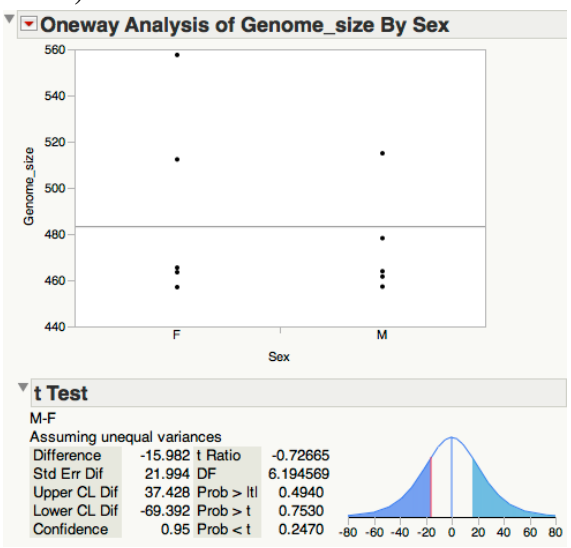
Fixed Effect Tests

Source	Nparm	DF	DFDen	F Ratio	Prob > F
Sex[Genus,Species]	9	9	94.02	5.9737	<.0001*

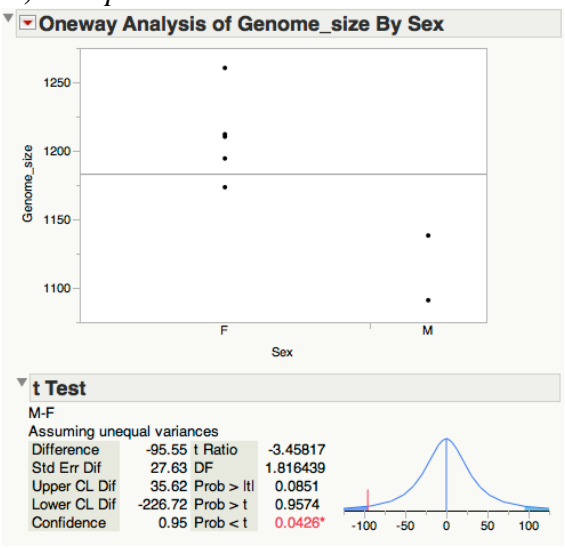
Text S2. Student's t results for sex differences in genome size

Sex differences in genome size were examined across species that had at least two estimates for each sex. The 30 Photuris specimens were not tested due to uncertainties in species identity. Two-tailed p-values were used in Benjamini-Hotchberg corrected analysis. A-G: the results for each species tested. H: the table of hypothesis testing results, including number of males and females and Benjamini-Hochberg corrected false discovery rates (FDR).

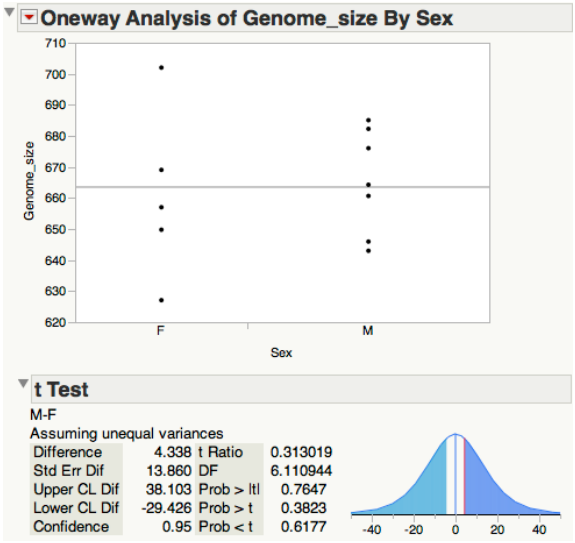
A) *Ld. atra*



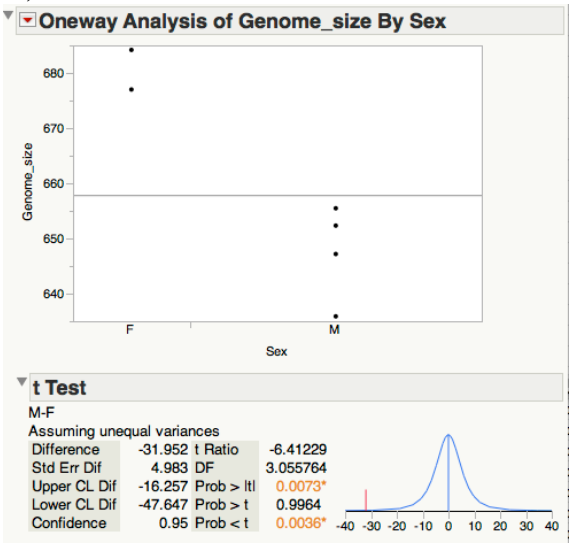
B) *Pa. sp. WAT*



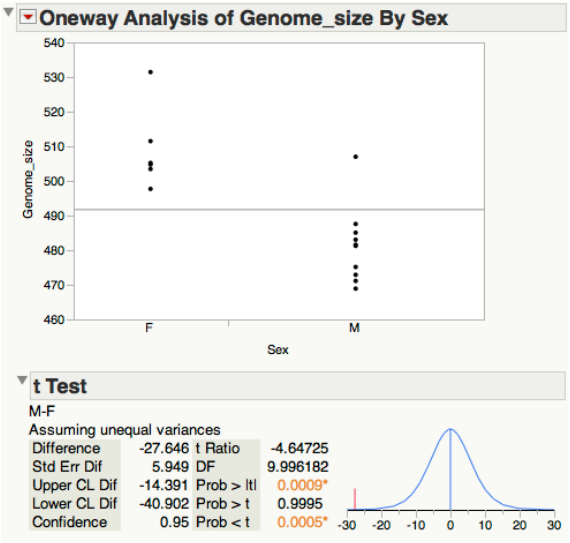
C) *Pn. carolinus*



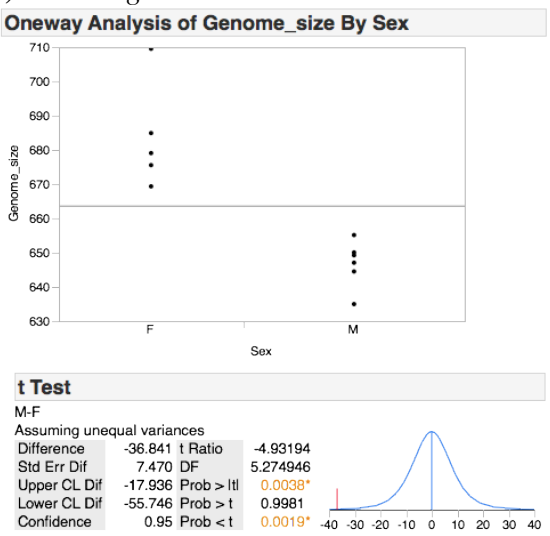
D) *Pn. curtatus*



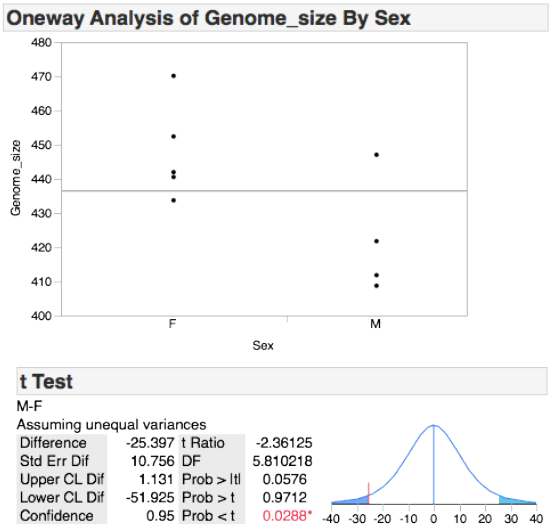
E) *Pn. macdermotti*



F) *Pn. marginellus*



G) *Pn. pyralis*



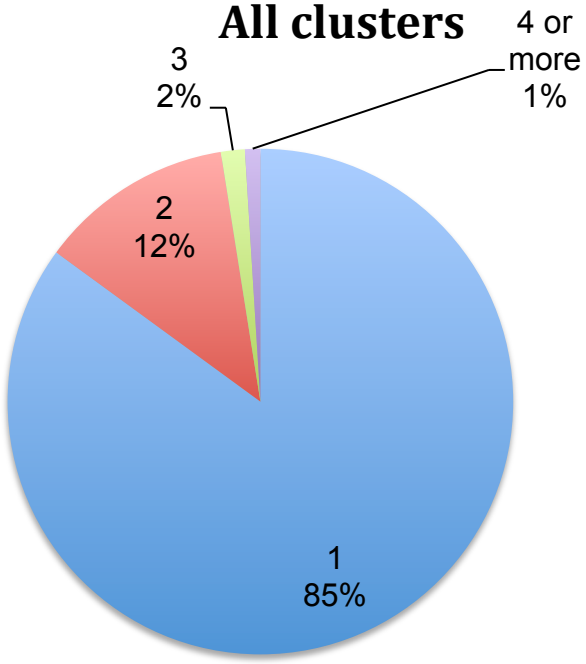
H) Student's t results after correcting for multiple comparisons

Species	Nmale	Nfemale	*p*	α	Comparison	B-H corrected FDR
Pn. macdermotti	10	5	0.0009*	0.05	7	0.007
Pn. marginellus	6	5	0.0038*	0.05	6	0.008
Pn. curtatus	5	2	0.0073*	0.05	5	0.010
Pn. pyralis	4	5	0.0576	0.05	4	0.013
Pa. sp. WAT	2	5	0.0851	0.05	3	0.017
Ld. atra	5	5	0.494	0.05	2	0.025
Pn. carolinus	7	5	0.7647	0.05	1	0.050

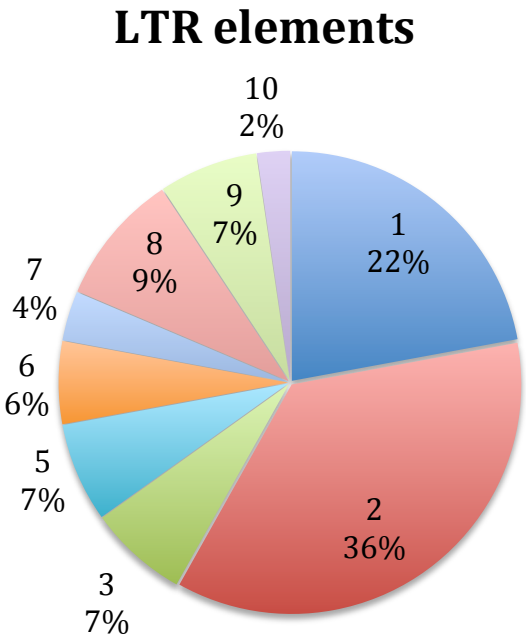
* Significant with B-H correction

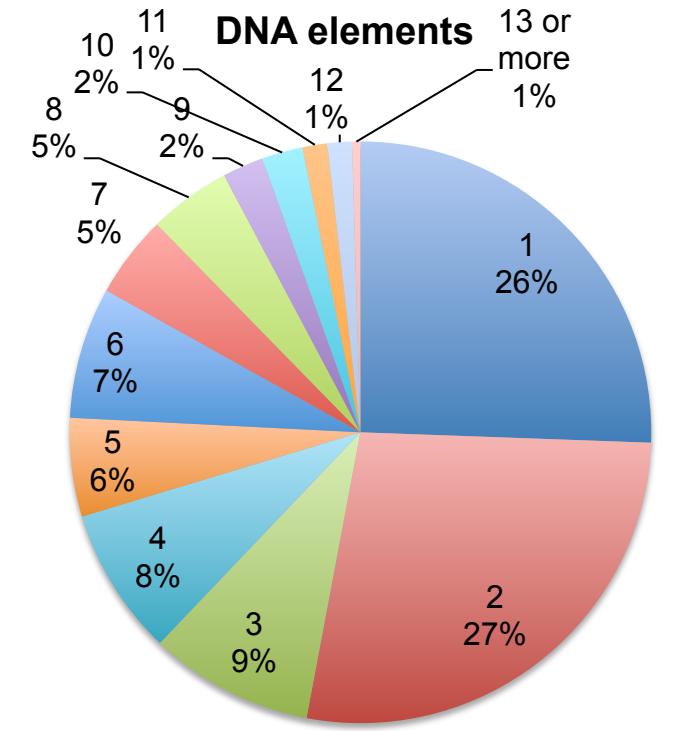
Test S3. Shared clusters

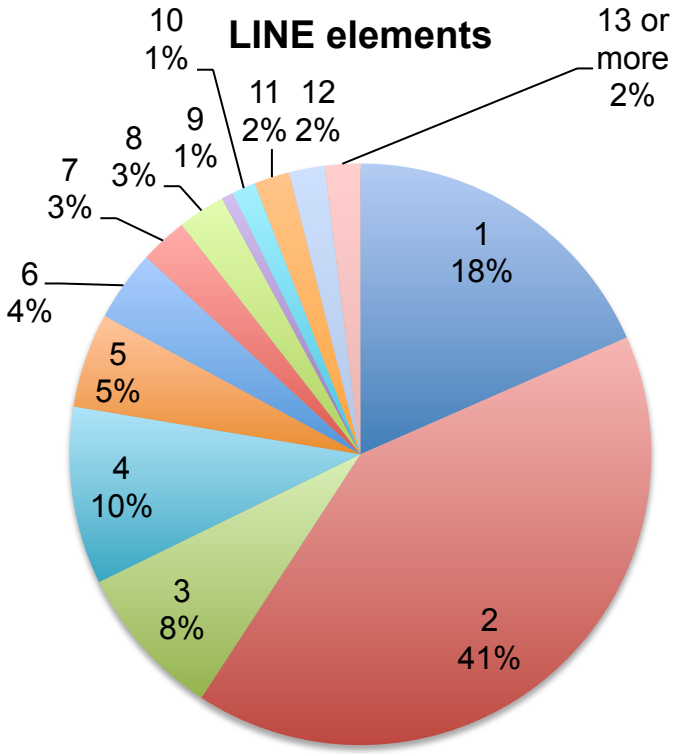
Pie charts show number of clusters that are unique (1) to, or shared (2-13 or more) among species. Percentage of total clusters shown below the number of species shared. Charts for all clusters and transposable element orders are shown. The table shows the top nine most-shared clusters and their classifications.



N	Cluster	Repeat
21	CL13	ribosomal
19	CL12	Simple_repeat
17	CL107	LINE
15	CL234	ribosomal
13	CL227	unknown
13	CL99	Simple_repeat
13	CL34	LINE
13	CL349	DNA
13	CL400	LINE




DNA elements
1
26%
2
27%
3
9%
4
8%
5
6%
6
7%
7
5%
8
5%
9
2%
10
2%
11
1%
12
1%
13 or
more
1%


LINE elements
1
18%
2
41%
3
8%
4
10%
5
5%
6
4%
7
3%
8
3%
9
1%
10
1%
11
2%
12
2%
13 or
more
2%

Text S4. Morphological differences between *Phausis* species investigated in this study

Genome size estimates for *Phausis* specimens from Watkinsville were almost two-fold larger than those for confirmed *Phausis reticulata* specimens. While the *Phausis* sp. WAT are morphologically similar to *Phausis inaccensa* in pronotum shape, the size and shape of the transparent parts of the pronotum, the form of the terminal segment, and that the eyes meet underneath reduced mouthparts, they conspicuously differ from *Phausis inaccensa* due to the presence of light organs (Fender 1966; Frick-Ruppert and Rosen 2008). The other sister taxon, *Phausis reticulata*, differs in pronotum, head, and terminal segment morphology, but has similar light organs. These morphological determinations were made using proxy specimens, since whole-bodies were used in the rest of the analysis. However, it is unlikely that proxy specimens were a different species than flow cytometery specimens because (1) mis-identification rate was low (only one mis-identified *Phausis inaccensa* male out of 55 total *Phausis* in the collection), (2) *Pa. inaccensa* males were not frequent in our sample from the night *Phausis sp.* WAT specimens were caught, and (3) most individuals were caught by tracking glowing males. The identification of a cryptic species in an investigation of genome size is not surprising—many investigations of genome size have discovered cryptic species (e.g. (Borkin, et al. 2001; Grishanin, et al. 2005; Panzera, et al. 2006).

APPENDIX C

Supplementary information for Sander *et al.*

To be submitted to *Molecular Ecology*.

Table S1. Specimens and Genbank information

Unique identifier (KSH), population (Code), Genbank accessions for Sanger-sequenced loci (LW: LW opsin, UV: UV opsin, LUC1: adult luciferase), and 3RAD status (x: used in 3RAD sequencing) for all 192 *Photinus pyralis* males included in this study. One individual from each population was excluded at random from 3RAD in order to include technical replicates. Details on the localities where specimens were collected and are in Table S1. Further details on 3RAD sequencing and quality control are given in Table S4.

KSH	Code	LW	UV	LUC1	3RAD
11313	AMNJ				x
11314	AMNJ				x
11315	AMNJ				x
11316	AMNJ				x
11317	AMNJ				
11318	AMNJ				X
11319	AMNJ				X
11320	AMNJ				X
11321	AMNJ				X
11322	AMNJ				X
11323	AMNJ				x
11324	AMNJ				x
11325	AMNJ				x
11326	AMNJ				x
11327	AMNJ				x
11328	AMNJ				x
9335	AMOH				x
9336	AMOH				x
9337	AMOH				x
9339	AMOH				x
9340	AMOH				x
9341	AMOH				x
9342	AMOH				x
9343	AMOH				x
9344	AMOH				x
9345	AMOH				x
9347	AMOH				x
9350	AMOH				x
9351	AMOH				x
9352	AMOH				
9353	AMOH				x
9354	AMOH				x
8121	ATGA				x
8125	ATGA				x

8164	ATGA	x
8165	ATGA	
8174	ATGA	x
8618	ATGA	x
8622	ATGA	x
8871	ATGA	x
8872	ATGA	x
8880	ATGA	x
8881	ATGA	x
8882	ATGA	x
8883	ATGA	x
11496	ATGA	x
11497	ATGA	x
11499	ATGA	x
8963	BYMS	
8964	BYMS	x
8965	BYMS	x
8966	BYMS	x
8967	BYMS	x
8969	BYMS	x
8970	BYMS	x
8971	BYMS	x
8972	BYMS	x
8973	BYMS	x
8974	BYMS	x
8976	BYMS	x
8977	BYMS	x
8978	BYMS	x
8979	BYMS	x
8981	BYMS	x
11451	DEMI	x
11452	DEMI	x
11454	DEMI	x
11455	DEMI	x
11456	DEMI	
11457	DEMI	x
11458	DEMI	x
11459	DEMI	x
11460	DEMI	x
11461	DEMI	x
11462	DEMI	x
11463	DEMI	x
11464	DEMI	x

11465	DEMI	x
11466	DEMI	x
11467	DEMI	x
11062	DETX	x
11063	DETX	x
11064	DETX	x
11065	DETX	
11066	DETX	x
11067	DETX	x
11068	DETX	x
11069	DETX	x
11070	DETX	x
11071	DETX	x
11072	DETX	x
11073	DETX	x
11075	DETX	x
11076	DETX	x
11077	DETX	x
11078	DETX	x
11533	HFTN	x
11534	HFTN	x
11536	HFTN	x
11537	HFTN	x
11538	HFTN	x
11539	HFTN	x
11541	HFTN	x
11542	HFTN	x
11544	HFTN	x
11546	HFTN	x
11547	HFTN	x
11548	HFTN	x
11549	HFTN	
11550	HFTN	x
11552	HFTN	x
11553	HFTN	x
10686	MANJ	x
10690	MANJ	x
10691	MANJ	x
10692	MANJ	
10694	MANJ	x
10695	MANJ	x
10696	MANJ	x
10697	MANJ	x

10698	MANJ				x
10700	MANJ				x
10701	MANJ				x
10702	MANJ				x
10703	MANJ				x
10704	MANJ				x
10705	MANJ				x
10706	MANJ				x
11169	SANC				x
11170	SANC				x
11171	SANC				x
11172	SANC				
11173	SANC				x
11174	SANC				x
11175	SANC				x
11176	SANC				x
11177	SANC				x
11179	SANC				x
11180	SANC				x
11181	SANC				x
11182	SANC				x
11183	SANC				x
11185	SANC				x
11187	SANC				x
9036	SLMO				x
9037	SLMO				x
9038	SLMO				x
9039	SLMO				x
9040	SLMO				x
9041	SLMO				x
9042	SLMO				x
9043	SLMO				x
9044	SLMO				x
9045	SLMO				
9046	SLMO				x
9048	SLMO	N/A	N/A	N/A	
9149	SLMO				x
9150	SLMO				x
9151	SLMO				x
9152	SLMO				x
11019	VATX				x
11020	VATX				x
11021	VATX				x

11022	VATX		x
11023	VATX		x
11024	VATX		
11030	VATX		x
11035	VATX		x
11038	VATX		x
11039	VATX		x
11040	VATX		x
11041	VATX		x
11042	VATX		x
11043	VATX	N/A*	x
11044	VATX		x
11045	VATX		x
11080	WYAR		x
11081	WYAR		x
11084	WYAR		x
11089	WYAR		x
11090	WYAR		
11091	WYAR		x
11092	WYAR		x
11093	WYAR		x
11094	WYAR		x
11095	WYAR		x
11096	WYAR		x
11097	WYAR		x
11098	WYAR		x
11099	WYAR		x
11100	WYAR		x
11101	WYAR		x

* Excluded from final analysis because of difficulty in PCR amplification and sequencing.

Table S2. Primers and PCR cycling conditions.

Long-wavelength (LW) and ultraviolet (UV) opsin, and adult luciferase (LUC1) were amplified from genomic DNA using touchdown PCR with external primers. The names of forward (F) and reverse (R) primers, their nucleotide sequences, and where they were first reported (source) are given.

A. External primers

Full-length loci were amplified from genomic DNA using external primers designed from flanking sequences. PCR conditions using Qiagen taq polymerase were: initial denaturation at 94°C for 3 minutes, then 3-step cycles of 94°C for 45 seconds, the appropriate annealing temperature for 1 minute, and 72°C for the appropriate extension time (Ext) given the length of the locus. There was a final extension at 72°C for 3 minutes. A touchdown protocol was used to increase specificity of the primers: annealing began at the initial annealing temperature (Ann) and was decreased by 1°C each cycle for the first 7 cycles, then maintained at 7°C below the Ann for the remaining 28 cycles (total=35 cycles).

Locus	Primer	Sequence	Source	Ann (°C)	Ext (min:sec)
LW	LoutF	CATGGTGGTCGTGTTAATG	This study	66	2:00
	LoutR	TAGCCTGCAAGGTTATATTTAG	This study		
UV	UVP_-113F	see source	Sander and Hall 2015	56	2:00
	Uphiz_R	see source	Sander and Hall 2015		
LUC1	pyrluc_beginF	GGAATTCCTTTGTGTTACATTCT	This study	62	2:30
	pluc_endR	AAAATTACCATTCATCAATTTGC	This study		

B. Internal primers

Both external and internal primers were used to sequence all loci bidirectionally at the Georgia Genomics Facility (Athens, GA).

Locus	Primer	Sequence	Source
LW	LWT2_329F	see source	Sander and Hall 2015
	LWT2_568R	see source	Sander and Hall 2015
	LWP_807F	see source	Sander and Hall 2015
	LWT2_1004R	see source	Sander and Hall 2015
UV	UVphmidH_F	see source	Sander and Hall 2015
	UVP_2352R	see source	Sander and Hall 2015
	UVP_997F	see source	Sander and Hall 2015

	UV_1134R	see source	Sander and Hall 2015
LUC1	pluc_m2F	GCGTTATTTATCGGAGTTGC	This study
	pluc_m1R	TAGGCTGCGAAATGTTCATA	This study
	pluc_m4F	TATGTAAACAATCCGGAAGC	This study
	pluc_m3R	TTCGTCCCAGTAAGCTATGTC	This study
	pluc_m3F	TATGTGGATTTCGAGTCGTC	This study
	pluc_m2R	AGGGATCGTAAAAACAGCTC	This study

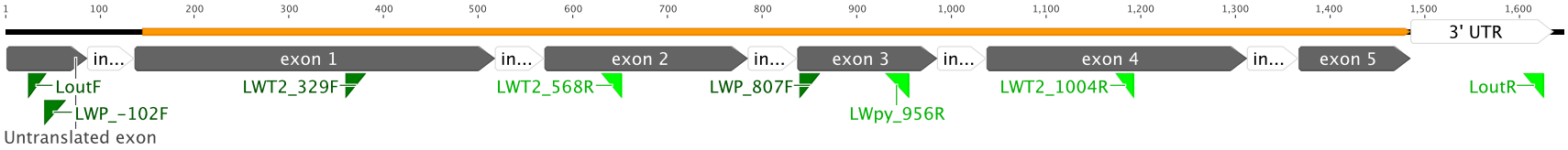
C. Secondary primers

Secondary primers were used to amplify specific exons that were difficult to sequence due to insertions/deletions in flanking introns. Most specimens did not require the use of secondary primers.

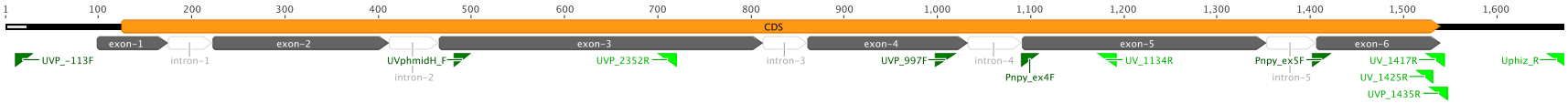
Locus	Primer	Sequence	Source
LW	LWpy_956R	TACAAATAGTGGTAAAAAGTACACG	This study
	LWP_-102F	TGTGAAGGTACATTCACTTGCAG	This study
UV	Pnpy_ex5F	TCAGAATGGAACTCCAAAAG	This study
	Pnpy_ex4F	GGCCAAAAAGATGAATGTAG	This study
	UV_1425R	see source	Sander and Hall 2015
	UVP_1435R	see source	Sander and Hall 2015
	UV_1417R	GTCGCAGCCGGTTCGGTCG	This study
LUC1	749r_Photinus	CCAAAACCGTGATGGAATGGAAC	This study

D. Primer placement

1. LW opsin



2. UV opsin



3. LUC1

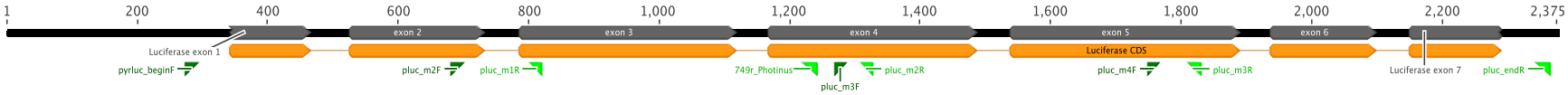


Table S3. Specimens used in COI tree

Unique number in KSH collection (KSH), taxon (Species), 4-letter locality code (Location), year of capture, and Genbank accession for each specimen used to generate the COI phylogeny. Genus abbreviations: Pn.= *Photinus*, Pt. = *Photuris*, Pg. = *Pyropyga*. Locations: Amwell, NJ (AMNJ); Amesville, OH (AMOH); Ashdown, AR (ASAR); Athens, GA (ATGA); Athens, Ohio (ATOH); Bucks County, PA (BCPA); Belcamp/Harford, MD (BHMD; Bayles Road, Bloomington, IN (BRIN); Byhalia, MS (BYMS); Caldwell City, TX (CCTX); Chicago, IL (CGIL); Charleston, IL (CHIL); Charlottesville, VA (CHVA); Cranbury/Middlesex, NJ (CMNJ); Cookeville, TN (COTN); Denison, TX (DETX); Douglas/Lawrence, KS (DLKS); Doylestown, PA (DOPA); Easton, PA (EAPA); Eureka, MO (EUMO); Gonzales, TX (GOTX); Guadelupe River, TX (GRTX); Great Smoky Mountains National Park (GSMNP); Harrison, OH (HAOH); Hickory Flats Branch, TN (HFTN); Jackson County, TN (JCTN); Kutztown, PA (KUPA); Mahwah, NJ (MANJ); Montgomery Bell State Park, TN (MBTN); Montua, OH (MOOH); Moody, TX (MOTX); Nashville, TN (NATN); Ontelaunee Lake, PA (ONPA); Renfrew, PA (REPA); Salisbury, NC (SANC); State College, PA (SCPA); St. Louis, MO (SLMO); Vanderpool, TX (VATX); Whittington, IL (WHIL); Washington, D.C. (WSDC); Winston-Salem, NC (WSNC); Wynne, AR (WYAR). Outgroups (OG**)** are indicated in bold. 2 letter location abbreviations are state abbreviations.

KSH	Species	Location	Year	Genbank
11273	*Pn. pyralis*	AMNJ	2013	
9330	*Pn. pyralis*	AMOH	2012	
9338	*Pn. pyralis*	AMOH	2012	
9349	*Pn. pyralis*	AMOH	2012	
11079	*Pn. pyralis*	ASAR	2013	
8025	*Pn. pyralis*	ATGA	2010	
8029	*Pn. pyralis*	ATGA	2010	
8175	*Pn. pyralis*	ATGA	2011	
8176	*Pn. pyralis*	ATGA	2011	
8819	*Pn. pyralis*	ATGA	2011	
8844	*Pn. pyralis*	ATGA	2011	
8846	*Pn. pyralis*	ATGA	2012	
8871	*Pn. pyralis*	ATGA	2012	
8872	*Pn. pyralis*	ATGA	2012	
8873	*Pn. pyralis*	ATGA	2012	
8880	*Pn. pyralis*	ATGA	2012	
8881	*Pn. pyralis*	ATGA	2012	
8882	*Pn. pyralis*	ATGA	2012	
8883	*Pn. pyralis*	ATGA	2012	
8892	*Pn. pyralis*	ATGA		
9311	*Pn. pyralis*	ATOH	2012	
9317	*Pn. pyralis*	ATOH	2012	
9318	*Pn. pyralis*	ATOH	2012	
11391	*Pn. pyralis*	BCPA	2013	

889	*Pn. pyralis*	BHMD	2001	KP121581.1[1]
9240	*Pn. pyralis*	BRIN	2012	
9028	*Pn. pyralis*	BYMS	2012	
975	*Pn. pyralis*	CCTX	2001	KP121582.1[1]
45	*Pn. pyralis*	CGIL	2001	
9160	*Pn. pyralis*	CHIL	2012	
10602	*Pn. pyralis*	CHVA	2012	
8470	*Pn. pyralis*	CHVA	2011	
8478	*Pn. pyralis*	CHVA	2011	
339	*Pn. pyralis*	CMNJ	2001	
8022	*Pn. pyralis*	COTN	2010	
11053	*Pn. pyralis*	DETX	2013	
289	*Pn. pyralis*	DLKS		
291	*Pn. pyralis*	DLKS		
292	*Pn. pyralis*	DLKS		
10802	*Pn. pyralis*	DOPA	2012	
10803	*Pn. pyralis*	DOPA	2012	
10804	*Pn. pyralis*	DOPA	2012	
8504	*Pn. pyralis*	DOPA	2011	
8509	*Pn. pyralis*	DOPA	2011	
8527	*Pn. pyralis*	DOPA	2011	
10825	*Pn. pyralis*	EAPA	2012	
9059	*Pn. pyralis*	EUMO	2012	
9077	*Pn. pyralis*	EUMO	2012	
10983	*Pn. pyralis*	GOTX	2013	
157	*Pn. pyralis*	GRTX	1998	
8365	*Pn. pyralis*	GSMNP	2011	
8366	*Pn. pyralis*	GSMNP	2011	
8369	*Pn. pyralis*	GSMNP	2011	KP121616.1[1]
9374	*Pn. pyralis*	HAOH	2012	
9384	*Pn. pyralis*	HAOH	2012	
9372	*Pn. pyralis*	HFTN	2012	
124	*Pn. pyralis*	IN	1999	
125	*Pn. pyralis*	IN	1999	
9020	*Pn. pyralis*	JCTN	2012	
9021	*Pn. pyralis*	JCTN	2012	
65	*Pn. pyralis*	KS		
9839	*Pn. pyralis*	KUPA	2012	
9852	*Pn. pyralis*	KUPA	2012	
10686	*Pn. pyralis*	MANJ	2012	
10690	*Pn. pyralis*	MANJ	2012	
10691	*Pn. pyralis*	MANJ	2012	
9025	*Pn. pyralis*	MBTN	2012	

17A	*Pn. pyralis*	MD		KP121624.1[1]
311	*Pn. pyralis*	MOOH	2001	
312	*Pn. pyralis*	MOOH	2001	
11046	*Pn. pyralis*	MOTX	2013	
11047	*Pn. pyralis*	MOTX	2013	
8852	*Pn. pyralis*	NATN	2012	
8857	*Pn. pyralis*	NATN	2012	
8862	*Pn. pyralis*	NATN	2012	
10045	*Pn. pyralis*	ONPA	2012	
9486	*Pn. pyralis*	REPA	2012	
9487	*Pn. pyralis*	REPA	2012	
11158	*Pn. pyralis*	SANC	2013	
63	*Pn. pyralis*	SCPA	2000	
102	*Pn. pyralis*	SLMO	1998	
9148	*Pn. pyralis*	SLMO	2012	
11021	*Pn. pyralis*	VATX	2013	
9109	*Pn. pyralis*	WHIL	2012	
72	*Pn. pyralis*	WI		
73	*Pn. pyralis*	WI		
8018	*Pn. pyralis*	WSDC	2010	
8019	*Pn. pyralis*	WSDC	2010	
322	*Pn. pyralis*	WSNC	2001	
323	*Pn. pyralis*	WSNC	2001	
324	*Pn. pyralis*	WSNC	2001	
325	*Pn. pyralis*	WSNC	2001	
11080	*Pn. pyralis*	WYAR	2013	
11526	*Pn. pyralis*			
11527	*Pn. pyralis*			
11528	*Pn. pyralis*			
11529	*Pn. pyralis*			
218	*Pn. pyralis*			
	Pn. pyralis			EU009313.1[2]
226	*Pn. concisus*	**OG**		KP121573.1[1]
9655	*Pn. carolinus*	**OG**		[3]
9660	*Pn. carolinus*	**OG**		[3]
9670	*Pn. carolinus*	**OG**		[3]
	Pt. quadrifulgens	**OG**		EU009310_1[2]
	Pg. decipiens	**OG**		KP121569.1[1]

[1] Stanger-Hall and Lloyd 2015
[2] Stanger-Hall et al. 2007
[3] Faust et al. 2012

Figure S1. Expanded COI phylogeny

Phylogeny with individual specimen names indicated. Values at nodes are Bayesian support values. Pt = *Photuris*, Pg = *Pyropyga*, Py = *Photinus pyralis*, cn= *Photinus consicus*, ca = *Photinus carolinus*.

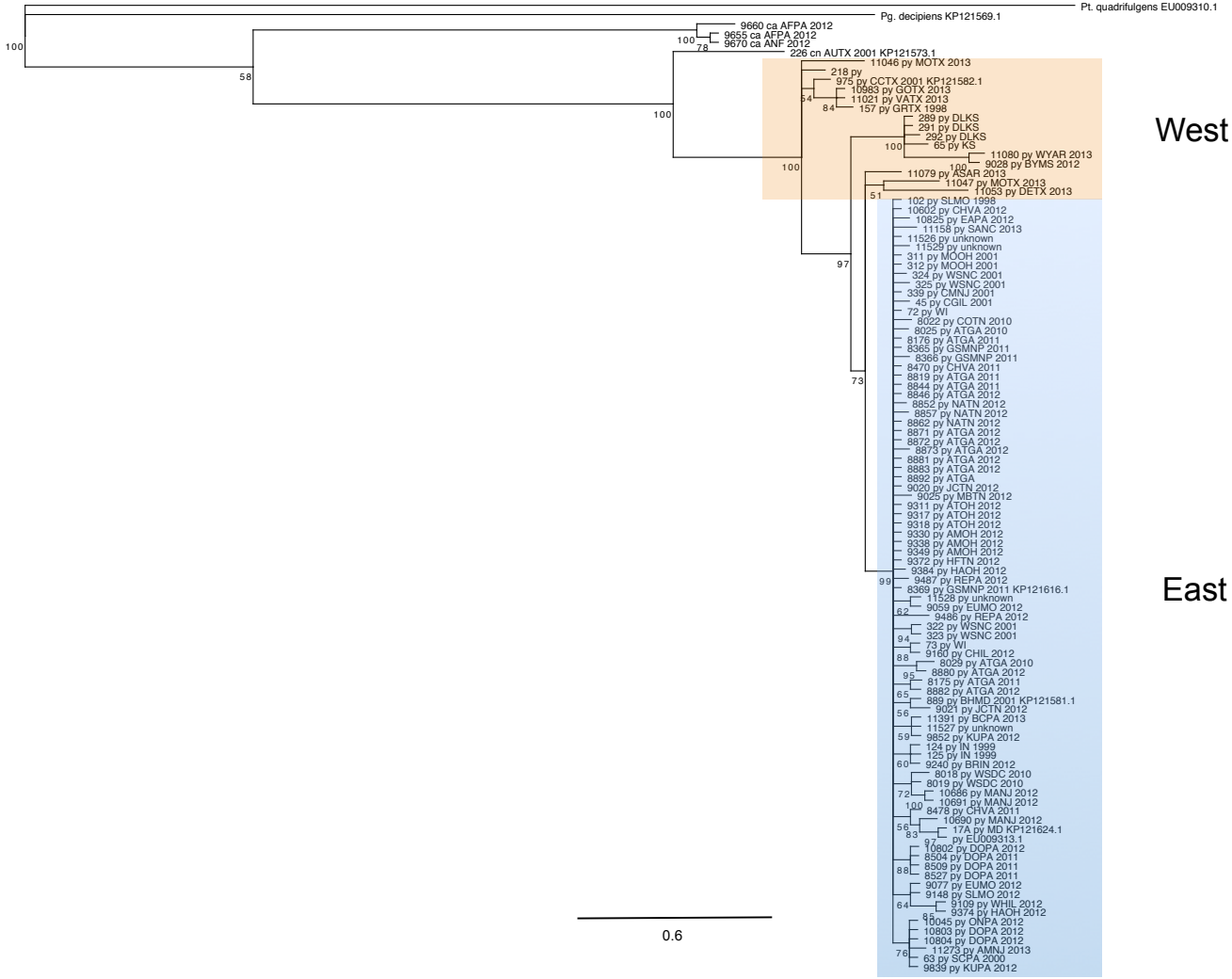
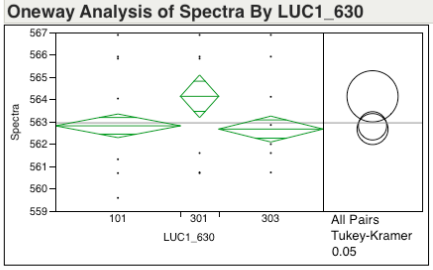


Figure S2. Results of ANOVA for selected LUC1 SNPs

A) Site 630

Fit Y by X Group

Oneway Analysis of Spectra By LUC1_630



Oneway Anova

Summary of Fit

Rsquare	0.045326
Adj Rsquare	0.032681
Root Mean Square Error	2.272086
Mean of Response	562.9519
Observations (or Sum Wgts)	154

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
LUC1_630	2	37.00954	18.5048	3.5845	0.0301*
Error	151	779.51842	5.1624		
C. Total	153	816.52795			

Means Comparisons

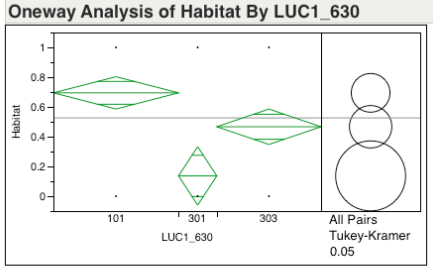
Comparisons for all pairs using Tukey-Kramer HSD

Connecting Letters Report

Level		Mean
301	A	564.14136
101	B	562.81986
303	B	562.67417

Levels not connected by same letter are significantly different.

Oneway Analysis of Habitat By LUC1_630



Oneway Anova

Summary of Fit

Rsquare	0.145694
Adj Rsquare	0.134379
Root Mean Square Error	0.466081
Mean of Response	0.525974
Observations (or Sum Wgts)	154

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
LUC1_630	2	5.594084	2.79704	12.8758	<.0001*
Error	151	32.802020	0.21723		
C. Total	153	38.396104			

Means Comparisons

Comparisons for all pairs using Tukey-Kramer HSD

Connecting Letters Report

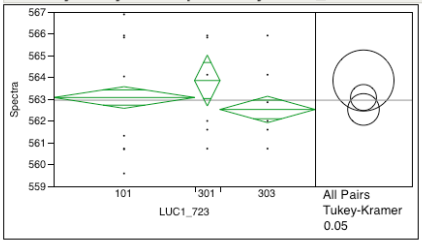
Level		Mean
101	A	0.69444444
303	B	0.46666667
301	C	0.13636364

Levels not connected by same letter are significantly different.

B) Site 723

Fit Y by X Group

Oneway Analysis of Spectra By LUC1_723



Oneway Anova

Summary of Fit

Rsquare	0.028893
Adj Rsquare	0.016031
Root Mean Square Error	2.291556
Mean of Response	562.9519
Observations (or Sum Wgts)	154

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
LUC1_723	2	23.59217	11.7961	2.2463	0.1093
Error	151	792.93578	5.2512		
C. Total	153	816.52795			

Means Comparisons

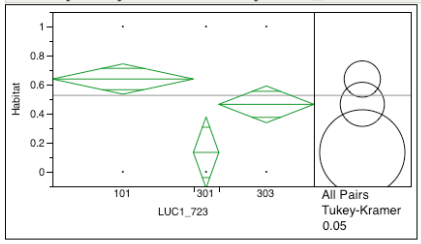
Comparisons for all pairs using Tukey-Kramer HSD

Connecting Letters Report

Level		Mean
301	A	563.85467
101	A	563.07530
303	A	562.52714

Levels not connected by same letter are significantly different.

Oneway Analysis of Habitat By LUC1_723



Oneway Anova

Summary of Fit

Rsquare	0.093175
Adj Rsquare	0.081164
Root Mean Square Error	0.480194
Mean of Response	0.525974
Observations (or Sum Wgts)	154

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
LUC1_723	2	3.577573	1.78879	7.7576	0.0006*
Error	151	34.818531	0.23059		
C. Total	153	38.396104			

Means Comparisons

Comparisons for all pairs using Tukey-Kramer HSD

Connecting Letters Report

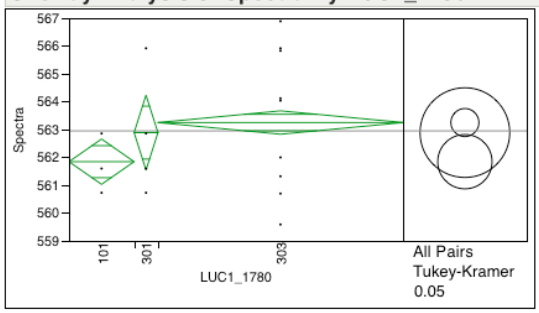
Level		Mean
101	A	0.63855422
303	A	0.46428571
301	B	0.13333333

Levels not connected by same letter are significantly different.

C) Site 1780

Fit Y by X Group

Oneway Analysis of Spectra By LUC1_1780



Oneway Anova

Summary of Fit

Rsquare	0.057409
Adj Rsquare	0.044924
Root Mean Square Error	2.257661
Mean of Response	562.9519
Observations (or Sum Wgts)	154

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
LUC1_1780	2	46.87579	23.4379	4.5983	0.0115*
Error	151	769.65216	5.0970		
C. Total	153	816.52795			

Means Comparisons

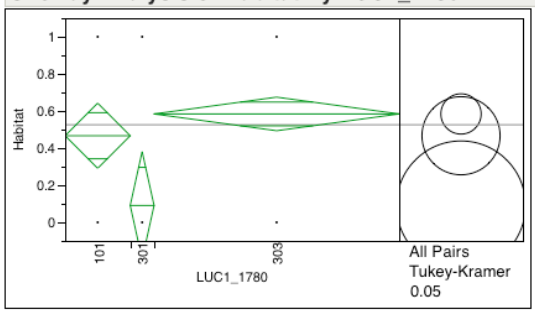
Comparisons for all pairs using Tukey-Kramer HSD

Connecting Letters Report

Level		Mean
303	A	563.25150
301	A B	562.89000
101	B	561.84600

Levels not connected by same letter are significantly different.

Oneway Analysis of Habitat By LUC1_1780



Oneway Anova

Summary of Fit

Rsquare	0.066908
Adj Rsquare	0.054549
Root Mean Square Error	0.487099
Mean of Response	0.525974
Observations (or Sum Wgts)	154

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
LUC1_1780	2	2.569019	1.28451	5.4138	0.0054*
Error	151	35.827085	0.23727		
C. Total	153	38.396104			

Means Comparisons

Comparisons for all pairs using Tukey-Kramer HSD

Connecting Letters Report

Level		Mean
303	A	0.58407080
101	A B	0.46666667
301	B	0.09090909

Levels not connected by same letter are significantly different.