Illumination at the Intersections of Genomics and Public Health: A Study of Opsins, SPurS, Cluster Machine, and Ancestry

Alice B. Popejoy

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Joseph Felsenstein, PhD, Chair

Stephanie Malia Fullerton, DPhil

Elizabeth Thompson, PhD

Program Authorized to Offer Degree:

Public Health Genetics

University of Washington

Abstract

Illumination at the Intersections of Genomics and Public Health:
A Study of Opsins, SPurS, Cluster Machine, and Ancestry in Genomics

Alice B. Popejoy

Chair of the Supervisory Committee: Joseph Felsenstein, PhD
Professor of Genome Sciences and Biology
Adjunct Professor of Computer Science and Statistics

Light is essential to life on planet Earth, yet researchers know very little about how biology and health are influenced by this ubiquitous source of energy. Opsins comprise a large and diverse group of light-sensitive proteins that are conserved across all major branches of the tree of life, but their precise function in humans remains unknown. Studying these proteins is an important step in understanding the impact of light on human health, perhaps leading to a deeper understanding of connectivity of the brain and central nervous system to other major tissue systems of the body. The aims of this doctoral dissertation include a detailed evolutionary genomics study of opsins in vertebrates; a methods-development project designed to uncover novel interspecific signals of natural selection; and a survey of under-represented ancestry groups in genomics research. While the approaches and goals of each chapter in this dissertation are distinct, the underlying theme of this research is to employ existing methods (or design new ones) to probe topics that are often ignored or taken for granted, and to challenge outdated scientific assumptions and paradigms.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

In the United States, it is often said that we "pull ourselves up by our bootstraps," but as I've heard quoted more and more recently by left-leaning politicians (and I agree): "First you need to buy the boots." In the spirit of concluding Chapter Four and upholding my own call to action, I want to acknowledge the many elements of privilege that have enabled me to get to this point in my career (in addition to being born white, heterosexual, and *cis*-gender). First and foremost, my parents and grandparents have provided the social and financial stability and opportunities to attend private elementary and secondary schools, including the academically rigorous Hamilton College, where my Biology advisor, Professor Jinnie Garrett, was supportive and instrumental to my development as a budding scientist. Secondly, I have received generous funding from the National Science Foundation (NSF) as a Graduate Research Fellow, and jointly with the Research Council of Norway for a Graduate Research Opportunities Worldwide (GROW) Fellowship; the National Institutes of Health and the UW Department of Biostatistics for multiple years of support on the Statistical Genetics Training Grant; and the UW Graduate and Professional Student Senate and all students for supporting my year of service as GPSS President.

Before I even considered pursuing a PhD, I had the good fortune of meeting Dr. Phoebe Starfield Leboy at the Association for Women in Science, who on her first day of meeting me made the case that I should abandon the idea of becoming a lawyer and get a PhD. It was she who first instilled in me the question: "Why not you?" Given the low percentage of PhD graduates who go on to earn a tenured professorship at a research institution (and even fewer for women), this is a near-daily mantra I have now adopted. From toughing out her pioneering days of being the first (and only) tenured female professor in the UPenn medical school for 20 years (where she co-opted a men's restroom by putting flower pots in the urinals), through to her final days battling ALS, Phoebe has massively inspired me to work for where I am today. The fact that I will have completed both a Certificate in Statistical Genetics and a doctoral degree program in five years while maintaining an active role in social justice initiatives (indeed, using science to further the

cause), serving as Graduate Student Body President, and becoming a mother with an active family life – I'm sure would have made her very proud.

Next, I would like to acknowledge the former core faculty members of the Public Health Genetics program, including two of my PhD Supervisory Committee members Malia Fullerton and Kelly Edwards, both of whom invested so much time and energy into creating a truly novel and important academic environment in which multi-disciplinary researchers and professionals like myself could thrive. I am grateful that Bruce Weir in the Department of Biostatistics is committed to keeping the PHG program going in the face of administrative hardship, and it is my hope that it continues to exist at the intersections of genomics, policy, law, social science, and bioethics. The interdisciplinary nature of PHG was what drew me to the program, and if not for the flexibility and diversity of curriculum, I'm not sure I would have ended up where I have.

Elizabeth Thompson was an early supporter of my success in the quantitative space of research, as I was new to statistics in general and she gave me the opportunity to learn statistical genetics on the fly with arguably too little prerequisite experience. I am grateful to the postdocs she introduced me to, particularly Alejandro Nato, who taught me the basics of command-line programming, including and starting with how to open a terminal window. Joshua Schraiber, a former postdoc in Josh Akey's lab in the Department of Genome Sciences (now an Assistant Professor at Temple University) was also instrumental in helping me gain ground in the more quantitative topics of population genetics.

The faculty member from whom I have objectively learned the most is Jim Thomas. He was the first professor to take me seriously when I pitched the idea of researching mysterious light-receptive proteins, and has spent countless hundreds of hours teaching me about gene family dynamics and evolutionary genomics. Joe Felsenstein, whose courses on population genetics and phylogenetic inference laid the foundation for my knowledge in these areas, has also been instrumental in helping me get to this point, agreeing (somewhat reluctantly) to be my PhD Supervisory Committee Chair and is a very supportive one, at that.

Finally, I am honored to thank and acknowledge my partner and fiancé, Jimmy Davidson, who has been a joyful addition to my life, not least of all because he has given me our daughter Kaia, now two years old. He has done an incredible job supporting me in pursuit of my degree by spending endless hours with her during the day and night while I work, and taking on the traditionally "female" household roles such as cooking, cleaning, and generally keeping everyone happy and sane. The privilege he has afforded me to work in an academic career is tremendous, and I will forever be grateful for his ceaseless support and positivity.

Moving forward, I will be joining the lab of Dr. Carlos Bustamante at Stanford University, where my postdoctoral responsibilities will involve spearheading a working group on ancestry in genomics, in association with the ClinGen project. I will also have the freedom to pursue my own research, and have planned future collaborations with many researchers in the United States, Estonia, Norway, and Denmark. Regarding immediate next steps, I plan to conduct an intra-species analysis of opsin variation to identify human-specific signals of natural selection. Additionally, I am collaborating with researchers at the Norwegian Centre for Mental Disorders Research (NORMENT) in Oslo on a project involving Vitamin D, to investigate relationships among opsin variation, psychiatric phenotypes, and levels of the nutrient.

In closing, I want to express my deepest appreciation for the opportunity to pursue a PhD in a field that is so flexible, dynamic and interesting. While I now feel like there is so much more that I don't know than when I started graduate school, the skills and training I have received here have prepared me for a dynamic career in academia. Given my interest in politics and social justice, some academic researchers have questioned my commitment to science; I've even been asked: "Are you sure you don't want to go back into politics?" In my view, the problem here is in the question. It is my belief that scientists have a responsibility to educate the public (yes, including politicians) about important issues that we face as a society. We also have a duty to our colleagues and students to make academia a more welcoming and inclusive atmosphere for women and minorities. As a future research faculty member, I am committed to doing both.

DEDICATION

This dissertation is dedicated to Phoebe Starfield Leboy.

I promise to always keep the fires burning.

Chapter 1. INTRODUCTION

As pioneers in a forward-thinking field at the intersection of genomics and society, graduates from the Institute for Public Health Genetics (IPHG) at the University of Washington are breaking new ground. We are integrating the social and quantitative aspects of genomics from multiple methodological angles, developing a rich pedagogy centered on ethical research. The foundation of our training spans two core knowledge areas: *Genomics in Public Health*[*] and *Implications of Genetics for Society*[†]. IPHG graduates typically focus on one of these two core knowledge areas, but I have conducted extensive training, research, and dissemination in both of them. In my particular course of study, I earned a Certificate in Statistical Genetics and conducted three major projects: 1) mapping the evolutionary history of light receptor genes across vertebrate species, 2) developing novel methods to detect theoretical evolutionary dynamics, and 3) quantifying the under-representation of non-Europeans as research participations while exploring the ethical, legal, social implications (ELSI) of diversity in genomics.

1.1 LIGHT AND BIOLOGY

My journey to uncover the mechanisms of interaction between light and biology began as an undergraduate at Hamilton College in 2008, when very little had been published on the genetic underpinnings of light detection in the body. I learned that organisms have biological light receptors both in the retina (which are responsible for color vision) and

[*] *Genomics in Public Health* encompasses individual disciplines that focus on public health applications of genomic data and research. Required coursework for a PhD in PHG includes biostatistics, bioinformatics, genetic epidemiology, pharmacogenomics, toxicogenetics, human genetics, and clinical genomics.

[†] *Implications of Genetics for Society* describes an exploration of how increasing use of genomic technology and research impacts people and populations. Disciplinary course work for a PhD in PHG includes public health, bioethics, social science, law, public policy, and health services or economics.

outside the eye, but the function of these so-called “non-visual” light receptors remains enigmatic. *Opsins* comprise a family of genes that encode light-sensitive proteins expressed throughout the body, most notably in the brain and central nervous system. It seems odd that humans should have light receptors in places such as heart muscle and kidney tissue, in addition to more sensible but puzzling places such as the epidermis, or skin cells. I am on a lifelong academic mission to grasp the evolutionary purpose of these genes, and hope to ultimately uncover their function in humans.

Chapter 2 of this dissertation documents complex evolutionary mechanisms of the opsin gene family, and demonstrates their inferred widespread functional importance based on genomic comparisons of distantly related species. In sum, the pressures of natural selection only act on stretches of the genome that have an impact on biology, specifically, affecting a species’ fitness. By surveying opsin protein sequences for signals of natural selection, I found that these genes are highly conserved over vast time scales, meaning that they are likely performing a vital function to many different species. It is also evident that some non-visual opsins may be involved in directed, or positive, selection, favoring variation that provides an evolutionary advantage for a particular species or lineage. While the adaptive importance of the non-visual opsins remains unclear, the results of this research lay the groundwork for future directions in the field.

1.2 METHODS DEVELOPMENT

During this scan for signals of selection, I noticed sites in aligned opsin protein sequences that appear to favor a small number of different amino acid residues that segregate between species types. For example, one residue appears to be conserved in all eutherian (placental) mammals while another is conserved in birds and lizards. I coined this observation a *shift in purifying selection* (*SpurS*), which indicates a site in the genome

that has been under strong purifying (or negative) selection to preserve a residue in one lineage, and then at one point in time a change occurred that was favored and conserved in another lineage. In order to identify these sites at the genome-wide level, I developed the purifying selection shift (Psi, or Ψ) statistic and wrote a program in Python to categorize sites based on their Ψ-values. The *SPurS* program is described in detail in **Chapter 3**, and is available as raw code through my GitHub account. It can also be utilized by user-interface through the University of Oslo Genome HyperBrowser analysis tool. The implications of this tool are potentially far-reaching, as it enables biologists with little computational or bioinformatics skills to identify precise candidate molecular drivers of morphological and behavioral differences between higher-order species groups.

Section 3.1 describes the Ψ statistic and theory, while section 3.2 illustrates the distribution of Ψ across a large set of conserved genes. Because SPurS sites likely occur by chance as a result of phylogenetic relationships between species, I compare the distribution of Ψ across real protein sequences to a set of sequences simulated under null conditions. Another way to distinguish true SPurS signals from those occurring by chance is to identify outliers among all sites sampled across the genome. These analyses and methods to determine statistical significance are described in **Chapter 4**.

The statistical challenge of identifying significant outliers is not trivial, and the issue of classification of true and false positives is of interest in both classical statistical methods and machine learning. In this dissertation, I describe a novel clustering method based on a conservative, collapsed version of *correspondence analysis*[‡] devised by Joseph Felsenstein. Working together, we developed an implementation algorithm which

---

‡ *Correspondence analysis* is a statistical clustering method related to principal components analysis (PCA), which decomposes the chi-square statistic over a contingency table of categorical variables. It is most popular in France, and has been applied prominently in social science analyses.

I coded in Python for use in tandem with *SPurS* and is flexible enough to be utilized for countless other purposes.

The basic idea of the algorithm is to separate or cluster groups of samples, such as genes or individuals, based on their value for a particular variable. A large contingency table, or matrix, is repeatedly collapsed in different combinations of ordered rows and columns to find the optimal chi-square value. The collapsed 2x2 chi-square that finds the maximum value for the whole matrix defines two clusters of input samples (collapsed rows) as significantly different from one another, relative to the variable of interest (collapsed columns). This method is explained in detail in **Chapter 4**, with SPurS as an example, to find clusters of genes that differ from one another with respect to their Ψ-values. The program, *ClusterMachine*, is publicly available in Python on my GitHub site and has the potential to be widely used for both supervised and unsupervised machine learning clustering problems and outlier identification.

1.3 ETHICAL, LEGAL AND SOCIAL IMPLICATIONS (ELSI)

While building a repertoire of skills in bioinformatics and statistical genetics, I realized a newfound potential for methods development that led to *SPurS* and *ClusterMachine*. This training in computational biology also opened many doors in terms of the type of large-scale analyses I could do to illuminate the lack of diversity in genomics research. As a public health geneticist, part of my duty is to ask social, ethical, and legal questions about the field of genomics, and use my quantitative skills to address them. My academic interest in the under-representation of minority populations is fueled by a personal passion for social justice activism as well as scientific thoroughness. There was a set of questions about interspecific variation in human opsins that I felt unable to address due to the dearth of genetic data on geographically and ancestrally diverse populations. Thus,

the *Genetics and Society* component of my dissertation is dedicated to quantifying diversity in genomics research. To summarize, I calculated the proportion of samples used in genome-wide association studies (GWAS) from each ancestral group represented in the GWAS Catalog and reported my findings in a *Nature* Comment with Malia Fullerton. **Chapter 5** details the methods and results of this study, as well as its broader impacts. In particular, I argue that discussions of diversity in genomics must pivot to include policy and societal solutions, in addition to methodological approaches to address population-specific genetic heterogeneity. This will require an honest examination of how lack of diversity is rooted in systemic, cultural, and historical racism and inequality.

1.4 SUMMARY

Each chapter in this dissertation represents a self-contained, independent research project. They are related through a thread of curiosity and inquiry during the course of interdisciplinary training across bioinformatics, genome sciences, statistical genetics, and the ethical, legal and social implications (ELSI) of genomics. The body of work in its entirety is an example of how training across diverse epistemologies of Public Health Genetics produces integrated research that incorporates evolutionary biology, theoretical and statistical methods development, computer science, bioethics, and social justice. Such cross-pollination of approaches and perspectives enhances the potential for creativity and innovation in research, which is revealed in the development of novel methods and theory in this dissertation. Moving forward, the need for integration of training in computational, biological, and ethical fields of inquiry will continue to grow, as the availability of vast amounts of data outpaces our ability to analyze, interpret, and assess their implications.

Chapter 2. ILLUMINATING HUMAN LIGHT RECEPTORS

Investigating the relationship between form and function is fundamental in biology and evolution. Various scales of inquiry (molecular, structural, organismal) inform the translation of physiology to biological mechanisms. Researchers often rely on functional studies of model organisms to elucidate this relationship, dissecting tissues and performing assays at various stages to understand pieces of the relevant pathways from form to function. While this is a highly valuable method for discovery science in basic biology, findings from model-system studies are not always transferrable to humans. For example, psychiatric or behavioral traits are likely to be species-specific, so a trait of interest in the human brain may not be sufficiently represented by most model organisms. As an alternative to this approach, certain non-invasive assays such as blood biomarker levels, genetic variant and gene expression profiles, brain imaging results, and superficial biopsies may be analyzed to identify biological pathways involved in the etiology of human traits and disease. Studies of this nature have been successful for traits and conditions that are straightforward to assay, but what about disease pathways or traits that do not have an easily measurable phenotype or obvious intermediate biomarker?

Although light is a ubiquitous source of energy for life on Earth, very little is known about interactions between light and human biology. I argue that this is because there are very few obvious *a priori* phenotypes to measure in association with light perception, and the exposure is difficult to measure and control due to ubiquitous environmental and artificial light stimuli. In this case, model organisms across the animal kingdom have been studied to unpack the role that light plays in biological pathways, such as studies of photobiology in amphibian skin cells (melanophores).[1] However,

investigators across diverse disciplines have found that organisms often interact with light in unique ways, including species-specific mechanisms and adaptations in response to light exposure. As such, there is a need to devise creative approaches for understanding the role that light plays in human biology, beyond the little that is already known in the field of circadian rhythm photo-entrainment.

Opsin genes encode opsin proteins – a family of G Protein-coupled Receptors (GPCRs) that have been demonstrated to be directly responsive to photons of light.[2,3,4] Some of these membrane-bound proteins are responsible for the translation of external light to visual sensory perception. Others that do not appear to have a role in visual perception may have several functions that have yet to be discovered. Gene expression studies have shown that opsins are present in the brain and central nervous system.[5] Thus, due to logistical and ethical challenges of conducting functional studies of light-sensitive, deep-brain proteins in living humans, computational and evolutionary biology offer advantageous approaches to learning about the role that these "non-visual" opsins may play in human biology and health.

By approaching the question of how light interacts with human biology through a lens of gene family evolution and tissue system expression, my dissertation research features aspects of vertebrate opsins from birth-death evolutionary dynamics across lineages to human-specific patterns of opsin gene expression. While this research is far-reaching methodologically, the goal is to systematically approach the question of how form translates to function in human opsins. The first step in such an undertaking is to demonstrate which opsins are present in humans, relative to other vertebrate species and lineages. Using a computational comparative evolution approach, I surveyed the genomes and proteomes of 29 species across a wide variety of vertebrate lineages (section 2.2) and

provide a detailed account of which opsins are present in which taxa. This account includes phylogenetic trees with bootstrap support values for internal branches based on 1000 replicate trees (Figures 2.3—2.8), and tables of orthologous genes across species with corresponding consensus trees for each opsin sub-family and maximum likelihood gene trees (Tables A-2—A-23 and Figures A-2.1—A-28 in the Appendix).

Once I identified the nine opsins that are definitively present in humans (Table 2.1), I sought to determine whether there is evidence of natural selection acting on these genes. While most opsins in humans are primarily under strong negative (purifying) selection across species, I found a few non-visual opsins (OPN3, OPN5, and RRH) to contain signals of positive selection using standard selection-identification methods (section 2.3). The OPN3 locus implicated in this analysis lies in an important functional region (light-sensitive ligand binding pocket) of the protein, indicating that variation in this gene may have played a role in environmental adaptation. I then investigated the role of opsins in species divergence, using a novel method I developed to test for lineage-specific shifts in purifying selection across a much larger variety of vertebrate species.

2.1 NON-VISUAL OPSINS: FROM FORM TO FUNCTION

Charles Darwin marveled at variation across species, using the visual system to illustrate his concept of natural selection as environmental pressure driving the adaptive fine-tuning of biological features.[6] He did this without the observation of genetic data, which today has permitted researchers to discover associations between variation in genetic sequences and variation in phenotypic traits. In theory, opsins represent an ideal model for investigating these associations because their reactivity to various wavelengths of light (their *spectral sensitivity*) is a quantitative trait, measurable in a laboratory setting.

In practice, researchers have been successful in mapping this phenotype primarily in visual opsins, or those found in rod and cone cells of the retina.[7] In 1987, Jeremy Nathans published a molecular genetics overview of visual opsins, elucidating the entire field as an uncharted path into advances in the study of vision.[8] Following Nathans' publication, he and many others showed that the absorption maximum (λ_{max}) of a pigment's spectral sensitivity shifts in response to amino acid substitutions introduced in functional regions responsible for ligand binding.[9,10,11,12,13] Subsequently, Yokoyama et al. elaborated on this mechanism of spectral tuning, and contributed a vast amount of experimental evidence demonstrating how specific amino acid changes have driven adaptive evolution of visual pigments in vertebrates.[14,15,16,17,18,19,20,21,22,23,24,25,26,27] In 1995, Chang et al. used a comparative phylogenetic approach to identify residues important in opsin wavelength regulation in the vertebrate visual opsins.[28] While these studies of spectral sensitivity of *visual* opsins have informed our understanding of the molecular mechanisms of adaptive evolution, there remains a gaping void in the literature on spectral sensitivity and other features of non-visual photopigments.

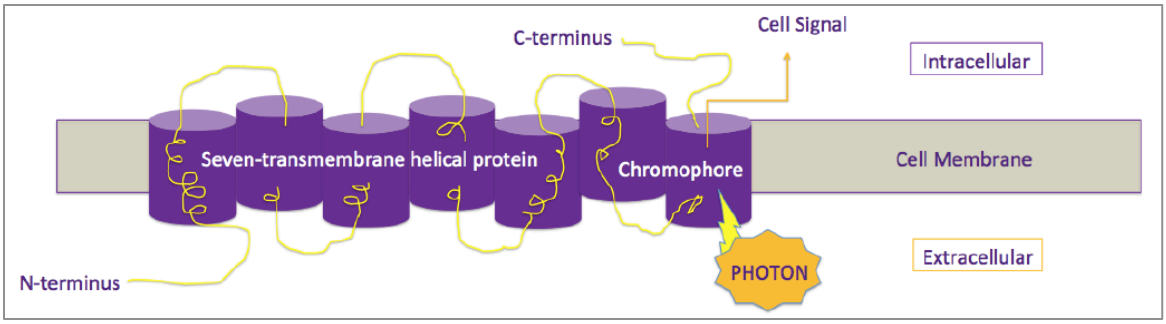


Figure 2.1. Membrane-bound structure of opsin proteins. Illustration of functional regions simplified and adapted from M.F. Whiting's representation: https://tinyurl.com/opsin-gif.

The structural model used to infer functional regions of genes in the superfamily of G Protein-Coupled Receptors (GPCRs) is the crystal structure of bovine rhodopsin,

published by Palczewski et al. (2000).[29] This model is used as a reference for functional regions in human opsins, assuming that all GPCRs are structurally similar. According to this rough structural model (Fig. 2.1), opsin proteins (like all other GPCRs) pass through the cell membrane seven times, with intra- and extra-cellular loops between the trans-membrane regions, an N-terminus outside the cell and a C-terminus inside the cell. The C-terminus structure is particularly difficult to resolve and as such, very little is known about it (UW Professor Fred Rieke, personal correspondence). In the seventh trans-membrane region of the protein, there is a chromophore-binding pocket that is responsible for binding the light-sensitive ligand retinal.[30] Previously cited studies by Yokoyama et al. have demonstrated that mutations introduced to this seventh trans-membrane region can shift the spectral sensitivity of opsins, providing functional evidence of its role in activation by a photon at a particular wavelength.

In personal correspondence with Dr. Yokoyama, I learned that researchers conducting these mutagenesis experiments in opsins have been unable to identify the spectral sensitivities corresponding to non-visual opsins, as *in vitro* gene expression methods used to manipulate and study visual opsins have not yet been successful on non-visual opsins. The reasons for this are unclear, and provide an opportunity for future research in this area. Furthermore, my reliance on personal communications with researchers familiar with the field is indicative of the dearth of strong and convincing findings in the literature related to non-visual opsins. While some evidence of their functional role exists in piecemeal across various model systems and natural populations, there remains to be presented a methodologically consistent review of all opsins found in humans.

Table 2.1 lists the nine opsin genes that are present in humans, classified by type according to Porter et al.'s 2012 review of opsin evolution, where "C-type" signifies

ciliary, "R-type" indicates rhabdomeric, and "Group 4" classifies opsins that fall outside the range of easily-categorized cell type in which the proteins are found.[31] Gene names and symbols may vary across publications; those presented here are the most common and recently cited: *Opn1sw, Opn1mw, Opn1lw*, and *Rrh* comprise the visual opsins; *Opn3, Opn4, Opn5, Rrh* and *Rgr* are considered "non-visual" opsins. These categories (visual vs. non-visual) are based on the cell type in which opsins are expressed; visual opsins are present in cells involved in detection of visible light. However, some of the "non-visual" opsins are also expressed in the retina, making those retinal ganglion cells "intrinsically photo-sensitive", thus the line between *visual* and *non-visual* is somewhat blurred.[32] Nevertheless, I have chosen to stick with this naming convention throughout.

Table 2.1. Summary of Human Opsins. Nine opsins are present in humans: four known to be involved in the visual pathway, and five with largely unknown function, often referred to as 'non-visual' (starred*). Classification of opsin type is based on Porter et al. (2012).

Opsin Type	Gene Symbol (chromosome)	Gene Name	Protein/Function
C-type	*Opn1sw* (7)	Short-wave-sensitive opsin	Blue cone receptor
C-type	*Opn1lw* (X)	Long-wave-sensitive opsin	Red cone receptor
C-type	*Opn1mw* (X)	Medium-wave-sensitive opsin	Green cone receptor
C-type	*Rho* (3)	Rhodopsin	Low-light (purple) rod receptor
C-type	**Opn3* (1)	Panopsin/Encephalopsin	Unknown
R-type	**Opn4* (10)	Melanopsin	Pupillary reflex/Circadian rhythm
Group 4	**Opn5* (6)	Neuropsin	Unknown
Group 4	**Rrh* (4)	Peropsin	Unknown / Photoisomerase
Group 4	**Rgr* (10)	Retinal GPCR	Unknown / Photoisomerase

The human *visual* opsin genes encode proteins in rod and cone cells in the retina, which are activated by photons.[33] While the function of non-visual opsins is largely mysterious, the basic biochemical mechanism of light-induced activation is known. Activation by a photon causes a conformational transformation of the ligand retinal from11-*cis* to all-*trans*,[34] setting off a phototransduction cascade, which translates an electrochemical signal into color vision in the brain.[35] The precise translation from signals in the brain to the perception of vision is an active area of study, and while fascinating, is not the topic of this dissertation. Rather, the purpose of this review is to demonstrate the importance of non-visual opsins relative to the comparatively well-understood visual opsins, by surveying their history of gene duplication, divergence, and conservation across species.

In the last two decades, over 2000 opsins have been discovered and sequenced across the animal kingdom, nine of which are present in humans, and five of which fall into the *non-visual* category.[36,37] Studies have suggested a primary biochemical function of some non-visual opsins in humans: photoisomerization, which converts the ligand retinal from an activated to a non-activated form;[38] others have described their reactivity to different types of G proteins.[39] One hypothesis is that some of the non-visual opsins serve as a sort of conformational 'switch' that allows the re-use of retinal, which is the ligand for other opsins that require the non-activated form.[40,41,42] This hypothesis is consistent with my findings, which show that certain opsins may be under very strong conservation and thus serve a fundamental biochemical purpose, whereas others may have more flexibility in their form to allow variability in function to arise.

In addition to this theory, researchers have conducted a few functional studies of human non-visual opsin proteins, including OPN3 (encephalopsin or panopsin), OPN4 (melanopsin), and OPN5 (neuropsin). However, some of the findings within studies are

inconclusive, and several findings disagree between studies. For example, some have implicated OPN4 in the circadian rhythm pathway,[43] and others have made the case that it potentially plays a role in Seasonal Affective Disorder (SAD).[44] However, a different group has reported both OPN3 and OPN4 are unnecessary in circadian rhythm entrainment in humans; instead reporting that OPN5 is responsible for photo-entrainment of circadian oscillators in the retina and cornea.[45] Furthermore, SAD is no longer considered a psychiatric condition independent from major depressive disorder in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V), casting doubt on epidemiological findings that use SAD as an outcome variable.[46] While many of these groups have briefly acknowledged that there may be genomic redundancy with regard to the opsins, meaning at least one or some are necessary but not all, this explanation has not been fully addressed.

Non-visual opsins were first revealed when Okano et al. (1994) published functional evidence of a "pineal photosensor" expressed in the chicken pineal gland, naming it *Pinopsin*.[47] After this initial discovery, pinopsin was identified as a photoreceptor in the pineal organs of birds and reptiles, and was reported absent in mammals.[48,49] When I first came across this finding in 2008, I followed it up with an undergraduate research project in Bioinformatics taught by Wei-Jen Chang at Hamilton College, and found multiple references to the pineal gland as "vestigial" in humans. Around that time, my uncle had a large brain tumor removed at the University of California San Francisco, and [although I can't recall his name] the surgeon assured me that pineal organs were unnecessary, and routinely removed in patients undergoing brain surgery. It seemed bizarre that a light-sensing organ in the geometric center of the brain

would have lost its function in humans, yet its form remained intact for millions of years, so I set out to understand the nature of light-sensing functions in the human body.

The pineal organ in humans is now widely considered an endocrine gland (producing melatonin), and thus it may play a role in circadian rhythm and reproductive hormone production;[50] however, it is directly involved in light detection in other species. Researchers have uncovered aspects of the complex relationship between physiology of light-sensitive organs such as the pineal and parapineal organs in birds, fish, and lizards, their evolutionary history, and function.[51] For example, the parapineal organ in some lizards and crocodiles is directly connected by a neural pathway to an external membrane that senses light for purposes of thermo-regulation, and the pineal organ in birds contains the circadian oscillator that entrains their circadian rhythm. [52,53] In contrast, the "master pacemaker" in humans, the hypothalamic suprachiasmatic nuclei (SCN), is considered responsible for circadian entrainment (as opposed to the pineal gland) so findings from model organism studies of the pineal and parapineal organs are not helpful for elucidating this system in humans.[54] Although there is indeed no ortholog of chicken pinopsin in humans, this does not rule out the possibility that other non-visual opsins are expressed there, nor does it rule out a light-sensing function of the human pineal gland. It also does not rule out the possibility that other opsin-containing tissues in the human brain, central nervous system, and other internal organs are also directly receptive to light.

Is it likely that photons interact directly with non-retinal tissue inside the body? Research on the near-infrared window of the electromagnetic spectrum suggests it is possible. In 1800, William Herschel discovered the existence of energy beyond the visible spectrum (specifically the near-infrared region, or NIR), somewhat by accident, as he measured temperature increases on a mercury thermometer associated with exposure

to different colors of light as dispersed by a prism.[55] Today, biomedical diagnostics and research take advantage of the fact that near-infrared light (700-2500 nm) can penetrate biological tissues such as skin and blood, e.g. in the use of near-infrared spectroscopy (NIRS).[56] Because the specific wavelengths of light that are absorbed by blood and water are lower than the wavelengths of NIR light, this region of the electromagnetic spectrum easily penetrates biological tissues and its photons scatter into the surrounding structures, typically when they stumble upon the cell nucleus and mitochondria. It is plausible, then, that certain wavelengths of light in the NIR penetrate the human body and photons from external energy sources are processed directly by opsins in our internal tissue systems. Viewing research on opsins through this lens, it seems reasonable that non-visual opsins may have retained light-sensing functions, even when expressed in tissues deep inside the body. It is also possible that the number of different opsins in different species and types of organisms is indicative of these various functions, as each protein is itself reactive to specific ranges of wavelengths along the non-visual electromagnetic spectrum.

Due to the preliminary nature of most functional hypotheses and contradictory findings in the literature about what opsins do, the precise role of human non-visual opsins remains unknown.[57] Whether there is even an important function for non-visual opsins in human biology is not yet confirmed. However, given the possibility of their light-sensing function, it is question worth asking. In the following sections, I demonstrate that non-visual opsins found in humans are functional and biologically important, by virtue of selective pressure that has shaped their duplication, divergence, and conservation over time. By providing a systematic review of the form of non-visual opsins in a gene family context across vertebrates, these findings offer a starting point for generating hypotheses about their function and importance in human biology.

2.2 BIRTH-DEATH DYNAMICS OF VERTEBRATE OPSIN EVOLUTION

The number and diversity of opsins varies greatly across species, and (like other GPCRs involved in sensory perception such as olfactory and chemosensory receptors[58,59]) opsins display a recurrent birth-death pattern of evolution, meaning they arise by duplication and subsequently diverge as substitutions accumulate. As reported by others and confirmed by my analyses, opsins comprise a monophyletic family of genes, meaning that they arose by a series of duplication events from a single original opsin. Each of these duplications was likely maintained in response to changes in environmental stimuli.[60] As these duplicate genes (paralogs) emerge, they accumulate substitutions that can shift the spectral sensitivity of the protein, allowing organisms to detect different wavelengths of light under certain conditions. Davies et al. (2012) describe the origins of this mechanism for adaptation to different photic, or light environments, as "spectral tuning" of the first photopigments that arose in agnathans (jawless fishes, such as lamprey and hagfish), allowing them to detect light in the Early Cambrian aquatic environment.[61]

When a particular adaptation is advantageous to a species, such as a newly duplicated and diverged opsin gene, it is kept at a nearly constant state over time by negative, or purifying, selection. Thus, orthologous genes that are observed in high conservation (those that have changed little across species over time since they existed as a single copy in a common ancestor) are presumed to be advantageous to the organisms harboring them. In contrast, duplicated genes that are not advantageous and/or serve no biological purpose are often inactivated (e.g., by protein truncating mutations) and eventually eliminated from the genome.[62] The recurrent process of gene duplication and loss (birth-death evolution) is responsible for the variable numbers of opsin genes found

in different lineages of vertebrate species. Figure 2.2 illustrates this variability in the number of gene copies present in each sub-family of opsins found across the various taxonomic groups. Section a. (Species Selection and Data Capture) and section b. (Orthologous Sequence Identification) of Chapter I in the Appendix, *Materials and Methods: Opsin Evolution*, details the approach used to obtain these results.

Briefly, methods I used to identify vertebrate opsins are comprised of five major steps: 1. Select pre-defined opsin proteins to serve as queries, 2. Search genomes and predicted protein sets of pre-selected vertebrate species for matches to query proteins, 3. Build phylogenetic trees of top search results (most probable opsins and closest non-opsin matches), 4. Manually prune trees to remove non-opsins and isoforms to determine unique orthologous copies across species, and 5. Repeat the process using different queries to ensure comprehensive capture of opsins in vertebrates. Once a set of opsins is determined across species (and all other related proteins fall outside a monophyletic clade of the opsin family), the number of opsins in the human genome can be determined by counting the number of human sequences in the final phylogenetic tree of opsins. Although most public databases of genomes and predicted protein sets report sequences that are orthologous to the human versions of each gene, an early finding of this research was that annotations are often incomplete or wrong. There are many factors that contribute to the precarious reliability of these annotations, including variability in gene names across databases and genome reference websites, genome assembly and alignment techniques or protein prediction algorithms that miss genes entirely, and orthologous gene calling programs that arbitrarily match non-orthologous genes due to missing or deleted genes in one species and lack of a better match in the database. All of these

methods and related issues are addressed in greater detail in the *Materials and Methods: Opsin Evolution* section of the Appendix, as previously mentioned.

As seen in Figure 2.2, there are many duplicated opsin genes in the teleost (bony) fish lineage, with comparatively fewer orthologues in lizards, birds and mammals, and the fewest present in lamprey. The highest average number of opsin orthologues observed for any lineage was 2.75 rhodopsin proteins per fish species, with 2-4 duplicated copies of *Rho* present in four fish species: zebrafish (3), fugu (2), platyfish (2), and cave fish (4).

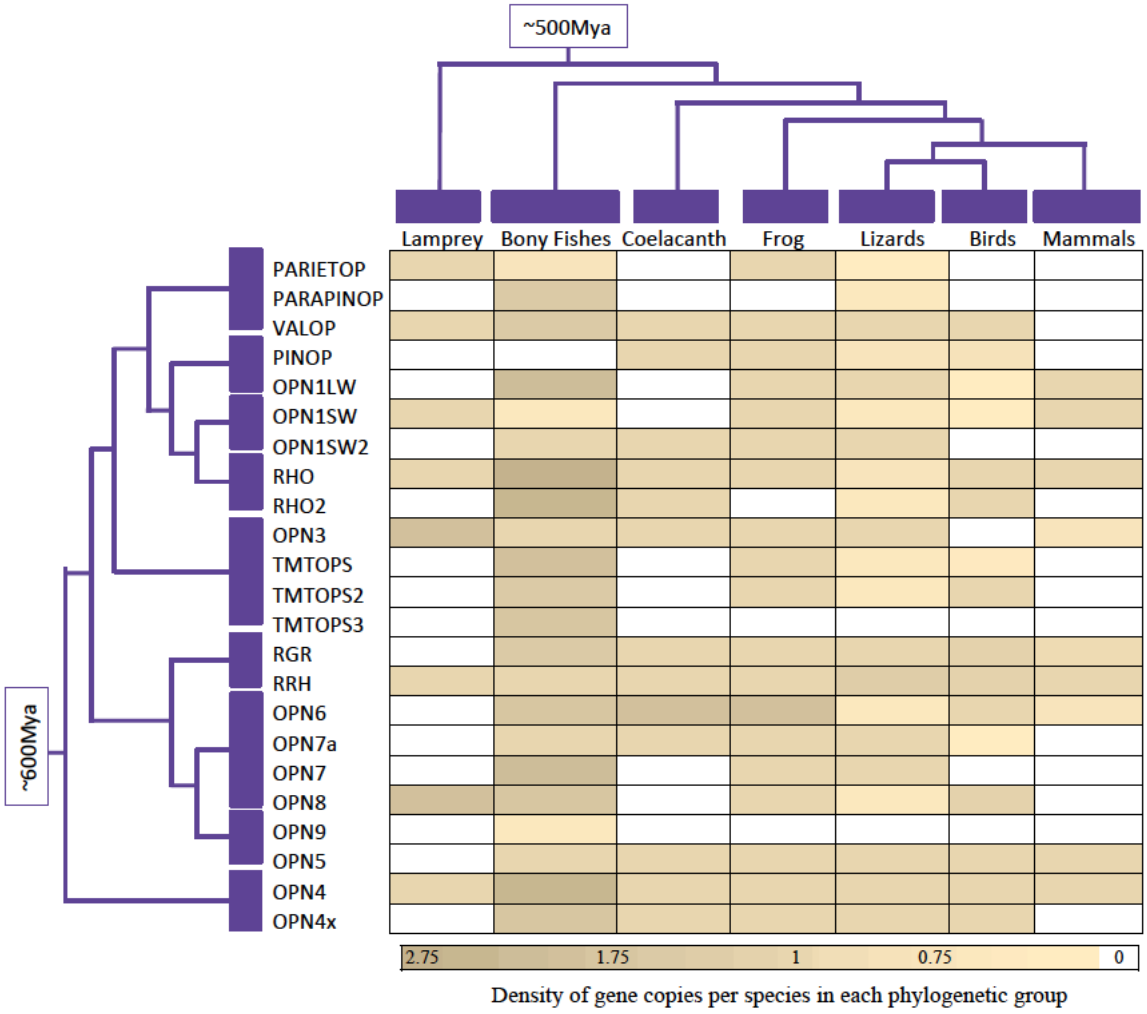


Figure 2.2. Heat map of vertebrate opsin orthologues. Each row represents a single gene, and each column represents a phylogenetic clade of species, ranging from 1 to 5 species per clade. Color density represents the number of gene copies (duplications) found in a phylogenetic clade, divided by the number of species in that clade (see

legend). A value of zero (white) indicates that there are no copies of a gene in any species of the clade (minimum) and a value of 2.75 indicates that the average number of gene copies for a given species in the clade is 2.75 (maximum). A density of 1 indicates that a single copy for a given gene is present per species, on average.

The overall abundance of opsins in fish makes sense from both an evolutionary and an ecological standpoint; a whole-genome duplication in teleost fishes has led to the common presence of two orthologous gene copies in fish for each gene in humans.[63] Additionally, the aquatic habitat of fish impacts their ability to see at various wavelengths of light, so it is plausible that multiple variations of each opsin gene are necessary in aquatic animals relative to a single copy in land-dwelling animals. In African cichlid populations, researchers have observed different combinations of photopigments in different species that are likely involved in sexual selection and speciation as well as survival in various photic (light) niches of lake environments.[64,65] This example provides an ecological context and adaptive evolutionary response to explain diverse and abundant visual opsins in fishes. However, it does not explain the trend for non-visual opsins unless non-visual opsins are directly light sensitive and performing a biological function.

In 2015, Davies et al. reported the presence of 42 distinct genes encoding light receptors in zebrafish, 32 of which they described as "nonvisual opsins" and demonstrating that they form functional photopigments with unique spectral sensitivities.[66] In my analysis, I identified 40 of these 42 proteins in zebrafish, and the two not found in zebrafish (parapinopsin-a and OPN9) were identified in other teleost fishes; likely missed in zebrafish due to sequence divergence or the use of non-fish sequences to capture orthologous genes. The authors of this zebrafish opsin study note that non-visual light detection is more complex than previously thought, and there are "significant biological implications" for investigating light detection in vertebrates.

Indeed, a large proportion of the opsins I identified (~31%) are annotated as "novel" or "unpredicted" with the source database having made no connection to opsins. Thus, there is a need to carefully characterize and correctly annotate opsin orthologs in the literature.

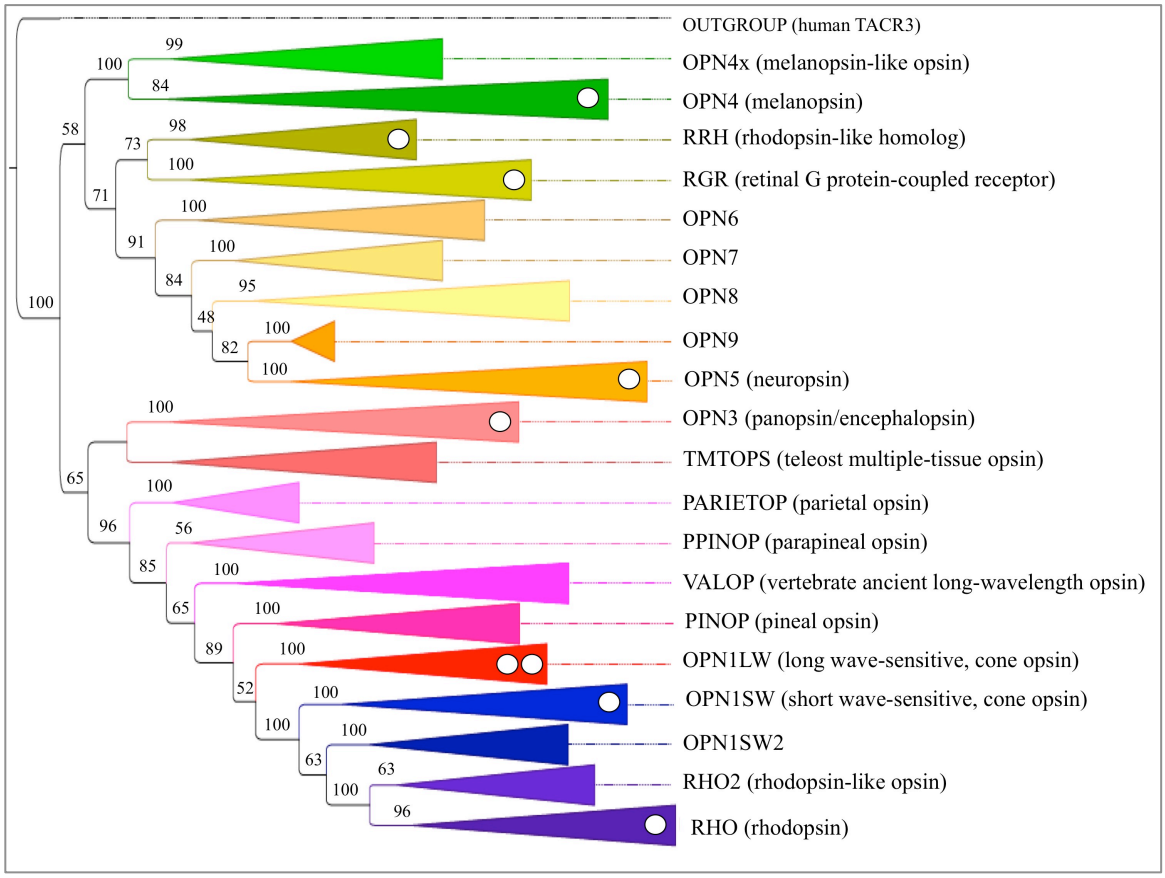


Figure 2.3. Opsin family majority rule consensus tree (constructed from 100 ML trees in J. Felsenstein's *Consense*, figure illustrated in A. Rambaut's *FigTree*). Consensus of 100 ML trees, each built from 413 protein sequences sampled across 29 vertebrate species; bootstrapped across amino acid sites with replacement. Node values reflect the *number* of bootstrap trees (out of 100) with the displayed branching pattern. Each of the opsin protein clades identified among vertebrates is represented by a unique color shade, and the five larger sub-families of opsins share a color palette: melanopsins (OPN4 and OPN4x) in green, photoisomerases (RRH and RGR) in yellow-green, neuropsins (OPN5, OPN6, OPN7, OPN7a, OPN8, and OPN9) in yellow-orange, panopsins (OPN3, TMTOPS, TMTOPS2, TMTOPS3) in salmon, other deep-brain opsins (PARIETOP, PINOP, PPINOP, and VALOP) in pink-magenta, and color vision opsins in their corresponding colors: red (OPN1LW), blue (OPN1SW, OPN1SW2), and purple (RHO, RHO2). Clades with white circles indicate the presence of a human protein; two circles in the OPN1LW clade show a recent duplication that created OPN1MW.

There are many opportunities for research on non-visual opsins, and there is a strong need for interdisciplinary studies that combine phylogenetic and computational evolutionary methods with functional investigations. In this study, I conducted a comprehensive phylogenetic analysis of opsins found in humans, and included all other opsins present in vertebrate species sampled among fish, lizards, birds, and amphibians. The materials used for this analysis and the methodology are detailed in the *Materials and Methods* section of the Appendix, complete with data tables (Tables A-2 through A-23) as well as consensus and maximum likelihood trees from each opsin sub-family (Figures A-2.1 through A-28) with bootstrap support values for individual protein-level topologies. Fig. 2.3 shows the macro-level opsin family consensus tree comprised of 20 protein clades, built from 100 PhyML[67] trees (each containing 413 protein sequences) in *Consense* from J. Felsenstein's PHYLIP[68] illustrated with A. Rambaut's FigTree.[69]

In total, I report on 23 distinct opsins that are present in multiple species, and collapse species- and lineage-specific duplications (particularly those in fish) into a single gene category, since the focus of this study is on human non-visual opsins. My findings are consistent with those of the most recent large-scale phylogenetic review of opsins (Porter et al., 2012), however my results provide much greater resolution about opsins found specifically in humans. The four opsin groups classified by Porter et al.: "C-type", "R-type", "Cnidops", and "Group 4" are too broad to be functionally meaningful in such a deeply divergent phylogeny, and it is difficult to ascertain from their text which of the sub-groups are present in humans. While this earlier study encompasses a greater diversity of species, I have limited the analysis here to vertebrates in order to obtain a clear view of the groups of photoreceptors that have evolved in and around mammals;

thus those that may be important in humans in their evolutionary context. Opsins found in humans are represented as white circles in the corresponding clades of Fig. 2.3.

Through a systematic process of collecting the most likely opsin proteins in vertebrates and verifying their orthology through multiple subsequent phylogenetic analyses, I identified 413 sequences from 29 different species that are situated evolutionarily in the monophyletic clade of opsins. Viewed through a birth-death evolutionary lens, the phylogeny shown in Fig. 2.3 can be seen as an illustration of 19 duplication and divergence events, each giving rise to a novel opsin protein. Support is generally high among these partitions, with the exception of the relationship between OPN4 and the other non-visual opsins (58/100). OPN4 often appears outside the two other major clades of opsins in maximum likelihood trees, but its placement with other non-visual opsins makes sense. There are additional duplications that have occurred within the protein clades shown, so there are in fact many more opsins than just the 20 categories in Fig. 2.3, most of which are present only in certain species and lineages.

There have also been gene loss events scattered throughout the tree that (together with duplications) have shaped the distribution of opsin proteins across these 29 vertebrate species, which become apparent at the sub-family level. Mammals have relatively few opsins overall, suggesting fewer duplications and greater gene loss in the mammalian lineage. Out of 23 opsin genes, eight are in the human genome: four Deep brain and visual opsins (*Opn1lw*, *Opn1mw*, *Opn1sw*, and *Rho*); two Photoisomerases (*Rrh* and *Rgr*); Panopsin (*Opn3*); Neuropsin (*Opn5*); and Melanopsin (*Opn4*). Based on this distribution of opsins and their phylogenetic relationships, it appears that at least one copy from each of the five sub-families is necessary for vertebrates; albeit fewer in mammals than birds and lizards, and the most in aquatic species including fish.

Five out of the eight opsin sub-families found in humans are represented in the lamprey genome, which suggests that these five are among the oldest of all the opsins: two visual (*Opn1sw*, blue cone opsin and *Rho*, low-light rod opsin); and three non-visual (*Opn3*, *Rrh*, and *Opn4*). Since lamprey is one of the most basal extant vertebrate species, it follows that the opsins present in both lamprey and humans are the oldest genes in the family, which thus provided the basis for subsequent duplications and losses over time. Lamprey also has three additional opsin genes that are likely ancient precursors to the modern-day family: parietal opsin (*Parietop*), vertebrate-ancient long wave-sensitive opsin (*Valop*), and opsin 8 (*Opn8*). The fact that these three are missing in mammals but *Opn8* is present in platypus supports gene loss of some of these early genes in mammals.

Genome sequencing and assembly issues are not trivial for comparative evolutionary genomics analyses, particularly in non-model organisms due to variable genome sequence quality. What appears to be a gene loss in one species may in fact be a missing chromosome segment in the genome assembly of that species. When searching for orthologous genes through protein sequence matches, it is also possible that incomplete or incorrect predicted protein sets contribute to apparent gene loss events. In order to protect against reporting these false positive gene loss events, I conducted both protein and gene searches for orthologues across species, and verified that a gene is missing from both the genome assembly and predicted protein set. While this method accounts for missing gene predictions, it does not resolve genome assembly errors.

The following sections profile the five major sub-families of opsins identified through this comprehensive phylogenetic approach: Deep brain and visual opsins, Panopsins, Photoismerases, Neuropsins, and Melanopsins. All five of these sub-groups are acknowledged in the Porter et al. (2012) review in various forms, with Deep

brain/visual and Panopsins branching in the "C-type category"; Melanopsins in the "R-type" category; and Neuropsins and Photoisomerases in the "Group 4" category. Presenting each of these sub-families in higher resolution, illustrating the individual genes that make up each group, and in which taxa they can be found, has never been done for this large variety of vertebrate species. The phylogeny of the species involved samples the vertebrate tree of life in every major taxonomic clade: teleost fishes, birds, lizards, amphibians, and mammals. It also includes taxa that can help date duplication and loss events such as lamprey and coelacanth, and taxa on long branches such as ostrich and platypus. In total, the 29 species sampled represent a broad array of vertebrate species while maintaining enough specificity within the major taxonomic clades to identify gene loss in individual species and lineages.

The figures in sections 2.2.1—2.2.5 are maximum likelihood trees (Figures 2.4—2.8), and the color legends show which species groups harbor copies of each gene. In order to provide a measure of confidence in the trees' topologies, bootstrap support values are given at each internal node. These values reflect the proportion of 1000 ML trees (constructed in PhyML by sampling amino acid sites in the multiple-species alignments with replacement) that have the same topology as the tree displayed. In order to run these large (time- and computationally-intensive) bootstrap analyses, 10 jobs in PhyML were run on each tree simultaneously, creating 100 bootstrapped trees per job. The resulting 1000 bootstrap replicate trees were then concatenated into a single multiple-tree file and analyzed in J. Felsenstein's *Consense* program to determine the overall bootstrap support values on each internal node for all 1000 trees.

Unlike Fig. 2.3, which is the majority rule consensus tree from *Consense* and its support values reflect the *number* of replicate trees out of 100 that share the shown

branching pattern, the trees used in the following sections are the trees with the highest likelihood value reported from PhyML, and support values show the proportion of 1000 replicate trees with shared topology. In most cases, the topologies of these ML trees agreed with the majority rule consensus tree, and are shown (rather than the consensus trees) so that branch lengths reflect the average number of substitutions per site.

2.2.1 *Deep Brain and Visual Opsins*

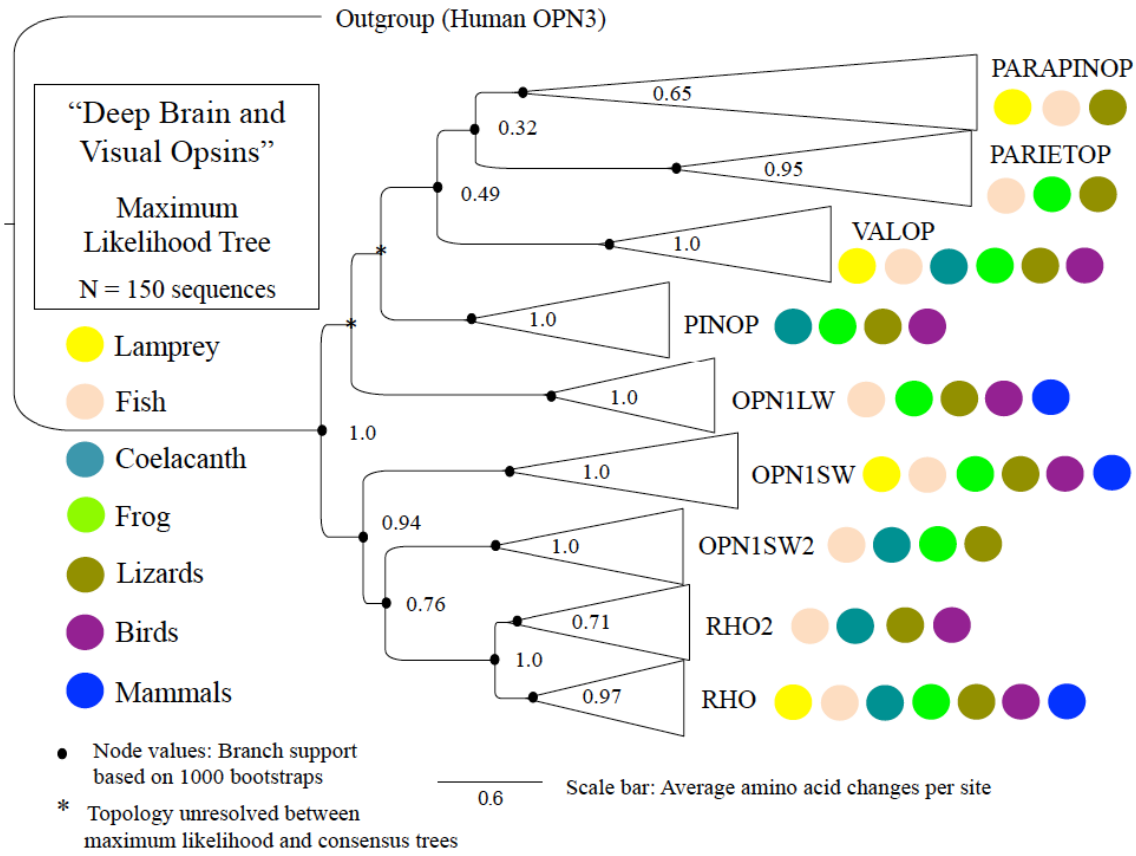


Figure 2.4. Deep brain and visual opsin maximum likelihood tree. These nine opsin genes form a monophyletic clade that includes the traditional "visual opsins" (rod and cone receptors). Every group of species sampled has at least three of these photopigments, and the color legend illustrates their distribution. Node values reflect the *proportion* of 1000 replicate trees sharing the shown branching pattern. Bootstrap (replicate) trees constructed in PhyML (N=150 sequences); amino acid sites were sampled with replacement for all 1000 trees. Support is high within and between each distinct protein clade, with the exception of the deep-brain opsins, which have some variable topology between the consensus and ML trees.

The deep brain and visual opsins comprise the largest category of opsins, some of which are involved in the visual pathway, and some of which are not. The phylogenetic relationship of these nine genes indicates that they are more closely related than any other genes, forming an opsin sub-family. Most are known functional photopigments found in the retina of the eye or other brain tissues. Mammals have three retinal photopigments: OPN1LW, OPN1SW (red and blue cone receptors, respectively) and RHO (low-light and peripheral vision rod receptor) but none of the deep-brain opsins. Deep brain opsins such as PARAPINOP and VALOP likely have ancient origins, as evidenced by their presence in lamprey and fish, suggesting that they have been lost in the mammalian lineage.

There is an entire field of research dedicated to the evolution of color vision that is tangential to my investigation of non-visual opsins at the gene family level. However, it is worth noting a major discrepancy between my findings and the literature. Contrary to studies that claim chimpanzees and gorillas have the same visual photoreceptors as humans due to their ability to distinguish between red and green in nature,[70] my results show only one copy of a long wave-sensitive opsin in both chimpanzee and gorilla. As shown in Figure A-7 (supplementary materials), the gene tree of *Opn1lw* indicates that chimpanzees have an ortholog of *Opn1lw* (red cone opsin) in humans, and gorillas have an ortholog of *Opn1mw* (green cone opsin) in humans. While it is possible that this is a novel finding, these tandem-duplicated photopigments on the X chromosome likely collapsed during sequencing or assembly of the chimpanzee and gorilla genomes.

2.2.2 *Panopsins*

The sub-family of opsins I have titled “Panopsins” include encephalopsins/panopsin (OPN3), and three sister versions of teleost multiple-tissue opsin (TMTOPS), which

together form a monophyletic clade of proteins with high bootstrap support. These opsins tend to be expressed in multiple tissue systems in species that have them, so the name *pan*opsin is both descriptive and referential. Since OPN3 is present in every species group sampled across vertebrates, including lamprey, it is likely the original panopsin, whilst TMTOPS proteins arose as subsequent paralogues. These latter proteins appear to have some importance in fish, birds, lizards, and amphibians, but are not necessary in all vertebrates since they are missing in the mammalian lineage; this protein likely arose early on and was subsequently lost in mammals.

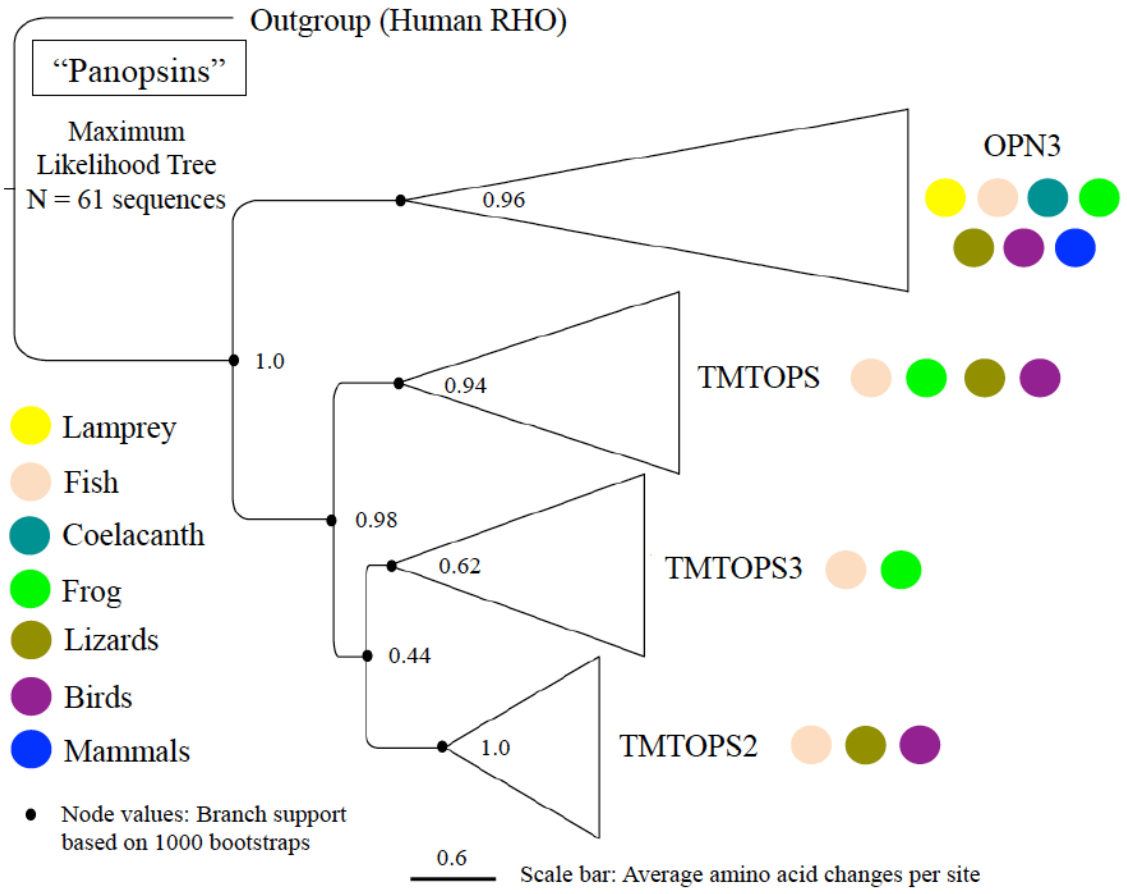


Figure 2.5. Panopsin maximum likelihood tree. Panopsins TMTOPS, TMTOPS2, and TMTOPS3 appear to be duplicated versions of OPN3 (the likely ancestral sequence) due to its prevalence across species. Node support values reflect the proportion of 1000 replicate ML trees (each with bootstrapped amino acid sites, N=61 protein sequences) with the displayed branching pattern.

In contrast to TMTOPS, OPN3 appears to be highly conserved, due to its ubiquitous nature across all species clades, including lamprey and coelacanth. However, some individual species have lost OPN3 (cow, brown bat, ostrich; see Fig.A-13 and Table A-11 in the Appendix), indicating that it is possibly under some variable selective pressures and not vital in all vertebrates. Section 2.3 will reveal a signal of positive selection in OPN3, which is consistent with the phylogenetic evidence for selective pressure.

2.2.3 *Neuropsins*

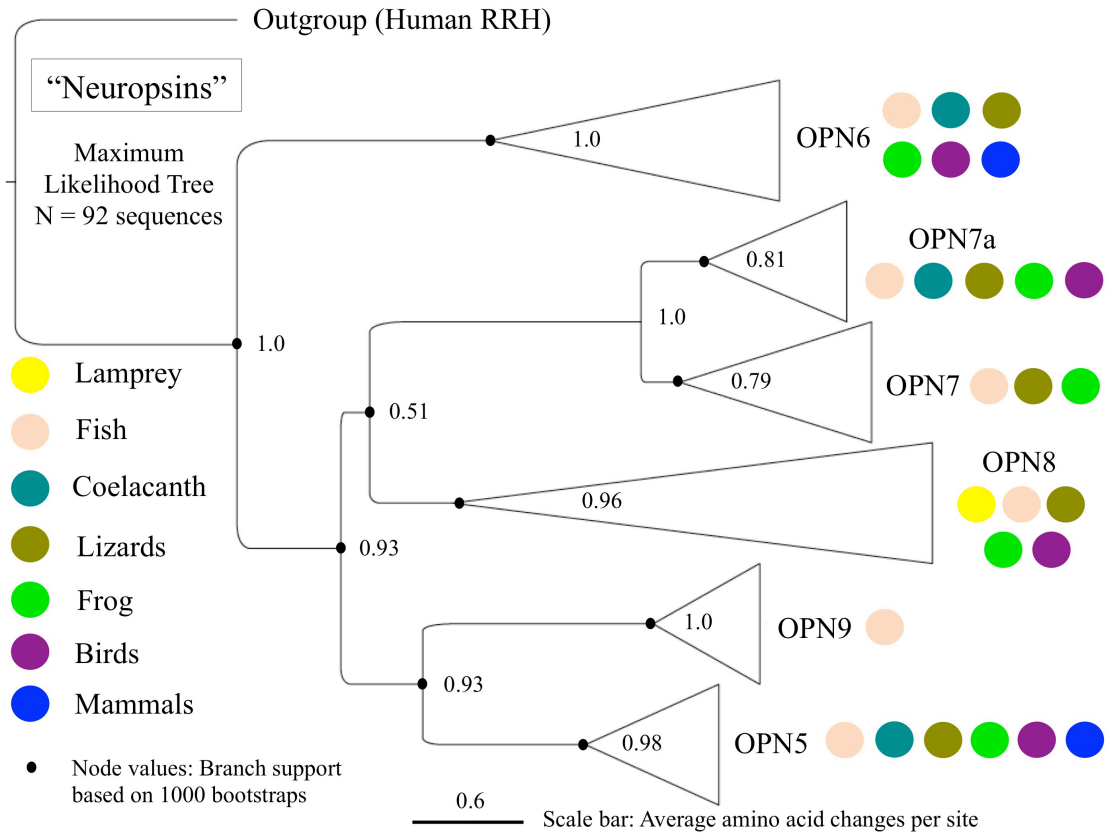


Figure 2.6. Neuropsin maximum likelihood tree. This sub-family is comprised of roughly six genes, and fish have many more paralagous versions, which are duplicates within the major clades that are collapsed on this tree. Node support values reflect the proportion of 1000 ML replicate trees (bootstrapped on amino acid sites, N=92 protein sequences) with shared branching pattern shown on this tree.

There are at least six distinct proteins in the monophyletic sub-family of "Neuropsins", which is comprised of OPN5 and its paralogues. The clade likely arose as a result of three

major duplication and divergence events, followed by subsequent gene births and deaths. The presence of OPN8 in the lamprey genome suggests that this was one of the earliest Neuropsins, conserved over time and across lineages, but was likely lost independently both in coelacanth and in the mammalian lineage, evidenced by its presence in birds and lizards. The phylogenetic positioning of OPN6 suggests this was possibly the original neuropsin from which the others arose through duplication, and was subsequently lost in several species including lamprey, and later in most mammals. The mammalian sequence in OPN6 comes from platypus, which appears to be the only mammal to have retained a copy. Multiple duplications have occurred across these protein clades (especially in fish); all copies are documented in Tables A-17—21; Figs A-21—25 in the Appendix.

2.2.4 *Photoisomerases*

Retinal G protein-coupled receptor (RGR) and Retinal epithelium-derived rhodopsin homolog, or peropsin (RRH) are the Photoisomerases of the vertebrate opsins. Both are present in mammals; RRH is present in every species sampled, and only lamprey and ostrich are missing RGR. As previously mentioned, the gene loss events are possibly artifacts of incomplete genome assemblies; however, it is also possible that RGR has been lost in lamprey and ostrich because it is somewhat redundant with RRH or other opsins, making its loss non-lethal in certain lineages. This is a profile of highly conserved genes, so it makes sense that their reported function is a basic biochemical one, as opposed to an adaptive mechanism responding to external light stimuli. While they may also function as light receptors, photoisomerases are likely biochemical converters that recycle the ligand for other light receptors. The specific accession numbers, gene trees

with internal topology for RRH and RGR are documented in Tables A-15—16 and Figures A-18—19 in the Appendix.

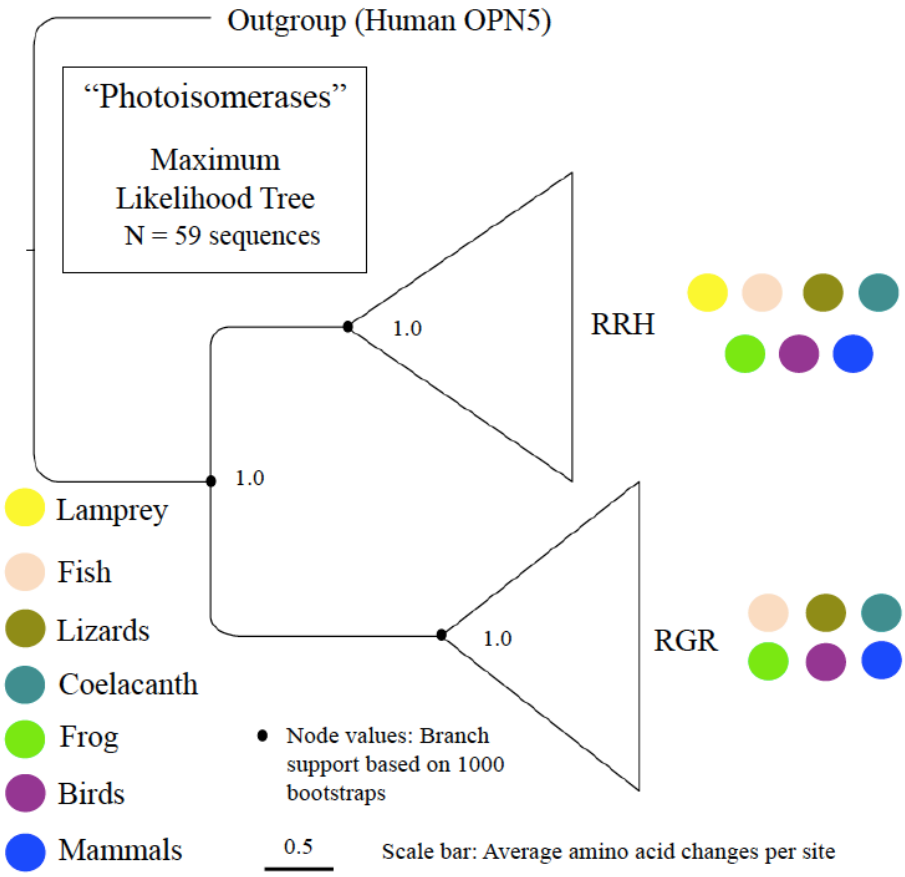


Figure 2.7. Photoisomerase maximum likelihood tree. Comprised of just two major genes, this opsin sub-family of photoisomerases is highly conserved across species types. Bootstrap support values derived from 1000 replicate ML trees, sampling amino acid sites with replacement (N=59 protein sequences). All 1000 replicates share this tree topology.

2.2.5 Melanopsins

Two major opsins are identified as "Melanopsins," including the original protein (OPN4) and multiple divergent forms of OPN4x. While all species sampled have at least one copy of OPN4 (and most fish have multiple), mammals do not have OPN4x. In contrast to other opsins, OPN4x does not appear to be monophyletic; however, the most

parsimonious explanation is that all of these OPN4x versions across the various lineages did at one point emerge from a single duplicated copy of OPN4, and that perhaps occurred long enough ago that they have greatly diverged from one another in the different lineages.

Due to the high level of conservation of OPN4 across all species, this is likely a very important functional protein. Zebrafish, platyfish, and cave fish all have three versions of this gene; suggesting it may be beneficial to have multiple copies in an aquatic environment (see Figures A-27—28 and Tables A-22—23 in the Appendix).

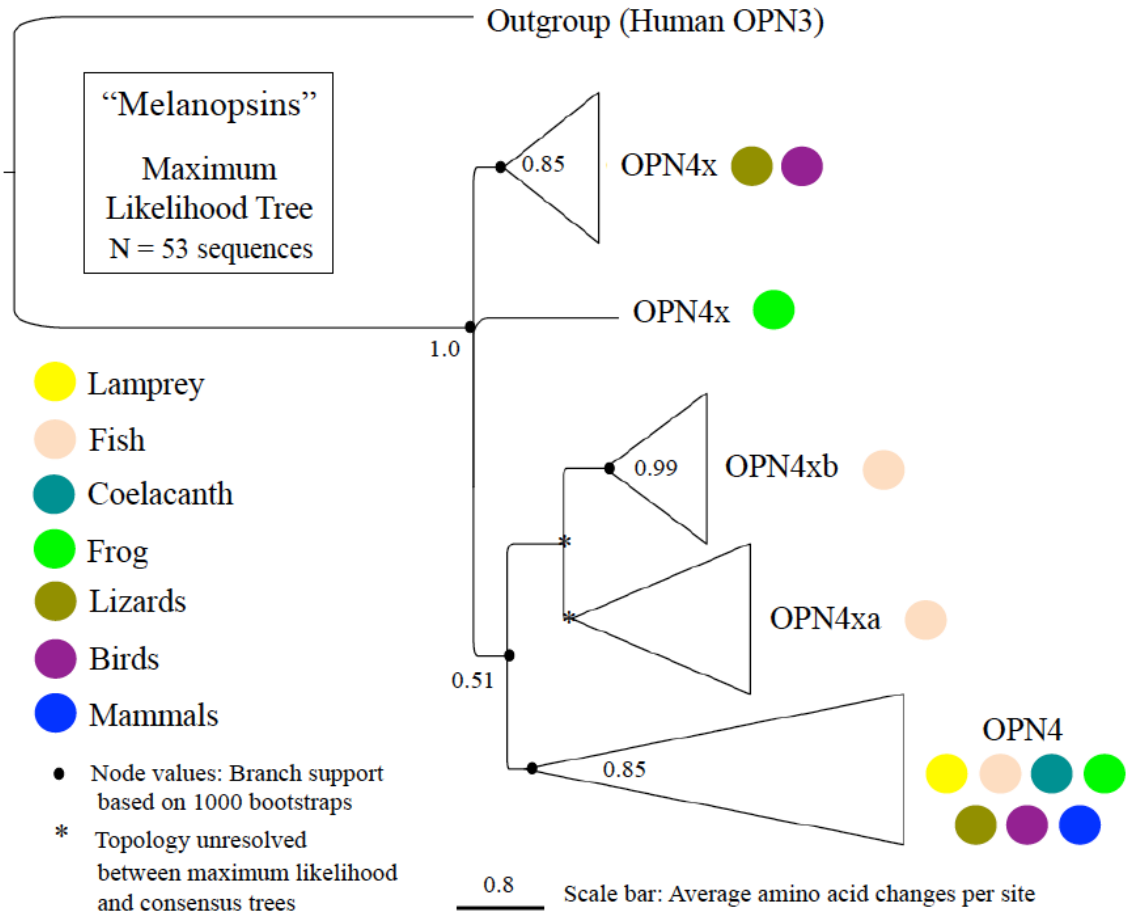


Figure 2.8. Melanopsin maximum likelihood tree. Bootstrap values derived from a consensus tree of 1000 ML replicates (N=53 protein sequences). Color legend illustrates the opsins present in each species group.

2.3 OPSINS UNDER NATURAL SELECTION IN VERTEBRATES

The birth-death evolutionary dynamics of opsins demonstrated in the previous section illustrate how versatile these proteins are, in terms of their distribution across phylogenetic clades and in light of their variable conservation across species. These results demonstrate that some opsins are important in different environmental niches, while others are less so, and a few of the opsins are crucial across all categories of vertebrate species. In order to drill down into the gene-level evolutionary characteristics within these opsins, it is then necessary to quantify their conservation. The mere presence of *groups* of opsins at the sub-family level in various species does not prove their relevance at the individual gene or protein level. In order to do this, I have conducted a scan of human opsins to determine whether or not natural selection appears to be acting on variation in these genes; and if so, what is the nature of this selective pressure.

In evolutionary and population genetics, computational methods are often used to assess selective constraint on a gene as an alternative to functional analyses using either in vitro assays or model organisms. While there is a certain level of controversy surrounding some methodologies that have been employed to detect positive selection in recent years,[71] it is widely accepted under Kimura's neutral theory (1968) that negative, or purifying selection, is implicit in the conservation of a gene over long periods of time.[72] This is intuitive because all regions of the genome in two species would diverge at roughly the same rate over time (assuming constant mutation rate and effective population size) without some force that maintains the identity of certain sequences in those diverging species. Thus, any divergence observed between DNA or protein sequences in different species should be the same across the whole genome or proteome, and should be proportional to the total divergence between the two species.

Under King and Jukes' (1969) theory of evolution, proteins under rigid structural and functional constraint are subject to stronger purifying selection than those under weak functional constraints, and thus would have a lower rate of amino acid substitutions than less selectively constrained proteins.[73] Thus, this further supports the idea that differences and similarities across protein sequences are a function of external pressure and relative functional importance. Additionally, according to Nei (2005), most functional genes in humans are under purifying selection because mutations that alter a necessary function are not likely to propagate in a population.[74] Assuming this logic, I surveyed the conservation of non-visual opsins across 24 divergent vertebrate species to determine if they have been under selective constraint throughout evolution as a way to infer their functional importance. Each analysis varied slightly in the number of taxa sampled, dependent on which species had a copy of the protein under investigation. Table 2.2 illustrates which species had available proteins to sample for each analysis, and where multiple copies were available, only one protein from each species was selected, based on having the highest percent identity with the human reference sequence.

Evolutionary evidence provides a starting point for investigating genes of unknown function. If non-visual opsins are highly conserved across species and have been under purifying selection throughout evolutionary time, this will provide a basis for further investigation into whether non-visual opsins are still under selection in humans. In this conservation analysis, I focused on the eight opsin genes (visual and non-visual) found in humans and conserved in other lineages, comparing them to orthologous sequences across a variety of species representing lineages of mammals, birds, amphibians, lizards, and fish. The justification for this selection of species is similar to that of the gene family-level analysis in section 2.2; a wide variety of species from

different lineages increases the sensitivity of the test, while limiting the analysis to vertebrates and including multiple taxa from each lineage increases specificity. All sequence data used in this analysis came from publicly available genome (coding) sequences and predicted protein sets from the Ensembl website[75] and the phylogeny of species included can be seen in Figure A-30 in the Appendix.

Table 2.2. Distribution of opsins among species sampled for conservation analysis. Although there are multiple versions of some proteins, relative to a single version in humans, only one protein per species was selected for comparison to its orthologous versions in other species. The protein sequence most closely related to human was preferred over more distantly related duplicate versions. Color shading corresponds to the evolutionary relationships between species, such that organisms closer to one another on the phylogenetic species tree share a color palette. Darker shading of a cell in the matrix indicates the presence of orthologous genes, and the number of squares indicates the number of orthologues in a species that correspond to a single human gene (per column). Lighter shading in a species row indicates an absence of the corresponding gene.

		Opn1lw	*Opn1sw*	*Rho*	*Opn3*	*Rgr*	*Rrh*	*Opn5*	*Opn4*
Lamprey	*Pteromyzon marinus*		□	□	□□		□		□
Zebrafish	*Danio rerio*	□□	□	□□□	□	□	□	□	□□□
Fugu	*Takifugu rubripes*	□		□□	□	□□	□	□	□
Platyfish	*Xiphophorus maculatus*	□□□	□	□□	□	□	□	□	□□□
Cave fish	*Astyanax mexicanus*	□□□		□□□□	□	□□	□	□	□□□
Coelacanth	*Latimeria chalumnae*			□	□	□	□	□	□
Frog	*Xenopus tropicalis*	□	□	□	□	□	□	□	□
Lizard	*Anolis carolinensis*	□	□	□	□	□	□□	□	□
Turtle	*Pelodiscus sinensis*	□	□	□	□	□	□	□	□
Chicken	*Gallus gallus*			□	□	□	□	□	□
Zebra finch	*Taeniopygia guttata*		□	□	□	□□	□□	□	□
Platypus	*Ornithorynchus anatinus*	□		□		□	□	□	□
Opossum	*Monodelphis domestica*	□	□	□	□		□	□	□
Armadillo	*Dasypus novemcinctus*			□	□	□	□	□	□
Elephant	*Loxodonta africana*	□	□	□	□	□	□	□	□
Mouse	*Mus musculus*	□	□	□	□	□	□	□	□
Gorilla	*Gorilla gorilla*	□	□	□	□	□	□	□	□
Human	*Homo sapiens*	□□	□	□	□	□	□	□	□
Chimpanzee	*Pan troglodytes*	□	□	□	□	□	□	□	□
Flying fox	*Pteropus vampyrus*	□	□	□	□	□	□	□	□
Brown bat	*Myotis lucifugus*	□	□	□		□	□	□	□
Dog	*Canis familiaris*	□	□	□	□	□	□	□	□
Dolphin	*Tursiops truncatus*	□	□	□	□	□	□	□	□
Cow	*Bos taurus*	□	□	□		□	□	□	□

The primary measure used to summarize conservation across species is dN/dS, or the ratio of non-synonymous substitutions to synonymous mutations. I calculated this

ratio using the CodeML package in Yang et al.'s Phylogenetic Analysis by Maximum Likelihood (PAML) program.[76,77,78] This measure provides evidence for selective constraint operating over long periods of evolutionary time and works best for species that are somewhat closely related. Because it is less reliable for species that are highly divergent, the outcome of this analysis is purely a hypothesis-generating tool that will drive the direction of subsequent analyses.

If the ratio of non-synonymous mutations (dN) to synonymous mutations (dS) is equal to one, the gene is inferred to be evolving neutrally because non-synonymous mutations are occurring at the same [neutral] rate as synonymous mutations and no selective pressure is acting on it to change the ratio. If dN/dS is significantly greater than one in a particular gene, then the gene is considered to be under positive, or Darwinian selection. If the ratio is smaller than one, the gene is considered to be under negative, or purifying selection. CodeML calculates this ratio from a multiple-species alignment of coding DNA sequences for a particular protein, and performs a likelihood ratio test (LRT) comparing two models of evolution; one where dN/dS is between zero and one, drawn from a beta distribution (M7) and another that includes this model but adds an additional site class that allows for a signal of positive selection at certain sites (M8). Only after there is a significant result of this test (sufficient evidence to reject the null hypothesis that the protein is either evolving neutrally or under purifying selection), can further analyses be conducted to pinpoint residues subjected to positive selection in the protein. The details and interpretation of these tests and are included in the *Materials and Methods: Opsin Evolution* section of the Appendix (p. 46).

The results of this analysis showed no evidence for positive selection among any of the visual pigments involved; *Rho*, *Opn1lw*, and *Opn1sw* all appear to be under strong

purifying (negative) selection with low average dN/dS values across sites. Similarly, there was no significant result for the non-visual pigment *Opn4* or the photoisomerase *Rgr*, which also appear to be under strong purifying selection with no evidence of positive selection on the protein. The remaining non-visual opsins, on the other hand, received a significant result for this first test; *Opn3*, *Opn5*, and *Rrh* all had sufficient evidence to reject the null hypothesis which excluded the possibility that dN/dS > 1. For these three genes, I then interpreted the results of a Bayes Empirical Bayes (BEB) analysis produced by CodeML on the same data, which provides the posterior probability that a particular site in the protein has a dN/dS (or omega) ratio greater than 1, signaling positive selection.[79] Out of the sites with dN/dS > 1 reported by BEB, I noted those with the posterior probability greater than 0.9, of which there were 17 across the three opsins.

Table 2.3. Summary results of Bayes Empirical Bayes (BEB) analysis in CodeML. The number of significant sites refers to sites in the alignment of each protein with a posterior probability > 0.95 that the omega value (dN/dS) is greater than one, indicating a signal of positive selection.

	Number of sites dN/dS >1	Number of significant sites (pp>0.90)	Location of significant sites (protein alignment)	Functional region of protein
Opn3	4	1	454	7th trans-membrane helical domain
Opn5	17	13	409-14; 417-20; 424; 429-432	C-terminus (intracellular)
Rrh	20	3	114; 116; 118	3rd trans-membrane helical domain

While it may be tempting to take these results at face value, there are a number of considerations that contribute to uncertainty in these findings and argue why they should be interpreted with caution. The most notable is the quality of the alignment and variable number of residues that are analyzed across sites. In some sites of the alignment many species have a gap such that residues from fewer taxa are analyzed, which contributes to a higher sensitivity and lower specificity; thus, increased false positives. One such case is the C-terminus region of *Opn5*. This is the functional region at the tail end of the protein

that gave a strong signal of positive selection from the BEB across many sites. However, the alignment in that region of the protein is not reliable because there are so many gaps across different taxa that the sequences are difficult to align (even by hand). As such, any results showing sites with positive selection in regions of a likely incorrect alignment cannot be interpreted with confidence.

Similarly, the "positive selection" results for *Rrh* are unconvincing. All sites in the alignment that are implicated in this protein are situated in a segment heavily inundated with gaps; only 6 out of 23 taxa in the alignment have a residue in this region, and they are from different taxonomic groups. The six species with sequences in this region are spread across mammals, fish, and reptiles, which makes the likelihood of this being a true insertion that was lost in the other 17 species less likely. In any case, the large amount of missing data alone should disqualify this region from being interpreted as having a true signal of positive selection.

Despite these negative results, the analysis of *Opn3* yielded more interesting findings. Regions of strong purifying selection (with little variation across species) flank the segment that showed a signal of positive selection in the BEB analysis of *Opn3*. In this case, the alignment is likely to be correct and thus one can have more confidence in interpreting the result. Additionally, the site of highest significance is in the seventh trans-membrane region of the protein, which has been implicated in previous studies to be involved in spectral tuning of photopigments. To verify its placement in this region of the protein, I obtained predicted trans-membrane helical regions using the Trans-membrane hidden Markov model (TM-HMM) prediction tool from the Center for Biological Sequence Analysis at the Technical University of Denmark.[80] Fig. 2.9 shows the TM-HMM plot of predicted helical domains in *Opn3* and the highlighted region

represents the seventh trans-membrane helical domain. Fig. 2.10 shows the multiple-species alignment segment that corresponds to the predicted seventh trans-membrane helical domain of OPN3, visualized in J. Thomas' program Bonsai.

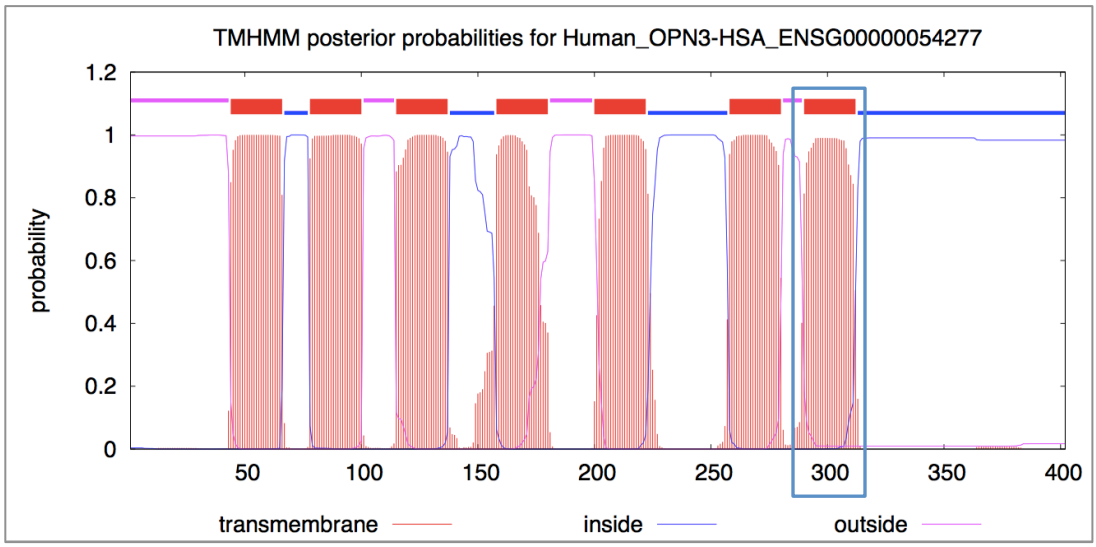


Figure 2.9. Predicted trans-membrane domains of OPN3. The helical domain outlined by a blue box in this TM-HMM plot corresponds to the region of the multiple-species alignment shown in Figure 2.10.

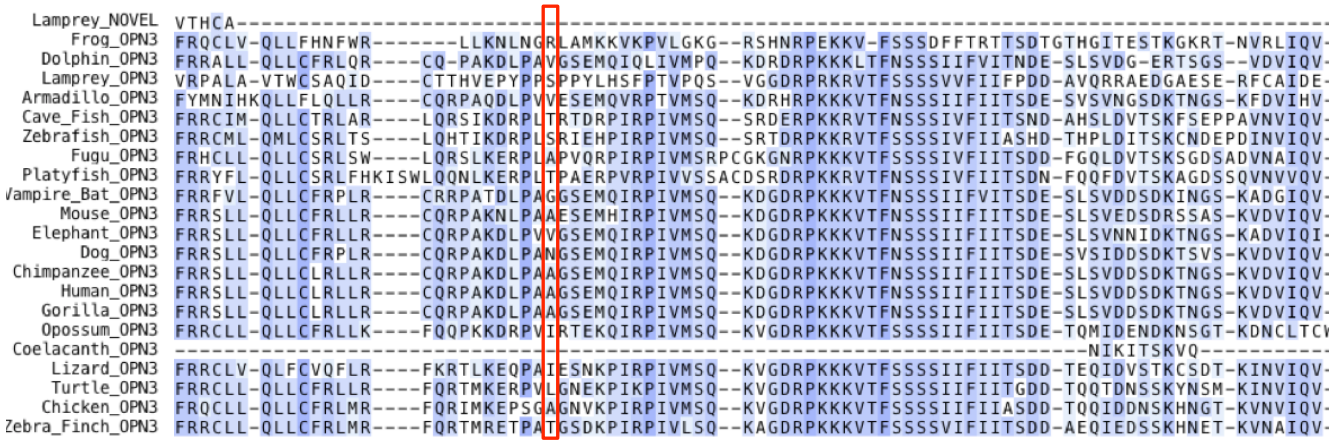


Figure 2.10. Multiple-species alignment of OPN3 sequences in seventh-TM helix. The column (amino acid site) outlined by a red box in the alignment corresponds to the site indicated by CodeML-BEB as likely under positive selection. Due to flanking by regions of greater conservation (darker blue) and strong alignment, this finding is likely a true positive signal. (Visualization from J. Thomas' Bonsai program.)

Utilizing structural information to interpret computational signals of natural selection in this way can help identify regions that are likely to be functionally important

for follow-up. For example, the trans-membrane regions and chromophore-binding pocket of opsins are likely to be areas of potential selective constraint. Thus, this finding of positive selection in the chromophore-binding pocket of *Opn3* is encouraging, and a novel finding worth reporting in the literature. It suggests that panopsin may be involved in evolutionary adaptation of species to environmental niches. This implies that it is indeed functional, and possibly related to sensory perception of stimuli.

2.4 EXPRESSION PROFILES OF HUMAN OPSINS

All findings reported thus far have been based on inter-species analyses across vertebrates, and although the opsins analyzed in the scan for natural selection are all present in humans, these findings alone do not prove that opsins are important in humans. Future research in this area will involve analyses of natural selection in opsins between and among human populations, but in the meantime we can learn about human-specific activity of opsins by looking at gene expression profiles of opsins. The increasing amount of expression data available in public databases make this type of analysis very practical and cost-effective, so I have conducted a survey of gene expression in human visual and non-visual opsins from three independent sources. The first is the Genotype Tissue Expression (GTEx) Portal, which hosts a plethora of rich data from tissue-specific gene expression data to variants that are associated with differential gene expression in human populations.[81] The second is the Human Protein Atlas (HPA), which compiles summaries of gene expression data across multiple platforms, and provides original content regarding tissue- and cell-specific gene expression in humans.[82] Finally, I consulted the RNA-Seq Transcriptome database of the human cerebral cortex hosted by the Barres Lab at Stanford[83] to get a sense of whether opsins are expressed in certain brain cell types.

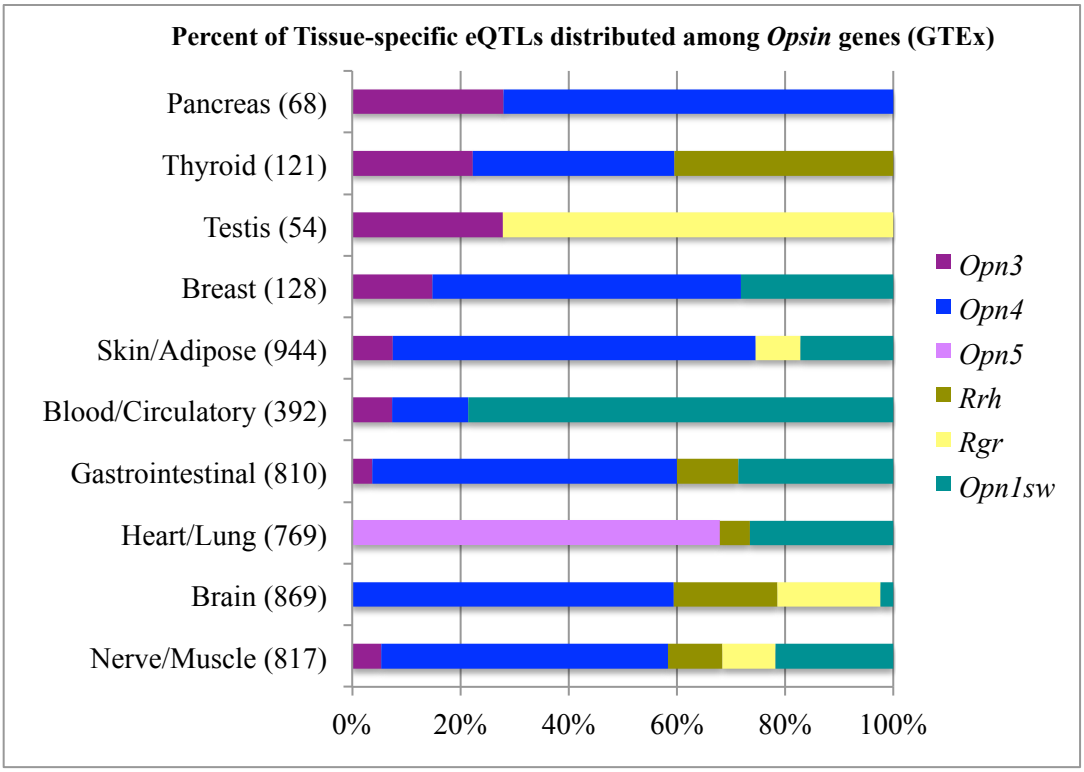


Figure 2.11. Expression trait loci (eQTLs) in opsins across 10 tissue types. These data were manually aggregated from the GTEx Portal website, combining tissue-specific eQTLs to form create 10 categories of tissue systems.

Recent studies have shown differential expression of non-visual opsins in human epidermal cells,[84] so I set out to determine whether there are specific variants in opsins that are driving differential expression, namely expression quantitative trait loci (eQTLs). The GTEx Project has found that eQTLs are common throughout the human genome, and their database seeks to combine genetic variant association results with genotype-linked expression data. The eQTLs reported in GTEx are validated by multiple methodologies including correction for multiple testing and significance thresholds that are specific to the type of eQTL and number of samples used to obtain the result. The results I present here are aggregated from all the opsins reported by GTeX to contain eQTLs in various tissue systems throughout the body. Fig. 2.11 illustrates the distribution of eQTLs in six

opsin genes across 10 categories of tissue types. *Rho, Opn1lw* and *Opn1mw* did not have any significant eQTLs in the database; thus *Opn1sw* is the only visual opsin represented.

In total, 4971 significant eQTLs were found in six opsin genes: 254 in *Opn3*, 2260 in *Opn4*, 522 in *Opn5*, 432 in *Rrh*, 362 in *Rgr*, and 1141 in *Opn1sw*. All of the eQTLs reported in *Opn5* were limited to tissue from the left ventricle of the heart, which is a surprising and somewhat puzzling place for a light-receptor called *neuropsin* to be. Similarly curious is the appearance of a large number of eQTLs in the blue cone photopigment (*Opn1sw*), which appear to impact gene expression in all the major tissue systems except for pancreas, thyroid, and testis. Most notably, it accounts for nearly 80% of the 392 eQTLs in opsins that affect the blood and circulatory system, but is least represented among the 869 eQTLs in brain, the organ situated most closely to where a retinal photopigment of this kind is supposed to be (the retina).

Another striking finding is that nearly half of the eQTLs in opsins are found in *Opn4* (melanopsin), and it is unclear whether this is due to an over-representation of studies on this gene, or a true enrichment of eQTLs. In any case, variants in this gene appear to be having an impact on several major tissue systems of the body, including the brain, nerves, skin, and the gastrointestinal tract. Again, the reason for this link to internal body tissues and variants in a non-visual photopigment remains to be determined. *Opn3* also has eQTLs distributed throughout most of the tissue categories, but in less abundance than *Opn4*. Finally, the photoisomerases *Rrh* and *Rgr* have eQTLs distributed mostly in different tissues, with variants in *Rrh* having an effect in thyroid, heart/lung, and gastrointestinal tract and those in *Rgr* impacting testis, skin/adipose tissue, brain, and nerve/muscle tissue. *Rrh* also has eQTLs in both brain and nerve tissue, and they appear to be in equal proportions to those found in *Rgr*.

The next resource I consulted for information about opsin activity in humans was the Human Protein Atlas, which combines expression data from its own dataset (HPA), GTEx, and the FANTOM5 dataset, which aggregates RNA data from the Cap Analysis of Gene Expression (CAGE) project. Data from all three of these resources are based on mRNA expression levels across all tissues, and specifically have data for brain, endocrine tissues, bone marrow & immune system, muscle tissues, respiratory system, liver & gall bladder, pancreas, gastrointestinal tract, kidney & urinary bladder, male reproductive system, breast & female reproductive system, adipose & soft tissue, and skin.

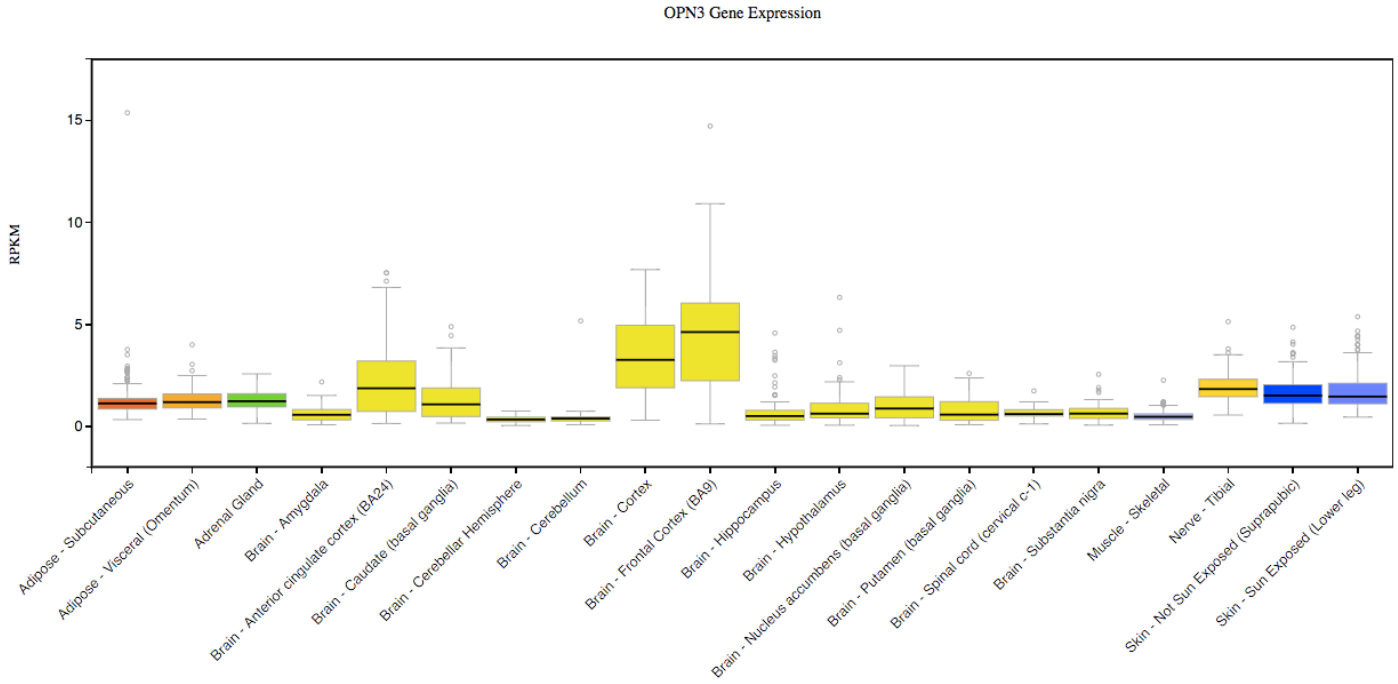


Figure 2.12. Expression of OPN3 across brain and other tissues. Brain tissues (yellow), adipose (orange), skin and muscle (blue), adrenal gland (green). GTEx data represented in reads per kilobase of transcript per million mapped reads (RPKM), sourced from the Human Protein Atlas (HPA).

I evaluated all nine human opsins using this aggregated dataset to try and unpack in which tissue systems of the body both visual and non-visual opsins have been found. By far the most highly expressed opsin across all sampled tissue system was panopsin (OPN3), which apparently lives up to its name in humans. It is most highly expressed in the brain, and most enriched specifically in the cerebral cortex. OPN1SW also had an impressive showing of ubiquitous expression across the major tissue systems (including the eye) although curiously it appeared most enriched in smooth muscle tissues such as in the gastrointestinal tract and arteries.

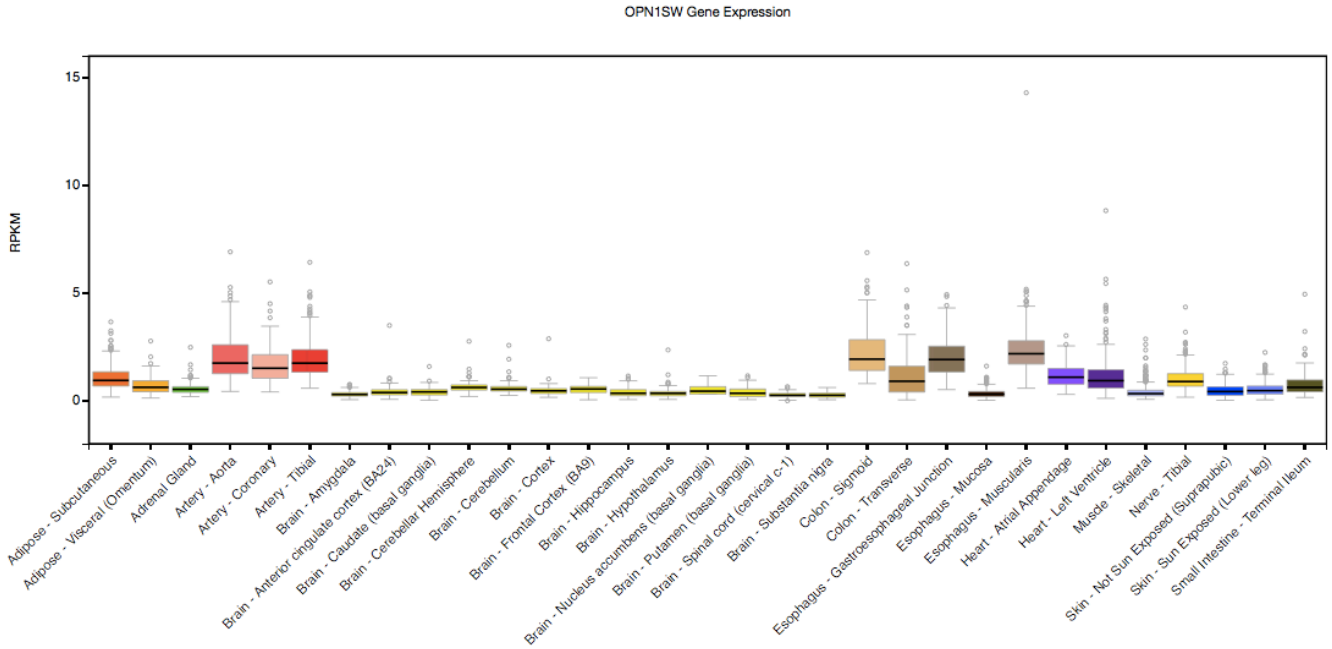


Figure 2.13. Expression of OPN1SW across brain and other tissues. Brain tissues (yellow), adipose (orange), arteries (red/pink), gastrointestinal tract (brown/gray), heart (purple), skin and muscle (blue), adrenal gland (green). GTEx data represented in reads per kilobase of transcript per million mapped reads (RPKM), sourced from the Human Protein Atlas (HPA).

Expression levels of OPN4, OPN5, RGR, and RRH were lower across the body (relative to OPN3 and OPN1SW), although all four of them showed strong expression in retinal tissue. Only the FANTOM5 database had signs of OPN1LW expression at

all, which was also located in the retina. OPN5 showed elevated expression in the left ventricle of the heart (in addition to the retina) but its expression was most observed in male reproductive tissues. OPN4 expression was most strongly observed in the retina, but it also showed enrichment in the brain (prominently in the basal ganglia), as well as heart and skeletal muscle and endocrine tissues. Between the photoisomerases, RRH was observed almost exclusively in the retina, although there was some elevated enrichment in the brain; while RGR showed among the highest expression levels of any of the proteins in the brain, most prominently in the cerebral cortex.

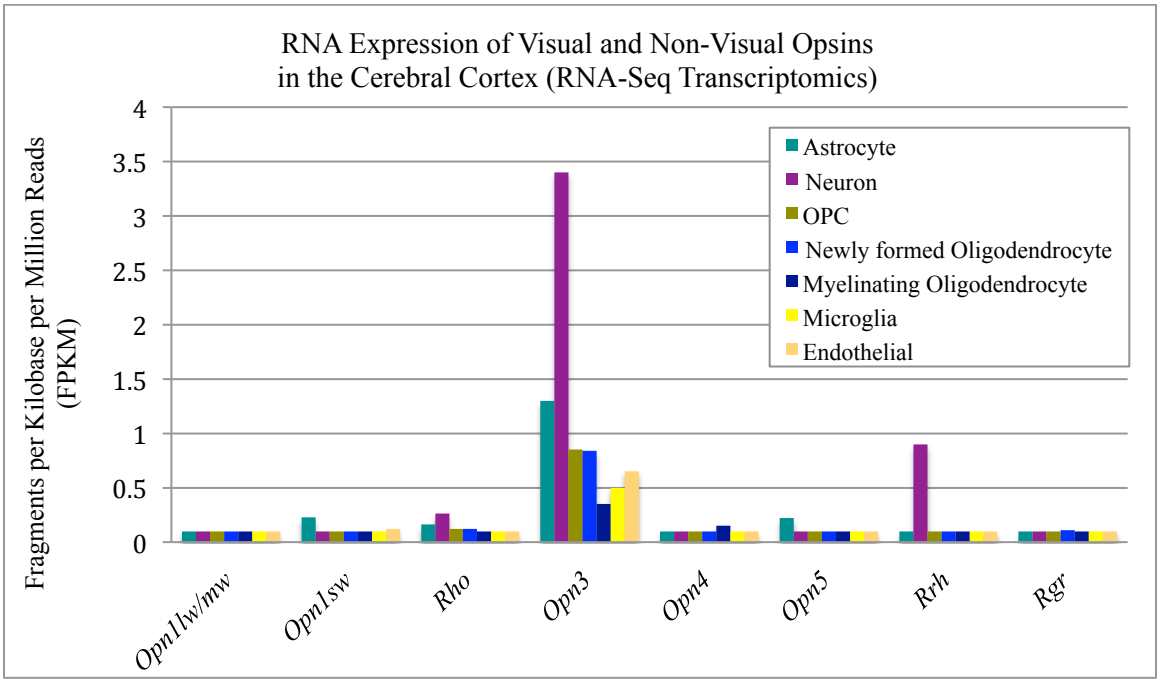


Figure 2.14. Opsin transcriptomics across cell types in the human cerebral cortex. The color legend indicates cell type in which each opsin is found via RNA-Seq, and the clusters of columns correspond to each gene. The results are presented on the y-axis in units of fragments per kilobase per million reads (FPKM). Genes with expression signals are *Opn3* (panopsin, a non-visual opsin) and *Rrh* (peropsin, a photoisomerase).

Throughout this investigation, I became more and more interested in the enrichment of opsins in the brain, and particularly the cerebral cortex. Thus, the final expression analysis I conducted involved a survey of opsin expression levels in specific

cell types in the cerebral cortex. Figure 2.14 shows the expression of each opsin, by cell type, in fragments per kilobase per million reads (FPKM) from the RNA-Seq Transcriptomics database hosted by the Barres Lab at Stanford University.

The results of this analysis were highly informative, as the expression levels of panopsin (OPN3) in all cell types of the cerebral cortex far outweighs any of the other opsins, and is particularly prominent in neuronal cells. These data are consistent with a trend of something interesting going on in OPN3. While, unfortunately, this research has not uncovered hard evidence of the precise function of OPN3 in humans, nor any of the other opsins, the combination of results from gene family phylogenetic analyses, natural selection scans, and gene expression profiles demonstrates that these proteins are indeed functional and biologically important across many tissue systems.

2.5 CONCLUSIONS AND HYPOTHESIS

It remains a mystery what light receptors could possibly be doing in heart muscle, reproductive tissue, and deep in the brain, but the fact is – humans have more non-visual light receptors, expressed more ubiquitously throughout the body, than visual photopigments. While it is possible that these non-visual opsins have been co-opted to perform functions other than detecting light in humans, evidence from phylogenetics, biochemistry, and functional studies suggest otherwise. Rather, it is perhaps more likely that there are interactions between external (or internal) light signals and biology, a relationship about which we still know very little.

One hypothesis I offer is that non-visual opsins may be the fundamental building blocks of a vast network of biophotonic activity throughout the body. Biophotons were first described in the 1990's, when researchers found experimentally

that plant and animal cells (including human epidermal) emit "ultraweak photons".[85] The presence of biophotons has more recently been demonstrated in the brains of humans, and proposed as the basis of an optical communication system through neurons, with myelinated axons serving as photonic waveguides.[86] Researchers at the Wuhan Institute for Neuroscience and Neuroengineering have implicated that human brains exhibit a spectral redshift in biophotonic activity, relative to other mammals, which may play a role in humans' higher cognitive processing abilities.[87] While this work is still preliminary, these authors and others claim that glutamate (an abundant neurotransmitter) can induce biophotonic activity, which would lead to photons of light playing a role in a quantum brain mechanism of advanced neural processing.

Viewing the human body as a holistic organism, it makes sense that such a system of activity would not be limited to the brain, but would radiate out into the central nervous system, vital organs, and skin. It seems to me that such a system would benefit from a network of light-receptive proteins expressed throughout the body. Thus, I propose that the non-visual opsins are genes responsible for light processing in the body, detecting photons emitted internally (or received from the environment), most likely at near-infrared wavelengths of electromagnetic energy. Given the deep phylogenetic history of the non-visual opsins beyond the first shared common ancestor of all extant vertebrate species, if my hypothesis is correct, then this system of light detection (and emission) could be fundamental to vertebrate biology. Finally, the signals of gene expression and positive selection in OPN3, in combination with its interspecies conservation and relatively close phylogenetic relationship to the visual opsins, suggest that *Opsin 3* may indeed be involved in an optical communication system connecting the brain, CNS, and major tissue systems through neuronal cells.

Chapter 3. SPurS: A NOVEL METHOD TO DETECT SHIFTS IN PURIFYING SELECTION

Biologists continue to debate the characteristics that define a unique species, and efforts to reveal molecular mechanisms of species divergence are ongoing.[88] While most existing methods involve comparisons of closely related species or natural populations of a single species, very few approaches have sought to investigate signals from distantly related species lineages. However, physiological and behavioral differences are more pronounced between more distantly related species, so this type of comparison may be better suited for detecting the molecular basis of observed differences between large groups of species. The recent sequencing of thousands of new species' genomes have made it possible to analyze signals of divergence over vast evolutionary time scales, which may elucidate the molecular underpinnings of major differences between lineages.

The aim of this chapter is to describe a novel method designed to detect *Shifts in Purifying Selection* (SPurS) and demonstrate its distribution across ~7500 highly conserved genes. I describe a rich dataset of multiple-species protein alignments across 76 mammals and non-mammals, with matched simulated alignments that allow for comparison to a null distribution. This also includes description of the pipeline used to conduct a SPurS analysis, as well as an online tool with built-in datasets that can be used by researchers with a range of backgrounds and bioinformatics skills.

A *Shift in Purifying Selection*, or SPurS, is the term I have coined to describe sites that are conserved within species groups and diverged between them. Thus, the *shift* from one residue to another between the two groups is conserved under purifying selection over millions of years. The statistic I developed to describe this phenomenon is loosely based on F_{ST}, a site-specific measure of divergence between populations of the same (or

closely related) species[89]. Rather than sample individuals from populations of a single species, SPurS analyses compare divergence between individual species belonging to the *same* category or clade, to divergence between species belonging to *different* categories. When these categories are phylogenetic clades, genes enriched for SPurS sites are candidates for loci underlying the functional and physiological differences that characterize divergence between the two lineages.

A1	x	z	x	y	x	y	y	y	x
A2	x	z	x	y	x	y	y	y	x
A3	x	z	x	y	x	y	x	z	x
A4	x	z	x	y	x	y	x	z	x
A5	x	z	x	y	x	y	x	z	x
B6	x	z	y	y	z	y	x	x	x
B7	x	z	x	y	z	y	x	x	x
B8	x	z	x	y	z	y	x	x	x
B9	x	z	x	y	z	y	z	z	x
B10	x	z	x	y	z	y	z	z	x

Figure 3.1. Conserved multiple-species alignment with SPurS site. Orthologous protein sequences (shared from a common ancestor) are aligned such that each row is a unique species (1-10) and each column is a site in the sequence with amino acid residues x, y, or z. The highlighted column illustrates a SPurS site, in which Clade A species have a conserved "x" and Clade B species have a "z".

Because SPurS sites are observed across deep evolutionary phylogenies, a single substitution that created the initial shift between groups of species is subsequently followed by purifying selection across all taxa after that distant point in time. These sites likely arise during brief periods of relaxed selective pressure or population bottlenecks, when substitutions are allowed to occur and persist. They may also arise when a novel substitution happens to be advantageous in a particular lineage at a particular point in time, and is conserved so long as conditions for its advantageousness persist.

The pattern of amino acid residues that segregate into two distinct clades or categories of species is not detected by methods designed to identify lineage-specific or recent positive selection between closely related species or within a single population compared to another. Similarly, methods to detect global purifying selection do not distinguish between sites on the basis of which residues are present in which categories of species (i.e. the species phylogeny). If there are such methods available, they have not been adopted into common practice. While others have developed methods to identify sites under purifying, positive, and diversifying selection across divergent species, the patterns observed in these cases are distinctly different than the one we describe.

For example, the dN/dS measure used to distinguish purifying and positive selection from a multiple-species alignment, most notably performed in Z. Yang's program PAML[90], identifies sites in coding sequences that have undergone multiple rounds of mutation and subsequent divergence. This method carefully takes into consideration transition/transversion rates at sites where both synonymous and non-synonymous nucleotide changes have occurred, and has the flexibility to consider site-rate heterogeneity. However, sites identified by this method as likely under "positive selection" in fact display a pattern of diversifying selection, in which a great variety of amino acids are present across the different species in an analysis. In the case of SPurS, there are typically only two residues present at a site across the entire multiple-species alignment (distributed in a clade-specific pattern), which does not provide enough of a diversifying signal for detection in PAML.

In addition to maximum likelihood approaches, Bayesian mixture models have been proposed to account for heterogeneity in evolutionary rates across sites and across lineages[91,92]. Site- and branch-rate methods are not applicable for identifying shifts in

purifying selection (SPurS) because rates do not vary across branches at SPurS sites since they are all under purifying selection, and also may not vary much across sites if the genes implicated are highly conserved. These methods are mostly intended to improve divergence estimates and phylogenetic tree topologies, which is not our objective.

3.1 Ψ STATISTIC FOR SPurS DETECTION

The concept of detecting within- and between- group differences in population genetics can be traced back to Sewall Wright's F-statistics in the 1920s[93]. The basic idea is to measure pairwise differences (heterozygosity) between individuals sampled from different populations to detect underlying structure. In this formulation of the SPurS statistic, I re-work the idea of pairwise heterozygosity between populations to measure the ratio of between- and within-group heterozygosity across large lineages in a phylogeny or other biological categories of species. This novel F_{ST}-like statistic to quantify shifts in purifying selection is called Ψ (Psi), conveniently an acronym for purifying selection shifts, and measures divergence between species from the same lineage group (e.g. within mammals) compared to divergence between species from different lineage groups (e.g., between mammals and lizards/birds). Ψ is calculated as a difference of *between*-clade heterozygosity and *within*-clade heterozygosity:

$$\Psi = \pi_B - \pi_W \tag{3.1}$$

where π_B is the ratio of the total number of pairwise differences in amino acid residues *between* species in different clades to the total number of comparisons made between clades,

$$\pi_B = \frac{\sum_{i\,\varepsilon\,C1,\ j\,\varepsilon\,C2}[\pi_{ij}]}{n_a n_b} \tag{3.2}$$

and π_W is the ratio of the total number of pairwise differences among species *within* the same clade, compared to the total number of comparisons made within clades,

$$\pi_W = \frac{\sum_{i<i'} \pi_{i,i'} + \sum_{j<j'} \pi_{j,j'}}{n_a(n_a-1)/2 + n_b(n_b-1)/2} \quad . \tag{3.3}$$

In both (3.2) and (3.3), n_a and n_b are the numbers of species analyzed from clades *a* and *b*, respectively. These are not always equal to the number of species from each clade in the multiple alignment; if a residue for a particular species is not present at the site being analyzed (represented as a gap in the alignment) then this taxon is not counted toward *n* in the relevant phylogenetic clade, or group of species defined as a clade for the purpose of the analysis.

In equation (3.2), π_{ij} represents a comparison of residues at site *S* between species *i* and *j* from two different phylogenetic clades (*a* and *b*, respectively). When the residues are the same, ($S_i = S_j$), $\pi_{ij} = 0$ and and when they are different ($S_i \neq S_j$), $\pi_{ij} = 1$. Thus, when $\sum \pi_{ij} = 0$, there is no heterozygosity between species from different clades ($\pi_B = 0$). When the number of divergent sites equals the total number of comparisons made, $\sum \pi_{ij} = n_a * n_b$, then $\pi_B = 1$, indicating complete heterozygosity between clades at this site, since all residues compared between phylogenetic groups are divergent from one another.

In equation (3.3), $\pi_{i,i'}$ represents a pairwise comparison of residues at site *S* from two species *i* and *i'* both sampled from clade *a*, and $\pi_{j,j'}$ represents a pairwise comparison of residues at site *S* from two species *j* and *j'* sampled from clade *b*. As above, when the residues compared between two species are the same ($S_i = S_{i'}$ or $S_j = S_{j'}$), there is no heterozygosity between them; thus $\pi_{i,i'} = 0$ or $\pi_{j,j'} = 0$. If the residues

between species are different ($S_i \neq S_{i'}$ or $S_j \neq S_{j'}$), then $\pi_{i,i'} = 1$ or $\pi_{j,j'} = 1$. If the number of divergent residues within clades *a* and *b* at site *S* is equal to the total number of comparisons among species within clades, then

$$\sum_{i=1}^{n_a-1} \pi_{i,i+1} + \sum_{j=1}^{n_b-1} \pi_{j,j+1} = n_a(n_a - 1)/2 + n_b(n_b - 1)/2 , \quad (3.4)$$

then $\pi_W = 1$, indicating that all residues from species within each clade are divergent from one another. This scenario is not likely when considering species within a phylogenetic clade, since more closely related species are more likely to share residues than harbor different ones.

The case of greatest interest, indicating a shift in purifying selection (SPurS), is one in which all species diverge between clades, and none diverge within clades. That is, when:

$$\sum_{i=1}^{n_a-1} \pi_{i,i+1} + \sum_{j=1}^{n_b-1} \pi_{j,j+1} = 0 \quad \text{and} \quad \sum_{i \,\varepsilon\, C1,\ j \,\varepsilon\, C2} [\pi_{ij}] = n_a * n_b ,$$

such that

$\pi_W = 0$ and $\pi_B = 1$; therefore

$$\Psi = 1 . \quad (3.5)$$

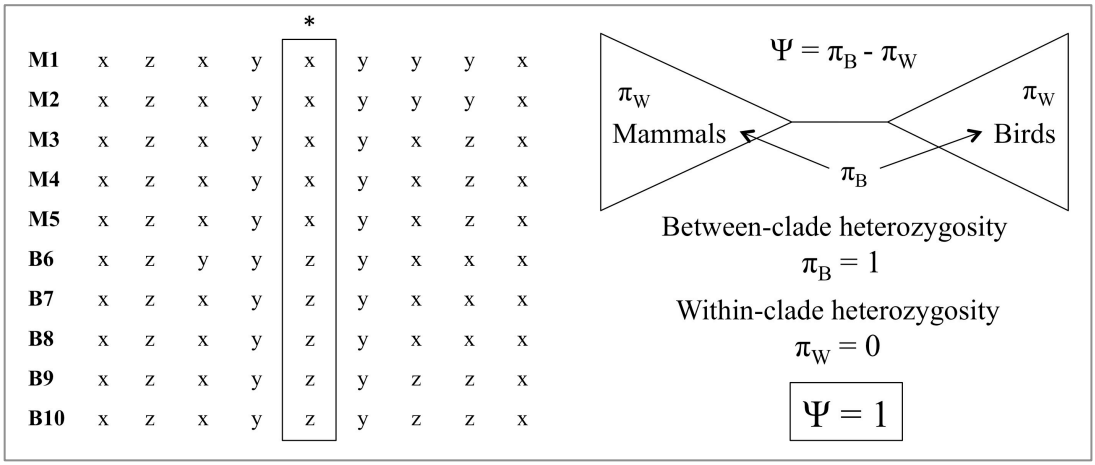


Figure 3.2. Calculating Ψ and detecting a shift in purifying selection. Multiple alignment of birds and mammals (left) and schematic of Ψ calculations in the case of a shift in

purifying selection (right). When Ψ = 1, there is complete conservation of an amino acid residue within each clade, and complete divergence between the two clades.

Sites in a multiple species alignment where $\Psi = 1$ (Fig. 3.2, equation 3.5) are characterized by two different residues that have been conserved within each of the defined phylogenetic clades or species groups. The various interpretations of other values of Ψ are outlined in Fig. 3.3, including when Ψ is less than, equal to, or greater than zero.

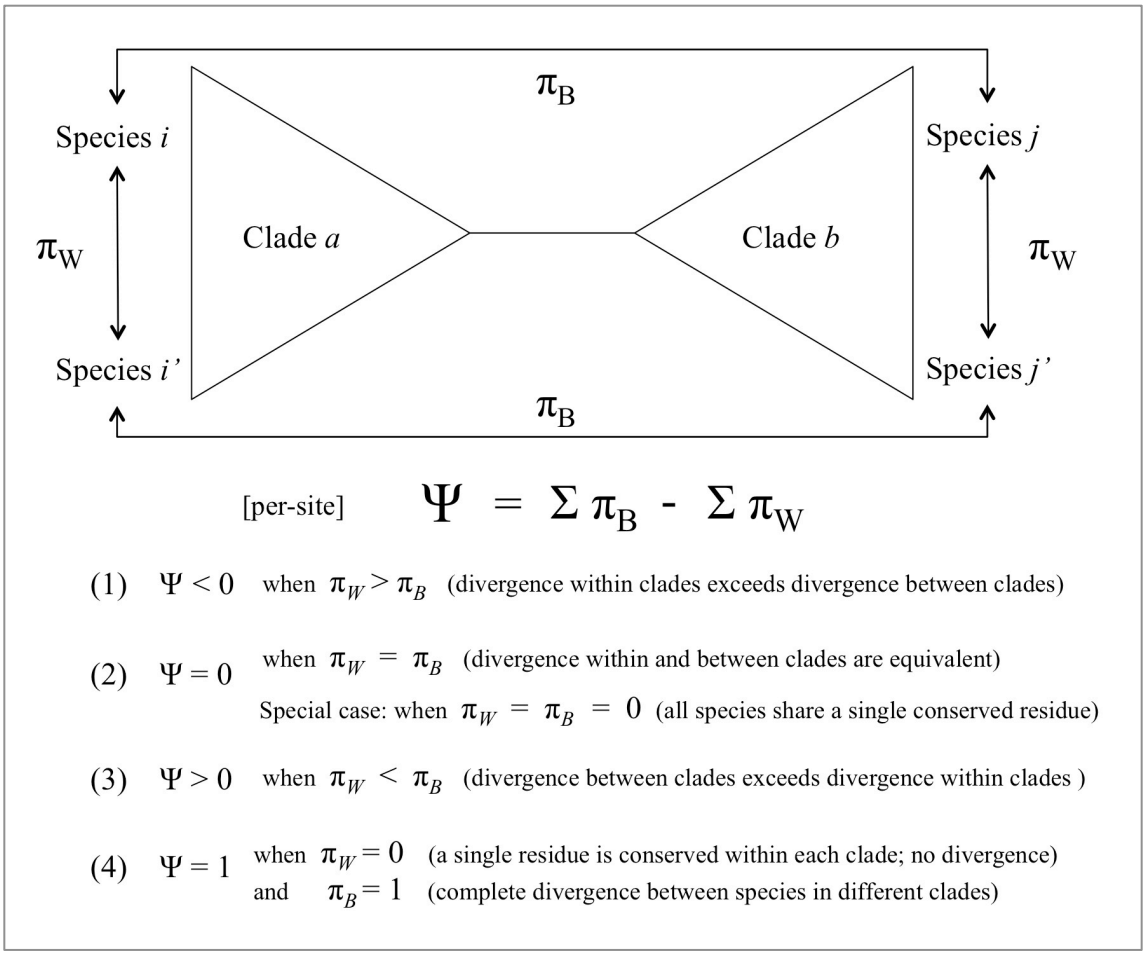


Figure 3.3. Diagram and Statistical Properties of the Ψ Statistic. The range of Ψ values and interpretation are described in cases (1) through (4), where each case represents a combination of divergence (heterozygosity, π) or conservation among species both within and between phylogenetic clades. The diagram above illustrates how Ψ is calculated, which is described mathematically in equations (3.1) to (3.5).

The obvious question that arose when developing a method to detect SPurS was why no one had done it before. Surely sites that show divergence between phylogenetic

clades, which are also highly conserved within those lineages, are likely to be associated with speciation events, and would be of interest for understanding biological differences between species. In discussions with J. Thomas and J. Felsenstein, we hypothesized that there have simply never before been sufficient genomic data on different types of species to make such a comparison statistically meaningful.

Patterns of divergence in a single gene that reflect divergence between species arise quite often by chance; that is, substitutions occur in the absence of any change in selective pressure, but appear to indicate such a change simply because the distribution of substitutions reflects the species phylogeny. For example, a multiple alignment of two distinct and highly divergent groups of species will most likely produce many sites with a pattern mimicking SPurS events; chimpanzee, gorilla, and human will almost always have the same residues compared to other mammals. So in order to determine whether SPurS are truly present, as opposed to sites that illustrate the same pattern by chance, the species included in a SPurS analysis must be somewhat distantly related. That is, there should be sufficient total branch length in the phylogenetic tree of the species included so that similarities between closely related species are not inflating the proportion of SPurS found. Thus, it has only been in the last few years that enough species' genomes have been sequenced at high quality to conduct such an analysis empirically.

3.2 GENOME-WIDE DISTRIBUTION OF SPURS SITES

In collaboration with Jim Thomas, I set out to determine the distribution of SPurS sites among genes conserved across 76 species that represent a broad sampling of vertebrates. Figure A-31 in the *Materials and Methods* section of the Appendix illustrates the full species phylogeny for taxa that are included in this analysis, and Table A-1 in the

Appendix lists all of these species' common and scientific names, as well as taxonomic categorization. All species data, including taxonomic information, species phylogeny (below) and protein sequence alignments, were provided by J. Thomas. I wrote the algorithms and programming required to conduct the analysis, and carried out all implementations of the program as well as subsequent statistical analyses.

As a novel statistical measure, the distribution of Ψ across the genome is yet unknown, as is the proportion of SPurS sites (where Ψ = 1). The proportion of observed SPurS sites will vary depending on the species chosen for analysis, the clades under comparison (and which genes are conserved among them), as well as which sites contain orthologous residues across those species in each gene. To estimate the genome-wide proportion of SPurS sites comparing mammals (including eutherian, or placental, and non-eutherian) to non-mammals (birds, lizards, crocodilians), I calculated Ψ for all sites where a sufficient number of species had a conserved residue. Sites with > 50% gaps across the alignment (or < 3 species residues per clade) were not given Ψ values; only the sites analyzed were counted toward the length of simulated alignments for each gene.

'Genome-wide' is a relative term; since every species' genome is unique, the entire genomes of various species will have variable numbers and combinations of genes. Thus, we selected a set of genes that are present in both humans and coelacanth for the genome-wide SPurS analysis. In theory, genes that have orthologous copies in both humans and the distantly related coelacanth also have the opportunity to be present in species descendant from this commonly shared lineage. All the species in this shared lineage are represented in a species tree made from concatenated exons (translated into protein) from across the genome in Figure A-31 in the Appendix. Gene loss is common across the vertebrate tree of life and many of the genes sampled did not have orthologous

copies in all 76 sampled taxa. Figure 3.4 illustrates the proportion of proteins (N=7484) that have various numbers of taxa (56—76) present in the multiple-species alignment.

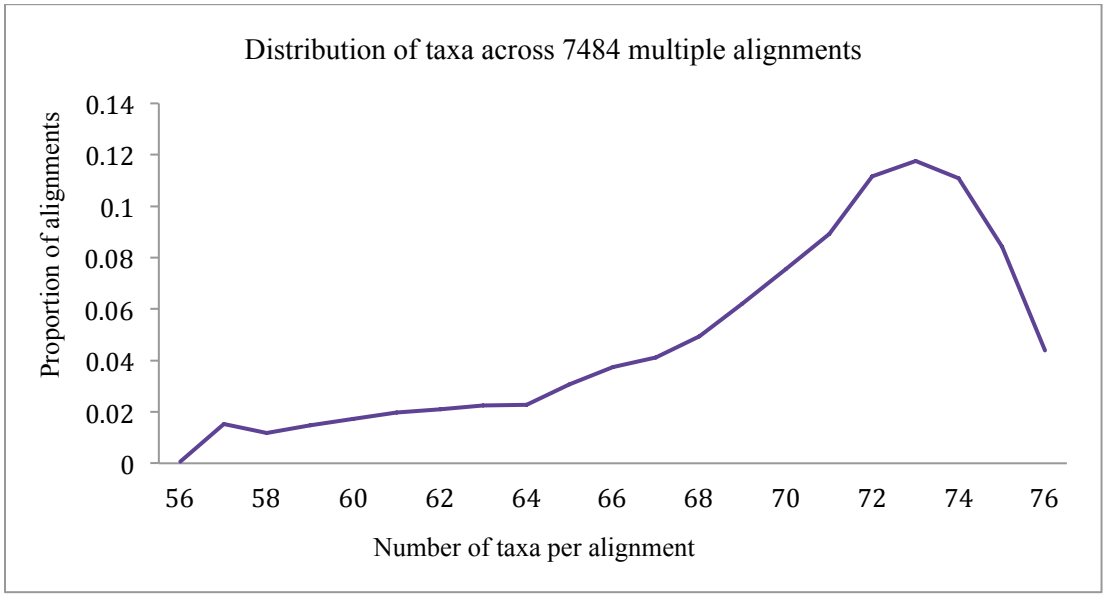


Figure 3.4. Number of taxa represented among all alignments. The number of taxa (species) with sequences orthologous to the human version of each gene is represented by the x-axis, and the proportion of genes sampled containing each respective number of sequences is on the y-axis. The minimum number of taxa included in an alignment is 57, and the maximum number of taxa included was 76 (the total number of species sampled).

3.2.1 *Simulated Dataset*

Mutation and genetic drift are confounding factors that may lead to a SPurS-like signal at a locus in the absence of any selective pressure. In this case, sites where one amino acid is observed in all species belonging to one phylogenetic clade and another amino acid is observed in all species belonging to a different clade occur simply because those happened to be the residues present in the respective founders of those clades. Depending on the particular sites, genes, species and clades being compared in a SPurS analysis, there will be a variable number of sites across genes that exhibit these SPurS-like qualities (where $\Psi = 1$). Therefore, it is necessary to simulate a dataset that is matched to the real data on the number of sites, genes, and species analyzed, using an empirical model of amino acid substitution without any true signal of shifts in purifying selection.

To deal with variable protein length and number of taxa from different parts of the evolutionary tree represented across the ~7K alignments, I created a tailored simulated dataset to match the real alignments in terms of species phylogeny, number of taxa, and alignment length. This process involved an initial simulation of 3 million sites of a multiple-species alignment under the JTT model of amino acid substitution, which estimates substitution rate parameters under the general time reversible model from real datasets.[94] This simulation was conducted in Seq-Gen[95] and conditioned on the species phylogeny of taxa included in the real data so that the simulated alignment contains all 76 species (and their respective protein sequences) that would be observed if the evolutionary processes driving their variation were only dependent on evolutionary relationships between those species and a general model of amino acid substitution that is uniform across branches and amino acid sites. The species phylogeny on which these simulated multiple-alignments were conditioned is found in Figure A-31 in the Appendix.

This species phylogeny was also multiplied by a total branch-length scalar designed to maximize the proportion of SPurS sites observed in the resulting simulated dataset. This provides a conservative null to which real genes can be compared in the search for SPurS enrichment. The scenario that truly maximizes the proportion of SPurS sites is a phylogeny that has long total branch lengths within clades being compared, and a short branch between the two clades. Our selection of species for this analysis was done with this in mind, particularly with the addition of non-eutherian mammals such as opossum and platypus (which reduces the branch length between mammals and non-mammals and increases the internal total branch lengths within the mammalian clade).

In order to arrive at this branch length scalar used to construct a conservative simulated “null” dataset to identify real genes enriched for SPurS sites, I created dozens

of simulated datasets at various scalars using the species phylogeny in Figure A-30 in the Appendix and determined empirically which scalar maximized the proportion of sites where $\Psi = 1$. Figure 3.5 illustrates the results of this process, with each bar representing a different dataset generated with a certain rate parameter. The final simulated dataset with 3M sites was created using a rate parameter of 0.2, and then sites from the full alignment were cut into individual simulated "genes", each matching a single gene in the real dataset with the precise species and number of sites that could be analyzed for SPurS.

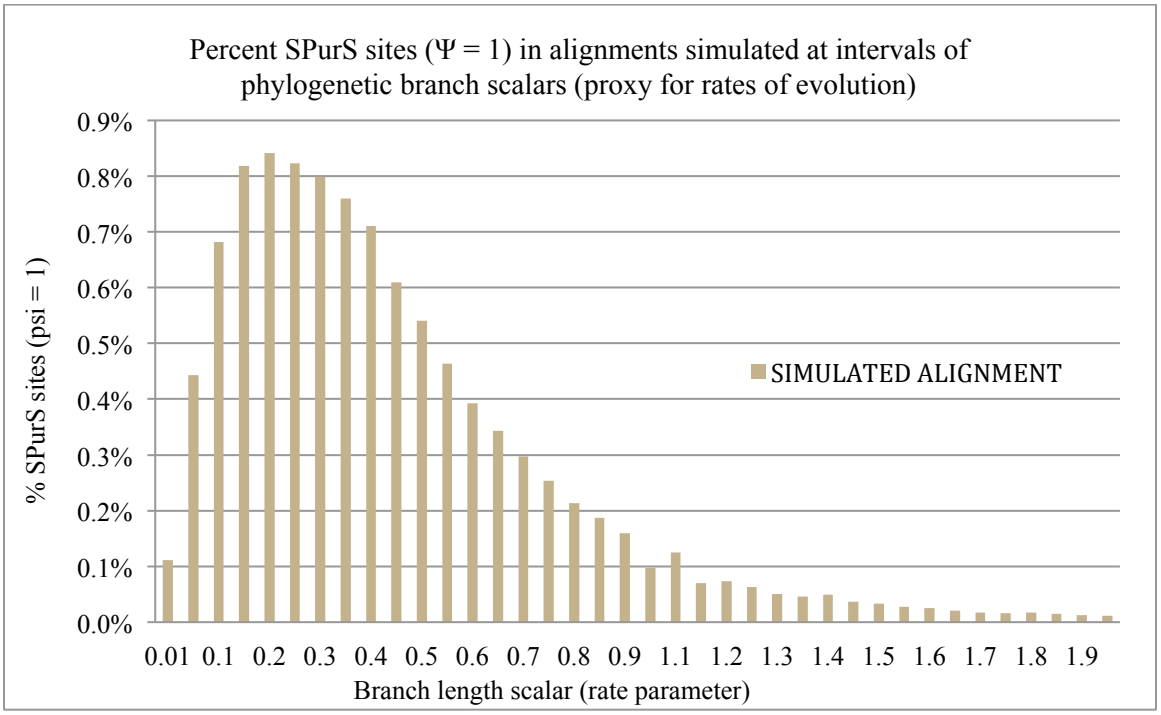


Figure 3.5 Simulated alignments to maximize SPurS sites. Simulated multiple-species alignments from Seq-Gen, conditioned on a real phylogenetic tree, at varying rate parameters. The maximum percentage of SPurS sites for any simulated alignment was between 0.8 and 0.9%, with the rate parameter 0.2. A denser distribution around the peak reveals a maximum of 0.84% SPurS sites, at a rate parameter of 0.23.

The choice of 0.2 as a branch-length multiplier reflects a desire to be conservative in the estimate of genes "enriched" for SPurS sites, relative to the simulated null expectation. That is, the branch length scalar of 0.2 maximized the probability of finding SPurS in the absence of a true selective phenomenon. Indeed, the simulated alignments

were *more* likely to contain SPurS sites than the real alignments, which is desirable for a conservative analysis since SPurS occur less frequently than expected. It then follows that any genes whose real protein alignments have a much larger proportion of SPurS sites than would be expected, based on the proportion of SPurS sites in its matched simulated alignment maximized for SPurS sites, is likely exhibiting a true signal of enrichment for shifts in purifying selection.

The maximum percentage of SPurS sites observed (comparing mammals and non-mammals) for any simulated dataset was 0.84%, with a rate parameter of 0.23. This means that using this rate parameter as a branch length scalar of simulated data maximizes the proportion of SPurS sites observed by chance, which is a highly conservative null distribution to draw from. Note that the percentage of SPurS sites observed when the branch length scalar was equal to 1 (matching the true branch lengths of the tree) was close to 0.1%. This suggests that the actual proportion of SPurS sites in the real data is likely much lower (on average) than the simulated data, which were constructed to maximize the occurrence of SPurS sites.

3.2.2 *Distribution of SPurS Sites*

For each of the 7484 conserved genes in our sample, I calculated the average Ψ across all sites. The median of these average Ψ across is 0.21; the minimum for any gene is -0.001 and the maximum is 0.67. The largest number of genes (2147, 29%) has an average Ψ across sites between 0.3 and 0.4, followed closely by average Ψ between 0.1 and 0.2 (2114, 28%). 1450 genes (19%) have average Ψ between 0 and 0.1, and 996 (13%) have average Ψ between 0.2 and 0.3. The minority of genes had Ψ in the tail ends of the distribution; only 5 genes had averages less than or equal to zero, 22 had averages

between 0.6 and 0.7. This concentration of average Ψ in the low end of the distribution indicates that the level of divergence between species between different clades or groups is only slightly larger than between species in the same clades or groups. Since these genes are highly conserved across the vertebrate tree of life, it makes sense that there is overall low divergence between species on average, across sites. The next step is to determine whether there is heterogeneity in Ψ between sites, and what the distribution of Ψ across sites looks like between real and simulated data.

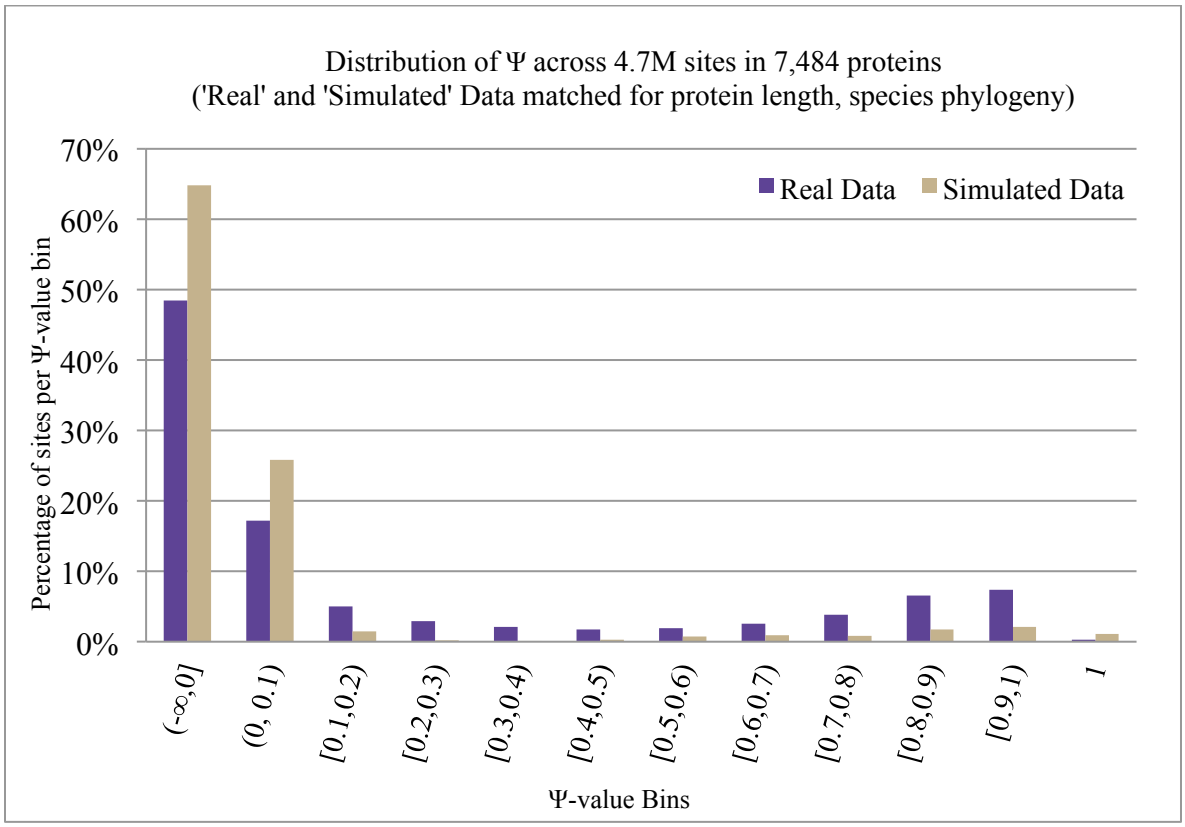


Figure 3.6 Genome-wide distribution of Ψ among real and simulated data. The y-axis represents the proportion of sites falling into ordered categorical bins of Ψ ranges (x-axis). Each simulated multiple-species alignment in the analysis was designed to maximize the proportion of Ψ values = 1 (branch length multiplied by 0.2) and matched to a real alignment on protein length and number of taxa.

Figure 3.6 illustrates the distribution of Ψ across sites with sufficient species (4.74M sites) for real and simulated multiple-species alignments. In the simulated dataset, the

total percentage of sites observed with Ψ = 1 was 1.1%. This was significantly higher than the real dataset (X^2 = 20250.08, p-value < 1.3e-311[§], df = 1), with 0.32% of all sites with Ψ = 1. This means that the occurrence of SPurS sites in real genes is much less likely than expected under the null. There also appears to be a larger proportion of sites in the real data (compared to simulated) in the upper tail of the distribution, where Ψ > 0.7. As such, I sought to determine whether certain genes are enriched for these sites with higher Ψ values, relative to all other genes and compared to the matched simulated genes.

If a subset of real genes contain a proportion of sites with Ψ = 1 that exceeds most other genes, and no such subset exists among the simulated data, then those genes may be interesting from an evolutionary perspective. For example, genes enriched for shifts in purifying selection between different species groups (mammals vs. non-mammals), may underlie the observed morphological differences between these two groups, thus are involved in the development of traits that give rise to divergence between lineages. These SPurS-enriched genes could also be involved in adaptation to changing environmental pressures, which favors lineage-specific conservation of certain amino acids relative to other lineages. The full extent of future applications of this novel method have yet to be realized, but I predict that the increasing availability of data will facilitate its use.

---

§ All p-values for chi-square tests were calculated in Python, using various built-in functions. The initial function used was chi2.df from scipy.stats, which has an upper limit of 17 significant digits in its calculation capacity. That means that any p-value smaller than 1e-17 automatically rounds to 0.0. In this version, the function chi2.sf from scipy.stats.distributions is used to calculate very small p-values, but this function also has an upper limit (311 significant digits). For chi-square values greater than 1424 (df=1), the resulting p-value is rounded to 0.0. Significance levels then estimated at p-value < 1e-311 refer to values whose number of significant digits exceeds the calculation capacity of the most precise available functions. The primary objective of these tests is to determine heterogeneity; so the lack of precision at these low p-values is not problematic.

3.3 SPurS TOOL IN THE GENOMIC HYPERBROWSER

The Python code for my SPurS program will be available for download with instructions on my GitHub account, but considering its potential useful application to a diverse array of biological questions, there may be researchers with minimal bioinformatics skills who would benefit from its use. As such, I have collaborated with Geir Kjetil Sandve and Diana Domanska from the Institute for Bioinformatics at the University of Oslo to integrate the SPurS program into a user-friendly tool on their Genomic Hyperbrowser GSuite platform for genomic data analysis. The Hyperbrowser is built and hosted on Galaxy, an open data-sharing platform designed for "accessible, reproducible, and transparent computational biomedical research" that allows users to freely access datasets and conduct analyses on their own datasets via an online user-interface.[96] The SPurS tool is currently available on the Genomic Hyperbrowser[**] with sample real and simulated alignments available for a handful of genes from the genome-wide SPurS analysis. There is also an option to upload new alignments, as well as revised control files, so that users may conduct their own analyses on independent datasets. Once the results of the genome-wide scan are published, all 7484 alignments and matched simulated datasets will be released through this platform, so that users will be able to search for genes of interest and conduct SPurS analyses on different combinations of species.

In its current version, there are 15 proteins available for analysis with matching simulated datasets, including melanopsin (OPN4) and a group of genes that are commonly investigated, according to several genomics blog posts on the subject. Users may select one gene at a time, and choose whether to compare mammals to birds, or

[**] https://hyperbrowser.uio.no/spurs

mammals to sauropsids (birds/lizards). There is also the option to create and upload an original control file, which can create new categories of species to compare. The output of this tool provides circular plots of Ψ values per site, ordered across all analyzed sites of the protein selected. It also produces a table and figures illustrating the distribution of sites among Ψ bins, including a histogram for both real and simulated data.

One of the benefits of this collaboration is the added feature of visualizing SPurS plots (conceptualized by me; designed in collaboration with and illustrated by D. Domanska in HTML), which allows for the immediate visual comparison of the distribution of SPurS across a protein. A play on associations with the name of this method, SPurS plots will produce circular graphs that resemble spurs on a boot, with regions of low Ψ interspersed by spikes where SPurS sites occur. While SPurS plots from simulated data also resemble real-life spurs, the idea is to visually compare the two plots and determine whether there appear to be more spikes in the real data relative to the matched simulated data. All results, including ordered vectors of Ψ for both real and simulated data, are available for download on the results page. Figure 3.7 illustrates the results of two SPurS analyses from the Hyperbrowser tool interface, one for FAM81A and the other for SLC1A3.

The resulting SPurS plots (Fig. 3.7) were calculated on multiple-sequence alignments of FAM81A and SLC1A3 across 76 species, with Ψ calculated between mammals and birds. Without any statistical analyses, an eyeball assessment of these two SPurS plots (simulated vs. real data) reveals a notable difference in the proportion of sites exhibiting high Ψ values for FAM81A, but not for SLC1A3. These observations are confirmed by a Pearson's Chi-square test, which reveals a statistically significant difference between the proportion of SPurS sites (where Ψ = 1) in FAM81A (p-value =

0.0001) and not for SLCA13 (p-value = 0.84). The interpretation of these findings is that FAM81A is enriched for SPurS sites, relative to the null, while SLCA13 is not. This means that FAM81A may be involved in physiological or functional differences between birds and mammals, and should be followed up as a potential mechanism of speciation. While these analyses must be done one at a time on the Hyperbrowser tool, users can download the entire pipeline from my GitHub account and replicate the analysis or conduct a new study on other protein sets.

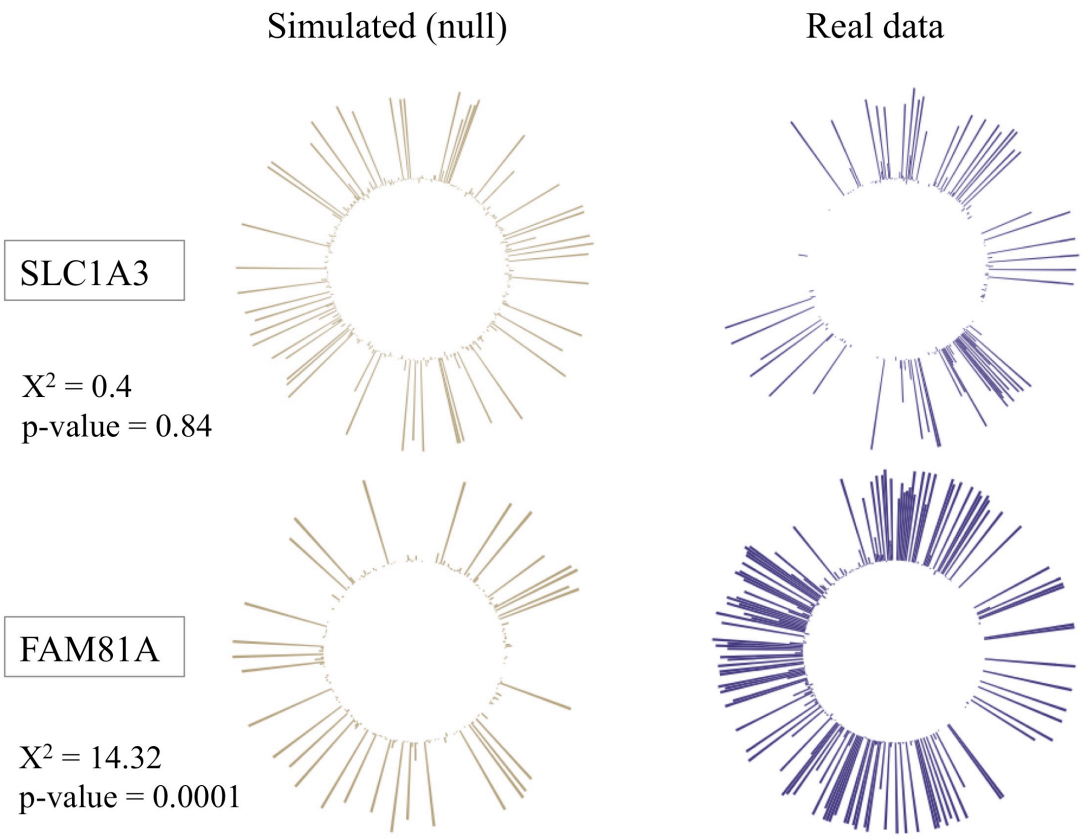


Figure 3.7. SPurS analysis results comparing mammals and birds. SPurS plots illustrate per-site Ψ along the sequences of two different real proteins (right) and their matched simulated datasets (left).

Plans are underway to integrate statistical analyses into the SPurS Hyperbrowser tool to provide users with greater ability to interpret the results of their SPurS analyses as well as other datasets of a similar nature. Future directions also include greater flexibility

with regard to data types and simulation options available. However, the first step is to publish the genome-wide SPurS results and make all data for that study publicly available on the Genomic Hyperbrowser for reproducibility, transparency, and additional species comparisons by the research community. The ultimate goal of this publication effort is to maximize the potential pool of SPurS users, so they are not inhibited by a lack of bioinformatics training, including non-computational life scientists and students. I anticipate that biologists with hypotheses about certain genes involved in speciation and divergence events may be interested in implementing these tools, and the user interface provided by the Genomic Hyperbrowser will provide enhanced ability to do so.

Chapter 4. CLUSTER MACHINE AND THE CONSERVATIVE CHI-SQUARE COLLAPSE

In order to determine whether a set of genes is significantly enriched for SPurS sites (Ψ = 1) compared to simulated data and all other genes, I developed a Python program called *ClusterMachine*, which implements an algorithm developed by Joe Felsenstein. According to Felsenstein, his algorithm is related to correspondence analysis, which [he notes] was introduced by H.O. Hirschfeld in 1935, reinvented by R.A. Fisher (1940) and E.J. Williams (1952), developed and popularized by J.P. Benzecri (1969), and described in a useful discussion by M.O. Hill in 1974.[97,98,99,100,101] The method of identifying axes of heterogeneity using principal components with weights as described by E.J. Williams[99] is one of the foundational concepts in this field, where the degrees of freedom for a likelihood ratio test are determined by adding up the total number of rows and columns, subtracted by 3 in order to account for various constraints (df = R + C – 3). In this formulation, all possible combinations of rows and columns are considered, with corresponding weights included in the calculation.

In contrast, the method developed by J. Felsenstein uses a more restrictive set of weights, such that rows and columns are collapsed and tested in a particular order and many fewer than all possible combinations are tested. Thus, in this more restrictive environment, the df threshold of R + C – 3 is conservative relative to its use in previous methodologies. I have called the core algorithm *Conservative Chi-Square Collapse (C3)* because it is based on collapsing a large contingency table into 2x2 chi-square tables and is conservative relative to correspondence analysis. The formal mathematical justification for this method will be described in a later publication co-authored by Felsenstein and myself, but here I provide an overview of how it is implemented in *ClusterMachine*.

The primary function of C3 is to separate data from a large contingency table or count matrix into two clusters that are significantly and maximally heterogeneous. A form of unsupervised machine learning, C3 uses the data in a matrix to identify orthogonal axes that maximize heterogeneity between two clusters of data. First, the matrix is collapsed into a 2x2 table at one set of vertical and horizontal axes, a chi-square test statistic is calculated, and the process is repeated at different orthogonal axes to find the maximum chi-square statistic of all 2x2 tables. Columns are ordered such that each column represents a range of values ("bins") and each row represents a single unit (gene).

Table 4.1. Sample *ClusterMachine* input matrix. Bins (columns) are ordered ranges of values for the variable of interest, min = 0, max = z. Rows represent individual units (1-n), each with a distribution of counts across the ordered columns. (Of course, columns can represent a variable of any range of values.) The C3 algorithm iteratively collapses this matrix into 2x2 chi-square tables and tests each one to find the global X^2 max.

	Bin 1 (0,1]	**Bin 2 (1,x]**	**. . .**	**Bin N (y,z)**
Unit 1	# observations in unit 1 with value in range (0,1]	# observations in unit 1 with value in range (1,x]	…	# observations in unit 1 with value in range (y,z)
Unit 2	# observations in unit 2 with value in range (0,1]	# observations in unit 2 with value in range (1,x]	…	# observations in unit 2 with value in range (y,z)
. . .	…	…	…	…
Unit n	# observations in unit n with value in range (0,1]	# observations in unit n with value in range (1,x]	…	# observations in unit n with value in range (y,z)

Applied to SPurS data, each row of the matrix is a gene, and each column represents a range of Ψ, such that each cell of the matrix is the number of sites that fall into a particular Ψ -bin. However, *ClusterMachine* is not tailored uniquely to SPurS data and can, in theory, be used to find heterogeneity in any large contingency table or count matrix. Table 4.1 illustrates the format of matrixes to be analyzed in *ClusterMachine*.

ClusterMachine is also flexible in terms of the format of input data. It will read in a count matrix with headers (as seen in Table 4.1, with option --MatrixFile) or a more

raw data file (with option --DataFile) where each row is comprised of a unit name, tab-separated from a subsequent vector or list of comma-separated observations. The program reads this data file and transforms it into a count matrix based on user-specified columns, or a list of ranges that the user desires as values delineating each column of the matrix (option --BinFile). Once the matrix is ready for analysis, *ClusterMachine* determines whether there is a simulated dataset to compare to input data (--SimCompare 'Y' and either --SimMatrix for a simulated matrix or --SimData for a simulated raw data file). In the context of a SPurS analysis, the real vs. simulated data formulation is meaningful. However, one could also replace the simulated dataset with a control dataset of another type (e.g., cases vs. controls) and the comparison would be similarly valid.

4.1 PARTIAL-COLLAPSE FOR OUTLIER DETECTION

When presented with two datasets, *ClusterMachine* will first conduct a *partial* C3 analysis by setting a horizontal axis *a priori* and searching for a vertical axis that maximizes heterogeneity between the two horizontal data clusters. For example, two input matrixes (one based on real, and the other on simulated data) will be collapsed into just two rows by adding up all the rows in each column bin. These rows represent the two datasets, one for real data, and the other simulated. This new matrix is then used as the input for a partial C3 analysis. Given these parameters, *ClusterMachine* considers only these two rows and then collapses adjacent columns with a moving vertical dividing line between them to create 2x2 tables. It then finds the combination of collapsed columns that maximizes the chi-square test statistic, such that the vertical axis represents the threshold of heterogeneity between the real and simulated data (two input rows).

Table 4.2. Collapsed matrix of real v. simulated data for partial C3 analysis. Each row represents a different dataset (originally Table 4.1), and each column represents a range of values that the observations of interest can take. In this case, the range of values is (0,z) and the *n* is the number of rows in each dataset. Each cell of the matrix is then the sum of all rows (per dataset) of the number of observations in each bin.

	Bin 1 (0,1]	**Bin 2 (1,x]**	**…**	**Bin N (y,z)**
Real Data	$\sum_1^n observations\ in\ range\ (0,1]$	$\sum_1^n observations\ in\ range\ (1,x]$	…	$\sum_1^n observations\ in\ range\ (y,z]$
Simulated Data	$\sum_1^n observations\ in\ range\ (0,1]$	$\sum_1^n observations\ in\ range\ (1,x]$	…	$\sum_1^n observations\ in\ range\ (y,z]$

Once *ClusterMachine* has created the matrix in Table 4.2 by pre-collapsing two input matrixes (each versions of Table 4.1) into a 2xC table where C is the number of column bins, it will conduct a series of 2x2 chi-square tests to identify maximally heterogeneous clusters of data. For each 2x2 chi-square test, degrees of freedom are calculated as the number of rows, plus the number of columns, minus three (df = R + C – 3). This is in fact the appropriate number of df for testing all combinations of rows and columns in the matrix, but since C3 has ordered columns and rows, and tests only a subset of all possible collapsed combinations, it is a conservative threshold for statistical significance.

The first 2x2 chi-square table in the case of a partial C3 between real and simulated data will set the first column (Bin 1 in Table 4.2) as Bin A, and then collapse all other columns in the same way as before (Bin 2 — Bin N in Table 4.2) into a new, aggregate Bin B. After calculation of the chi-square test statistic and p-value on this 2x2 table, *ClusterMachine* will record these values and shift data from the second column in the original matrix (Bin 2 in Table 4.2) into the aggregate Bin A, while removing it from aggregate Bin B. The next chi-square test will be conducted on this new 2x2 table, results compared to the first test, and so on. Only the maximum chi-square of this iterative process will be ultimately reported, along with the p-value and cut-off bin that ultimately

produced the highest X^2 value with a significant p-value. Because the number of rows in this version of the partial C3 will always be 2, and the number of columns in the original dataset stays constant, df for each of these tests is $C - 1$ (because df = R + C – 3; R = 2).

Conducted on SPurS data, the results of this partial C3 analysis showed significant heterogeneity between real and simulated data (X^2 = 865868.24, p-value <1.3e-311, df = 11), and this maximum chi-square was identified with Bin 2 as the cut-off. Looking at the distribution of Ψ across real and simulated data (Fig. 3.6), these results make perfect sense. The greatest difference between real and simulated data with regard to the proportion of sites in certain Ψ bins is clearly between the first two columns and the remaining bins. Thus, the fact that *ClusterMachine* identified these clusters as maximally heterogeneous between the two datasets is encouraging.

What this axis of heterogeneity reveals is that a large proportion of sites across these ~7500 genes are highly conserved across all species sampled, or that a similar amount of divergence is present among all species, regardless of their phylogenetic groupings. Recall from Fig. 3.3 that low Ψ values indicate sites in a multiple-species alignment where divergence among species within a clade is similar to divergence between species in different clades. When this value is zero, there is often no divergence between any of the species sampled. A finding of this nature is characteristic of unsupervised machine learning approaches, where seemingly obvious features of the data are highlighted by significant statistical findings that were not anticipated or sought after.

Useful in its own right, this functionality is not particularly helpful for assessing the statistical significance of SPurS signals between real and simulated data. What is perhaps more relevant is the ability of *ClusterMachine* to register a user-specified cut-off for collapsing bins horizontally, such that it will compare input data relative to a certain

range of values. In the case of assessing significant heterogeneity between real and simulated data relative to SPurS signals, the option --SelectBin can be used to specify Bin N ($\Psi = 1$) as the cut-off bin. Table 7 illustrates this pre-specified 2x2 chi-square table based on real and simulated data analyzed in my genome-wide SPurS analysis, and the resulting chi-square test was significant ($X^2 = 20250.1$, p-value <1e-311, df = 1). Rows and columns are pre-specified for this analysis, so it is a 2x2 chi-square with df = 1.

Table 4.3. Pre-collapsed 2x2 table of real and simulated data. Columns are collapsed into SPurS and non-SPurS sites; rows are aggregated for each dataset.

	$\Psi < 1$	$\Psi = 1$
Real Data	4729697	15183
Sim Data	4692967	51913

The difference between real and simulated data with regard to their distribution of SPurS sites was reported in section 3.2.2, and a pre-specified chi-square as described here was the method used to obtain that result. Next, I used *ClusterMachine* to determine whether there is heterogeneity among genes with regard to SPurS signals in both real and simulated datasets. Significant enrichment of SPurS sites in some genes, relative to others, is determined on the basis of heterogeneity identified by a partial C3 analysis (using the $\Psi = 1$ bin as a pre-specified collapse point in the matrix). The basic function of the algorithm is the same for this implementation as in the previously described partial C3 between real and simulated data, but its search for a vector of heterogeneity is horizontal, rather than vertical. That is, columns are pre-collapsed based on a user-specified threshold, but rows remain intact from the original input matrix. This analysis can be done both on real and simulated datasets, but each dataset must be analyzed

independently, as *ClusterMachine* is looking for heterogeneity among individual units *within* a dataset, as opposed to between them.

Table 4.4. R x 2 Table for a partial C3 in *ClusterMachine*. In this partially-collapsed table, the original input matrix has been collapsed horizontally into two columns, such that each cell in the first column (Bin 1) is an aggregate of cell counts in bins to the left of a pre-specified bin threshold (*T*) and cells in the second column (Bin 2) aggregate cell counts in bins to the right of bin threshold *T* (where *N* is the total number of bins in the original matrix and *n* is the total number of units, or rows).

	Bin 1 (0,x]	**Bin 2 (x,z)**
Unit 1	$\sum_{1}^{T-1} cell\ counts\ for\ Unit\ 1$	$\sum_{T}^{N} cell\ counts\ for\ Unit\ 1$
Unit 2	$\sum_{1}^{T-1} cell\ counts\ for\ Unit\ 2$	$\sum_{T}^{N} cell\ counts\ for\ Unit\ 2$
...	...	...
Unit *n*	$\sum_{1}^{T-1} cell\ counts\ for\ Unit\ n$	$\sum_{T}^{N} cell\ counts\ for\ Unit\ 3$

In contrast to the partial C3 comparison between two datasets, in which the input matrix is pre-collapsed vertically into two rows, this version of the partial C3 involves a horizontal pre-collapse of columns. In the former case, the columns to be iterated over for multiple chi-square collapse-and-test steps are already ordered by nature of the matrix. In this case, the rows need to be re-shuffled after the initial column collapse, so that they are ordered in ascending or descending order of the proportion of values falling into the testing column. For example, the partially collapsed matrix used to test for SPurS-enriched genes has two columns ($\Psi < 1$ and $\Psi = 1$). The number of rows equals the number of genes ($R = n = 7484$), so the degrees of freedom for all subsequent chi-square tests equals: R + C – 3 = 7483. Again, this is a conservative threshold for significance

since not all possible combinations of rows and columns are under consideration. In order for the algorithm to identify heterogeneous clusters of genes relative to their enrichment of SPurS sites, the rows need to be ordered by the proportion of observations falling into Bin 2. Thus, as in the previous supervised C3, the algorithm will crawl down the ordered rows of the partially collapsed matrix, collapsing them into 2x2 tables as it goes, and identifying the maximum chi-square test statistic among all tests. If there is significant heterogeneity, the program will report which genes are in the two maximally heterogeneous clusters, and provide a matrix file of each cluster for subsequent analyses.

When I conducted this analysis on simulated data (matched to the genome-wide SPurS alignments), *ClusterMachine* reported no significantly heterogeneous data clusters, which indicates uniformity in the proportion of SPurS sites among simulated alignments; most of these simulated proteins have roughly 1% SPurS sites. In contrast, the partial C3 found significant heterogeneity in the proportion of SPurS sites ($\Psi = 1$) among real alignments ($X^2 = 17271.32$, p-value <1e-311, df = 7483). In the first partial C3 run, 613 genes out of 7484 were identified as significantly enriched for SPurS sites, relative to all other genes with a range of 0.89–6.53% sites having $\Psi = 1$. The fact that these genes were identified in the real dataset as enriched for SPurS sites, while no such heterogeneity was detected across matched simulated datasets implies that there is some biological phenomenon driving differential proportions of SPurS sites across genes.

In a subsequent partial C3 run of these 613 SPurS site-enriched genes, *ClusterMachine* found 79 genes with a significantly higher proportion of SPurS sites relative to the original 613 SPurS-enriched genes ($X^2 = 778.82$, p-value = 1.75e-05, df=612) with a range of 2.28—6.53% SPurS sites (Ψ=1). However, a third run of C3 on these final 79 genes showed no evidence of heterogeneity among them relative to the

proportion of SPurS sites. Figs. 4.1 and 4.2 showcase the differences in distribution of SPurS sites between subsets of genes identified by *ClusterMachine.*

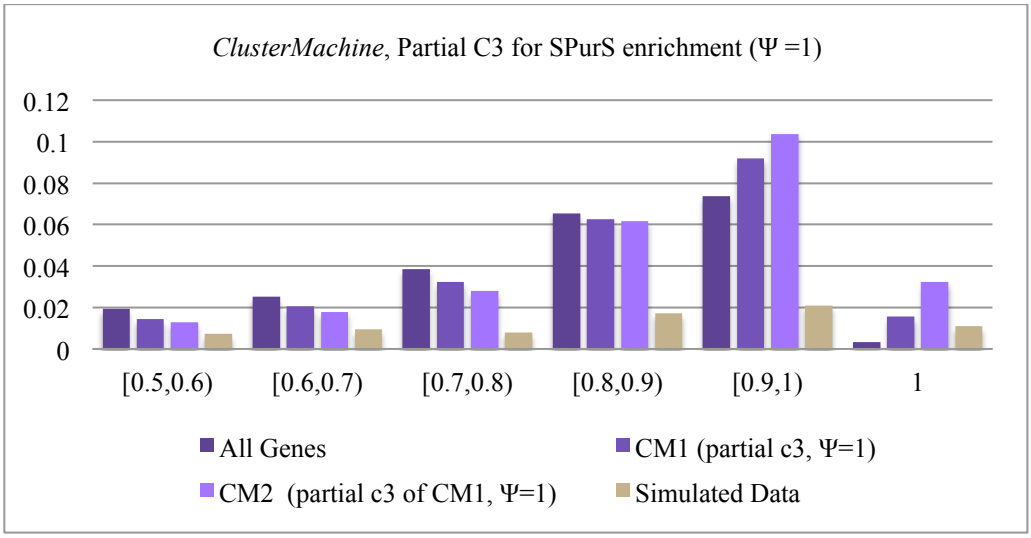


Figure 4.1. Shifting distribution of Ψ values between subsets of genes identified by *ClusterMachine.* The first and second iterations of partial C3 identified SPurS-enriched genes, represented by CM1 (*ClusterMachine* run #1) and CM2 (*ClusterMachine* run #2) in varying shades of purple. Simulated data are shown in gold, and all data are also represented in deep purple to provide a comparison to the distribution of sites observed in the gene clusters identified by subsequent iterations of *ClusterMachine* runs.

Gene Name	Sites Ψ =1
TMIE	6.54%
FAM81A	6.52%
ARVCF	6.50%
WNT2B	5.68%
GMPPA	5.48%
TMEM121	5.13%
RHOV	5.11%
PACSIN3	4.48%
CHAC1	4.39%
OLFML3	4.34%

Table 4.5. Genes with highest percentage of SPurS sites ($\Psi = 1$). Top 10 out of 613 enriched.

The results in Fig. 4.2 demonstrate the specificity of a supervised C3 approach. While the overall distribution of Ψ in these overlapping sets of genes does not differ by much, the differences between them with regard to sites in the Ψ =1 bin is striking (Fig. 4.1). Table 4.5 shows the top 10 most SPurS-enriched genes identified, and I will conduct follow-up studies to learn about what role these genes might play in species differences.

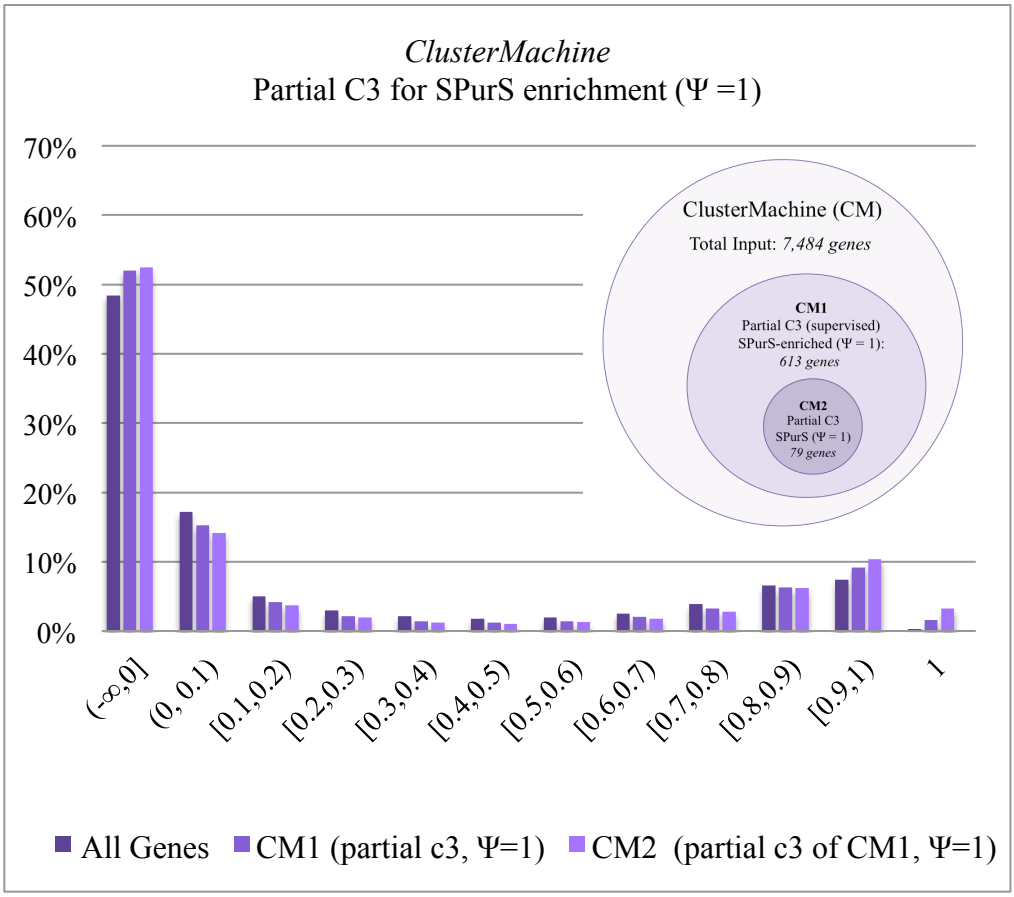


Figure 4.2. Distribution of Ψ across sites among genes identified by two partial C3 runs in *ClusterMachine.* Shows the process of iterating through genes to identify the most significantly enriched for SPurS sites (top right) and the overall distribution of psi-values.

This targeted analysis is useful for identifying proteins that are SPurS-enriched relative to all others in the dataset, and in theory can be applied to any large contingency table or count matrix from which one wishes to extract empirical outliers. For example, gene expression data have an inherent structure where genes are ordered in rows and respective relative expression values represent the amount of protein that is found in a variety of tissues in the body. If one were interested in identifying clusters of gene expression that were significantly heterogeneous across a range of tissues, then *ClusterMachine* could identify gene clusters that are differentially expressed relative to pre-specified or un-specified groups of tissues.

4.2 UNSUPERVISED C3 FOR DETECTING HETEROGENEITY

While the previously described cluster analyses are informative when there is an *a priori* hypothesis about the type of heterogeneity one is interested in finding, an unsupervised approach will identify clusters of data without any pre-specifications. The unsupervised C3 algorithm is essentially a combination of the two partial implementations (vertical partial C3 and horizontal partial C3) across all possible orthogonal axes treated as collapse-points for columns ordered by ranges of a variable of interest, and rows ordered relative to the proportion of values in the collapsed testing column. The input matrix is a complete dataset, with no pre-collapsed columns or bins. *ClusterMachine* then performs a series of 2x2 chi-square tests, treating each column bin as the cut-off threshold (horizontal collapse, Table 4.4), re-ordering the rows accordingly and conducting a partial C3 analysis on every ordered combination of rows (vertical collapse, Table 4.2). The vertical and horizontal axes of the 2x2 table that maximizes the chi-square test statistic over the entire contingency table are then orthogonal axes defining data clusters that are maximally heterogeneous over the whole matrix space. For those familiar with Principal Components Analysis (PCA), this can be thought of as roughly equivalent to the axes of the first two principal components.

In an effort to identify heterogeneity among genes relative to their distribution of Ψ, I conducted an unsupervised C3 in *ClusterMachine* on the full matrix of SPurS data, for both real and simulated data. In both real and simulated datasets, the first iteration of this analysis identified Bin 1 ($\Psi \leq 0$) as the vertical axis of maximal heterogeneity, with roughly half of genes (N=3490) falling into the cluster of genes enriched for Ψ less than or equal to zero (X^2 = 500168.21, p-value < 1.3e-311, df = 7483). Referring back to Fig.

3.6, these clusters can be observed in the difference in height between columns in the first bin compared to all others in the graph. What this allows us to do, however, is identify genes that are driving this heterogeneity in the distribution of sites across Ψ bins. In the biological context of SPurS data, genes enriched for sites with $\Psi \leq 0$ are likely to be highly conserved across taxa (no divergence between species of any taxonomic group), or the level of conservation across taxa is the same within taxonomic groups as between taxonomic groups (heterozygosity within clades is ≥ heterozygosity between clades). Refer to cases (1) and (2) in Fig. 3.3 for a mathematical refresher on this interpretation.

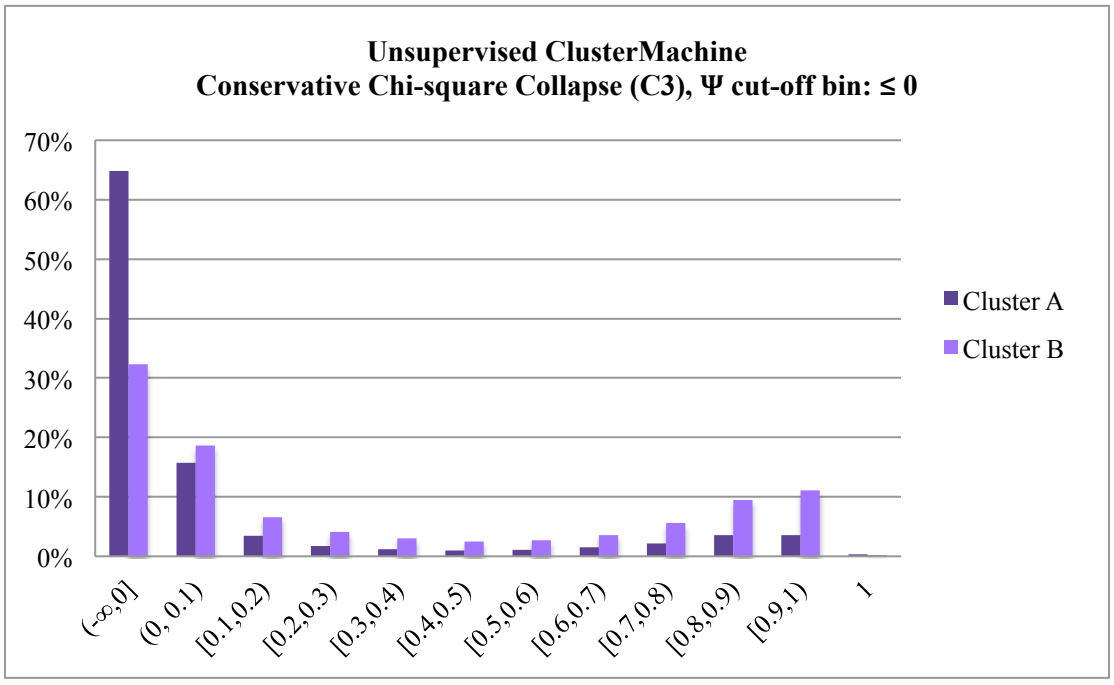


Figure 4.3. Differential distribution of sites across Ψ bins between gene clusters identified by an unsupervised C3. In dark purple, Cluster A represents genes enriched for sites with $\Psi \leq 0$, and the lighter purple, Cluster B, represents the remaining genes.

In light of this context, the cluster of genes identified by *ClusterMachine* as maximally heterogeneous compared all others relative to $\Psi \leq 0$ is likely to be informative. Since we know that genes under strong purifying selection are more likely than others to be functionally important, genes that are identified in this cluster of high

conservation, or equal divergence between and within phylogenetic clades, may play a role in fundamental biological processes across species. Figure 4.3 illustrates the difference in overall distribution of sites among Ψ bins between genes identified in the enriched cluster (Cluster A) and all others (Cluster B).

Although the vertical vector of heterogeneity that separates these two gene clusters is based on the proportion of sites in Bin 1 ($\Psi \leq 0$), there is a net shift in the distribution of the two clusters, such that the cluster of genes with sites enriched for $\Psi \leq 0$ are also much less likely to have sites with higher Ψ. This is likely due to the fact that different genes are under variable selective pressures, so the vast number of sites with $\Psi \leq 0$ is concentrated in highly conserved genes. In contrast, genes that are less concentrated with highly conserved sites may be more flexible evolutionarily, allowing more sites with high Ψ (indicating a possible shift in purifying selection between groups).

4.3 GENE CLUSTER INTERPRETATION AND BRAIN GENES

My initial hypothesis about genes enriched for SpurS sites was that these may be more likely than other genes to be involved in the brain, under the assumption that variation in "brain genes" might be responsible for differences between types of species, such as mammals and sauropsids. However, this turned out to be false. Fellow graduate student Saurabh Srinivasan at the Norwegian Centre for Mental Health Disorders (NORMENT) at the University of Oslo shared with me a list of 6175 genes that are expressed throughout the body, especially enriched in the brain and determined by experts in psychiatric genetics to be involved in important brain functions.

In order to determine whether these "brain genes" are over-represented among SPurS-enriched genes, I calculated the percentage of genes in outlier clusters identified

by *ClusterMachine* that are found on this list of brain genes. The SPurS-enriched gene cluster identified by a partial C3 conditioned on $\Psi = 1$ actually has a slightly lower proportion of brain genes (39%) than all genes combined (41%).

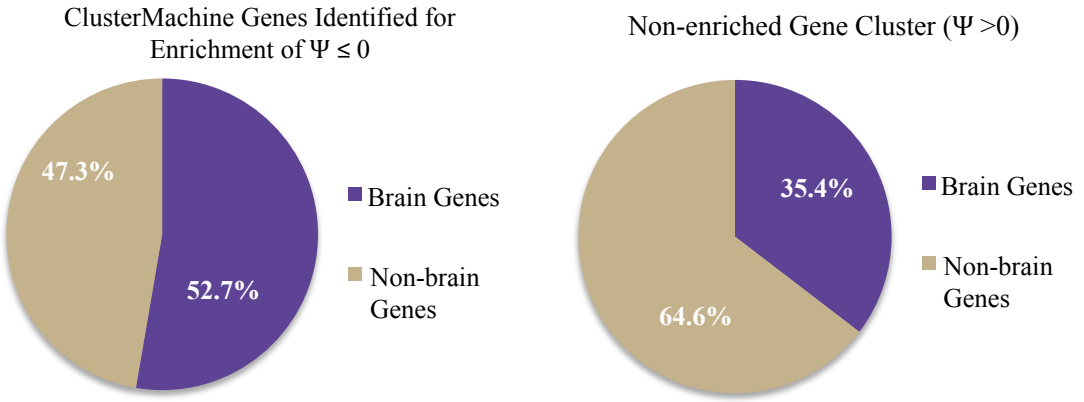


Figure 4.4. Proportion of brain genes among clusters from unsupervised C3.

In contrast, gene clusters identified by the unsupervised C3 analysis showed differential proportions of brain genes, relative to one another. The distribution of sites across all Ψ bins for the two clusters resulting from *ClusterMachine* is illustrated in Fig. 4.3. As discussed, the genes in Cluster A are likely to be highly conserved across species, so their relatively higher proportion of brain genes (52.7%) relative to genes in Cluster B (35.4%) may be indicative of the ubiquitous importance of brain genes. That is, species across divergent phylogenetic clades such as mammals and birds/lizards are not so different when it comes to our brain genes; rather, purifying selection conserves these genes across species to preserve crucial, basic functions of cognitive processing and brain development.

4.4 DISCUSSION AND CONCLUSIONS

This exercise in supervised and unsupervised clustering on SPurS data is a proof of concept for future analyses using the Conservative Chi-square Collapse algorithm, and the results are promising. *ClusterMachine* has been shown to identify orthogonal axes of heterogeneity in a large dataset, separating clusters of genes in such a way that appears to be biologically meaningful. It is also malleable to a variety of different approaches, with the ability to conduct both targeted and unsupervised clustering analyses. The flexibility in implementation and simplicity of design lends itself to widespread applicability across different datasets, projects, and research disciplines. *ClusterMachine* was originally designed as a statistical package to analyze SPurS results, but it is now clear that its potential applications are far-reaching. It may be worthwhile to integrate *ClusterMachine* as an independent tool on the Genomic Hyperbrowser, which could be used on its own or in tandem with SPurS.

While the involvement of SPurS-enriched genes in the development of morphological and behavioral differences remains to be determined, the analysis of SPurS results in *ClusterMachine* has provided a statistical framework for determining whether genes are truly enriched for SPurS sites, relative to other genes and simulated data. Biological data analyses could greatly benefit from a flexible tool based on simple statistics that allows researchers to identify statistically significant outliers. In the same way that principal component analysis (PCA) has been used extensively across genetic studies, I believe that the Conservative Chi-Square Collapse (C3) also has the potential for broader impacts. My hope is that many other researchers, whether involved in genomics or not, will find the *ClusterMachine* statistical package useful in a variety of applications, and I anticipate future work in this area to refine and explore new ways to use *ClusterMachine*.

Chapter 5. WHOSE GENOMES MATTER?

A reference to the Black Lives Matter movement (which confronts acts of police brutality committed disproportionately against people of color in the United States), the title of this chapter is meant to evoke a sense of racial injustice in the genomics community. Originally proposed as the title of an article describing an analysis I conducted on the proportion of ancestry groups represented in genome-wide association studies (GWAS), this title was subsequently rejected and finally replaced by "Genomics is failing on diversity". While provocative in nature at its suggestion that some genomes may matter more than others, it is meant to probe researchers in genomics to consider the ethical and societal implications of who is included (and indeed who is excluded) from genomics research. There is a plethora of scientific, practical, and historical reasons that individual researchers choose to study primarily populations of European descent. However, the culmination of these individual decisions has lead to a decades-long stretch of genomics research that has effectively laid the groundwork for precision medicine based on a small sub-set of genomic variation that exists in the global human population. This tension between individual decisions based on statistical power and availability of data (among other reasons) and the ethical imperative of inclusivity expressed by social scientists and activists is one that has yet to be fully reconciled in the community.

In this chapter, I provide a personal reflection on teaching ancestry in genomics, briefly outlining the historical and sociological context surrounding this complex and controversial topic. I also describe the methodology and results of an analysis I conducted to calculate the proportion of various ancestry groups that are represented in GWAS and a copy of the final publication containing this analysis is included in the Appendix.

5.1 INTRODUCTION TO RACE AND ANCESTRY IN GENOMICS

Teaching a large lecture hall full of undergraduates from diverse cultural and ethnic backgrounds about race and ancestry at first seemed daunting and overwhelming. Despite having some elusive ancestral roots in the Choctaw nation of Oklahoma and a sense of cultural connection to native and indigenous peoples, a predominantly Nordic heritage dominates my outward appearance, and I have been brought up in an environment of great economic and cultural [white] privilege.

When planning my presentation for the class, I set aside these personal associations with race and ancestry, drawing instead on my technical knowledge of genomics and ancestry-informative markers. Through the lens of an academic scientist, I carefully delineated the difference between ancestry – a biological trait that can be traced through markers in the genome – and race, which I described as a social and cultural construct relating to individual and group identity. While this distinction is technically accurate from a scientific perspective, the rigid construction of race and ancestry in genomics is too simplistic, and also dismissive of the perspectives in the room, which have been shaped by personal experiences and cultural context. It is not only difficult and confusing for students to learn about the relationship between race and biology in a few bullet points, but it may also be damaging to their own personal views on the subject; representing a missed opportunity for deeper understanding.

In truth, the concept of race as a biological phenomenon has its roots in ancient times, with references to biological determinism (the idea that behavior is pre-determined by biology) as a justification for the superiority of certain individuals over others in Aristotle's *Politics* (384—322 BCE).[102] The pseudo-science known as *scientific racism* has subsequently been developed, documented,[103] and strategically employed to justify

systemic inequality and injustices against people of color throughout history, including colonialism, slavery, discrimination, segregation, and eugenics.[104]

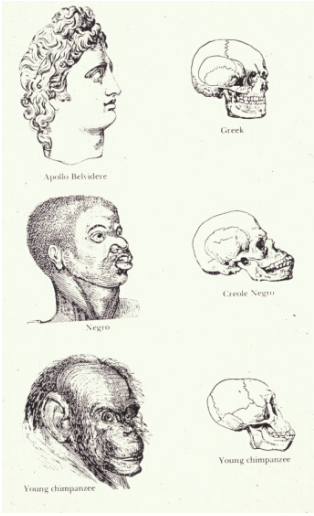


Figure 5.1 Illustration of cranial differences btwn. 'races'. *Indigenous Races of the Earth.*

Although there is no scientific evidence supporting these antiquated biological theories of race, the effects of *racism* can, indeed, play a role in the development of diseases and health outcomes.[105] Thus, it would be misleading and factually incorrect to state that there is definitively no connection between biology and racial identity. However, once we start down this road it becomes difficult to tease apart the effects of health outcomes as a result of environmental and socio-economic factors that are *associated* with race and cultural identity, and the actual impact of *racism* itself. In fact, it is common practice in biomedical research to use self-identified race/ethnicity (or subjectively identified by health professionals in a research or medical record) as a proxy for the myriad factors that are associated with race and can have an impact on health or response to certain medications. While it may sometimes be practical and useful to do so, it is an imperfect measure that is both scientifically and ethically inappropriate in certain scenarios. Where we need to be going as a research community is toward the precise identification of environmental factors and genomic variants (and their interactions) that produce differential health outcomes between individuals and populations, and accurately account for these variables directly in statistical analyses. This will have compound benefits of being more scientifically accurate and informative, while moving science even further away from its dark history of justifying structural, systemic inequality through biological distinction of races.

One important element of this work will be to disentangle true discoveries of associations between diseases or traits and genomic variants that tend to segregate in certain populations, and false associations that may be confounded by environmental factors that are also associated with populations of particular ancestry. In order to achieve this goal, there needs to be sufficient participation of individuals from groups of diverse ethnic, socio-economic, cultural, and ancestral backgrounds who are willing to share rich environmental, medical and genomic information with researchers.

This has been a key issue, and rightfully so; with such cases of research misconduct plaguing communities of color as the well-known Tuskegee Syphilis experiment (which followed unknowingly infected African American men in Alabama for decades without treatment for the sake of science) and a lawsuit in which Arizona State University was court-ordered to compensate the Havasupai tribe $700,000 for psychological and social damages relating to the use of their genetic data without proper informed consent.[106] However, it is commonly overstated how anti-research participation attitudes in communities of color are the primary drivers of disparity in genomics. Research has shown otherwise; in a study that directly measured this phenomenon, there was no difference in willingness to participate between racial and ethnic groups.[107] While there are residual and current trust issues between scientists and communities of color, there are additional systemic factors that contribute to the low numbers of research participants from certain ancestral groups.

Researchers often cite a variety of scientific or practical reasons to prefer working with samples of European ancestry, including availability of data in existing databases, statistical power and sample size, and methodological constraints such as an historical reliance on linkage disequilibrium (LD), or associations between allele frequencies of

variants sharing segments of chromosomes that share a common ancestral history. These methods have been designed an implemented primarily on individuals with European ancestry, so the particular characteristics of European genomes have been woven into the research infrastructure of genetic association studies. However, growing recognition of the importance of a diverse evidence base, as well as methodological advances in the analysis of heterogeneous (including admixed) populations, make continuing neglect of minority populations untenable. Investigators and funding agencies should deprioritize GWAS in Europeans in favor of replication and discovery in existing non-European datasets, accentuate efforts to recruit and phenotype research participants from underrepresented populations, and work to reduce systemic racial biases in the academy that prevent the advancement of investigators from diverse backgrounds. Longer term, building partnerships with minority communities, increasing stakeholder participation in the planning and follow-up of genetic research to address *their* needs and concerns, and recruitment, retention, and promotion of minority researchers will be key to reducing endemic cultural and racial biases in scientific research institutions, which have contributed to the persistent ancestry bias observed in GWAS.

5.2 GENOMICS IS FAILING ON DIVERSITY

In 2009, Need & Goldstein quantified the under-representation of non-European populations in genome-wide association studies (GWAS) and Bustamante et al. (2011) made the case for increased representation of minorities in genomics research based on the previous analysis.[108, 109] Although there have been efforts to increase the representation of non-Europeans in large sequencing projects such as the 1000 Genomes Project and the Human Genetic Diversity Project (HGDP), there has not been a

comprehensive follow-up to assess the success of this call to inclusion since the publication of these two papers.

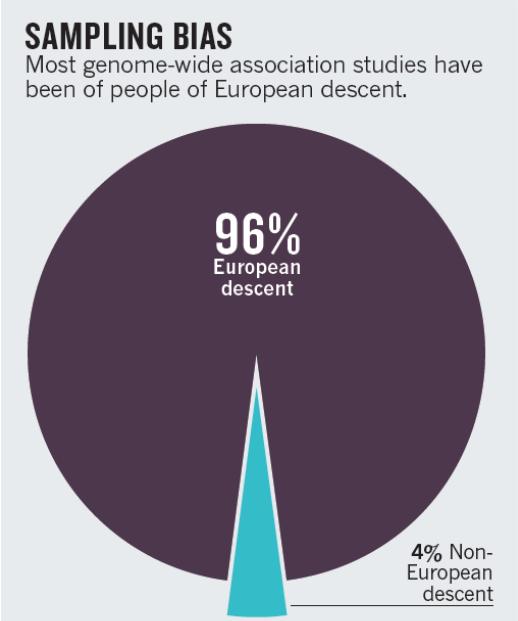


Figure 5.2 Pie chart from Bustamante et al.'s 2011 *Nature* Commentary

To conduct such a study, I surveyed the Catalog of Published Genome-Wide Association Studies (GWAS Catalog[110]) to update the 2009 data pie chart (Fig. 5.2) with data from 2016. To query the GWAS Catalog, I wrote a Python parser (tailored to the list and description of studies that can be downloaded from the GWAS Catalog website[††]), which identifies the number of independent publications in the Catalog, and reports the distribution of ancestral populations sampled across all studies. It also produces lists of the number of individuals from each unique study population, and can be tailored to produce lists of PubMed ID numbers of studies with different ancestral populations. The program and documentation are available upon request, but may require manual updates as additional populations are added, as it was designed to parse particular phrases and patterns of phrases currently found in the Catalog.

As sample descriptions are not standardized across studies, decisions needed to be made about how to categorize each sample with respect to broader ancestry bins for the aggregate analysis. These decisions were based on ancestry of origin and/or geographic

†† The GWAS Catalog was originally hosted by the National Human Genome Research Institute (NHGRI) but has since moved to the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) and is manually curated on an ongoing basis. https://www.ebi.ac.uk/gwas/docs/downloads

proximity. For example, Afrikaners from South Africa are included in the "European ancestry" bin because of their European (Dutch) ancestral origins, which is discordant from their current residence on the continent of Africa. Singapore and Malaysia are included in the "Asian ancestry" bin while the Philippines and other island nations are included in the "Pacific Islander" bin based on the geographic connection or separation from the continental landmass of Asia. A complete list of sample descriptions from the GWAS Catalog and ancestral categories can be found on page 52 of the Appendix.

Table 5.1 Number and proportion of GWAS participants by ancestral group in 2016, conducted by A.B. Popejoy, and in 2009 as reported by Need & Goldstein. Ancestral categories were selected based on those presented in the 2009 analysis and do not reflect the choices of the author.

	African ancestry only	**Arab & Mid.E.[a] only**	**Asian ancestry only**	**European & Jewish Only**	**Hispanic & L.A.[b] only**	**Native Peoples only**	**South Pacific Isl.[c] only**	**Mixed & multiple groups[d]**	**Total***
2016									
Group-based No. Studies	58 (2.31%)	3 (0.12%)	349 (13.90%)	1461 (58.18%)	19 (7.57%)	5 (0.20%)	9 (0.36%)	484 (19.28%)	2388 (95.10%)*
Total Samples	1045224 (3.08%)	27040 (0.08%)	4795132 (14.13%)	27435555 (80.82%)	184265 (0.54%)	17929 (0.05%)	94043 (0.28%)	347928 (1.02%)	33947116 (100%)
2009									
Group-based No. Studies	1 (0.27%)	0 (0.00%)	26 (6.97%)	322 (86.33%)	3 (0.80%)	2 (0.54%)	1 (0.27%)	**Mixed[d] Ancestry** 11 (2.95%)	366 (98.12%)*
Total No. Samples	9840 (0.57%)	0 (0.00%)	52877 (3.18%)	1581776 (96.37%)	1019 (0.06%)	1102 (0.06%)	2622 (0.15%)	92437 (5.32%)	1741673 (105.7%)*

[a] Arab and Middle Eastern; [b] Hispanic and Latin American; [c] South Pacific Islander; [d] Mixed group-based *studies* refer to studies conducted on multiple single-ancestry groups, including original and replication samples; mixed ancestry *samples* refer to individuals with mixed (more than one) ancestry.

* Totals are >100% due to samples overlapping ancestral categories, and <100% because studies are compared to the total pool, including those without ancestry information.

As in the 2009 analysis conducted by Goldstein and Need, the numbers and proportions of samples presented here need to be understood not as true individual samples, but as potential replicates from the same cohorts and databases (sampling with replacement, and a growing population to choose from over time). It is possible that a large portion of the GWAS conducted simply re-sampled the same individuals from a

few large databases for different studies. As such, the numbers of individuals reported in both 2009 and 2016 represent the number of times any individual from a particular ancestral background was sampled, not the number of times independent individuals were sampled from those populations. Table 5.1 presents the number and percent changes in GWAS participants per ancestral group in the last seven years, and categories are defined so as to provide direct comparisons between the two studies. Note that the 2009 analysis is based on 1.7M samples. As of August 2016, 2511 studies were included in the GWAS Catalog with nearly 35 million participant samples analyzed.[111] This is a >2000% increase in the number of individual samples studied in under a decade, with ample opportunity for increased sampling of previously under-represented populations.

As shown in Figure 5.3, the proportion of individuals of non-European descent among GWAS participants is 16% greater than it was seven years ago. Taken at face value, this is a laudable achievement. However, the majority of the growth in diversity is attributable to an increase in Asian study participants, demonstrating an unequal distribution of improvement across ancestry groups. The breakout bar graph (shades of green) shows that the largest slice of the pie representing non-Europeans (14%) is comprised of four different categories of Asian ancestral populations: Asian (China, Hong Kong, Japan, Korea, Mongolia, Taiwan and "Asia"), East Asian (referred to as "East Asia" in the GWAS Catalog), South Asian (Bangladesh, India, Nepal, and "South Asia"), and South East Asian (Malaysia, Singapore, Thailand, Vietnam, and "South East Asia"). By looking up each Asian-only GWAS in PubMed, I ascertained the host country of the institution affiliated with each study's first author, and found that 323 of the 349 Asian studies (~93%) were conducted in Asian countries by Asian researchers. In order to assess the inclusion of underserved populations in the United States, these studies were

removed and the resulting figure (Fig.5.3-C) shows that GWAS on non-majority populations have increased by only 2% in the last seven years.

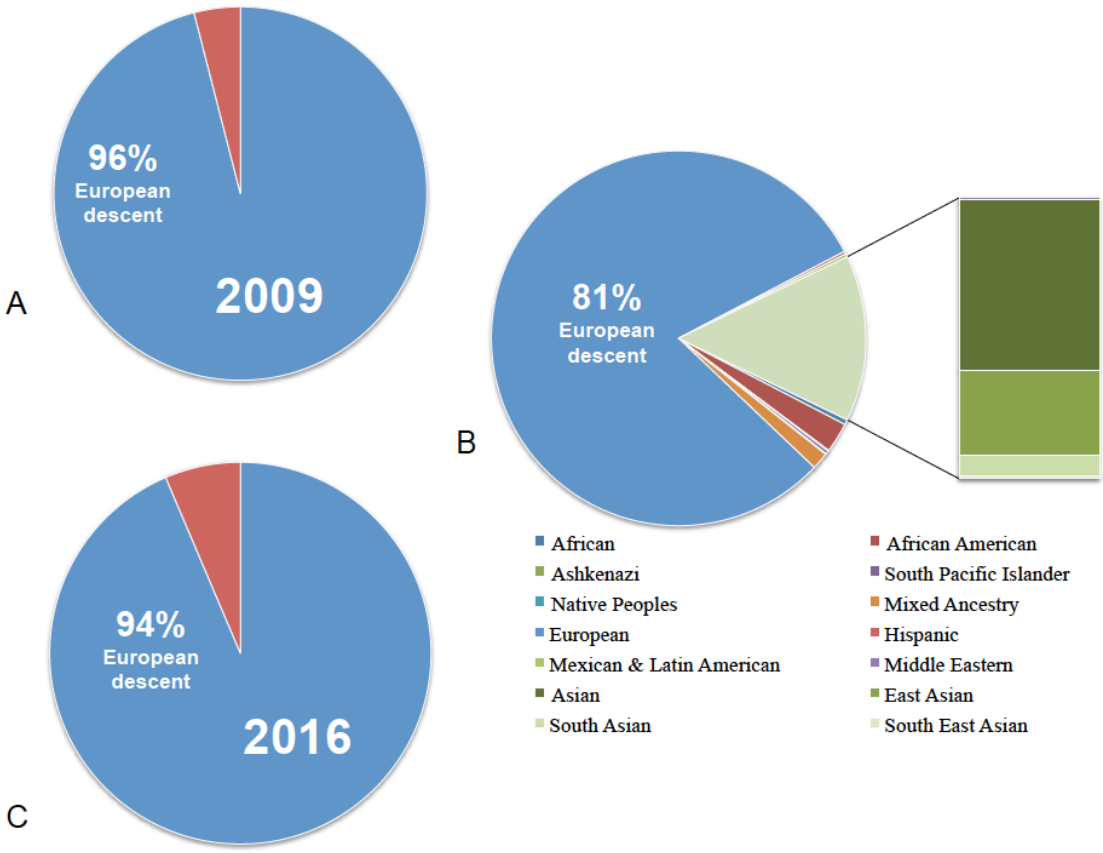


Figure 5.3 Change in Diversity of GWAS Sample Participants, 2009 vs. 2016. (A) Figure of sampling bias from Bustamante et al. (2011) showing data from Goldstein & Need (2009). The majority of GWAS before 2009 included European study samples (1,581,776 individuals with European ancestry; 65,907 individuals with non-European ancestry across 366 total studies). (B) Ancestry of GWAS participants March 2005 to August 2016, excluding 113 study samples (708,721 individuals) with no available race, ethnicity, or ancestry data. Total number of studies with ancestry information is N = 2,511 with n = 33,947,116 individuals. Note that these may not be unique individuals, if study samples were used multiple times by different investigators and published independently. Bar graph break-out represents the "Asian ancestry" category, which comprises 14% of all samples. (C) GWAS participants of European and non-European descent 2005–2016, excluding studies sampling only Asian ancestry, most of which are conducted by investigators in the same country of origin as the samples.

In the Catalog today at least 94% of GWAS participants from non-Asian countries are whites of European ancestry. Individuals from the most vulnerable and traditionally

underserved, under-represented populations including those of African and Latin American ancestry, Hispanics, and Native or indigenous peoples *combined* still represent <4% of all samples analyzed. Moreover, while there was a 2.5% increase in the proportion of African ancestry individuals and an increase of <1% in Hispanics and Latin Americans, the proportion of Native peoples (including Native Americans as well as other indigenous populations across the world) has actually *decreased* slightly since 2009. With such a large increase in the number of GWAS performed in the last seven years, this lack of growth in the diversity of samples analyzed is remarkable and deeply disconcerting.

When looking only at GWAS, the call-to-action for increased diversity in genomics appears to have fallen on deaf ears. However, advances in technology have led the field in new directions, including whole exome and genome sequencing. Among these new samples, attempts at greater inclusion and aggregation across cohorts have been modestly more successful. The Exome Aggregation Consortium (ExAC), for example, hosts data on genetic variants from >60,000 samples, representing 8.6% African, 9.5% Latino, and 60.4% European ancestry.[112] The National Heart, Lung, and Blood Institute (NHLBI) Trans-Omics for Precision Medicine (TOPMed) whole genome sequencing project includes 50% European American, 30% African American, 10% Hispanic/Latino, and 8% Asian participants (unpublished data).

These efforts to generate deeply sequenced data from diverse populations represent important resources for functional and/or clinical inference, but phenotypes are often missing or inconsistently available for such datasets, which include many fewer samples than those represented in the GWAS Catalog. Thus, it remains a significant concern that the more numerous and better phenotyped GWASed cohorts are so skewed.

We simply do not know how many of the thousands of statistically significant variant-disease associations observed in European ancestry populations replicate in other groups. We are also missing important opportunities to discover novel gene-disease associations in minority populations, inhibiting a full understanding of disease biology as well as risk prediction. Moreover, even where the evidence base is more diverse (i.e. in rapidly accruing exome collections) emerging data suggest that patients of non-European ancestry are still more likely to experience healthcare inequality as a result of variants of unknown significance[3] including, for example, more false positive tests[7] as a result of insufficiently diverse reference databases and control cohorts. In short, current incremental efforts have been inadequate and without urgent efforts to rectify imbalances in both analyses and recruitment, important healthcare disparities are likely to increase.

If the next seven years of genomics research persists on its current trajectory, the message broadcast by the scientific and medical genomics community to the rest of the world will be a harmful and misleading one: European descendants' genomes matter the most. New research initiatives, such as the Precision Medicine Initiative (PMI) Cohort Program, aim to bring thousands of new, ethnically, socioeconomically, and geographically diverse participants into biomedical [including genomic] research. Such efforts, which must aim for broad inclusion to ensure that that the benefits and risks of research are equitably distributed,[113] will nevertheless take considerable time and effort. Even while such efforts are underway, more can and must be done to begin to correct evidentiary biases, beginning with the prioritization of funding applications that propose novel or replication studies on African American, Hispanic/Latino, and Native American populations, and deprioritizing or declining to renew research on existing disease cohorts of European ancestry. Until the amount of genomic and phenotypic data from each major

ancestral population across the world becomes large enough to conduct amply powered GWAS in all of them, equality in clinical genomics and precision medicine will not be achieved. If this means that all future sequencing efforts on new participants for known diseases and GWAS efforts on existing data must exclude European ancestral populations in order to enrich currently under-represented groups, then that may be a logical and necessary course of action.

In many cases, prioritizing discovery efforts will require new recruitment; in those cases funding agencies will need to provide training and resources to help investigators build community partnerships. This includes building trust, two-way communication, and inclusion of participants as stakeholders rather than removed study subjects. While successful minority community partnerships have certainly been built with white European investigators at the helm, research recruitment efforts in communities of color have been most successful when conducted by investigators with matched race/ethnicity or in partnership with historically minority-serving institutions.[114] Therefore, I strongly recommend that efforts to diversify the pool of genomics samples also be bolstered with efforts to recruit, retain, and promote faculty and researchers of color in academic and non-academic institutions. Diversification of research-intensive biomedical departments, which are comprised of <4% African American, Hispanic, and Native American tenured and tenure-track faculty [combined] is particularly crucial.[115] This is a long-term and lofty ambition, given persistent systemic racial biases that prevent equitable distribution of NIH research grants and hiring and promotion practices that continue to favor white scientists.[116] Implicit bias training and awareness can lay the foundation for culture-shift, and should be considered a pre-requisite for recruitment and retention.

In light of recently elevated social and political awareness due to racial tensions igniting around the world, the time is now to acknowledge how systemic racial biases and injustices have also impacted biomedical research. Using the distribution of sample populations included in GWAS as a window into how these endemic issues have impacted genomic medicine over the last decade and a half, I have demonstrated that the ship is steering toward the shore. But a radical shift in thinking, funding priorities, and policies must occur before we are out of the storm.

The little progress we have seen in the diversity of GWAS participants raises the fundamental question of why there is such a predominant focus on those of European ancestry. In the review process for our commentary, one reviewer claimed that switching gears to GWAS in non-European ancestral groups would mean "losing all the progress we have made" toward discovering genetic mechanisms of disease. While I respectfully disagreed with the reviewer in communication with the editor, this sentiment is deeply disturbing, and dismisses the myriad opportunities that lie in emerging methodologies designed to extract novel findings from admixed and diverse populations.

5.2.1 *Broader Impacts*

Our commentary, "Genomics is failing on diversity" achieved a high level of notoriety in academic circles and in the media. It was referenced in talks at the American Society of Human Genetics meeting the week of its release, and soon after at a related research summit at the National Institutes of Health. The article has received over 20 citations in the online research-sharing platform *ResearchGate*, 40 citations on Google*Scholar*, and is the second suggested hit on a regular Google search with the query words: "genomics is".

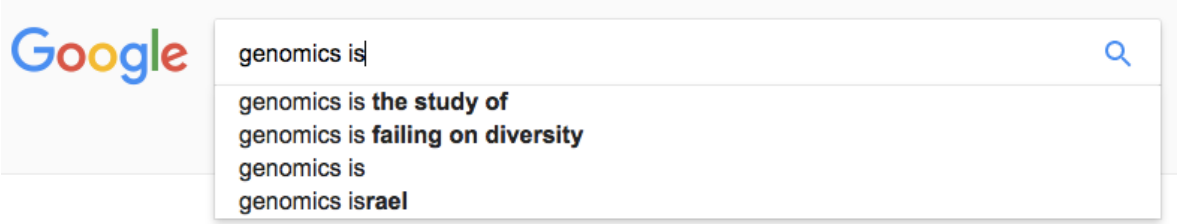


Figure 5.4 Screen shot from the Google search engine. *Opera* was the Internet browser used, which does not collect cookies on past searchers or save Google search history; query conducted in Oslo, Norway.

In August 2017, our article was highlighted in the popular news blog "FiveThirtyEight", a feature of *The New York Times* online, which was picked up by the *LinkedIn* news feed and other popular media outlets. In the last year, the story has been covered by *Pacific Standard*, *MIT Technology Review*, *Frontline Genomics*, and numerous other medical, genomics, and technology blogs such as *Bio-IT World*, the *Genetic Literacy Project, Genomes Web 2.0 and Bioethics,* and *GenomeWeb Daily News*.[117,118,119,120,121,122]

Early on in the outreach process, I made a concerted effort to raise awareness and spark discussion of the diversity issue in academic departments, in hopes of inspiring individual investigators to conduct research on more diverse population samples. I gave different departmental talks at the University of Washington and a presentation at the Genomics Salon for graduate students in Genome Sciences. I also gave phone and in-person interviews to news media outlets including the *UW Daily* and Health Sciences *Newsbeat*, National Public Radio (NPR), *Med City News,* and the German daily newspaper *Süddeutsche Zeitung*.[123,124,125] Ultimately, the impact of all this publicity and conversation remains to be seen. While many researchers are keen to discuss and acknowledge the topic of diversity in genomics, others seem to view it as a non-issue, preferring not to discuss it, most likely because the choice of population samples is seen

as a methodology-based decision about study design, not an ethical quandary about inequality that needs to be resolved at the individual researcher level.

In my experience in public policy and advocacy, tough conversations with those who disagree about systemic cultural and historical issues of inequality are crucial to raising awareness about the root causes of disparity. Working at the Association for Women in Science (AWIS) from 2010-2012 taught me that providing data to demonstrate social issues is the most effective way to change people's minds, particularly scientists, about controversial issues. I also learned that issues for individual women in science are often not simply one-off concerns, or problems related to personal choices and personality; they are often deep-seated, systemic social and cultural biases that disadvantage women, on average, compared to men.

On the topic of biased ancestry samples in GWAS, I believe that one of the key issues contributing to the dearth of research participants from diverse ancestral backgrounds is directly related to the lack of racially diverse academic researchers conducting the studies. By 2012, <4% of tenured and tenure-track faculty were of African American, Native American, or Hispanic descent; a strikingly similar figure to the 6% among research participants in 2016.[126] Anecdotally, many of the grant-funded programs currently underway to increase the representation of diverse groups among research participants are spearheaded by, or directly involve diverse faculty members. Recruitment efforts in communities of color have also been most successful when conducted by investigators with matched race/ethnicity or in partnership with historically minority-serving institutions.[127] It seems obvious that there is a connection between the lack of diversity among investigators and study participants, but surprisingly this is a very uncommon and unpopular topic of conversation; most likely because it is a difficult

problem to solve, and perhaps because it somehow strikes more closely to home for academic researchers.

It is not always comfortable to have these discussions, and some will have a higher tolerance for dealing with emotions of others (which inevitably get involved). But it is important, and is precisely what I did with my undergraduate students in their discussion section, several days after the lecture in which I gave a cut-and-dry scientific definition of race and ancestry in genomics. In preparation for these smaller group discussions, I consulted a fellow PhD student at UW studying Minority Studies and Multiculturalism, Gregory Diggs-Yang to see how I could approach the topic in a more inclusive way. He explained that concepts of race and ancestry are loaded with emotional baggage for some students, and not others. "Some of us [people of color] are at a calculus level when discussing these issues, while others are still learning arithmetic," he gently explained, and that it may be beneficial to start a conversation about race and ancestry by asking students what *they* believe these words mean. Indeed, this exercise was well worthwhile, as I learned that many white students had never really considered their own racial identity or even thought much about the concept of ancestry and biology. Others had a wealth of information to share, including insights and personal stories highlighting the difficulty of navigating micro-aggressions (the experience of frequent but subtle, often unintentional slights by others based on racial biases or stereotypes) and dealing with discrepancies between their own self-perceived cultural racial identity and outward appearance perceived and acted upon by others.

For fellow students of arithmetic (or perhaps pre-algebra) on these issues, I highly recommend "Unpacking the Invisible Knapsack" by Peggy McIntosh as an introduction to white privilege and the myriad ways that white people enjoy freedom that is not shared

among people of color.[128] An important step towards progress in creating more inclusive societal structures is for the white majority population and leaders of governmental and cultural institutions to acknowledge white privilege and its ill effects. This has a very important impact on discussions about systemic racism and inequality; first, it opens a conversation about ways in which structures and systems disadvantage people of color, and subsequently provides opportunities to discuss how those systems might be changed or improved to benefit more diverse groups of people, be they from different cultural, ethnic, or socio-economic backgrounds.

Part of the work I will be conducting as a postdoctoral scholar in Carlos Bustamante's lab at Stanford University will be to spearhead a working group for the Clinical Genomics Consortium (ClinGen). The goal will be to determine when it is appropriate to test individuals and genomic regions for ancestry in clinical and research contexts. Another project I intend to work on relates to the indigenous people of northern Norway, Sweden, Finland, and the Kola Peninsula of Russia: the Sámi. This population has been subjected to similar policies on assimilation as the Native Americans in the United States, and the issue of genetic research in their community is equally sensitive. As I work with researchers at the University of Oslo to describe the population genetics of Norway, I will carefully consider the ethical, legal, and social implications (ELSI) of the work we are undertaking, particularly with regard to the Sámi population's interests.

With the ever-increasing interconnectedness of our world, questions about how people and populations are impacted by genomics will continue to arise. Data privacy, patient access to genetic information, informed consent for research participants, and many more issues will become increasingly relevant moving forward. A PhD in Public Health Genetics has prepared me for this future, and I'm looking forward to taking it on.

BIBLIOGRAPHY

[1] Rollag, M.D., Provencio, I., Sugden, D., Green, C.B. Cultured amphibian melanophores: a model system to study melanopsin photobiology. *Methods Enzymol.* (2000) 316:291—309.

[2] Whitmore, D., Foulks, N.S., Sassone-Corsi, P. Light acts directly on organs and cells in culture to set the

[2] Whitmore, D., Foulks, N.S., Sassone-Corsi, P. Light acts directly on organs and cells in culture to set the vertebrate circadian clock. *Nature* (2000) 404:87—91.

[3] Friedmann, D., Hoagland, A., Berlin, S., Isacoff, E.Y. A spinal opsin controls early neural activity and drives a behavioral light response. *Curr. Bio.* (2015) 25(1):69—74.

[4] Koyanagi, M., Takada, E., Nagata, T., Tsukamoto, H., Terakita, A. Homologs of vertebrate Opn3 potentially serve as a light sensor in nonphotoreceptive tissue. *PNAS* (2013) 110(13):4998—5003.

[5] Blackshaw, S. and S.H. Snyder. Encephalopsin: A Novel Mammalian Extraretinal Opsin Discretely Localized in the Brain. *J. Neuroscience* (1999) 19(10):3681—3690.

[6] Darwin, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (1895) John Murray, London: pp.186.

[7] Lamb, T., Collin, S., and Pugh, E. Evolution of the vertebrate eye: opsins, photoreceptors, retina and eye cup. *Nature Reviews* (2007) Vol. 8: 960—975.

[8] Nathans, J., Molecular Biology of Visual Pigments. *Ann. Rev. Neurosci.* (1987) Vol. 10: 163—94.

[9] Chan, T., Lee, M., Sakmar, P. Introduction of Hydroxyl-bearing Amino Acids Causes Bathochromic Spectral Shifts in Rhodopsin: Amino acid substitutions responsible for red-green color pigment spectral tuning. *Journal of Biological Chemistry* (1992) Vol. 267, No. 14: 9478—9480.

[10] Nathans, J. Determinants of visual pigment absorbance: Identification of the retinylidene Schiff's base counterion in bovine rhodopsin. *Biochemistry* (1990) Vol. 29: 9746—9752.

[11] Oprian, D., Asenjo, A., Lee, N., and Pelletier, S. Design, chemical synthesis, and expression of genes from the three human color vision pigments. *Biochemistry* (1991) Vol. 30: 11367—11372.

[12] Sakmar, T., Franke, R., Khorana, H. Glutamic acid-113 serves as the retinylidene Schiff base counterion in bovine rhodopsin. *Proc. Natl. Acad. Sci. USA* (1989) Vol. 86: 8309—8313.

[13] Zhukovsky, E., and D. Oprian. Effect of carboxylic acid side chains on the absorption maximum of visual pigments. *Science* (1989) Vol. 246: 928—930.

[14] Yokoyama, S. Amino Acid Replacements and Wavelength Absorption of Visual Pigments in Vertebrates. *Mol. Biol. Evol.* (1995) Vol. 12, No. 1: 53—61.

[15] Yokoyama, S. Molecular genetic basis of adaptive selection: Examples from color vision in vertebrates. *Annual Review of Genetics* (1997) Vol. 31: 315—336.

[16] Yokoyama, S., and F. Radlwimmer. The "five-sites" rule and the evolution of red and green color vision in mammals. *Mol. Biol. Evol.* (1998) Vol. 15: 560—567.

[17] Yokoyama, S., Radlwimmer, F. and Blow, N. Ultraviolet pigments in birds evolved from violet pigments by a single amino acid change. *Proceedings of the National Academy of Sciences* (2000) Vol. 97: 7366—7371.

[18] Yokoyama, S., and T. Tada. The spectral tuning in the short wavelength-sensitive type 2 pigments. *Gene* (2003) 306: 91—98.

[19] Hiramatsu, C., Radlwimmer, F.B., Yokoyama, S. Kawamura, S. Mutagenesis and reconstitution of middle-to-long wavelength-sensitive visual pigments of New World monkeys for testing the tuning effect of residues at sites 229 and 233. *Vision Research* (2004) Vol. 44: 2225—2231.

[20] Takahashi, Y., and S. Yokoyama. Genetic basis of spectral tuning in the violet-sensitive visual pigments of African clawed frog, *Xenopus laevis*. *Genetics* (2005) Vol. 171: 1153-1160.

[21] Yokoyama, S., Starmer, W.T., Takahashi, Y. Tada, T. Tertiary structure and spectral tuning of UV and violet pigments in vertebrates. *Gene* (2006) 365: 95-103.

[22] Yokoyama, S., and T. Tada. "The evolution of ultraviolet vision in vertebrates." *Evolution of Nervous Systems* (2007) T. H. Bulloch, ed., Elsevier, Amsterdam: pp. 349—353.

[23] Takenaka, N. and S. Yokoyama. Mechanisms of spectral tuning in the RH2 pigments of Tokay gecko and American chameleon. *Gene* (2007) Vol. 399: 26—32.

[24] Yokoyama, S., Takenaka, N. Blow, N. A novel spectral tuning in the short wavelength-sensitive (SWS1 and SWS2) pigments of bluefin killifish (*Lucania goodei*). *Gene* (2007) Vol. 396: 196—202.

[25] Yokoyama, S., Tada, T. Yamato, T. Modulation of the absorption maximum of rhodopsin by amino acids in the C-terminus. *Photochem. Photobiol.* (2007) Vol. 83: 236—241.
[26] Yokoyama, S., Yang, H., and Starmer, D. Molecular basis of spectral tuning in the red- and green-sensitive (M/LWS) pigments in vertebrates. *Genetics* (2008) 179: 2037—2043.
[27] Yokoyama, S. Evolution of dim-light and color vision pigments. *Annu. Rev. Genomics Hum. Genet.* (2008) Vol. 9: 259—282.
[28] Chang, B., Crandall, K., Carulli, J., Hartl, D. Opsin Phylogeny and Evolution: A Model for Blue Shifts in Wavelength Regulation. *Molecular Phylogenetics and Evolution* (1995) Vol. 4, No. 1: 31—43.
[29] Palczewski, K., Kumasaka, T., Hori, T., Behnke, C., Motoshima, H., Fox, B., Le Trong, I., Teller, D., Okada, T., Stenkamp, R., Tamamoto, M., Miyano, M. Crystal structure of rhodopsin: A G Protein-Coupled Receptor. *Science* Vol. 289: 739–745.
[30] Davis, W.I.L., Collin, S.P., Hunt, D.M. Molecular ecology and adaptation of visual photopigments in craniates. *Molecular Ecology* (2012) 21:3121—3158.
[31] Porter, M., Blasic, J., Bok, M., Cameron, E., Pringle, T., Cronin, T., Robinson, P. Shedding new light on opsin evolution. *Roc. R. Soc. B* (2012) Vol. 279: 3—14.
[32] Provencio, I. and D.M. Warthen. Melanopsin, the photopigment of intrinsically photosensitive retinal ganglion cells. *WIREs Membrane Transport and Signaling* (2012) 1:228—237.
[33] Shichida, T. and T. Matsuyama. Evolution of opsins and phototransduction. *Phil. Trans. R. Soc. B* (2009) Vol. 364: 2881—2895.
[34] Hubbard, R., and A. Kropf. The action of light on rhodopsin. *Proc. Natl. Acad. Sci. USA* (1958) Vol. 44: 130—139.
[35] Chabre, M. and P. Deterre. Molecular mechanism of visual transduction. *Eur. J. Biochem.* (1989) Vol. 179: 255—266.
[36] Terakita, A. The opsins. *Genome Biology* (2005) Vol. 6: 213.
[37] Terakita, A., Kawano-Yamashita, E., Koyanagi, M. Evolution and diversity of opsins. *WIREs Membr. Transp. Signal* (2012) Vol. 1, No. 1:104—111.
[38] Shichida, et al. Evolution of opsins and phototransduction… (2009) [See Ref. 33]
[39] Terakita, A., Kawano-Yamashita, E., Koyanagi, M. Evolution and diversity of opsins. *WIREs Membrane Transport and Signaling* (2012) 1:104—111.
[40] Davies, W.L., Hankins, M.W., Foster, R.G. Vertebrate ancient opsin and melanopsin: divergent irradiance detectors. *Photochemical and Photobiological Sciences* (2010) The Royal Society of Chemistry.
[41] Terakita, et al. Evolution and diversity of opsins… (2012) [See Ref. 37]
[42] Koyanagi, et al. Homologs of vertebrate Opn3… (2013) [See Ref. 4]
[43] Panda, S., Sato, T.K., Castrucci, A.M., Rollag, M.D., DeGrip, W.J., Hogenesch, J.B., Provencio, I., Kay, S.A. Melanopsin (Opn4) Requirement for Normal Light-Induced Circadian Phase Shifting. *Science* (2002) 298:2213—2216.
[44] Roecklein, K.A., Wong, P.M., Miller, M.A., Donofry, S.D., Kamarck, M.L, Brainard, G.C. Melanopsin, photosensitive ganglion cells, and seasonal affective disorder. *Neuroscience and Biobehavioral Reviews* (2013) 37: 229—39.
[45] Buhr, E.D., Yue, W.W.S., Ren, X., Jiang, Z., Liao, H.w.R., Mei, X., Vemaraju, S., Nguyen, M.T., Reed, R.R., Lang, R.A., Y, K.W., Van Gelder, R.N. Neuropsin (OPN5)-mediated photoentrainment of local circadian oscillators in mammalian retina and cornea. *PNAS* (2015) 112(42):13093—13098.
[46] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V),* American Psychiatric Association, Arlington, VA, USA, 2016.
[47] Okano, T., Yoshizawa, T., Fukada, Y. Pinopsin is a chicken pineal photoreceptive molecule. *Nature* (1994) Vol. 372: 94—97.
[48] Taniguchi, Y., Hisatomi, O., Yoshida, M., Tokunaga, F. Pinopsin expressed in the retinal photoreceptors of a diurnal gecko. *FEBS Letters* (2001) Vol. 496: 69—74.
[49] Frigato, E., Vallone, D., Bertolucci, C., Foulkes, N. Isolation and characterization of melanopsin and pinopsin expression within photoreceptive sites of reptiles. *Naturwissenschaften* (2006) Vol. 93: 379—395.
[50] Macchi, M. and J. Bruce. Human pineal physiology and functional significance of melatonin. *Front Neuroendocrinol.* (2004) 25 (3–4):177–95.

[51] Vigh, B., Manzano, M.J., Zadori, A., Frank, C.L., Lukats, A., Rohlich, P., Szel, A., David, C. Nonvisual photoreceptors of the deep brain, pineal organs and retina. *Histology and Histopathology* (2002) 17: 555—590.
[52] Bagnara, J.T. and M.E. Hadley. Endocrinology of the amphibian pineal. *Am. Zool.* (1970) 10:201—16.
[53] Vigh, et al. (2002)
[54] Guido, M.E., Garbarino-Pico, E., Contin, M.A., Valdez, D.J., Nieto, P.S., Verra, D.M., Acosta-Rodriguez, V.A., de Zavalia, N., Rosenstein, R.E. Inner retinal circadian clocks and non-visual photoreceptors: Novel players in the circadian system. *Progress in Neurobiology* (2010) 92: 484—504.
[55] Davies, A.M.C. William Herschel and the discovery of near infrared. Spectroscopy Europe (2000) 12(2): 10—16.
[56] Smith, A.M., Mancini, M.C., Nie, S. Bioimaging: Second window for *in vivo* imaging. *Nature Nanotechnology* (2009) 4:710—11.
[57] Shichida, et al. Evolution of opsins and phototransduction… (2009) [See Ref. 33]
[58] Nozawa, M. and M. Nei. Evolutionary dynamics of olfactory receptor genes in Drosophila species. *PNAS* (2007) 104(17): 7122—7127.
[59] Thomas, J.H. and H.M. Robertson. The Caenorhabditis chemorceptor gene families. *BMC Biology* (2008) 6:42.
[60] Davies, et al. Vertebrate ancient opsin and melanopsin…(2012) [See Ref. 40]
[61] Davies, W.I.L., Collin, S.P., Hunt, D.M. Molecular ecology and adaptation of visual photopigments in craniates. *Molecular Ecology* (2012) 21: 3121—3158.
[62] Nei, M. and A.P. Rooney. Concerted and Birth-and-Death Evolution of Multigene Families. *Annu. Rev. Genet.* (2005) 39:121—152.
[63] Glasauer, S.M. and S.C. Neuhauss. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol. Genet. Genomics* (2014) 289(6): 1045—60.
[64] Spady, T.C., Parry, J.W.L, Robinson, P.R., Hunt, D.M., Bowmaker, J.K., Carleton, K.L. Evolution of the Cichlid Visual Palette through Ontogenetic Subfunctionalization of the Opsin Gene Arrays. *Mol. Biol. Evol.* (2006) 23(8):1538—1547.
[65] Terai, Y., Seehausen, O., Sasaki, T., Takahashi, K., Mizoiri, S., Sugawara, T., Sato, T., Watanabe, M., Konijnendikj, N., Mrosso, H.D.J., Tachida, H., Imai, H., Shichida, Y., Okada, N. Divergent Selection on Opsins Drives Incipient Speciation in Lake Victoria Cichlids. *PLoS Biology* (2006) 4(12): e433.
[66] Davies, W.I.L., Tamai, T.K., Zheng, L., Fu, J.K., Rihel, J., Foster, R.G., Whitmore, D., Hankins, M.W. An extended family of novel vertebrate photopigments is widely expressed and displays a diversity of function. *Genome Research* (2015) 25:1666—1679.
[67] Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* (2010) 59(3):307-21.
[68] Felsenstein, J. 2009. PHYLIP (Phylogeny Inference Package) version 3.7a. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
[69] FigTree: A Tree Figure Drawing Tool. Version 1.4.2, 2006-2014. Andrew Rambaut, Institute of Evolutionary Biology, University of Edinburgh.
[70] Jacobs, G.H. Evolution of colour vision in mammals. *Phil. Trans. R. Soc. Lond. B. Biol. Sci.* (2009) 364(1531):2957—2967.
[71] Nei, M., Suzuki, Y., Nozawa, M. The Neutral Theory of Molecular Evolution in the Genomic Era. *Annu. Rev. Genomics Hum. Genet.* (2010) Vol. 11: 265—289.
[72] Kimura, M. Evolutionary rate at the molecular level. *Nature* (1968*a*) Vol. 217: 624—626.
[73] King, J., and T. Jukes. Non-Darwinian evolution. *Science* (1969) Vol. 164: 788—798.
[74] Nei, M. Selectionism and Neutralism in Molecular Evolution. *Mol. Biol. Evol.* (2005) Vol. 22, No. 12: 2318—2342.
[75] Flicek, P., Ridwan Amode, M., Barrell, D., Beal, K., Konstantinos, B., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., García Girón, C., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A., Keenan, S., Kulesha, E., Martin, F., Maurel, T., McLaren, W., Murphy, D., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S., Vullo, A., Wilder, S., Wilson, M., Zadissa, A., Aken, B., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T., Kinsella, R., Muffato, M.,

Parker, A., Spudich, G., Yates, A., Zerbino, D., Searle, S. Ensembl 2014 *Nucleic Acids Research* (2014) Vol. 42

[76] Nei, M., and T. Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* (1986) Vol. 3:418—426.

[77] Bielawski, J., and Yang, Z. Maximum likelihood methods for detecting adaptive protein evolution. Nielsen, R, (ed.) *Statistical Methods in Molecular Evolution* (2005) Springer-Verlag, New York: pp.103–124.

[78] Yang, Z., Nielsen, R., Goldman, N., Pedersen, A. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. Genetics (2000) Vol. 155: 431—449.

[79] Yang, Z., Wong, W., Nielsen, R. Bayes Empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* (2005) 22:1107—1118.

[80] Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* (2001) 305(3):567—80.

[81] Lonsdale et al. The Genotype-Tissue Expression (GTEx) Project. *Nature Genetics* (2013) 45(6):580—85.

[82] Uhlén M et al. Tissue-based map of the human proteome. *Science* (2015) 347(6220):1260419.

[83] Zhang, Y., Chen, K., Sloan, S.A., Bennett, M.L., Scholze, A.R., O'Keefe, S., Phatnani, H.P., Guarnieri, P., Caneda, C., Ruderisch, N., Deng, S., Liddelow, S.A., Zhang, C., Daneman, R., Maniatis, T., Barres, B.A., Wu, J.Q. RNA-Seq transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *Journal of Neuroscience* (2014) 34(36):11929—11947.

[84] Haltaufderhyde, K., Ozdeslik, R., Wicks, N., Najera, J., Oancea, E. Opsin Expression in Human Epidermal Cells. *Photochemistry and Photobiology* (2015) 91: 117—123.

[85] Niggli, H.J. Artificial sunlight irradiation induces ultraweak photon emission in human skin fibroblasts. *Journal of Photochemistry and Photobiology B: Biology* (1993) 18(2-3): 281—285.

[86] Zarkeshian, P., Kumar, S., Tuszynski, J., Barclay, P., Simon C. Are there optical communication channels in the brain? *arXiv: 1708.08887 [physics.bio-ph],* Submitted August 23, 2017.

[87] Wang, Z., Wang, N., Li, Z., Xiao, F., Dai, J. Human high intelligence is involved in spectral redshift of biophotonic activities in the brain. *PNAS* (2017) 113(31):8753—8758.

[88] Shapiro, J., Leducq, J.B., Mallet, J. What Is Speciation? *PLoS Genetics* (2016) 12(3): e1005860.

[89] Holsinger, K.E. and B.S. Weir. "Genetics in geographically structured populations: defining, estimating and interpreting FST". *Nat Rev Genet* (2009) 10 (9): 639–650.

[90] Yang, Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586-1591

[91] Lartillot, N. and H. Philippe. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Mol. Biol. Evol.* (2014) 21(6):1095—1109.

[92] Soltis, P.S., Soltis, D.E., Savolainen, V., Crane, P.R., Barraclough, T.G. Rate heterogeneity among lineages of tracheophytes: Integration of molecular and fossil data and evidence for molecular living fossils. *PNAS* (2002) 99(7): 4430—35.

[93] Wright, S. Genetical structure of populations. *Nature* (1950) 166 (4215): 247–9.

[94] Jones, D.T., Taylor, W.R., Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *CABIOS* (1992) 8:275—282.

[95] Rambaut, A. and Grassly, N. C. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* (1997) 13: 235-238.

[96] Afgan, E., et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research* (2016) 44(W1): W3-W10 doi:10.1093/nar/gkw343

[97] Hirschfeld, H.O. A connection between correlation and contingency. *Proc. Camb. Phil. Soc.* (1935) 31: 520—524.

[98] Fisher, R.A. The precision of discriminant functions. *Ann. Eugen. Lond.* (1940) 10:422—429.

[99] Williams, E.J. The use of scores for the analysis of association in contingency-tables. *Biometrika* (1952) 39: 274—289.

[100] Benzécri, J.P. Statistical analysis as a tool to make patterns emerge from data. *In Methodologies of Pattern Recognition* (S. Watanabe, ed., 1969), pp. 35—60. New York: Academic.

[101] Hill, M. O. Correspondence analysis: a neglected multivariate method. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* (1974) 23(3): 340-354.

[102] Baker, Earnest. 1950. The Politics of Aristotle. London: Oxford University Press.

[103] Nott, J.C., Gliddon, G.R., Maury, L.F., Pulszky, F.A., Meigs, J.A. Indigenous races of the earth (1857) [Accessed online from the Library of Congress, August 2017.]

[104] Patricia Hill Collins, *Black feminist thought: knowledge, consciousness, and the politics of empowerment* (2nd ed., 2000), Glossary, p. 300: "Scientific racism was designed to prove the inferiority of people of color"; Simon During, *Cultural studies: a critical introduction* (2005), p. 163: "It [*sc. scientific racism*] became such a powerful idea because ... it helped legitimate the domination of the globe by whites"; David Brown and Clive Webb, *Race in the American South: From Slavery to Civil Rights* (2007), p. 75: "...the idea of a hierarchy of races was driven by an influential, secular, scientific discourse in the second half of the eighteenth century and was rapidly disseminated during the nineteenth century". (*Wikipedia citation.*)

[105] 2012 National Healthcare Disparities Report. June 2013. Agency for Healthcare Research and Quality, Rockville, MD. US Department of Health and Human Services (US-DHHS) http://archive.ahrq.gov/research/findings/nhqrdr/nhdr12/index.html

[106] Mello, MM and LE Wolf. The Havasupai Indian Tribe Case — Lessons for Research Involving Stored Biologic Samples. *New England Journal of Medicine* (2010) 363:204-207.

[107] Wendler, D., Kington, R., Madans, J., Van Wye, G., Christ-Schmidt, H., Pratt, L.A., Brawley, O.W., Gross, C.P., Emanuel, E. Are Racial and Ethnic Minorities Less Willing to Participate in Health Research? *PLoS Medicine* (2005) 3(2): e19.

[108] Need, A., and Goldstein, D. Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics* (2009) Vol. 25, No. 11: 489—494.

[109] Bustamante, C.D., Burchard, E.G, De La Vega, F.M. Genomics for the World. *Nature* Vol. 475, No. 7355: 163—165.

[110] Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., Parkinson, H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* (2014) Vol. 42 (Database issue): D1001-D1006.

[111] Burdett, T. (EBI), Hall, P.N. (NHGRI), Hastings, E. (EBI) Hindorff, L.A. (NHGRI), Junkins, H.A. (NHGRI), Klemm, A.K. (NHGRI), MacArthur, J. (EBI), Manolio, T.A. (NHGRI), Morales, J. (EBI), Parkinson, H. (EBI), and Welter, D. (EBI). The NHGRI-EBI Catalog of published genome-wide association studies. Available atwww.ebi.ac.uk/gwas. Accessed 5 August 2016, v1.0.

[112] Lek, M., Karczewski, K., Minikel, E, Samocha, K.E., Banks, E., et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* (2016) 536:285—291.

[113] Cohn, E., Henderson, G., Appelbaum, P.S. for the Working Group on Representation and Inclusion in Precision Medicine Studies. Distributive justice, diversity, and inclusion in precision medicine: what will success look like? *Genetics in Medicine* (August 2016).

[114] Yancey, A.K., Ortega, A.N., Kumanyika, S.K. Effective Recruitment and Retention of Minority Research Participants. *Annu. Rev. Public Health* (2006) 27(1):1—28.

[115] Leboy, P.S., and J.F. Madden. Limitations on Diversity in Basic Science Departments. *DNA and Cell Biology* (2012) 31(8):1—7.

[116] Check Hayden, E. Racial bias continues to haunt NIH grants: Minorities still less likely to win biomedical funding. *Nature News* (2015) 527: 286—87.

[117] "The Inexcusable Lack of Diversity in Genetic Studies" (Michael White, *Pacific Standard* Oct 4 2016).

[118] "Solving the Lack of Diversity in Genomic Research" (Emily Mullin, *MIT Technology Review* 2016).

[119] "Further bias in personal genomics" (Editor's note, *Frontline Genomics* 2016).

[120] "Why racial diversity in genetics studies matters in patient care" (Elizabeth Newbern, *Genetic Literacy Project* 2016).

[121] "Personal Genomics Open Access Datasets Even More European-Biased Than Scientific Literature?" (Genome Diary, *Genomes, Web 2.0 and Bioethics* 2016).

[122] "Among the Multitude: A Look at the Complexity of Diversity in Genomics" (Benjamin Ross, *Bio-IT World* Oct 13 2016).

[123] "Genomics continues to fail on diversity" (Timothy Kenney, *UW Daily* Nov 14 2016).

[124] "Human genome studies lack diversity" (Leila Gray, *HSNewsBeat* Oct 27 2016).

[125] "Genomics Medicine Ireland raises $40M for large-scale genomics research" (Juliet Preston, *Med City News* 2016).

[126] Leboy, P.S., and J.F. Madden. Limitations on Diversity… (2012) [See Ref. 115].

[127] Yancey, A.K., Ortega, A.N., Kumanyika, S.K. Effective Recruitment and Retention of Minority Research Participants. *Annu. Rev. Public Health* (2006) 27(1):1—28.

[128] McIntosh, Peggy. Unpacking the Invisible Knapsack. *Working Paper 189*. "White Privilege and Male Privilege: A Personal Account of Coming To See Correspondences through Work in Women's Studies" (1988).

APPENDIX

Table of Contents

Materials and Methods: Opsin Evolution

Species Selection and Data Capture

The species selected for this study represent a broad spectrum of vertebrates, with one or a few specimens from each major branch of the phylogenetic tree of vertebrate species. Lamprey (*Pteromyzon marinus*) was included as the most distal species in the tree relative to the other vertebrates; zebrafish (*Danio rerio*), fugu (*Takifugu rubripes*), platyfish (*Xiphophorus maculatus*), and cave fish (*Astyanax mexicanus*) were included as representatives of various bony (teleost) fish families and coelacanth (*Latimeria chalumnae*) was also represented [evolutionarily distant from the teleost fishes]; birds included chicken (*Gallus gallus*), hummingbird (*Calypte anna*), ostrich (*Struthio camelus*), penguin (*Aptenodytes forsteri*), and zebra finch (*Taeniopygia guttata*); lizards included American alligator (*Alligator mississippiensis*), crocodile (*Alligator sinensis*), Anolis lizard (*Anolis carolinensis*), turtle (*Pelodiscus sinensis*), and there was one amphibian: frog (*Xenopus tropicalis*). Both eutherian and non-eutherian mammals were included, the latter represented by opossum (*Monodelphis domestica*) and platypus (*Ornithorynchus anatinus*) and the former represented by armadillo (*Dasypus novemcinctus*), little brown bat (*Myotis lucifugus*), cow (*Bos taurus*), chimpanzee (*Pan troglodytes*), dog (*Canis familiaris*), elephant (*Loxodonta africana*), gorilla (*Gorilla gorilla*), human (*Homo sapiens*), mouse (*Mus musculus*), dolphin (*Tursiops truncatus*), and large flying fox (*Pteropus vampyrus*).

In total, 29 taxa were included in the phylogenetic assessment of the presence of orthologous opsin genes across vertebrate lineages. All protein sequences used for the positive selection analysis (and most sequences for the comparative evolution analysis) were downloaded as database files from the Ensembl website (release version 84, March 2016)[i] and compiled into BLAST searchable databases using the makeblastdb function in the BLAST+ suite of command-line tools available through the National Center for Biotechnology Information (NCBI) website.[ii] Additional genomes and predicted protein sets for species not included in the Ensembl database were obtained from BGI (hummingbird, ostrich, and penguin) and from RefSeq (alligator and crocodile).[iii,iv] Due to the dearth of whole-genome assemblies matched to predicted protein sets,

the scan for positive selection included only species whose data were obtained from Ensembl.

Table A-1 provides detailed information on all taxa included in the SPurS analysis (Section B).

Table A-1. Species Sampled for Genome-Wide SPurS Analysis. The taxa used in the assessment of the genome-wide distribution of shifts in purifying selection (SPurS) are included in this table. Each row is a taxon; where the first column is the code name used in the species tree, with the clade group identified in parentheses (B = bird, M = mammal, S = other sauropsid, or non-mammal). The remaining columns provide the common and scientific names of each taxon sampled, and their taxonomic categories. Starred taxa were included in the scan for positive selection. All species information was provided by Jim Thomas.

Species Code Name	Scientific Name	Common Name	Superclass/ Class	Subclass/ Superorder/ Order/Suborder	Superfamily/ Family/Subfamily
Aplat (B)	*Anas_platyrhynchos*	duck	Galloanserae	Anseriformes	Anatidae
Ggal (B)	**Gallus_gallus*	chicken	Galloanserae	Galliformes	Phasianidae
Chavoc (B)	*Charadrius_vociferus*	killdeer	Neoaves	Charadriiformes	Charadriidae
Clivi (B)	*Columba_livia*	pigeon	Neoaves	Columbiformes	Columbidae
Cuccan (B)	*Cuculus_canorus*	cuckoo	Neoaves	Cuculiformes	Cuculidae
Tauery (B)	*Tauraco_erythrolophus*	turaco	Neoaves	Cuculiformes	Musophagidae
Eurhel (B)	*Eurypyga_helias*	sunbittern	Neoaves	Eurypygiformes	Eurypygidae
Fcher (B)	*Falco_cherrug*	falcon	Neoaves	Falconiformes	Falconidae
Opihoa (B)	*Opisthocomus_hoazin*	hoatzin	Neoaves	Opisthocomiformes	Opisthocomidae
Zalb (B)	*Zonotrichia_albicollis*	sparrow	Neoaves/A.1	Passeriformes	Emberzidae
Phumi (B)	*Pseudopodoces_humilis*	ground_tit	Neoaves/A.1	Passeriformes	Paridae
Manvit (B)	*Manacus_vitellinus*	manakin	Neoaves/A.1	Passeriformes	Pipridae
Mundu (B)	*Melopsittacus_undulatus*	parrot	Neoaves/A.1	Psittaciformes	Psittaculidae
Halleu (B)	*Haliaeetus_leucocephalus*	eagle	Neoaves/A.2	Accipitriformes	Accipitridae
Colstr (B)	*Colius_striatus*	mousebird	Neoaves/A.2	Coliiformes	Coliidae
Picpub (B)	*Picoides_pubescens*	woodpecker	Neoaves/A.2	Piciformes	Picidae
Apavit (B)	*Apaloderma_vittatum*	trogon	Neoaves/A.2	Trogoniformes	Trogonidae
Nipnip (B)	*Nipponia_nippon*	ibis	Neoaves/B	Ciconiiformes	Threskiornithidae
Egrgar (B)	*Egretta_garzetta*	egret	Neoaves/B	Pelecaniformes	Ardeidae
Padel (B)	*Pygoscelis_adeliae*	penguin	Neoaves/B	Sphenisciformes	Spheniscidae
Chapel (B)	*Chaetura_pelagica*	swift	Neoaves/C	Apodiformes	Apodidae
Calann (B)	*Calypte_anna*	hummingbird	Neoaves/C	Apodiformes	Trochilidae
Strcam (B)	*Struthio_camelus_australis*	ostrich	Paleognathae	Struthioniformes	Struthionidae
Tingut (B)	*Tinamus_guttatus*	tinamu	Paleognathae	Tinamiformes	Tinamidae
Alsin (S)	*Alligator_sinensis*	crododilian	Sauropsida	Crocodilia	
Chasi (M)	*Chrysochloris_asiatica*	golden_mole	Eutheria	Afrotheria	Chrysochloridae
Lafr (M)	**Loxodonta_africana*	elephant	Eutheria	Afrotheria	Elephantidae
Eedw (M)	*Elephantulus_edwardii*	elephant_shrew	Eutheria	Afrotheria	Macroscelididae
Oafe (M)	*Orycteropus_afer_afer*	aardvark	Eutheria	Afrotheria	Orycteropodidae
Etel (M)	*Echinops_telfairi*	tenrec	Eutheria	Afrotheria	Tenrecidae
Tman (M)	*Trichechus_manatus _latirostris*	manatee	Eutheria	Afrotheria	Trichechidae
Cfam (M)	**Canis_lupus_familiaris*	dog	Eutheria	Carnivora	Caniformia/Canidae
Fcat (M)	*Felis_catus*	cat	Eutheria	Carnivora	Feliformia/Felidae
Btau (M)	**Bos_taurus*	cow	Eutheria	Cetartiodactyla	Bovidae
Cbact (M)	*Camelus_bactrianus*	camel	Eutheria	Cetartiodactyla	Camelidae

Oorc (M)	*Orcinus_orca*	orca	Eutheria	Cetartiodactyla	Cetacea/Odontoceti
Sscr (M)	*Sus_scrofa*	pig	Eutheria	Cetartiodactyla	Suidae
Minnat (M)	*Miniopterus_natalensis*	yang_bat	Eutheria	Chiroptera/Yangochiroptera	Vespertilionidae
Mluc (M)	**Myotis_lucifugus*	yang_bat	Eutheria	Chiroptera/Yangochiroptera	Vespertilionidae
Palec (M)	*Pteropus_alecto*	yin_bat	Eutheria	Chiroptera/Yinpterochiroptera	Pteropodidae
Galvar (M)	*Galeopterus_variegatus*	colugo	Eutheria	Dermoptera	Cynocephalidae
Eeur (M)	*Erinaceus_europaeus*	hedgehog	Eutheria	Eulipotyphla	Erinaceidae
Sara (M)	*Sorex_araneus*	shrew	Eutheria	Eulipotyphla	Soricidae
Ccri (M)	*Condylura_cristata*	mole	Eutheria	Eulipotyphla	Talpidae
Ocun (M)	*Oryctolagus_cuniculus*	rabbit	Eutheria	Lagomorpha	Leporidae
Opri (M)	*Ochotona_princeps*	pika	Eutheria	Lagomorpha	Ochotonidae
Ecab (M)	*Equus_caballus*	horse	Eutheria	Perissodactyla	Equidae
Mjav (M)	*Manis_javanica*	pangolin	Eutheria	Pholidota	Manidae
Cjac (M)	*Callithrix_jacchus*	NW_monkey	Eutheria	Primate/Haplorhini	Callitrichidae
Mmul (M)	*Macaca_mulatta*	OW_monkey	Eutheria	Primate/Haplorhini	Cercopithecidae/ Cercopithecinae
Hsap (M)	**Homo_sapiens*	human	Eutheria	Primate/Haplorhini	Hominidae
Tsyr (M)	*Carlito_syrichta*	tarsier	Eutheria	Primate/Haplorhini	Tarsiidae
Mmur (M)	*Microcebus_murinus*	mouse_lemur	Eutheria	Primate/Strepsirrhini	Cheirogaleidae
Ogar (M)	*Otolemur_garnettii*	galago	Eutheria	Primate/Strepsirrhini	Galagidae
Hgla (M)	*Heterocephalus_glaber*	naked_mole_rat	Eutheria	Rodentia	Bathyergidae
Cpor (M)	*Cavia_porcellus*	guinea_pig	Eutheria	Rodentia	Caviomorpha/Caviidae
Jjac (M)	*Jaculus_jaculus*	jerboa	Eutheria	Rodentia	Dipodidae
Dord (M)	*Dipodomys_ordii*	kangaroo_rat	Eutheria	Rodentia	Heteromyidae
Cgri (M)	*Cricetulus_griseus*	hamster	Eutheria	Rodentia	Muroidea/Cricetidae
Moch (M)	*Microtus_ochrogaster*	vole	Eutheria	Rodentia	Muroidea/Cricetidae
Mmus (M)	**Mus_musculus*	mouse	Eutheria	Rodentia	Muroidea/Muridae
Rnor (M)	*Rattus_norvegicus*	Norway rat	Eutheria	Rodentia	Muroidea/Muridae
Nangal (M)	*Nannospalax_galili*	blind_mole_rat	Eutheria	Rodentia	Muroidea/Spalacidae
Stri (M)	*Ictidomys_tridecemlineatus*	ground_squirrel	Eutheria	Rodentia	Sciuridae
Tbelc (M)	*Tupaia_chinensis*	tupaia	Eutheria	Scandentia	Tupaiidae
Dnov (M)	*Dasypus_novemcinctus*	armadillo	Eutheria	Xenarthra	Dasypodidae
Shar (M)	*Sarcophilus_harrisii*	Tasmanian_devil	Metatheria	Dasyuridae	Dasyuridae
Mdom (M)	**Monodelphis_domestica*	opossum	Metatheria	Didelphimorphia	Didelphidae
Oana (M)	**Ornithorhynchus_anatinus*	platypus	Monotremata	Monotremata	Ornithorhynchidae
Gekjap (S)	*Gekko_japonicus*	gekko	Sauropsida	Squamata/ Bifurcata/Gekkota	Gekkonidae
Acar (S)	**Anolis_carolinensis*	lizard	Sauropsida	Squamata/Iguania	Dactyloidae
Thasir (S)	*Thamnophis_sirtalis*	snake	Sauropsida	Squamata/ Serpentes	Colubridae
Pmol (S)	*Python_bivittatus*	snake	Sauropsida	Squamata/ Serpentes	Pythonidae
Promuc (S)	*Protobothrops_ mucrosquamatus*	snake	Sauropsida	Squamata/ Serpentes	Viperidae
Cpic (S)	*Chrysemys_picta_bellii*	turtle	Sauropsida	Testudines/ Cryptodira	Emydidae
Pesin (S)	**Pelodiscus_sinensis*	turtle	Sauropsida	Testudines/ Cryptodira	Trionychidae

Orthologous Sequence Identification

One important challenge of interspecific species comparisons is the correct identification of orthologous genes, meaning that the human version of a particular opsin is compared to the same version of that gene in a different species (1:1 ortholog). Figure A-1 (below) illustrates the difference between 1:1 orthologues and non-orthologous genes (or 1:many, in the case of recent species-specific gene duplications and divergence).

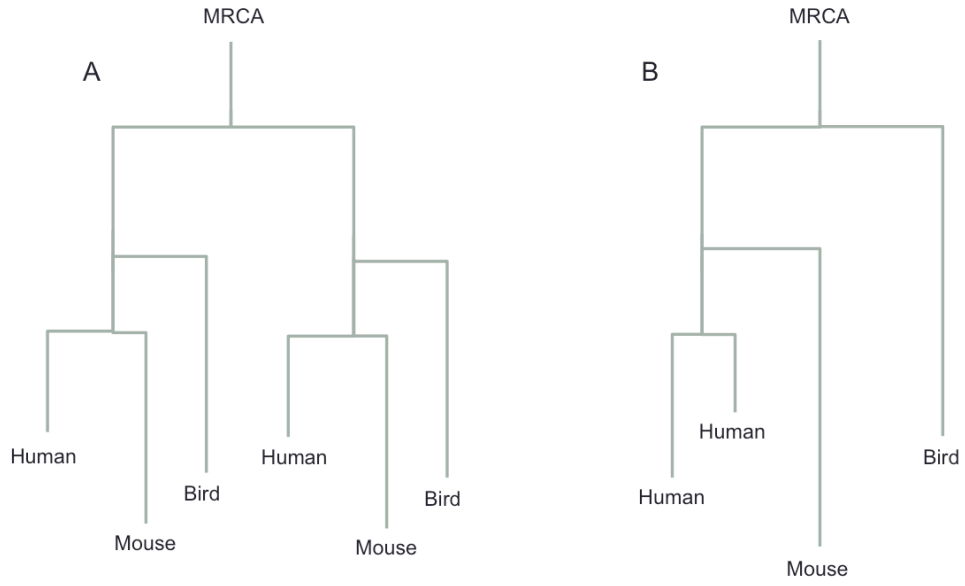


Figure A-1. Phylogenies illustrating orthologues (A) and non-orthologues (B) show that orthologous gene trees match the species tree, whereas trees with non-orthology contain multiple sequences from a single species in the same clade. MRCA: most recent common ancestor, where branching events represent divergence of sequences and species over time.

In order to conduct comparative evolutionary analyses on genes in different species, it is of crucial importance to identify sequences that are truly orthologous, meaning they descended from the same ancestral gene copy in a common ancestor. While Ensembl and other websites report orthologous sequences relative to human genes, these annotations can be misleading and are quite often incorrect. There is also the issue of multiple isoforms of a gene (of which there are many in opsins), and it is often unclear to which human isoform the other species' sequences are matched. Therefore, it is necessary to undertake a tedious but critical step when conducting inter-species analyses to empirically identify orthologous sequences.

For closely related species such as within eutherian (placental) mammals, the PhyOP pipeline developed by Goodstadt & Ponting (2006) can be used to verify the orthologous

relationships; however, this pipeline cannot be used for distantly related species whose synonymous sites are saturated with mutations because the method is based on the rate of mutations at synonymous sites.[v] Similarly, the OrthoMaM database can be used to find orthologues in mammals, but cannot be used to determine orthologous relationships in more distant species.[vi] Thus, I used a custom approach developed in collaboration with Jim Thomas to systematically identify orthologous opsins across 29 distantly related vertebrate species.

Our approach involves the use of a phylogenetic framework beginning at the [macro] gene family level, narrowing down to the [micro] gene tree level, and then re-building the family tree based on sub-family and gene phylogenies. The first step of this macro-micro method involves the identification of sequences most closely matching a query, based on sequence identity and length of the alignment between them (E-value and Bit score in BLAST). Then, sequences branching together in a clade on the macro family tree are selected to construct gene trees, on which one can then calculate branch lengths, boostrap values for local tree topology, and verify the monophyletic relationship of the sequences within a micro (sub-family) clade. Finally, the macro tree is pruned based on verified sequences in each micro tree, and a large consensus tree is built from several bootstrap trees to estimate the macro-level tree topology.

The first list of protein sequences compiled for the macro framework tree contained all the top hits from every BLAST search using the human opsin proteins as queries, which amounted to 610 sequences. This list contained many non-opsin proteins by design, so that no true opsins would be missed. I aligned the sequences using the standard methods in MUSCLE,[vii] converted the output alignment to a .phylip file format using a custom Python script written by J. Thomas, and constructed a maximum-likelihood tree in PhyML.[viii] This large tree provided a phylogenetic blueprint from which to determine the protein sequences that should be included in each individual gene tree, and a subsequent opsin family tree to visualize the relationships between opsins. Visualizing each clade in the tree containing a human query sequence, I collected all sequences that branched within the same clade and constructed individual gene trees for each of the human opsins using the same approach. The resulting gene trees with bootstraps based on 1000 replicate trees can be found in the following section on opsin sub-families.

To be sure that all the relevant protein sequences for each opsin gene were obtained, I conducted a tblastn search in parallel to the blastp searches, using the same human protein query to search nucleotide databases of each species and comparing the number of hits to the blastp results. If the number of unique hits was the same between the two searches, I assumed that the results from querying the predicted protein set were complete. However, if a tblastn search showed matching results for a human protein in a species where there was no matching result in the blastp search, I conducted further tblastn searches in that species using every human exon, and then translated the resulting nucleotide sequences into proteins for the subsequent analyses. Since protein sets are often predicted based on the genome assembly, it is more likely that the prediction algorithm missed a protein and thus will be present in the assembly but not the protein set. However, genome assemblies are notoriously incomplete, so it is always possible that orthologous sequences were missed entirely, due to missing segments of the assemblies in regions of genome that happen to contain our genes of interest.

Some sequences for which there was not enough information to be properly placed in the phylogeny (obtained through concatenation of just a few exons, for example) needed to be thrown out because of their unreliable branching relationship to the rest of the sequences. The remaining sequences that branch together in a clade with the query protein are orthologous. In the case of multiple sequences from a single species that branch next to one another in the same clade as the human query, one sequence was randomly chosen as the representative from that species.

Finally, all sequences from each sub-family tree were concatenated into a macro-level opsin family tree. After conducting this entire analysis, I realized that some orthologous copies of non-human opsins may be missing due to the human-query capture system, so I conducted additional BLAST searches on all species using chicken or zebrafish queries for each sub-family of opsins not found in humans, and repeated the entire tree-building process with a complete set of 413 sequences. I estimated bootstrap values for the consensus topology within sub-families (micro) based on 1000 PhyML trees with *consense* from J. Felsenstein's PHYLIP program[ix] and 100 bootstrap phylogenies of the macro opsin family tree. The resulting macro consensus tree is Figure 3 in the main text of this dissertation, and ML micro trees are in the following section.

Opsin Sub-Families: Phylogenetic Trees and Ortholog Tables

As mentioned in the main text, I identified five major groups of opsins, comprised of 20 sub-families within the opsin family of light-sensitive GPCRs. I hesitate to claim a definitive number of unique opsin genes across vertebrates, since it is impossible to determine whether closely related sequences from the same species branching together in a topology are isoforms or true paralogs, without independently verifying annotated gene identification numbers. However, it is clear from the phylogeny that there are 23 distinct vertebrate opsins that are represented across multiple species – a hallmark of an independent gene that distinguishes it from a recent lineage- or species-specific duplication. In this section, I profile each of these 23 distinct opsins across vertebrates in ortholog tables with sequence accession numbers. I have also combined opsins from each of the five major sub-families and provide the resulting maximum likelihood trees with bootstrap values based on consensus topologies of 1000 ML bootstrap phylogenies.

Ortholog tables. Each ortholog table is labeled with the most commonly annotated name and symbol, and contains the sequence accession numbers of opsins branching together in a monophyletic clade. Each of these ortholog clades contains representative protein sequences from multiple species, which branch in patterns that often match the species phylogeny. When there are entire groups of species that do not contain orthologous sequences for a particular gene, the rows of that group are collapsed into a single row representing the entire category. For example, no bird or mammal that was sampled has parietal opsin protein (PARIETOP) so the table contains just two rows for these categories of species, labeled "BIRDS" and "MAMMALS", referring to "many" species, and have an "X" in lieu of a gene name. Gene names cited as "NOVEL" or "UNPRED" are sequences that have not been annotated in their source database, and those with different names than the title of the table have most likely been incorrectly annotated.

Consensus and ML trees. For each sub-family of opsins, I found the tree with the maximum likelihood (out of 1000 bootstraps) and the protein tree representing each gene is an expanded/collapsed version of the sub-family tree to highlight the corresponding clade. Bootstrap values are based on the consensus tree of these 1000 trees (*Consense*), and an asterisk is present at nodes where the topologies of the consensus and ML trees disagree.

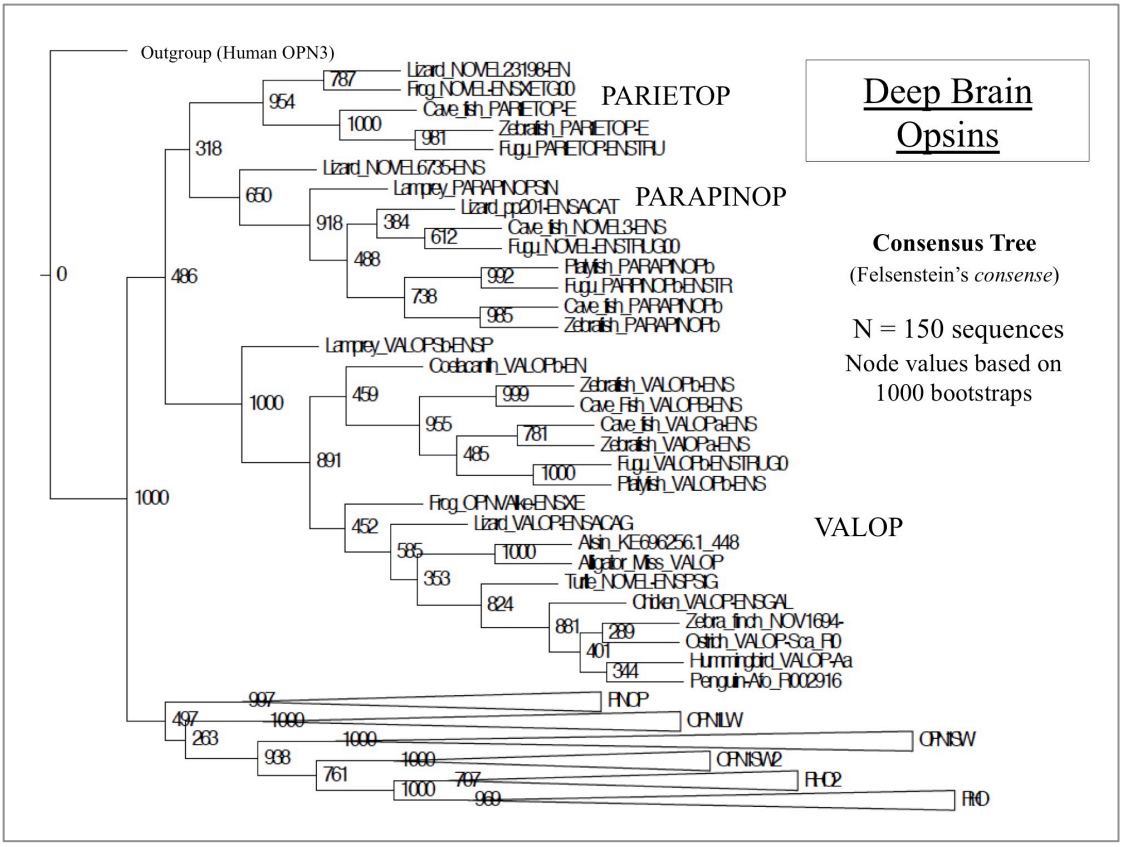


Figure A-2.1. Consensus tree of deep brain and visual opsins. Highlights internal branch support for deep brain, non-visual opsins parietal opsin, parapinopsin, and vertebrate ancient long wave-sensitive opsin.

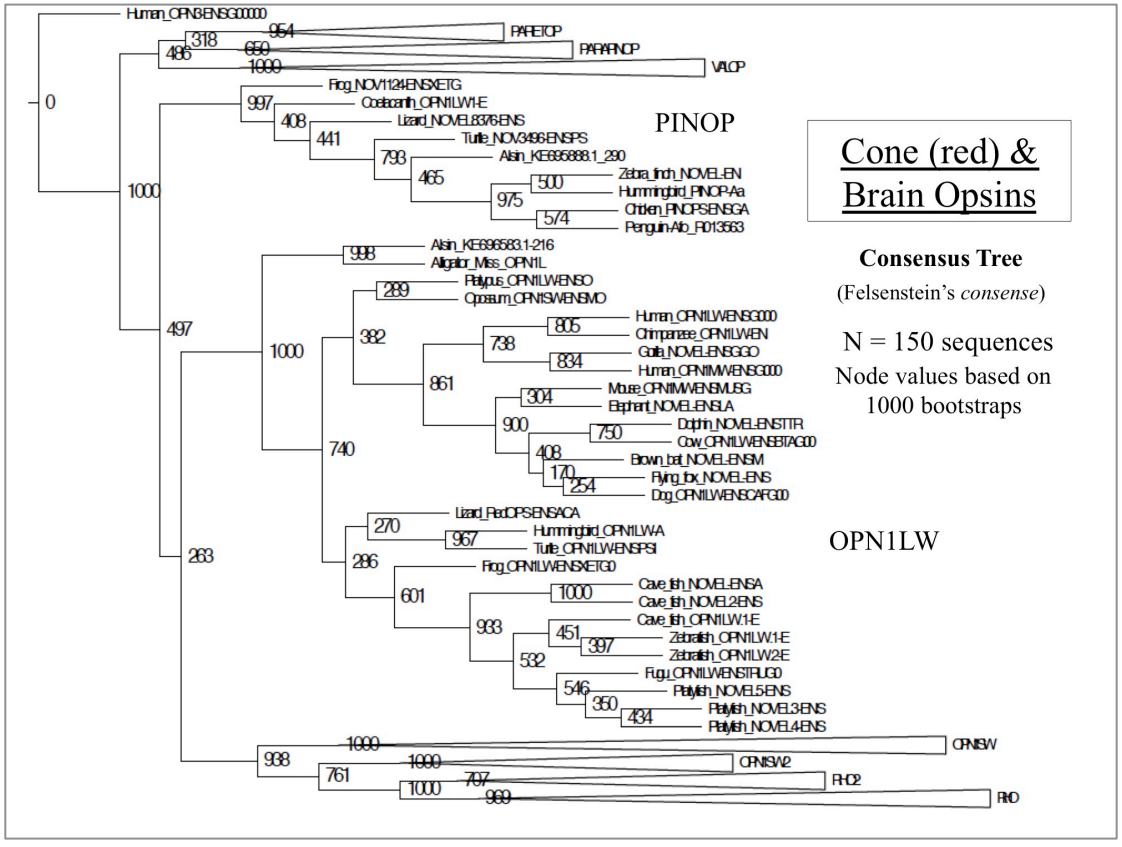


Figure A-2.2. Consensus tree of deep brain and visual opsins. Highlights internal branch support for pineal opsin (PINOP) and long wave-sensitive opsin 1 (OPN1LW), a red cone photopigment.

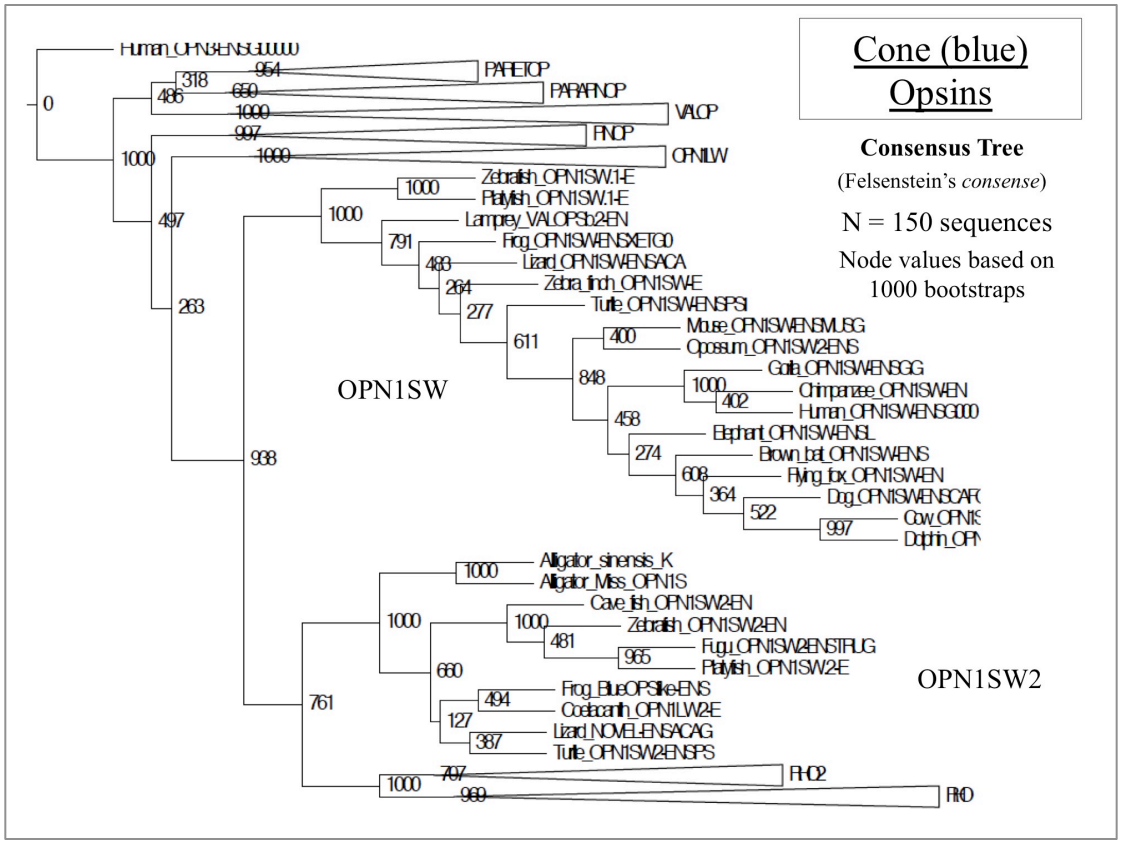


Figure A-2.3. Consensus tree of deep brain and visual opsins. Highlights internal branch support for blue cone opsins short wave-sensitive opsin 1 (OPN1SW) and short wave-sensitive opsin 1 2 (OPN1SW2).

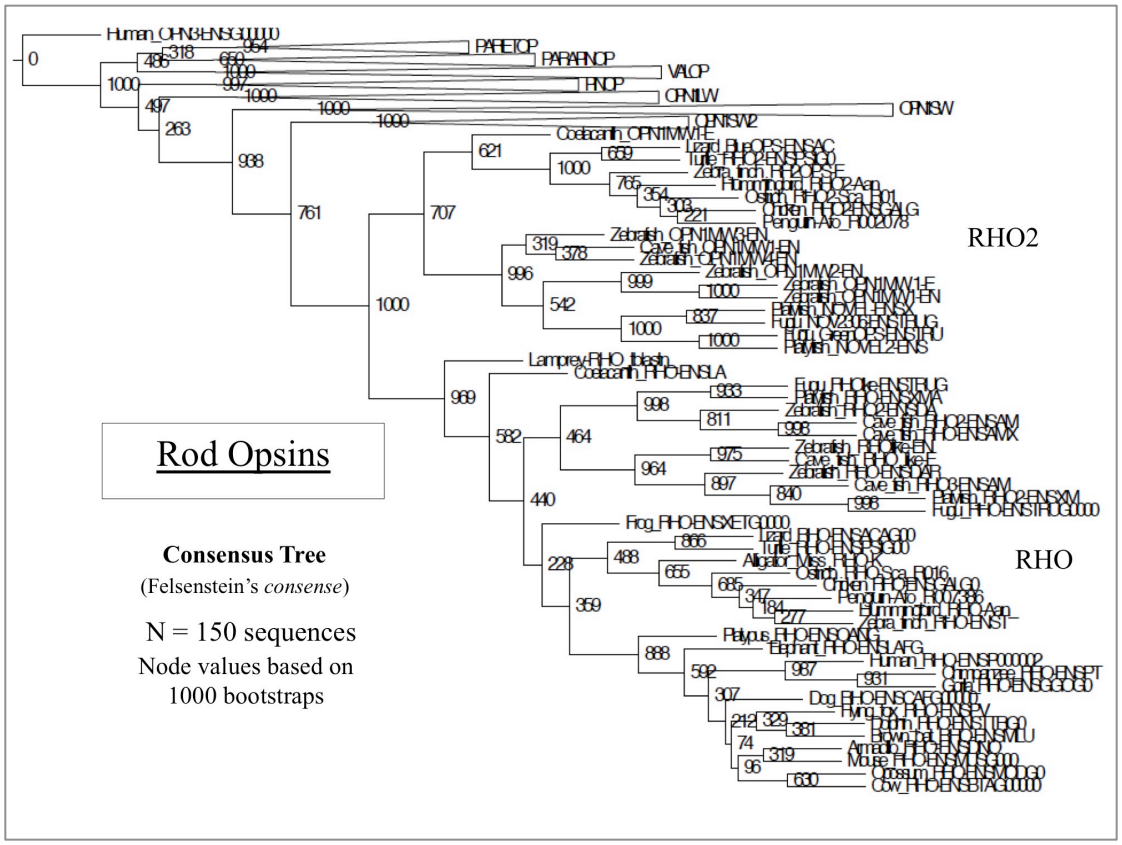


Figure A-2.4. Consensus tree of deep brain and visual opsins. Highlights internal branch support for rod opsins rhodopsin (RHO) and rhodopsin-like protein (RHO2).

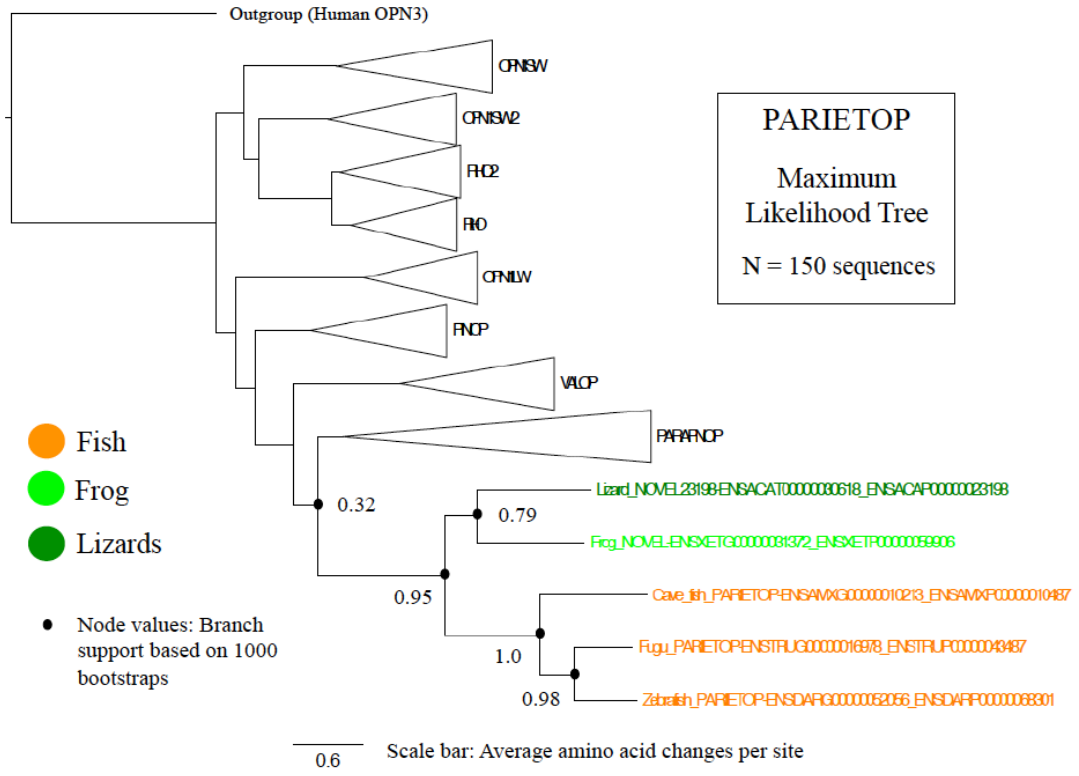


Figure A-3. Parietal opsin phylogeny. The PARIETOP clade of the deep brain and visual opsin ML tree is expanded to show topology and support for each node, based on a consensus tree of 1000 bootstraps. Only fish, frog, and lizard have a copy of this gene, and they have not been annotated in frog or lizard.

Table A-2. Parietal opsin orthologs. Catalog of sequences found in the PARIETOP gene clade of deep brain and visual opsins, represented only in fish, lizard, and frog; absent in all mammals and birds.

PARIETAL OPSIN (PARIETOP)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	X	-	-
Zebrafish	*D. rerio*	PARIETOP	ENSDARG00000052056	ENSDARP00000068301
Fugu	*T. rubripes*	PARIETOP	ENSTRUG00000016978	ENSTRUP00000043487
Platyfish	*X. maculatus*	X	-	-
Cave fish	*A. mexicanus*	PARIETOP	ENSAMXG00000010213	ENSAMXP00000010487
Coelacanth	*L. chalumnae*	X	-	-
Frog	*X. Tropicalis*	NOVEL	ENSXETG00000031372	ENSXETP00000059906
Lizard	*A. carolinensis*	NOVEL	ENSACAG00000028626	ENSACAP00000023198
Turtle	*P. sinensis*	X	-	-
Crocodile	*A. sinensis*	X	-	-
Alligator	*A. mississippiensis*	X	-	-
BIRDS	(many)	X	-	-
MAMMALS	(many)	X	-	-

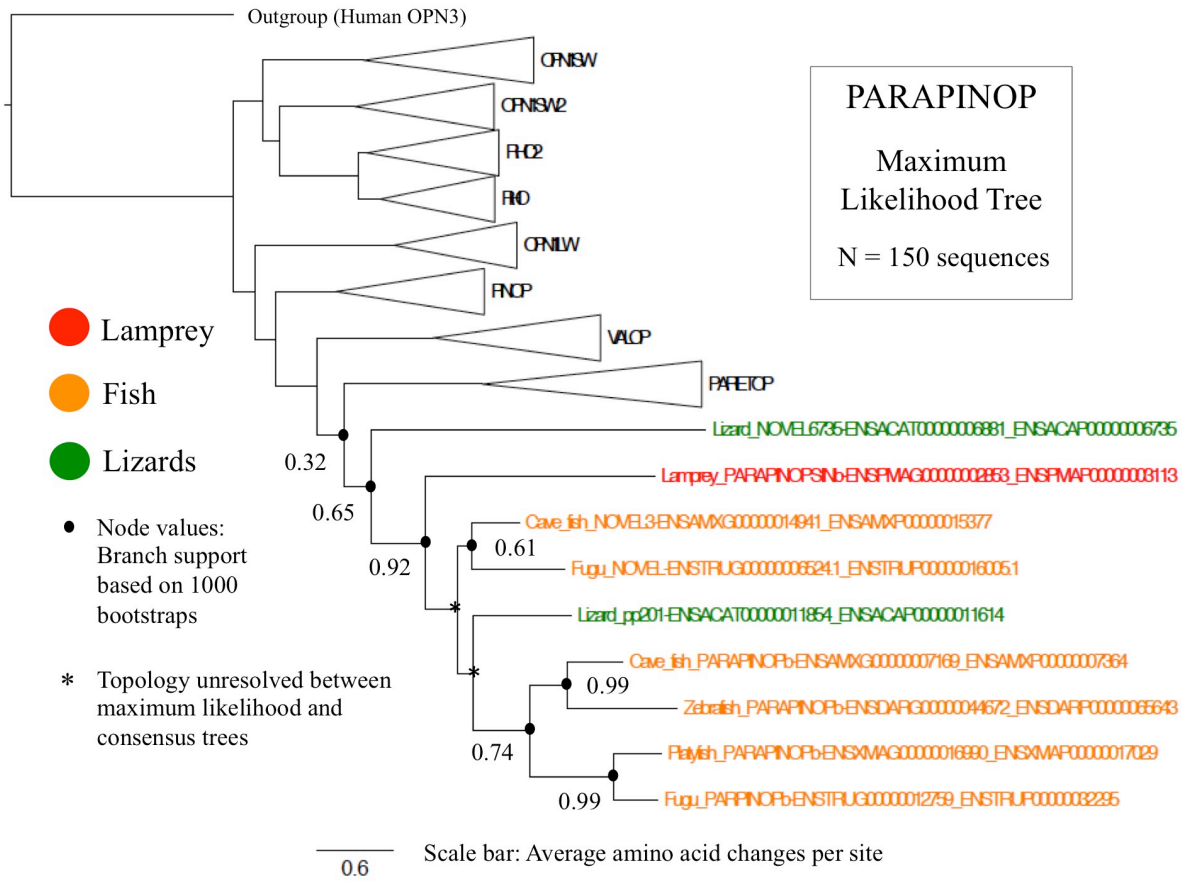


Figure A-4. Parapinopsin phylogeny. The PPINOP clade of the deep brain and visual opsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. Lamprey has one ortholog of this protein, while some fishes and lizard have two copies.

Table A-3. Parapineal opsin orthologs. Catalog of sequences found in the PPINOP clade of deep brain and visual opsins, represented in lamprey, fish, and lizard; absent in all mammals and birds.

PARAPINEAL OPSIN (PPINOP)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	PPINOPb	ENSPMAG00000002853	ENSPMAP00000003113
Zebrafish	*D. rerio*	PPINOPb	ENSDARG00000044672	ENSDARP00000065643
Fugu	*T. rubripes*	NOVEL	ENSTRUG00000006524.1	ENSTRUP00000016005.1
		PPINOPb	ENSTRUG00000012759	ENSTRUP00000032295
Platyfish	*X. maculatus*	PPINOPb	ENSXMAG00000016990	ENSXMAP00000017029
Cave fish	*A. mexicanus*	NOVEL	ENSAMXG00000014941	ENSAMXP00000015377
		PPINOPb	ENSAMXG00000007169	ENSAMXP00000007364
Coelacanth	*L. chalumnae*	X	-	-
Frog	*X. tropicalis*	X	-	-
Lizard	*A. carolinensis*	pp201	ENSACAT00000011854	ENSACAP00000011614
		NOVEL	ENSACAG00000006891	ENSACAP00000006735
Turtle	*P. sinensis*	X	-	-
CROCODILIANS	(two)	X	-	-
BIRDS	(many)	X	-	-
MAMMALS	(many)	X	-	-

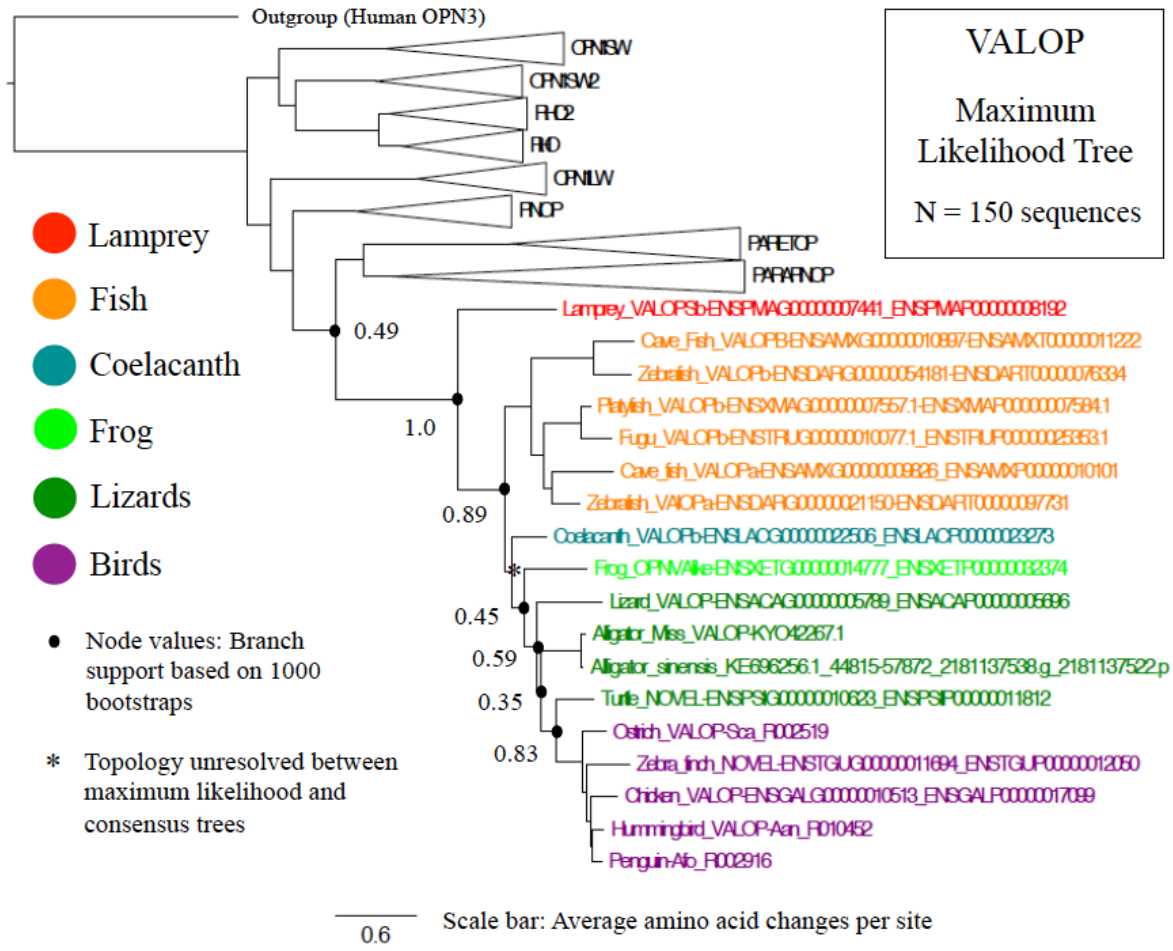


Figure A-5. Vertebrate ancient long-wave sensitive opsin phylogeny. The VALOP clade of the deep brain and visual opsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. All non-mammalian species sampled have at least one copy of this protein, and two copies are present in cave fish and zebrafish.

Table A-4. Vertebrate ancient long-wave sensitive opsin orthologs. Fish and coelacanth appear to be the only types of species with annotations of this protein (thus "NOVEL" and "UNPRED" names), despite its presence in all other species except for mammals. This was likely one of the first evolved opsins, which was subsequently lost early on in the mammalian lineage.

VERTEBRATE ANCIENT LONG-WAVELENGTH OPSIN (VALOP)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	VALOPb	ENSPMAG00000007441	ENSPMAP00000008192
Zebrafish	*D. rerio*	VALOPa	ENSDARG00000021150	ENSDARP00000088502
		VALOPb	ENSDARG00000054181	ENSDARP00000070808
Fugu	*T. rubripes*	VALOPb	ENSTRUG00000010077.1	ENSTRUP00000025353.1
Platyfish	*X. maculatus*	VALOPb	ENSXMAG00000007557.1	ENSXMAP00000007584.1
Cave fish	*A. mexicanus*	VALOPb	ENSAMXG00000010897.1	ENSAMXP00000011222.1
		VALOPa	ENSAMXG00000009826	ENSAMXP00000010101

Coelacanth	*L. chalumnae*	VALOPb	ENSLACG00000022506	ENSLACP00000023273
Frog	*X. tropicalis*	NOVEL	ENSXETG00000014777	ENSXETP00000032374
Lizard	*A. carolinensis*	NOVEL	ENSACAG00000005789	ENSACAP00000005696
Turtle	*P. sinensis*	NOVEL	ENSPSIG00000010623	ENSPSIP00000011812
Crocodile	*A. sinensis*	KE696256.1	2181137538.g	2181137522.p
Alligator	*A. mississippiensis*	KYO42267.1	N/A	N/A
Ostrich	*S. camelus*	UNPRED	Sca_R002519	N/A
Chicken	*G. gallus*	TCTN3	ENSGALG00000010513	ENSGALP00000017099
Hummingbird	*C. anna*	UNPRED	Aan_R010452	N/A
Zebra finch	*T. guttata*	NOVEL	ENSTGUG00000011694	ENSTGUP00000012050
Penguin	*A. forsteri*	UNPRED	Afo_R002916	N/A
MAMMALS	(many)	X	-	-

"Pineal Opsin (PINOP)"

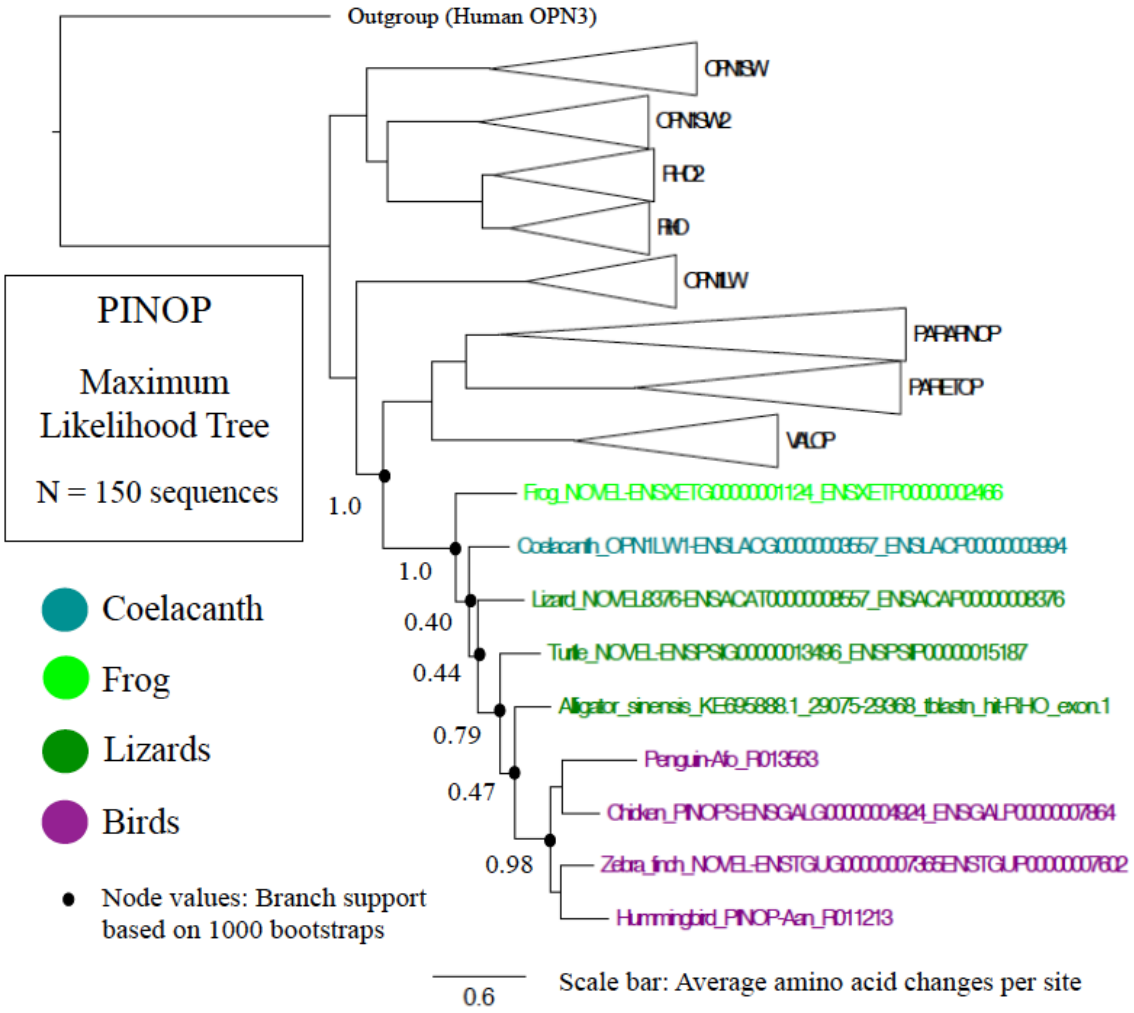


Figure A-6. Pinopsin phylogeny. The PINOP clade of the deep brain and visual opsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. Neither fish nor mammals have a copy of this gene; however, it is present in coelacanth, frog, lizard, turtle, and crocodile. All birds except ostrich also have a copy. Thus, it possibly arose on the reptilian lineage after the split from a common ancestor with fish, and was subsequently lost in the mammalian linage and ostrich. It is also possibly just missing from the ostrich genome assembly. Unfortunately, due to the lack of closely related sister species clades, it is difficult to ascertain whether this is a true gene loss on the branch to ostrich or a genome sequencing/assembly error.

Table A-5. Pinopsin orthologs. The only species with an annotated pinopsin was chicken, and although coelacanth has an ortholog of this protein, it was incorrectly annotated as "OPN1LW1". Most other reptiles and birds have a copy of this protein, but they are annotated as "NOVEL" or "UNPREDICTED".

PINEAL OPSIN, PINOPSIN (PINOP)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	X	-	-
BONY FISHES	(many)	X	-	-
Coelacanth	*L. chalumnae*	OPN1LW1	ENSLACG00000003557	ENSLACP00000003994
Frog	*X. Tropicalis*	NOVEL	ENSXETG00000001124	ENSXETP00000002466
Lizard	*A. carolinensis*	NOVEL	ENSACAT00000008557	ENSACAP00000008376
Turtle	*P. sinensis*	NOVEL	ENSPSIG00000013496	ENSPSIP00000015187
Crocodile	*A. sinensis*	UNPRED	KE695888.1	(tblastn hit from *RHO*)
Alligator	*A. mississippiensis*	X	-	-
Ostrich	*S. camelus*	X	-	-
Chicken	*G. gallus*	PINOPS	ENSGALG00000004924	ENSGALP00000007864
Hummingbird	*C. anna*	UNPRED	Aan_R011213	N/A
Zebra finch	*T. guttata*	UNPRED	ENSTGUG00000007365	ENSTGUP00000007602
Penguin	*A. forsteri*	UNPRED	Afo_R013563	N/A
MAMMALS	(many)	X	-	-

"Long Wave-sensitive Opsin 1 (OPN1LW)"

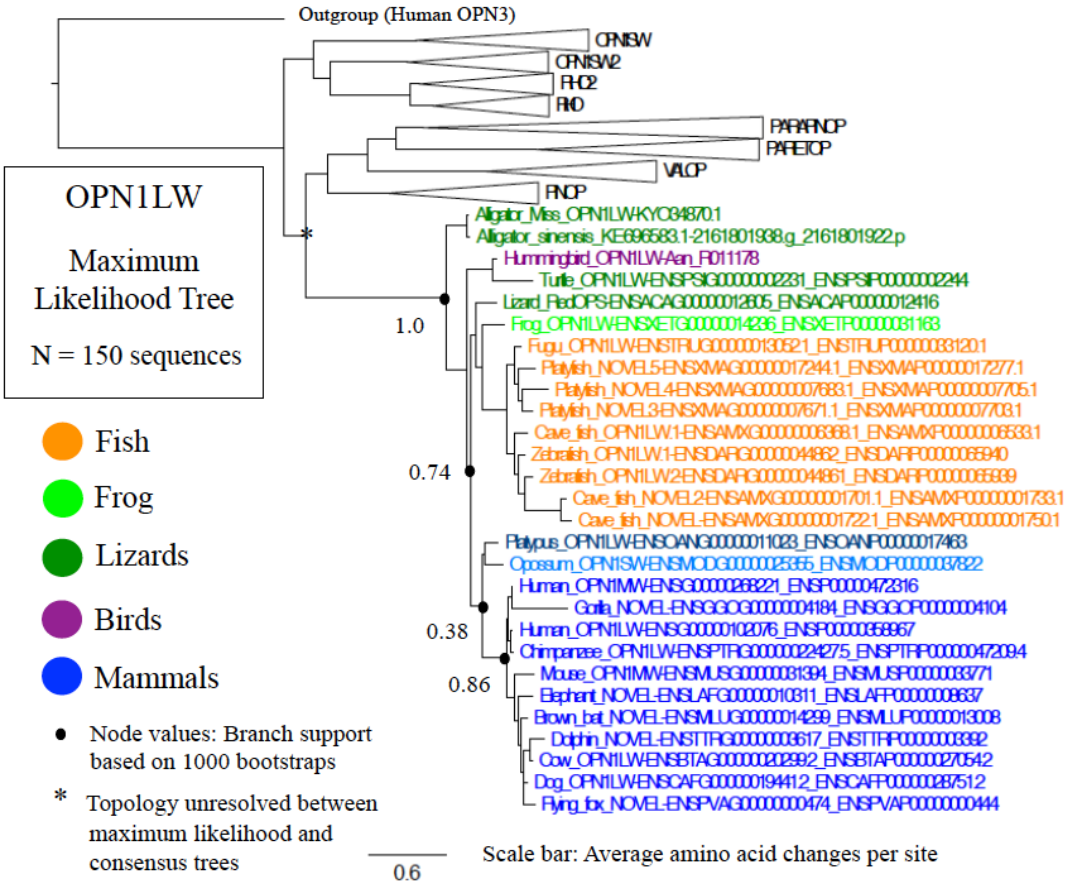


Figure A-7. Long wave-sensitive opsin 1 phylogeny. The OPN1LW clade of the deep brain and visual opsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. Support is high for the separation of primate sequences from other mammals, as well as between human-chimp (OPN1LW) and human-gorilla orthologs (OPN1MW).

Table A-6. Long wave-sensitive opsin 1 orthologs. All fish sampled had at least one copy of OPN1LW, with several duplicated versions in three out of four. Among birds, OPN1LW was detected only in hummingbird, whereas a copy was found in all other reptiles. Humans have two copies, while armadillo is missing the gene and all other mammals have a single copy.

LONG WAVE-SENSITIVE OPSIN 1 (OPN1LW)

Species short-hand	Species Name	Annotated Name	Gene Accession Number	Protein Accession Number
Lamprey	*P. marinus*	X	-	-
Zebrafish	*D. rerio*	OPN1LW.2	ENSDARG00000044861	ENSDARP00000065939
		OPN1LW.1	ENSDARG00000044862	ENSDARP00000065940
Fugu	*T. rubripes*	OPN1LW	ENSTRUG00000013052.1	ENSTRUP00000033120.1
Platyfish	*X. maculatus*	NOVEL	ENSXMAG00000007671.1	ENSXMAP00000007703.1
		NOVEL	ENSXMAG00000017244.1	ENSXMAP00000017277.1
		NOVEL	ENSXMAG00000007683.1	ENSXMAP00000007705.1
Cave fish	*A. mexicanus*	NOVEL	ENSAMXG00000001722.1	ENSAMXP00000001750.1
		NOVEL	ENSAMXG00000001701.1	ENSAMXP00000001733.1
		OPN1LW.1	ENSAMXG00000006368.1	ENSAMXP00000006533.1
Coelacanth	*L .chalumnae*	X	-	-
Frog	*X. tropicalis*	OPN1LW	ENSXETG00000014236	ENSXETP00000031163
Lizard	*A. carolinensis*	RedOPS	ENSACAG00000012605	ENSACAP00000012416
Turtle	*P. sinensis*	NOVEL	ENSPSIG00000002231	ENSPSIP00000002244
Crocodile	*A. sinensis*	KE696583.1	2161801938.g	2161801922.p
Alligator	*A. mississippiensis*	KYO34870.1	N/A	N/A
Ostrich	*S. camelus*	X	-	-
Chicken	*G. gallus*	X	-	-
Hummingbird	*C. anna*	UNPRED	Aan_R011178	N/A
Zebra finch	*T. guttata*	X	-	-
Penguin	*A. forsteri*	X	-	-
Platypus	*O. anatinus*	OPN1LW	ENSOANG00000011023	ENSOANP00000017463
Opossum	*M. domestica*	OPN1SW	ENSMODG00000025355	ENSMODP00000037822
Armadillo	*D. novemcinctus*	X		
Elephant	*L. africana*	NOVEL	ENSLAFG00000010311	ENSLAFP00000008637
Mouse	*M. musculus*	OPN1MW	ENSMUSG00000031394	ENSMUSP00000033771
Gorilla	*G. gorilla*	NOVEL	ENSGGOG00000004184	ENSGGOP00000004104
Human	*H. sapiens*	OPN1LW	ENSG00000102076	ENSP00000358967
		OPN1MW	ENSG00000268221	ENSP00000472316
Chimpanzee	*P. Troglodytes*	OPN1LW	ENSPTRG00000022427.5	ENSPTRP00000047209.4
Flying fox	*P. vampyrus*	NOVEL	ENSPVAG00000000474	ENSPVAP00000000444
Brown bat	*M. lucifugus*	NOVEL	ENSMLUG00000014299	ENSMLUP00000013008
Dog	*C. familiaris*	OPN1LW	ENSCAFG00000019441.2	ENSCAFP00000028751.2
Dolphin	*T. truncatus*	NOVEL	ENSTTRG00000003617	ENSTTRP00000003392
Cow	*B. taurus*	OPN1LW	ENSBTAG00000020299.2	ENSBTAP00000027054.2

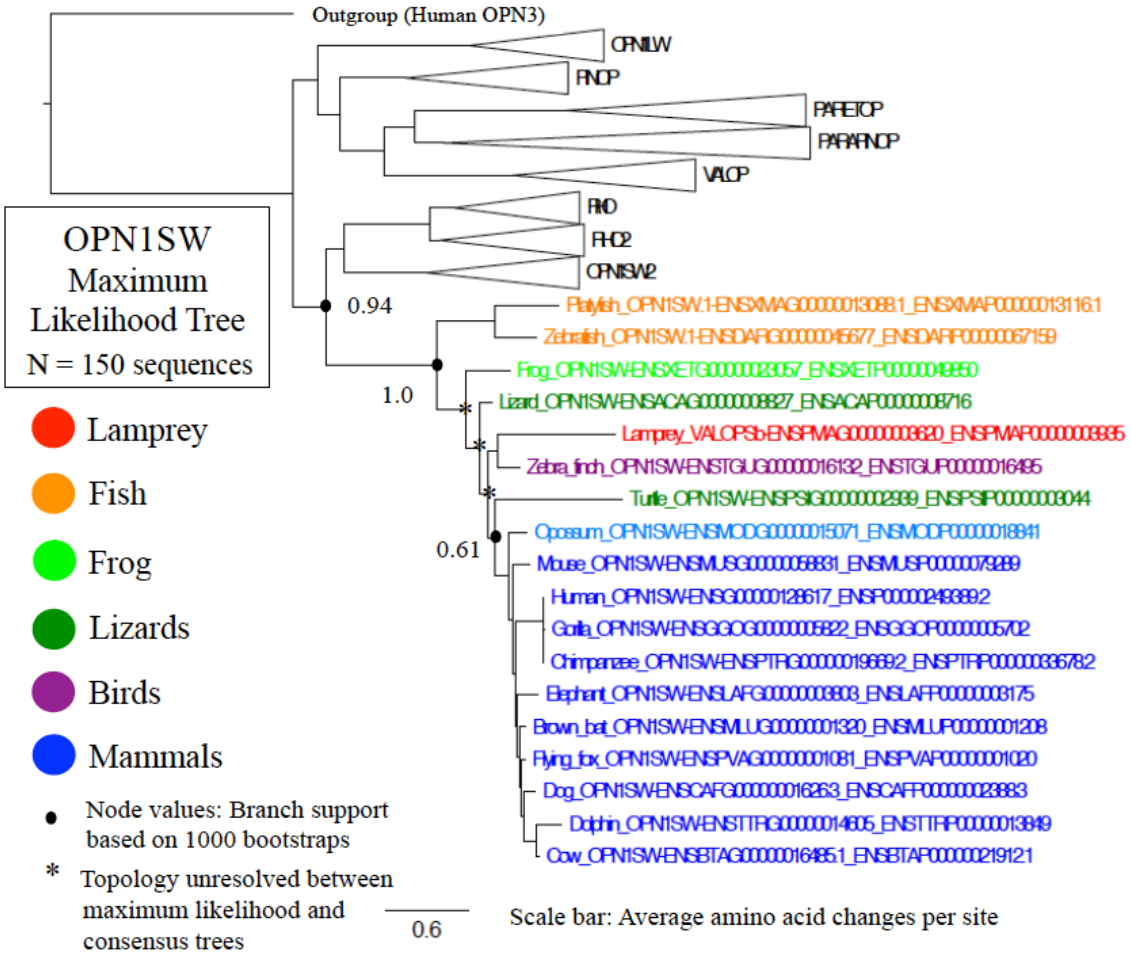


Figure A-8. Short wave-sensitive opsin 1 phylogeny. The OPN1SW clade of the deep brain and visual opsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. All categories of species are found to have orthologs of OPN1SW, but several species in each category are missing the gene.

Table A-7. Short wave-sensitive opsin 1 orthologs. Gene loss of OPN1SW seems to be common across species in different lineages, with orthologs present in a single bird out of four, two fish out of four, one bird out of three, and three of the remaining reptiles. Additionally, armadillo and platypus are missing this gene, while there is a single copy in all other mammals.

SHORT WAVE-SENSITIVE OPSIN 1 (OPN1SW)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	VALOPSb	ENSPMAG00000003620	ENSPMAP00000003935
Zebrafish	*D. rerio*	OPN1SW.1	ENSDARG00000045677	ENSDARP00000067159
Fugu	*T. rubripes*	X	-	-
Platyfish	*X. maculatus*	OPN1SW.1	ENSXMAG00000013088.1	ENSXMAP00000013116.1
Cave fish	*A. mexicanus*	X	-	-
Coelacanth	*L .chalumnae*	X	-	-
Frog	*X. tropicalis*	OPN1SW	ENSXETG00000023057	ENSXETP00000049850
Lizard	*A. carolinensis*	OPN1SW	ENSACAG00000008827	ENSACAP00000008716
Turtle	*P. sinensis*	OPN1SW	ENSPSIG00000002939	ENSPSIP00000003044
Crocodile	*A. sinensis*	X	-	-

Alligator	*A. mississippiensis*	X	-	-
Zebra finch	*T. guttata*	OPN1SW	ENSTGUG00000016132	ENSTGUP00000016495
OTHER BIRDS	(many)	X	-	-
Platypus	*O. anatinus*	X	-	-
Opossum	*M. domestica*	OPN1SW	ENSMODG00000015071	ENSMODP00000018841
Armadillo	*D. novemcinctus*	X	-	-
Elephant	*L. africana*	OPN1SW	ENSLAFG00000003803	ENSLAFP00000003175
Mouse	*M. musculus*	OPN1SW	ENSMUSG00000058831	ENSMUSP00000079289
Gorilla	*G. gorilla*	OPN1SW	ENSGGOG00000005822	ENSGGOP00000005702
Human	*H. sapiens*	OPN1SW	ENSG00000128617	ENSP00000249389.2
Chimpanzee	*P. troglodytes*	OPN1SW	ENSPTRG00000019669.2	ENSPTRP00000033678.2
Flying fox	*P. vampyrus*	OPN1SW	ENSPVAG00000001081	ENSPVAP00000001020
Brown bat	*M. lucifugus*	OPN1SW	ENSMLUG00000001320	ENSMLUP00000001208
Dog	*C. familiaris*	OPN1SW	ENSCAFG00000001626.3	ENSCAFP00000002388.3
Dolphin	*T. truncatus*	OPN1SW	ENSTTRG00000014605	ENSTTRP00000013849
Cow	*B. taurus*	OPN1SW	ENSBTAG00000016485.1	ENSBTAP00000021912.1

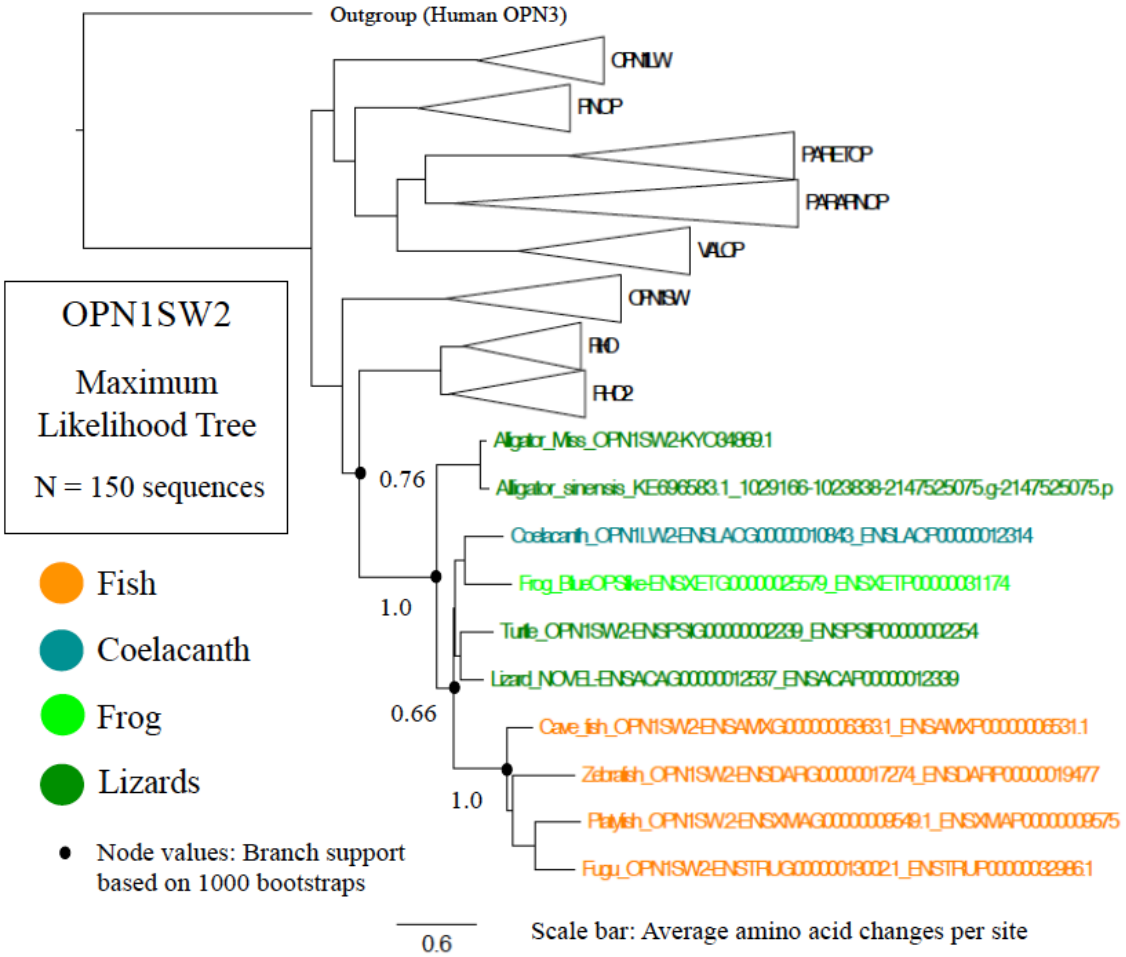


Figure A-9. Short wave-sensitive opsin 1, 2 phylogeny. The OPN1SW2 clade of the deep brain and visual opsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. Fishes, coelacanth, frog, and all the non-bird reptiles have a copy of this gene, suggesting it may be advantageous in an aquatic environment.

Table A-8. Short wave-sensitive opsin 1, 2 orthologs. All five of the species that are missing OPN1SW in the fish and crocodilian clades have a copy of OPN1SW2: coelacanth, alligator, crocodile, cave fish, and fugu. This suggests that perhaps only one version of this gene is necessary in certain habitats, while the conservation of both copies in frog, lizard, turtle, and two fish species suggests the importance of both in other environmental habitats or species niches.

SHORT WAVE-SENSITIVE OPSIN 1, 2 (OPN1SW2)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	X	-	-
Zebrafish	*D. rerio*	OPN1SW2	ENSDARG00000017274	ENSDARP00000019477
Fugu	*T. rubripes*	OPN1SW2	ENSTRUG00000013002.1	ENSTRUP00000032986.1
Platyfish	*X. maculatus*	OPN1SW.2	ENSXMAG00000009549.1	ENSXMAP00000009575.1
Cave fish	*A. mexicanus*	OPN1SW2	ENSAMXG00000006363.1	ENSAMXP00000006531.1
Coelacanth	*L .chalumnae*	OPN1LW2	ENSLACG00000010843	ENSLACP00000012314
Frog	*X. tropicalis*	BlueOPSlike	ENSXETG00000025579	ENSXETP00000031174
Lizard	*A. carolinensis*	NOVEL	ENSACAG00000012537	ENSACAP00000012339
Turtle	*P. sinensis*	NOVEL	ENSPSIG00000002239	ENSPSIP00000002254
Crocodile	*A. sinensis*	KE696583.1	2147525075.g	2147525075.p
Alligator	*A. mississippiensis*	UNPRED	KYO34869.1	N/A
BIRDS	(many)	X	-	-
MAMMALS	(many)	X	-	-

"Rhodopsin (RHO)"

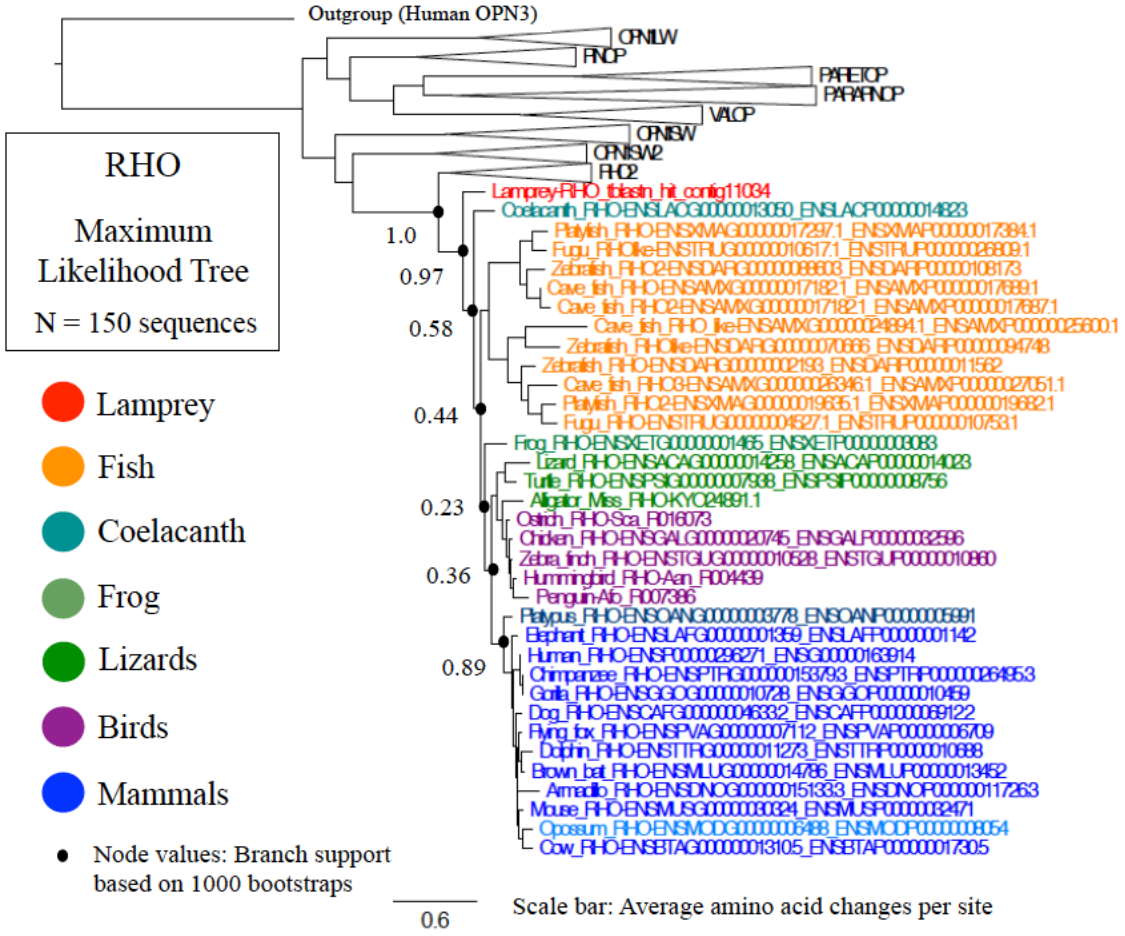


Figure A-10. Rhodopsin phylogeny. The RHO clade of the deep brain and visual opsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. This gene is highly conserved across all lineages, and the gene tree is similar to the species tree.

Table A-9. Rhodopsin orthologs. Every species sampled had at least one copy of RHO, with the exception of crocodile. Due to the high level of conservation across all other species, and its presence in alligator, it is likely that the appearance of a missing gene is due to an incomplete genome sequence or assembly. Alternatively, it may have been lost in the crocodile lineage and nowhere else. All fish have at least two copies of this gene, demonstrating that they have retained both copies from the whole-genome duplication, and it is likely advantageous to have multiple versions of RHO in an aquatic environment.

RHODOPSIN (RHO)

Species short-hand	Species Name	Annotated Name	Gene Accession Number	Protein Accession Number
Lamprey	*P. marinus*	UNPRED	tblastn hit contig11034	N/A
Zebrafish	*D. rerio*	RHO	ENSDARG00000002193	ENSDARP00000011562
		RHO2	ENSDARG00000088603	ENSDARP00000108173
		RHOlike	ENSDARG00000070666	ENSDARP00000094748
Fugu	*T. rubripes*	RHO	ENSTRUG00000004527.1	ENSTRUP00000010753.1
		RHOlike	ENSTRUG00000010617.1	ENSTRUP00000026809.1
Platyfish	*X. maculatus*	RHO	ENSXMAG00000017297.1	ENSXMAP00000017384.1
		RHO2	ENSXMAG00000019635.1	ENSXMAP00000019682.1
Cave fish	*A. mexicanus*	RHO	ENSAMXG00000017182.1	ENSAMXP00000017689.1
		RHOlike	ENSAMXG00000024894.1	ENSAMXP00000025600.1
		RHO2	ENSAMXG00000017182.1	ENSAMXP00000017687.1
		RHO3	ENSAMXG00000026346.1	ENSAMXP00000027051.1
Coelacanth	*L. chalumnae*	RHO	ENSLACG00000013050	ENSLACP00000014823
Frog	*X. tropicalis*	RHO	ENSXETG00000001465	ENSXETP00000003083
Lizard	*A. carolinensis*	RHO	ENSACAG00000014258	ENSACAP00000014023
Turtle	*P. sinensis*	RHO	ENSPSIG00000007938	ENSPSIP00000008756
Crocodile	*A. sinensis*	-	-	-
Alligator	*A. mississippiensis*	KYO24891.1	N/A	N/A
Ostrich	*S. camelus*	UNPRED	Sca_R016073	N/A
Chicken	*G. gallus*	RHO	ENSGALG00000020745	ENSGALP00000032596
Hummingbird	*C. anna*	UNPRED	Aan_R004439	N/A
Zebra finch	*T. guttata*	RHO	ENSTGUG00000010528	ENSTGUP00000010860
Penguin	*A. forsteri*	UNPRED	Afo_R007386	N/A
Platypus	*O. anatinus*	RHO	ENSOANG00000003778	ENSOANP00000005991
Opossum	*M. domestica*	RHO	ENSMODG00000006488	ENSMODP00000008054
Armadillo	*D. novemcinctus*	RHO	ENSDNOG00000015133.3	ENSDNOP00000011726.3
Elephant	*L. africana*	RHO	ENSLAFG00000001359	ENSLAFP00000001142
Mouse	*M. musculus*	RHO	ENSMUSG00000030324	ENSMUSP00000032471
Gorilla	*G. gorilla*	RHO	ENSGGOG00000010728	ENSGGOP00000010459
Human	*H. sapiens*	RHO	ENSG00000163914	ENSP00000296271
Chimpanzee	*P. troglodytes*	RHO	ENSPTRG00000015379.3	ENSPTRP00000026495.3
Flying fox	*P. vampyrus*	RHO	ENSPVAG00000007112	ENSPVAP00000006709
Brown bat	*M. lucifugus*	RHO	ENSMLUG00000014786	ENSMLUP00000013452
Dog	*C. familiaris*	RHO	ENSCAFG00000004633.2	ENSCAFP00000006912.2
Dolphin	*T. truncatus*	RHO	ENSTTRG00000011273	ENSTTRP00000010688
Cow	*B. taurus*	RHO	ENSBTAG00000001310.5	ENSBTAP00000001730.5

"Rhodopsin 2 (RHO2 or RH2)"

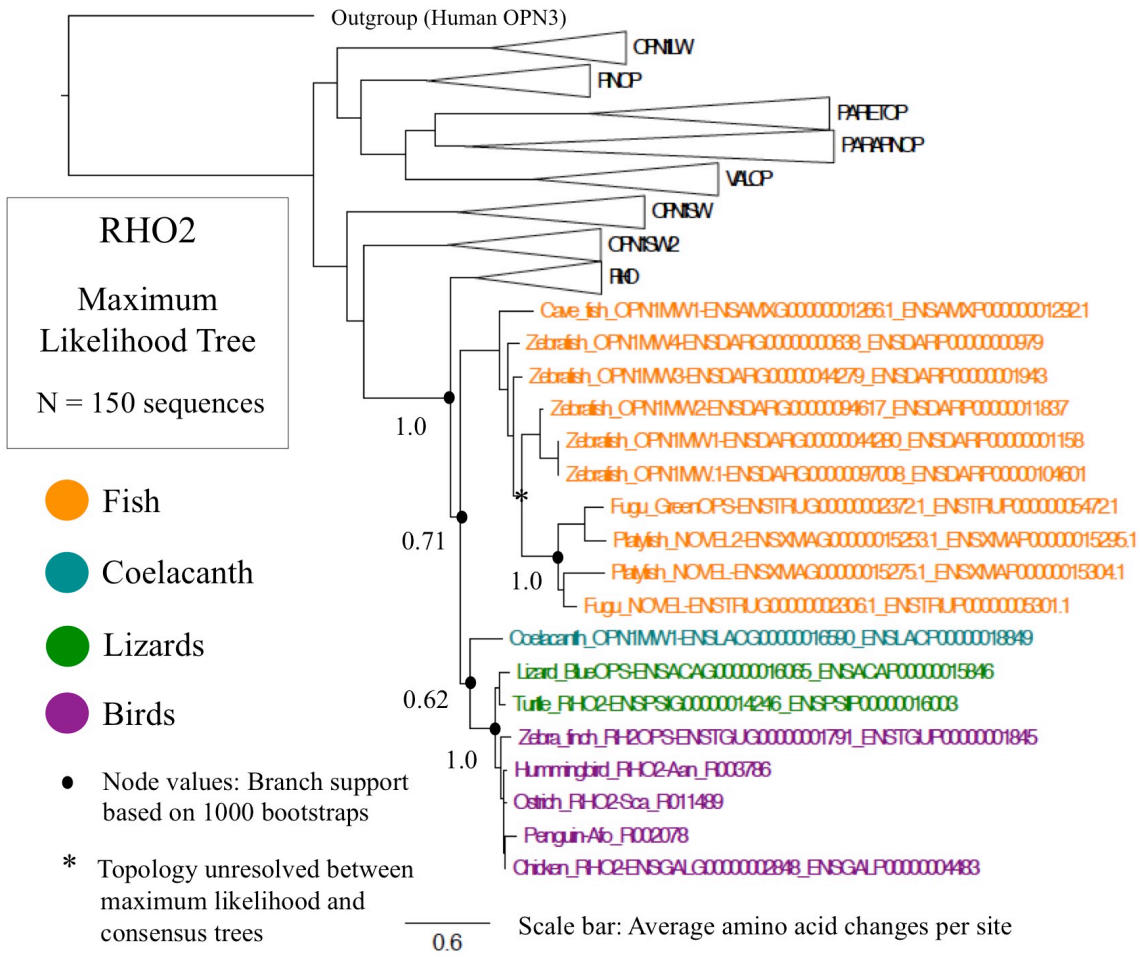


Figure A-11. Rhodopsin 2 phylogeny. The RHO2 clade of the deep brain and visual opsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. All sequences in this clade are monophyletic, with strong branch support demonstrating that they indeed came from an initial common ancestral gene, most likely a duplicate version of RHO, its closest relative. RHO2 has multiple copies with annotations varying widely from OPN1MW in zebrafish to GreenOPS in fugu and BlueOPS in lizard.

Table A-10. Rhodopsin 2 orthologs. Similar to its sister gene, *RHO, RHO2* is present in all fish and birds; however, it is missing in frog, crocodile and alligator, and is missing entirely in mammals. Zebrafish has five copies of *RHO2*; the other fish have at least one. It is possible that the other fish also have an expanded suite of versions of this gene, but their genomes have been less intensively analyzed in comparison to zebrafish and as such are not present in the database.

RHODOPSIN-2 (RHO2)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	-	-	-
Zebrafish	*D. rerio*	OPN1MW1	ENSDARG00000044280	ENSDARP00000001158
		OPN1MW2	ENSDARG00000094617	ENSDARP00000011837
		OPN1MW3	ENSDARG00000044279	ENSDARP00000001943
		OPN1MW4	ENSDARG00000000638	ENSDARP00000000979
		OPN1MW.1	ENSDARG00000097008	ENSDARP00000104601
Fugu	*T. rubripes*	GreenOPS	ENSTRUG00000002372.1	ENSTRUP00000005472.1

		NOVEL	ENSTRUG00000002306.1	ENSTRUP00000005301.1
Platyfish	*X. maculatus*	NOVEL	ENSXMAG00000015275.1	ENSXMAP00000015304.1
		NOVEL	ENSXMAG00000015253.1	ENSXMAP00000015295.1
Cave fish	*A. mexicanus*	OPN1MW1	ENSAMXG00000001266.1	ENSAMXP00000001292.1
Coelacanth	*L. chalumnae*	OPN1MW1	ENSLACG00000016590	ENSLACP00000018849
Frog	*X. tropicalis*	-	-	-
Lizard	*A. carolinensis*	BlueOPS	ENSACAG00000016065	ENSACAP00000015846
Turtle	*P. sinensis*	RHO2	ENSPSIG00000014246	ENSPSIP00000016003
Crocodile	*A. sinensis*	-	-	-
Alligator	*A. mississippiensis*	-	-	-
Ostrich	*S. camelus*	UNPRED	Sca_R011489	N/A
Chicken	*G. gallus*	GreenOPS	ENSGALG00000002848	ENSGALP00000004483
Hummingbird	*C. anna*	UNPRED	Aan_R003786	N/A
Zebra finch	*T. guttata*	RH2OPS	ENSTGUG00000001791	ENSTGUP00000001845
Penguin	*A. forsteri*	UNPRED	Afo_R002078	N/A
MAMMALS	(many)	-	-	-

PANOPSINS

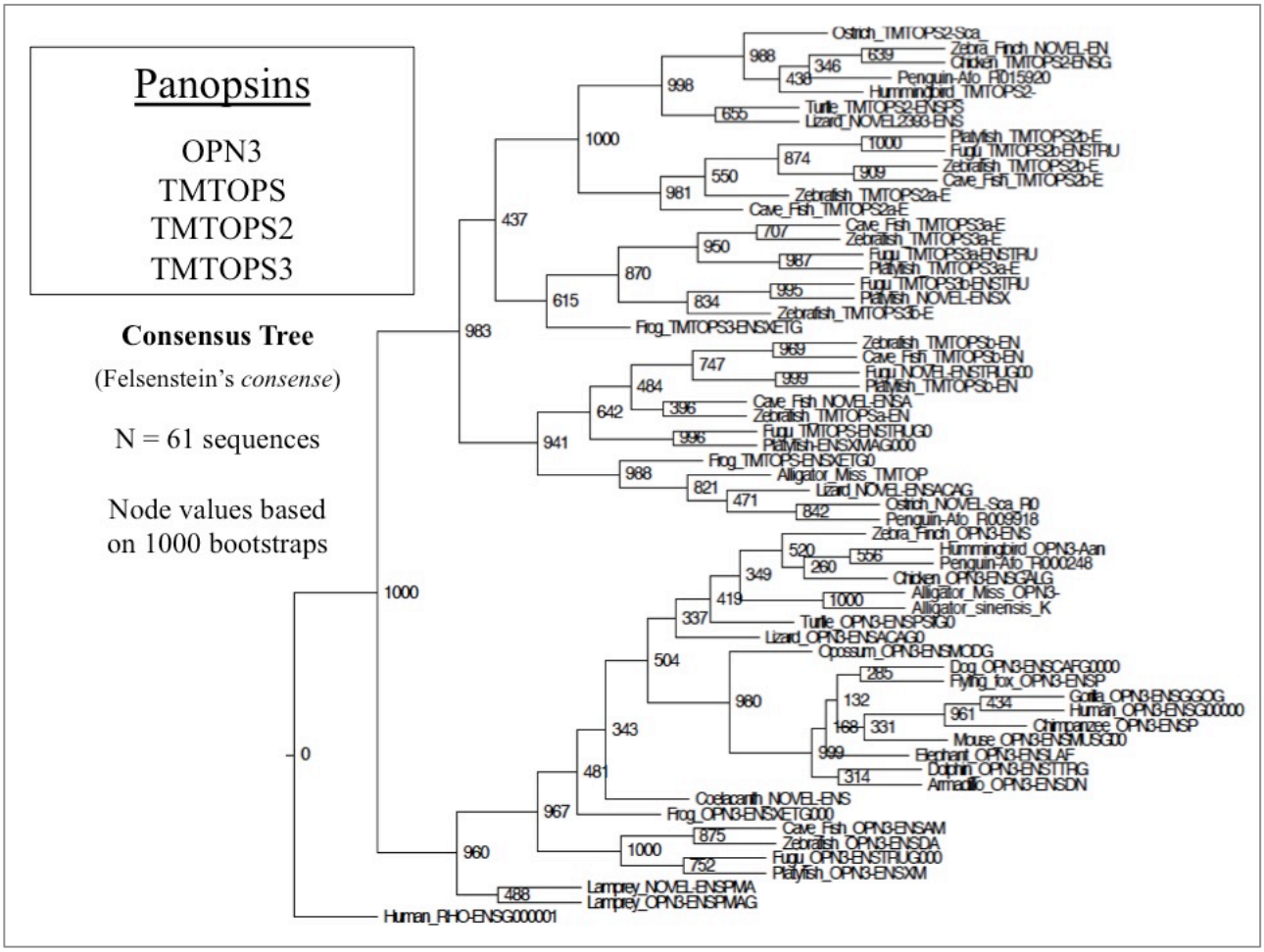


FIGURE A-12. Consensus tree for the "Panopsin" sub-family. Bootstrap values reflect the proportion of trees generated with the branching pattern shown among 1000 bootstrap trees. The number of sequences used over the whole tree was N = 61. Consensus tree obtained from Felsenstein's *consense* program, using 1000 maximum likelihood trees generated in *PhyML*

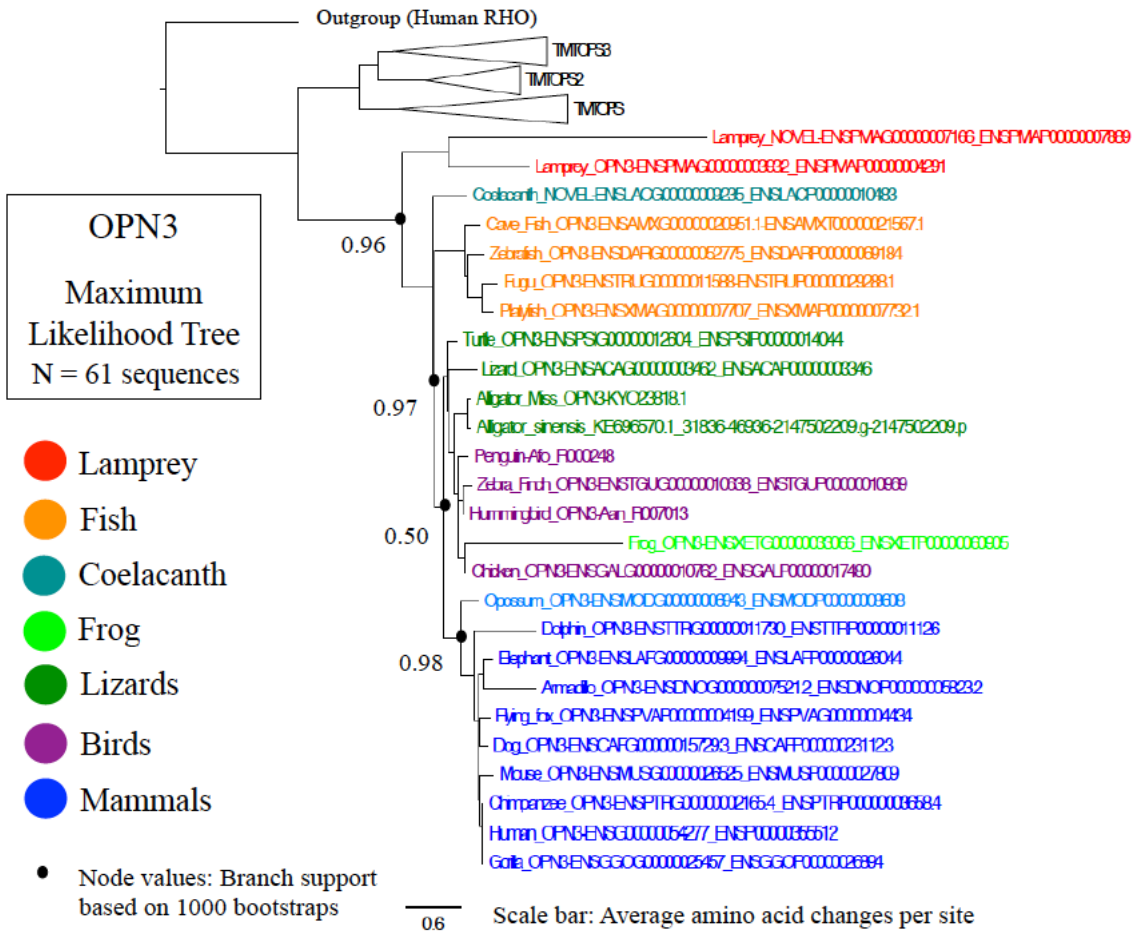


Figure A-13. Panopsin/encephalopsin phylogeny. The OPN3 clade of the panopsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. All categories of species have a representative in this gene clade, including lamprey and coelacanth, demonstrating its high level of conservation across species.

Table A-11. Panopsin/encephalopsin orthologs. All species other than ostrich, platypus, brown bat, and cow have a single copy of this gene, while lamprey has two paralagous (duplicated) versions. This is a profile of a highly conserved gene, with an unusual pattern of loss across various lineages and species.

PANOPSIN/ENCEPHALOPSIN (OPN3)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	NOVEL	ENSPMAG00000007166	ENSPMAP00000007889
		OPN3	ENSPMAG00000003932	ENSPMAP00000004291
Zebrafish	*D. rerio*	OPN3	ENSDARG00000052775	ENSDARP00000069184
Fugu	*T. rubripes*	OPN3	ENSTRUG00000011588	ENSTRUP00000029288.1
Platyfish	*X. maculatus*	OPN3	ENSXMAG00000007707	ENSXMAP00000007732.1
Cave fish	*A. mexicanus*	OPN3	ENSAMXG00000020951.1	ENSAMXP00000021567.1
Coelacanth	*L .chalumnae*	OPN3	ENSLACG00000009235	ENSLACP00000010483
Frog	*X. tropicalis*	OPN3	ENSXETG00000033066	ENSXETP00000060905
Lizard	*A. carolinensis*	OPN3	ENSACAG00000003462	ENSACAP00000003346
Turtle	*P. sinensis*	OPN3	ENSPSIG00000012604	ENSPSIP00000014044
Crocodile	*A. sinensis*	KE696570.1	2147502209.g	2147502209.p

Alligator	*A. mississippiensis*	KYO23818.1	N/A	N/A
Ostrich	*S. camelus*	X	-	-
Chicken	*G. gallus*	OPN3	ENSGALG00000010762	ENSGALP00000017480
Hummingbird	*C. anna*	UNPRED	Aan_R007013	N/A
Zebra finch	*T. guttata*	OPN3	ENSTGUG00000010638	ENSTGUP00000010969
Penguin	*A. forsteri*	UNPRED	Afo_R000248	N/A
Platypus	*O. anatinus*	X	-	-
Opossum	*M. domestica*	OPN3	ENSMODG00000006943	ENSMODP00000008608
Armadillo	*D. novemcinctus*	OPN3	ENSDNOG00000007521.2	ENSDNOP00000005823.2
Elephant	*L. africana*	OPN3	ENSLAFG00000009994	ENSLAFP00000026044
Mouse	*M. Musculus*	OPN3	ENSMUSG00000026525	ENSMUSP00000027809
Gorilla	*G. gorilla*	OPN3	ENSGGOG00000025457	ENSGGOP00000026894
Human	*H. sapiens*	OPN3	ENSG00000054277	ENSP00000355512
Chimpanzee	*P. Troglodytes*	OPN3	ENSPTRG00000002165.4	ENSPTRP00000003658.4
Flying fox	*P. vampyrus*	OPN3	ENSPVAG00000004434	ENSPVAP00000004199
Brown bat	*M. lucifugus*	X	-	-
Dog	*C. familiaris*	OPN3	ENSCAFG00000015729.3	ENSCAFP00000023112.3
Dolphin	*T. truncatus*	OPN3	ENSTTRG00000011730	ENSTTRP00000011126
Cow	*B. taurus*	X	-	-

"Teleost Multiple-Tissue Opsin (TMTOPS)"

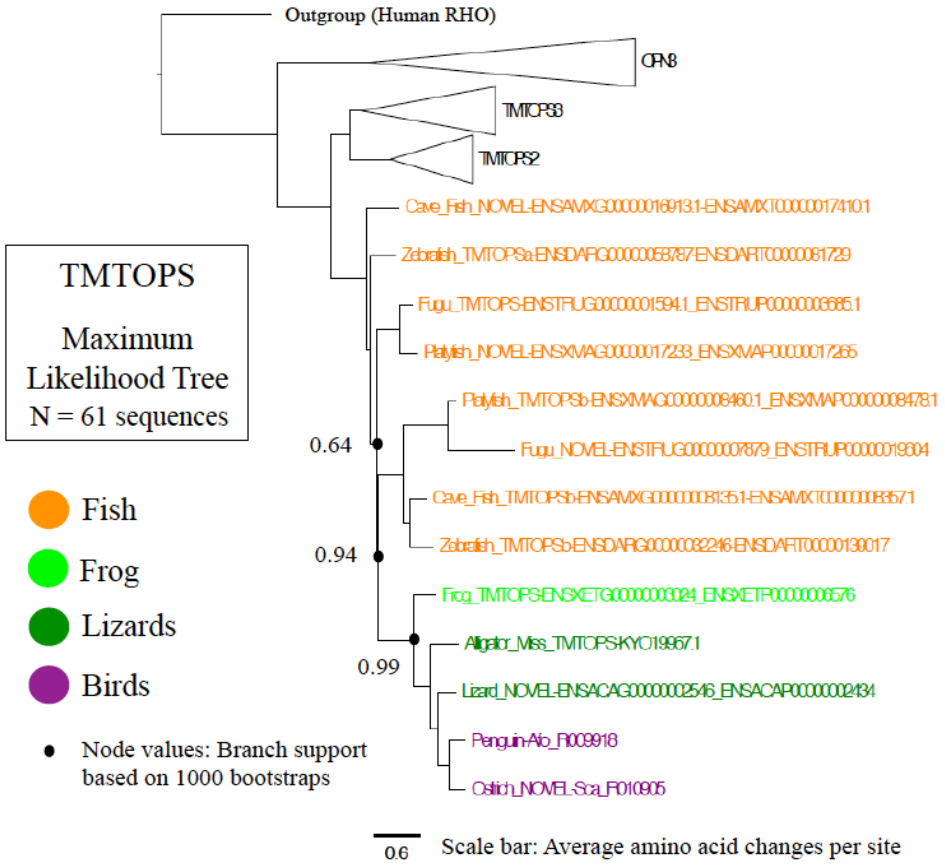


Figure A-14. Teleost multiple-tissue opsin phylogeny. The TMTOPS clade of the panopsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. Two copies are found in fish, with one copy each in frog, alligator, lizard, penguin, and ostrich.

Table A-12. Teleost multiple-tissue opsin orthologs. Two paralagous copies of this gene are conserved in all the fish sampled, while a single copy is present in frog, lizard, alligator, ostrich, and penguin. The gene is not present in mammals, lamprey, coelacanth, or most of the birds, suggesting it may have lineage-specific functions and can be useful in an aquatic environment. It also suggests that this protein is not vital for survival in an aquatic environment, since coelacanth and lamprey do not have copies. It likely arose as a duplicate version of OPN3, its sister protein, before the fish-reptilian common ancestral lineage divergence and was subsequently lost in several lineages over time.

TELEOST MULTIPLE-TISSUE OPSIN (TMTOPS)

Species short-hand	Species Name	Annotated Name	Gene Accession Number	Protein Accession Number
Lamprey	*P. marinus*	X	-	-
Zebrafish	*D. rerio*	TMTOPSa	ENSDARG00000058787	ENSDARP00000076168
		TMTOPSb	ENSDARG00000032246	ENSDARP00000037563
Fugu	*T. rubripes*	TMTOPS	ENSTRUG00000001594.1	ENSTRUP00000003685.1
		NOVEL	ENSTRUG00000007879	ENSTRUP00000019604
Platyfish	*X. maculatus*	TMTOPSb	ENSXMAG00000008460.1	ENSXMAP00000008478.1
		NOVEL	ENSXMAG00000017233	ENSXMAP00000017265
Cave fish	*A. mexicanus*	TMTOPSb	ENSAMXG00000008135.1	ENSAMXT00000008357.1
		NOVEL	ENSAMXG00000016913.1	ENSAMXP00000017410.1
Coelacanth	*L. chalumnae*	X	-	-
Frog	*X. tropicalis*	TMTOPS.2	ENSXETG00000003024	ENSXETT00000006576
Lizard	*A. carolinensis*	NOVEL	ENSACAG00000002546	ENSACAP00000002434
Turtle	*P. sinensis*	X	-	-
Crocodile	*A. sinensis*	X	-	-
Alligator	*A. mississippiensis*	UNPRED	KYO19967.1	N/A
Ostrich	*S. camelus*	NOVEL	Sca_R010905	N/A
Chicken	*G. gallus*	X	-	-
Hummingbird	*C. anna*	X	-	-
Zebra finch	*T. guttata*	X	-	-
Penguin	*A. forsteri*	NOVEL	Afo_R009918	N/A
MAMMALS	(many)	X	-	-

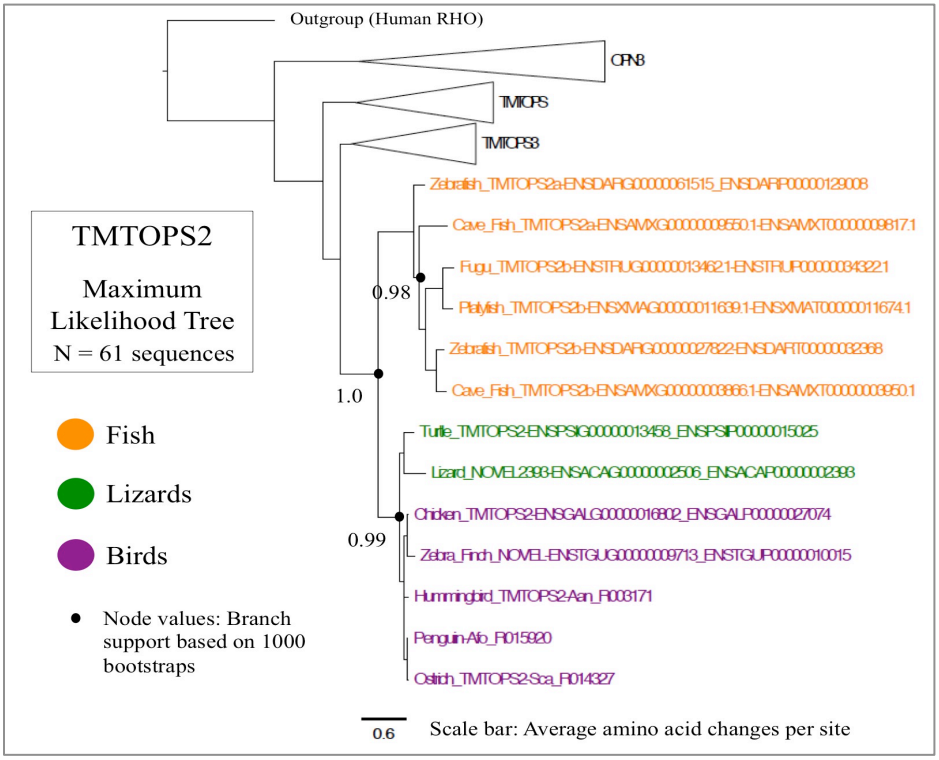


Figure A-15. Teleost multiple-tissue opsin 2 phylogeny. The TMTOPS2 clade of the panopsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. Fish, bird, and lizard lineages are represented in this clade.

Table A-13. Teleost multiple-tissue opsin 2 orthologs. Half of the fish retained two copies, while the other fishes and lizard, turtle, as well as all five birds have a single ortholog of this protein.

TELEOST MULTIPLE-TISSUE OPSIN 2 (TMTOPS2)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	X	-	-
Zebrafish	*D. rerio*	TMTOPS2b	ENSDARG00000027822	ENSDARP00000030023
		TMTOPS2a	ENSDARG00000061515	ENSDARP00000129008
Fugu	*T. rubripes*	TMTOPS2b	ENSTRUG00000013462.1	ENSTRUP00000034322.1
Platyfish	*X.maculatus*	TMTOPS2b	ENSXMAG00000011639.1	ENSXMAP00000011660.1
Cave fish	*A. mexicanus*	TMTOPS2a	ENSAMXG00000009550.1	ENSAMXP00000009817.1
		TMTOPS2b	ENSAMXG00000003866.1	ENSAMXP00000003950.1
Coelacanth	*L. chalumnae*	X	-	-
Frog	*X. tropicalis*	X	-	-
Lizard	*A. carolinensis*	NOVEL	ENSACAG00000002506	ENSACAP00000002393
Turtle	*P. sinensis*	NOVEL	ENSPSIG00000013458	ENSPSIP00000015025
CROCODILIANS	(two)	X	-	-
Ostrich	*S. camelus*	NOVEL	Sca_R014327	N/A
Chicken	*G. gallus*	NOVEL	ENSGALG00000016802	ENSGALP00000027074
Hummingbird	*C. anna*	NOVEL	Aan_R003171	N/A
Zebra finch	*T. guttata*	NOVEL	ENSTGUG00000009713	ENSTGUP00000010015
Penguin	*A. forsteri*	NOVEL	Afo_R015920	N/A
MAMMALS	(many)	X	-	-

"Teleost Multiple-Tissue Opsin 3 (TMTOPS3)"

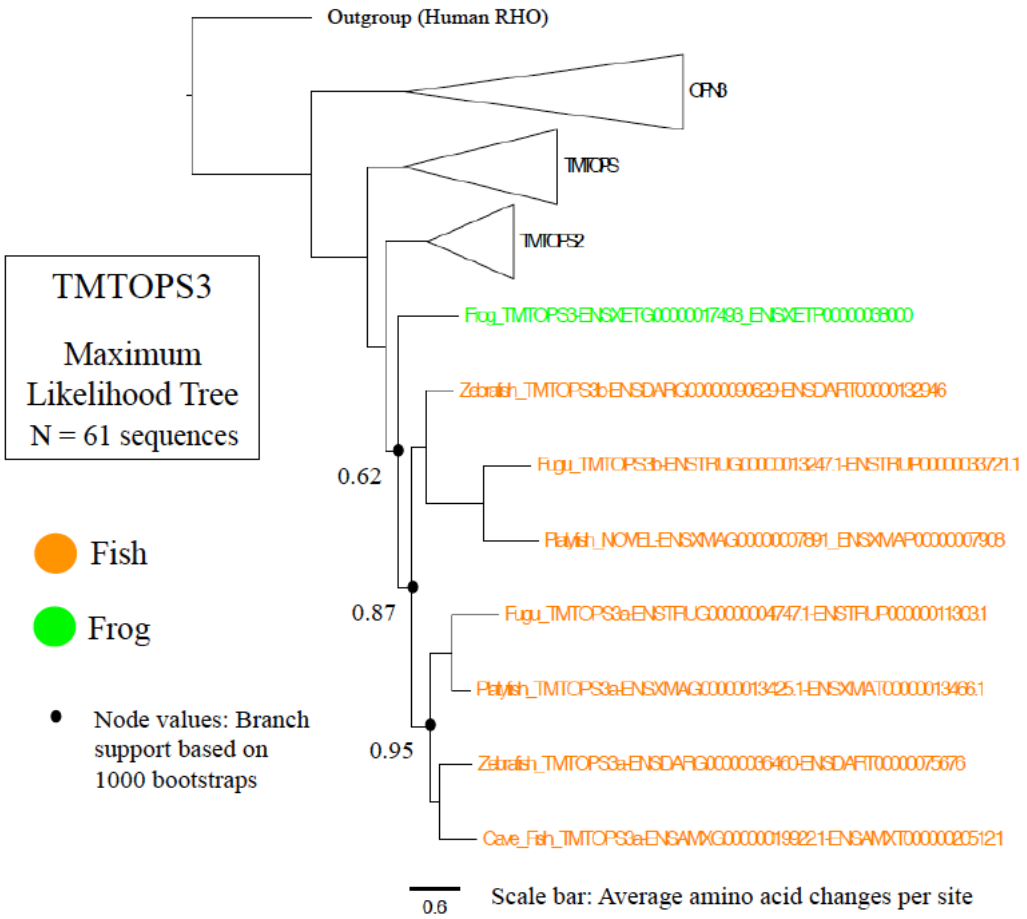


Figure A-16. Teleost Multiple-Tissue Opsin 3 phylogeny. The TMTOPS3 clade of the panopsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. Fish and frog have copies of this protein.

Table A-14. Teleost Multiple-Tissue Opsin 3 ortholog table. Sister to TMTOPS and TMTOPS2, only fish and frog appear to have TMTOPS3; two copies were found in fish and one in frog.

TELEOST MULTIPLE-TISSUE OPSIN 3 (*TMTOPS3*)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	X	-	-
Zebrafish	*D. rerio*	TMTOPS3a	ENSDARG00000036460	ENSDARP00000070158
		TMTOPS3b	ENSDARG00000090629	ENSDARP00000105609
Fugu	*T. rubripes*	TMTOPS3a	ENSTRUG00000013247.1	ENSTRUP00000033721.1
		TMTOPS3b	ENSTRUG00000004747.1	ENSTRUP00000011303.1
Platyfish	*X. maculatus*	TMTOPS3a	ENSXMAG00000013425.1	ENSXMAP00000013450.1
		NOVEL	ENSXMAG00000007891	ENSXMAP00000007908
Cave fish	*A. mexicanus*	TMTOPS3a	ENSAMXG00000019922.1	ENSAMXP00000020512.1
Frog	*X. tropicalis*	TMTOPS	ENSXETG00000017493	ENSXETP00000038000
LIZARDS	(many)	X	-	-
BIRDS	(many)	X	-	-
MAMMALS	(many)	X	-	-

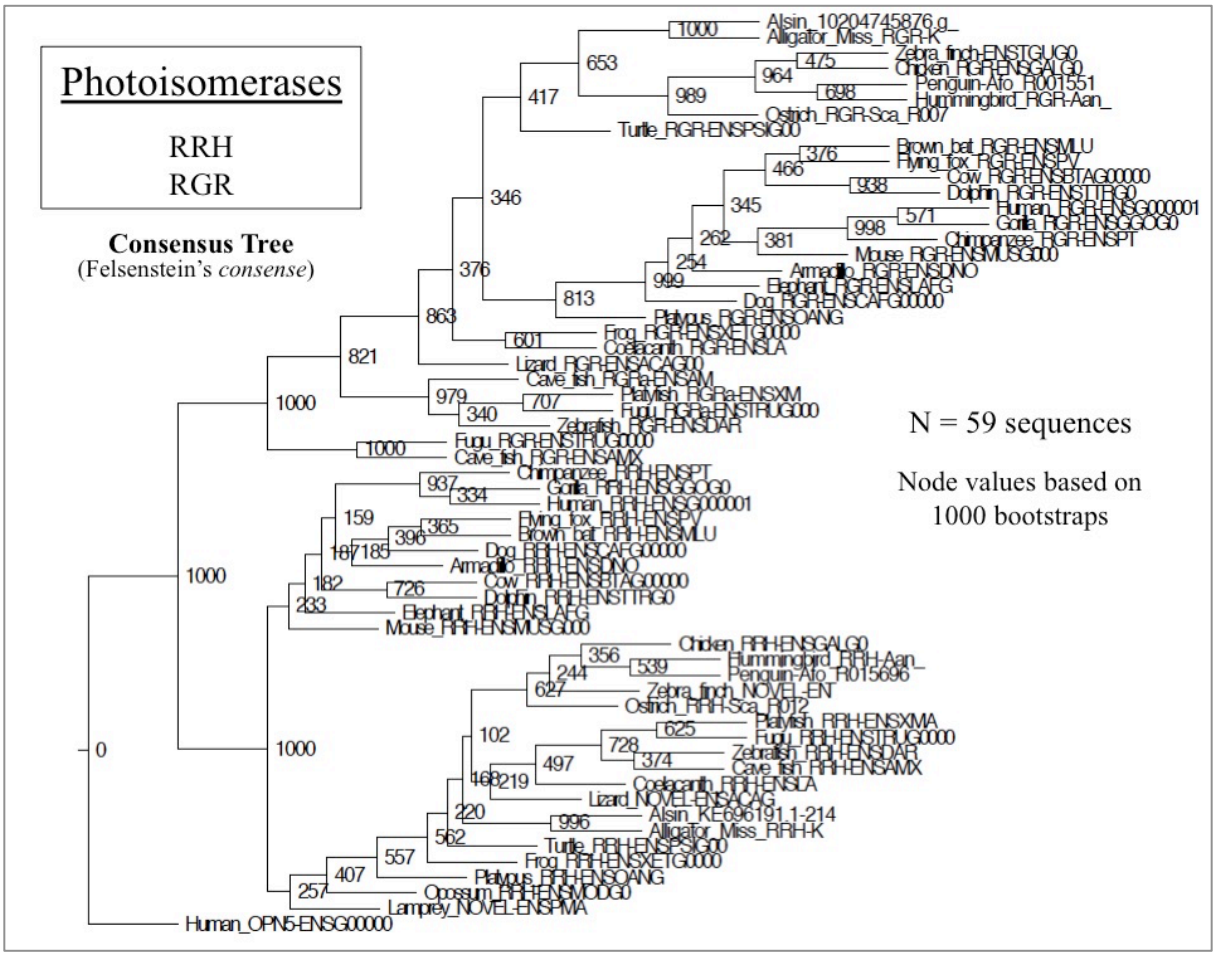


Figure A-17. Consensus tree for the "Photoisomerase" sub-family. Bootstrap values reflect the proportion of trees generated with the branching pattern shown among 1000 bootstrap trees. The number of sequences used over the whole tree was N = 59. Consensus tree obtained from Felsenstein's *consense* program, using 1000 maximum likelihood trees generated in *PhyML*.

"Retinal G Protein-Coupled Receptor (RGR)"

Table A-15. Retinal G protein-coupled receptor orthologs. All species sampled had at least one copy of RGR except for opossum and lamprey; two copies are conserved in some fish. This is thus a highly conserved protein that is (on its own) likely advantageous, but not essential, for life.

RETINAL G PROTEIN-COUPLED RECEPTOR (RGR)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	X	-	-
Zebrafish	*D. rerio*	RGR	ENSDARG00000054890	ENSDARP00000071634
Fugu	*T. rubripes*	RGRa	ENSTRUG00000009604.1	ENSTRUP00000024127.1
		RGR	ENSTRUG00000011877.1	ENSTRUP00000030014.1
Platyfish	*X. maculatus*	RGR	ENSXMAG00000003015.1	ENSXMAP00000003023.1
Cave fish	*A. mexicanus*	RGRa	ENSAMXG00000012172.1	ENSAMXP00000012519.1
		RGR	ENSAMXG00000004323.1	ENSAMXP00000004427.1
Coelacanth	*L. chalumnae*	RGR	ENSLACG00000013333	ENSLACP00000015146
Frog	*X. tropicalis*	RGR	ENSXETG00000005627	ENSXETP00000012410
Lizard	*A. carolinensis*	RGR	ENSACAG00000015384	ENSACAP00000015111

Turtle	*P. sinensis*	RGR	ENSPSIG00000003741	ENSPSIP00000003961
Crocodile	*A. sinensis*	UNPRED	10204745876.g	10204745860.p
Alligator	*A. mississippiensis*	KYO27724.1	N/A	N/A
Ostrich	*S. camelus*	UNPRED	Sca_R007528	N/A
Chicken	*G.gallus*	RGR	ENSGALG00000002299	ENSGALP00000003608
Hummingbird	*C. anna*	UNPRED	Aan_R012485	N/A
Zebra finch	*T. guttata*	NOVEL	ENSTGUG00000005992	ENSTGUP00000006166
		NOVEL	ENSTGUG00000014211	ENSTGUP00000014596
Penguin	*A. forsteri*	UNPRED	Afo_R001551	N/A
Platypus	*O. anatinus*	RGR	ENSOANG00000012634	ENSOANP00000019980
Opossum	*M. domestica*	X	-	-
Armadillo	*D. novemcinctus*	RGR	ENSDNOG00000004879.3	ENSDNOP00000003771.3
Elephant	*L. africana*	RGR	ENSLAFG00000014674	ENSLAFP00000012288
Mouse	*M. musculus*	RGR	ENSMUSG00000021804	ENSMUSP00000022338
Gorilla	*G. gorilla*	RGR	ENSGGOG00000012798	ENSGGOP00000012484
Human	*H. sapiens*	RGR	ENSG00000148604	ENSP00000350823
Chimpanzee	*P. troglodytes*	RGR	ENSPTRG00000002695.5	ENSPTRP00000004704.5
Flying fox	*P. vampyrus*	RGR	ENSPVAG00000004643	ENSPVAP00000004396
Brown bat	*M. lucifugus*	RGR	ENSMLUG00000016974	ENSMLUP00000015474
Dog	*C. familiaris*	RGR	ENSCAFG00000015930.4	ENSCAFP00000023432.2
Dolphin	*T. truncatus*	RGR	ENSTTRG00000015072	ENSTTRP00000014288
Cow	*B. taurus*	RGR	ENSBTAG00000015681.5	ENSBTAP00000020822.5

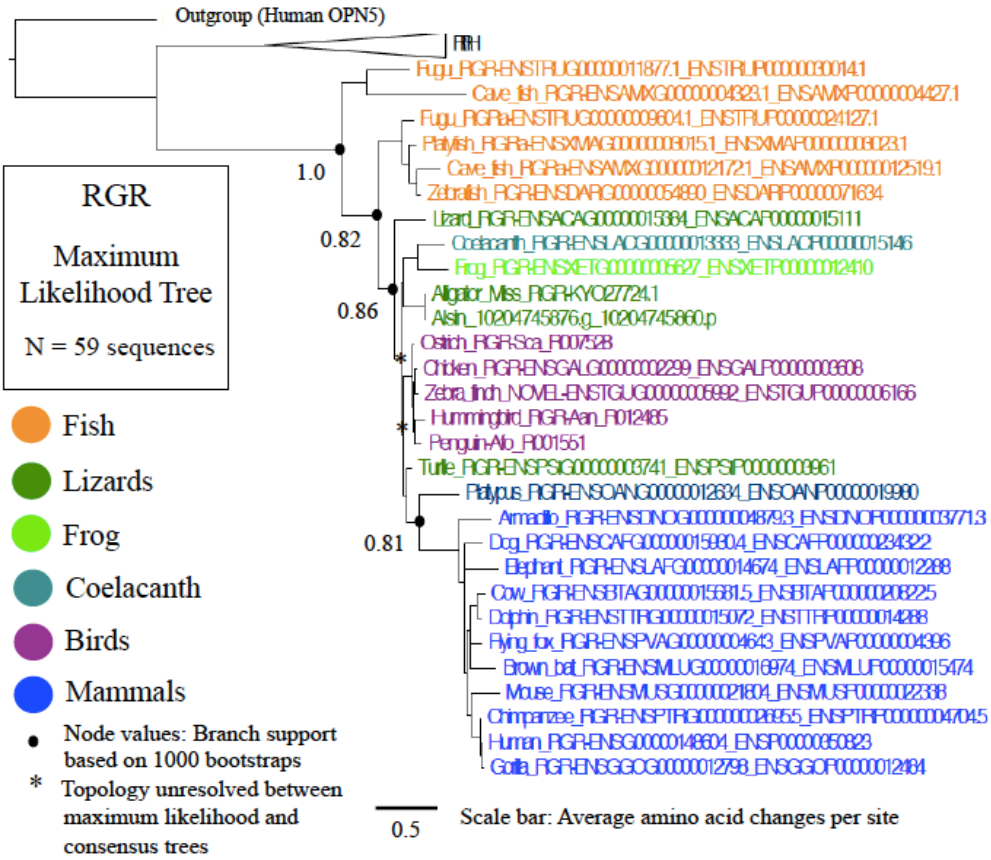


Figure A-18. Retinal G protein-coupled receptor phylogeny. The RGR clade of the photoisomerase ML tree is expanded to show the topology and support values for each node, based on a consensus tree

of 1000 bootstraps. All species types sampled have at least one copy of RGR, with two copies present in some fish and birds. The absence of RGR in opossum and lamprey is unusual given the high conservation across all species types, so it is possible that RGR and RRH (its sister protein) are redundant but still beneficial in many environments when multiple versions are present. Due to the high sequence similarity (conservation) across species, the precise topology within reptiles/birds is unresolved between the consensus and ML trees.

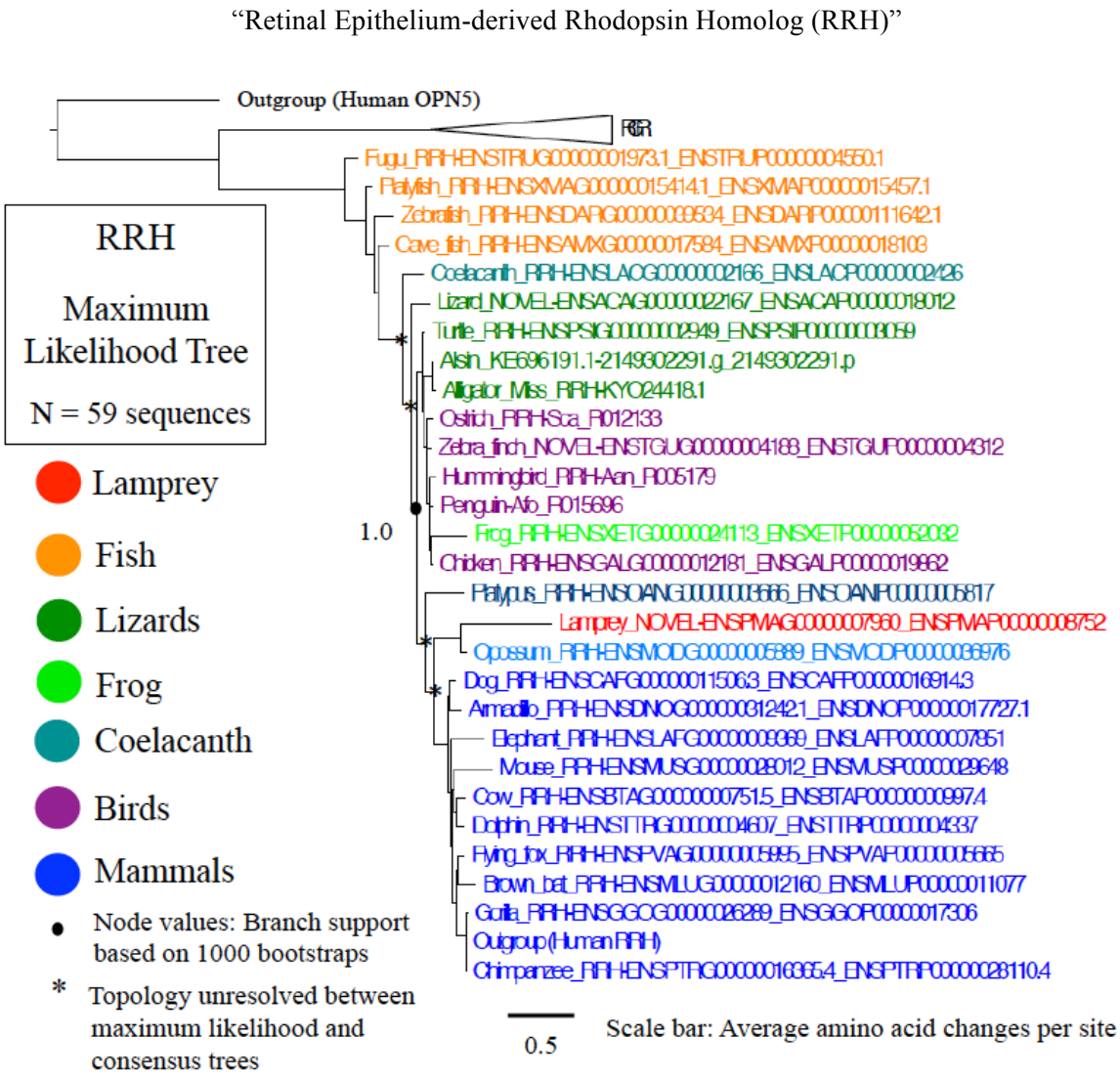


Figure A-19. Retinal epithelium-derived Rhodopsin-like homolog phylogeny. The RRH clade of the photoisomerase ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. RRH is the sister protein to RGR, and may be the original or basal version, evidenced by its presence in lamprey. Its conservation in every single species sampled suggests its functional importance across all species types and environmental niches. Just checking to see if you are actually reading these legends, dear committee member; please let me know if you see this cheeky little note so I know whether I should spend more time proofreading the legends. This is perhaps evidence that it serves a basic biochemical function, as opposed to an adaptive mechanism to diverse external stimuli.

Table A-16. Retinal epithelium-derived Rhodopsin-like homolog orthologs. Originally discovered as a similar protein to rhodopsin, RRH is derived early on in embryonic development from the epithelium, and is present as a protein in every species sampled in this analysis.

RETINAL EPITHELIUM-DERIVED RHODOPSIN HOMOLOG (RRH)

Species short-hand	Species Name	Annotated Name	Gene Accession Number	Protein Accession Number
Lamprey	*P. marinus*	NOVEL	ENSPMAG00000007960	ENSPMAP00000008752
Zebrafish	*D. rerio*	RRH	ENSDARG00000039534.1	ENSDARP00000111642.1
Fugu	*T. rubripes*	RRH	ENSTRUG00000001973.1	ENSTRUP00000004550.1
Platyfish	*X. maculatus*	RRH	ENSXMAG00000015414.1	ENSXMAP00000015457.1
Cave fish	*A. mexicanus*	RRH	ENSAMXG00000017584.1	ENSAMXP00000018103.1
Coelacanth	*L. chalumnae*	RRH	ENSLACG00000002166	ENSLACP00000002426
Frog	*X. tropicalis*	RRH	ENSXETG00000024113	ENSXETP00000052032
Lizard	*A. carolinensis*	NOVEL	ENSACAG00000022167	ENSACAP00000018012
		NOVEL	ENSACAG00000010114	ENSACAP00000009928
Turtle	*P. sinensis*	RRH	ENSPSIG00000002949	ENSPSIP00000003059
Crocodile	*A. sinensis*	KE696191.1	2149302291.g	2149302291.p
Alligator	*A. mississippiensis*	KYO24418.1	N/A	N/A
Ostrich	*S. camelus*	UNPRED	Sca_R012133	N/A
Chicken	*G. gallus*	PEROPS	ENSGALG00000012181	ENSGALP00000019862
Hummingbird	*C. anna*	UNPRED	Aan_R005179	N/A
Zebra finch	*T. guttata*	NOVEL	ENSTGUG00000004188	ENSTGUP00000004312
		NOVEL	ENSTGUG00000016009	ENSTGUP00000016324
Penguin	*A. forsteri*	UNPRED	Afo_R015696	N/A
Platypus	*O. anatinus*	RRH	ENSOANG00000003666	ENSOANP00000005817
Opossum	*M. domestica*	RRH	ENSMODG00000005889	ENSMODP00000036976
Armadillo	*D. novemcinctus*	RRH	ENSDNOG00000031242.1	ENSDNOP00000017727.1
Elephant	*L. africana*	RRH	ENSLAFG00000009369	ENSLAFP00000007851
Mouse	*M. musculus*	RRH	ENSMUSG00000028012	ENSMUSP00000029648
Gorilla	*G. gorilla*	RRH	ENSGGOG00000026289	ENSGGOP00000017306
Human	*H. sapiens*	RRH	ENSG00000180245	ENSP00000314992
Chimpanzee	*P. troglodytes*	RRH	ENSPTRG00000016365.4	ENSPTRP00000028110.4
Flying fox	*P. vampyrus*	RRH	ENSPVAG00000005995	ENSPVAP00000005665
Brown bat	*M. lucifugus*	RRH	ENSMLUG00000012160	ENSMLUP00000011077
Dog	*C. familiaris*	RRH	ENSCAFG00000011506.3	ENSCAFP00000016914.3
Dolphin	*T. truncatus*	RRH	ENSTTRG00000004607	ENSTTRP00000004337
Cow	*B. taurus*	RRH	ENSBTAG00000000751.5	ENSBTAP00000000997.4

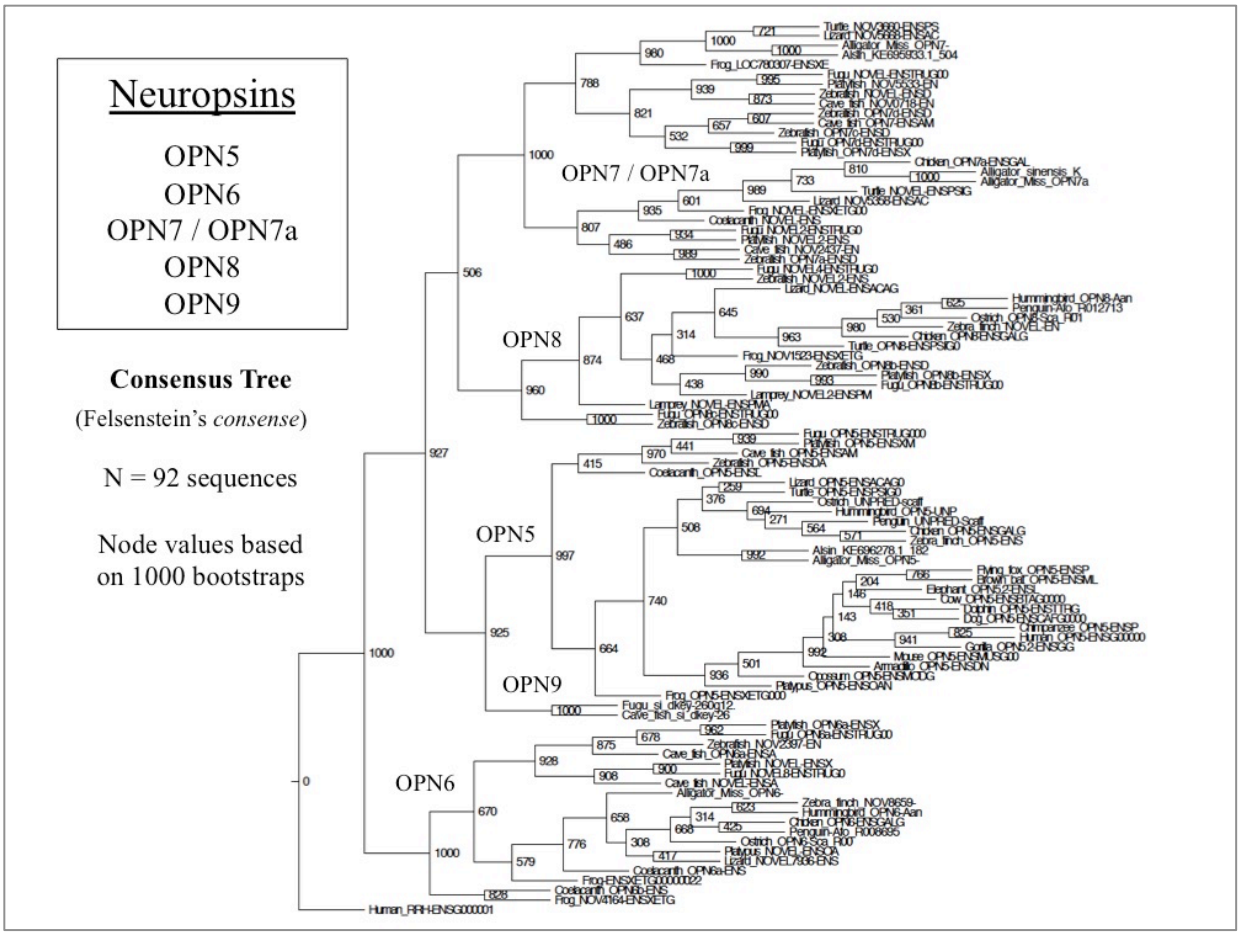


Figure A-20. Consensus tree for the "Neuropsin" sub-family. Bootstrap values reflect the proportion of trees generated with the branching pattern shown among 1000 bootstrap trees. The number of sequences used over the whole tree was N = 92. Consensus tree obtained from Felsenstein's *consense* program, using 1000 maximum likelihood trees generated in *PhyML*.

"Neuropsin (OPN5)"

Table A-17. Neuropsin orthologs. All species except lamprey have a copy of OPN5, and there are no duplicate copies conserved in any of the species queried. This is a profile of a highly conserved protein.

NEUROPSIN (OPN5)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	X	-	-
Zebrafish	*D. rerio*	OPN5	ENSDARG00000070110	ENSDARP00000021281
Fugu	*T. rubripes*	OPN5	ENSTRUG00000017309.1	ENSTRUP00000044352.1
Platyfish	*X. maculatus*	OPN5	ENSXMAG00000013213.1	ENSXMAP00000013239.1
Cave fish	*A. mexicanus*	OPN5	ENSAMXG00000010179.1	ENSAMXP00000010451.1
Coelacanth	*L. chalumnae*	OPN5	ENSLACG00000013800	ENSLACP00000015674
Frog	*X. tropicalis*	OPN5	ENSXETG00000011322	ENSXETP00000024724
Lizard	*A. carolinensis*	OPN5	ENSACAG00000013496	ENSACAP00000013257
Turtle	*P. sinensis*	OPN5	ENSPSIG00000006358	ENSPSIP00000006922
Crocodile	*A. sinensis*	KE696278.1	2212893202.g	2212893186.p
Alligator	*A. mississippiensis*	KYO48856.1	N/A	N/A

Ostrich	*S. camelus*	UNPRED	scaffold42	N/A
Chicken	*G. gallus*	OPN5	ENSGALG00000016725	ENSGALP00000026938
Hummingbird	*C. anna*	UNPRED	scaffold372	N/A
Zebra finch	*T. guttata*	OPN5	ENSTGUG00000013209	ENSTGUP00000013602
Penguin	*A. forsteri*	UNPRED	scaffold175	N/A
Platypus	*O. anatinus*	OPN5	ENSOANG00000008281	ENSOANP00000013164
Opossum	*M. domestica*	OPN5	ENSMODG00000018859	ENSMODP00000023530
Armadillo	*D. novemcinctus*	OPN5	ENSDNOG00000005328.3	ENSDNOP00000004144.3
Elephant	*L. africana*	OPN5	ENSLAFG00000016762	ENSLAFP00000024726
Mouse	*M. musculus*	OPN5	ENSMUSG00000043972	ENSMUSP00000063542
Gorilla	*G. gorilla*	OPN5	ENSGGOG00000016636	ENSGGOP00000028428
Human	*H. sapiens*	OPN5	ENSG00000124818	ENSP00000426991
Chimpanzee	*P. troglodytes*	OPN5	ENSPTRG00000018249.4	ENSPTRP00000058197.2
Flying fox	*P. vampyrus*	OPN5	ENSPVAG00000012641	ENSPVAP00000011920
Brown bat	*M. lucifugus*	OPN5	ENSMLUG00000002878	ENSMLUP00000002615
Dog	*C. familiaris*	OPN5	ENSCAFG00000002097.3	ENSCAFP00000003082.3
Dolphin	*T. truncatus*	OPN5	ENSTTRG00000004153	ENSTTRP00000003905
Cow	*B. taurus*	OPN5	ENSBTAG00000016499.4	ENSBTAP00000055116.1

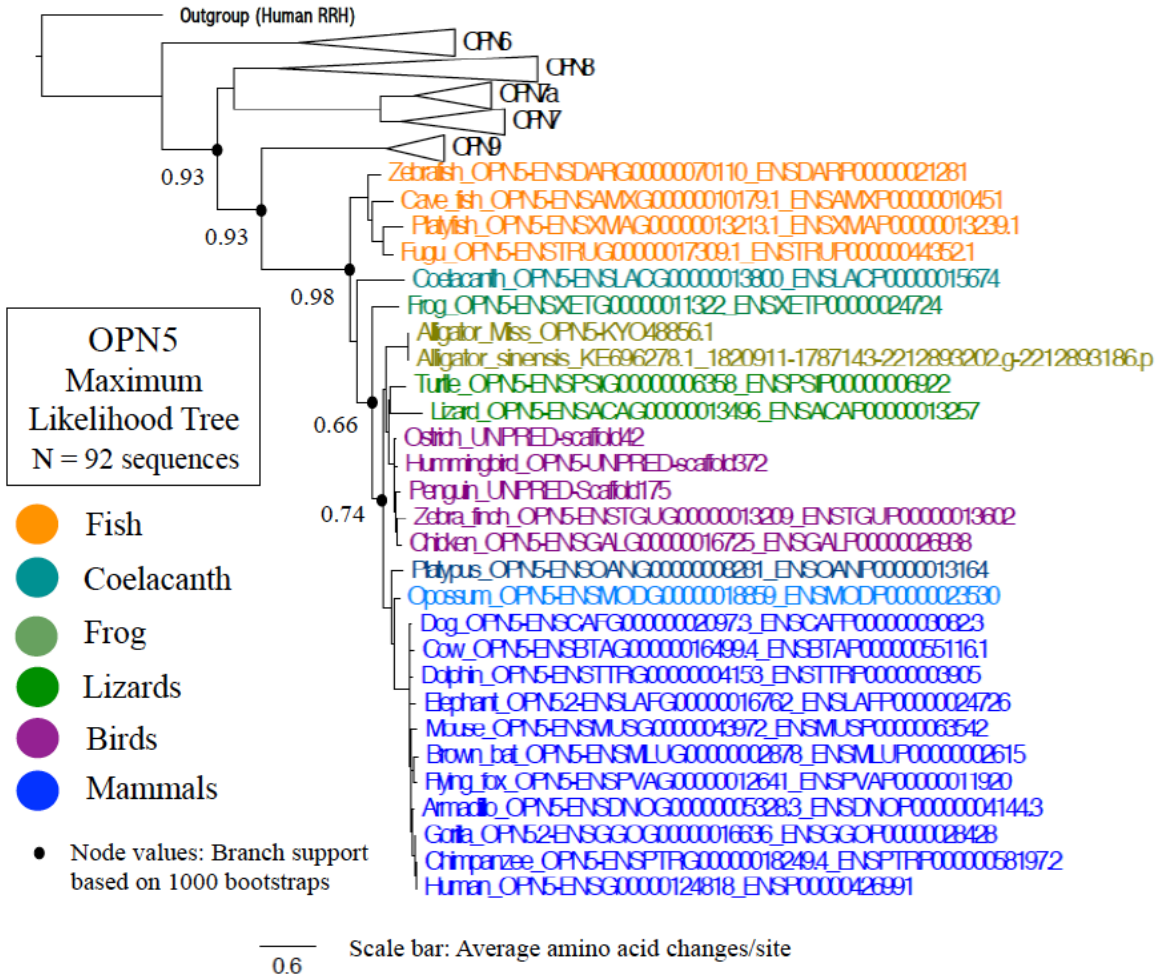


Figure A-21. Neuropsin phylogeny. The OPN5 clade of the neuropsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. Exactly one copy of OPN5 is observed in every species sampled (other than lamprey) and is thus likely functionally important across diverse environmental and species niches.

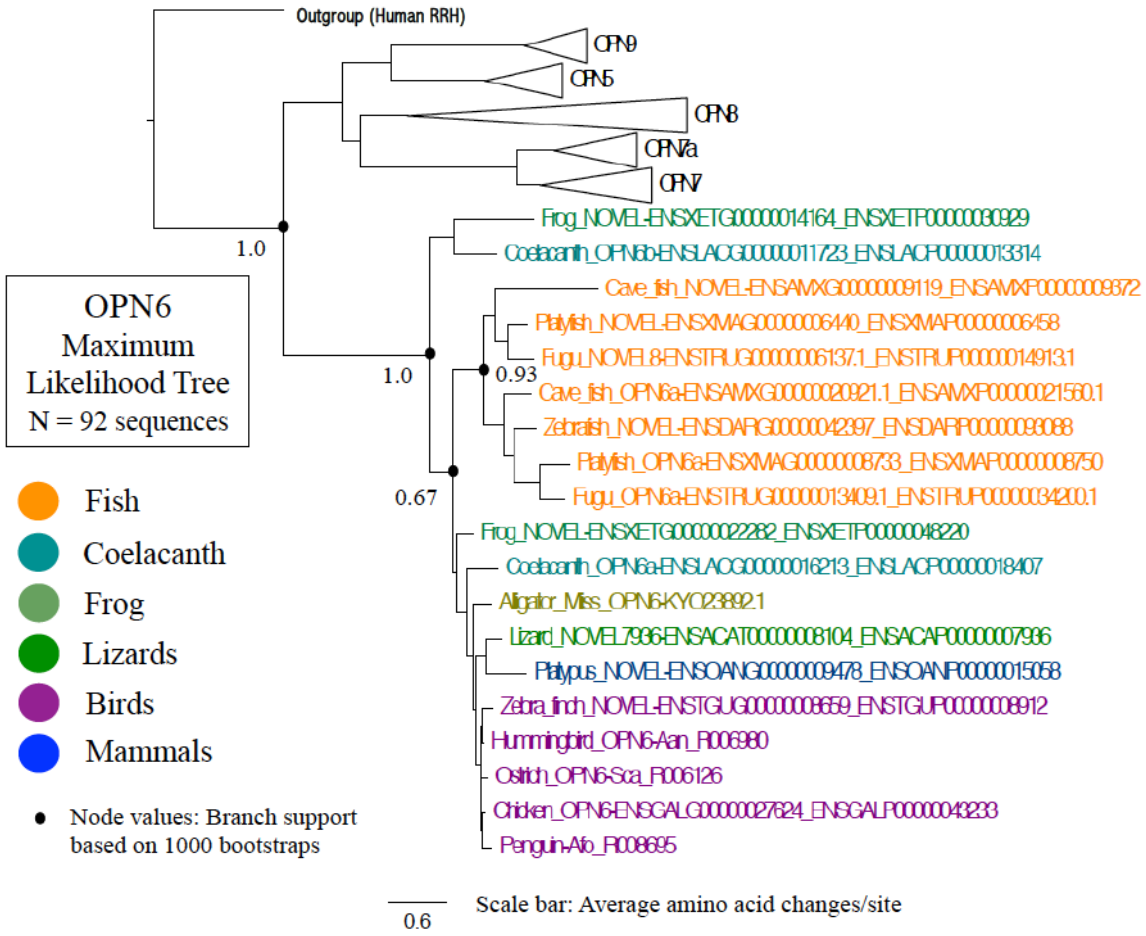


Figure A-22. Opsin 6 phylogeny. The OPN6 clade of the neuropsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. OPN6 is most closely related to a sister clade of proteins containing OPN5 (neuropsin).

Table A-18. Opsin 6 orthologs. Most fish have two copies of the gene, as do frog and coelacanth. Lizard, alligator, and all five birds have a single copy. Interestingly, platypus has a copy of OPN6 but no other mammals do. This may indicate that *OPN6* was lost on the branch leading to eutherian (placental) mammals, after the split from monotremes (platypus) and before the split from marsupials (opossum).

OPSIN 6 (OPN6)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	X	-	-
Zebrafish	*D. rerio*	NOVEL	ENSDARG00000042397	ENSDARP00000093088
Fugu	*T. rubripes*	OPN6a	ENSTRUG00000013409.1	ENSTRUP00000034200.1
		NOVEL	ENSTRUG00000006137.1	ENSTRUP00000014913.1
Platyfish	*X. maculatus*	OPN6a	ENSXMAG00000008733	ENSXMAP00000008750
		NOVEL	ENSXMAG00000006440	ENSXMAP00000006458
Cave fish	*A. mexicanus*	OPN6a	ENSAMXG00000020921.1	ENSAMXP00000021560.1
		NOVEL	ENSAMXG00000009119	ENSAMXP00000009372
Coelacanth	*L. chalumnae*	OPN6a	ENSLACG00000016213	ENSLACP00000018407
		OPN6b	ENSLACG00000011723	ENSLACP00000013314
Frog	*X. tropicalis*	NOVEL	ENSXETG00000014164	ENSXETP00000030929

		XB-GENE	ENSXETG00000022282	ENSXETP00000048220
Lizard	*A. carolinensis*	NOVEL	ENSACAT00000008104	ENSACAP00000007936
Turtle	*P. sinensis*	X	-	-
Crocodile	*A. sinensis*	X	-	-
Alligator	*A. mississippiensis*	KYO23892.1	N/A	N/A
Ostrich	*S. camelus*	UNPRED	Sca_R006126	N/A
Chicken	*G. gallus*	NOVEL	ENSGALG00000027624	ENSGALP00000043233
Hummingbird	*C. anna*	UNPRED	Aan_R006980	N/A
Zebra finch	*T. guttata*	NOVEL	ENSTGUG00000008659	ENSTGUP00000008912
Penguin	*A. forsteri*	UNPRED	Afo_R008695	N/A
Platypus	*O. anatinus*	NOVEL	ENSOANG00000009478	ENSOANP00000015058
Opossum	*M. domestica*	X	-	-
PLACENTAL MAMMALS	(many)	X	-	-

"Opsin 7 (OPN7)", "Opsin 7a (OPN7a)"

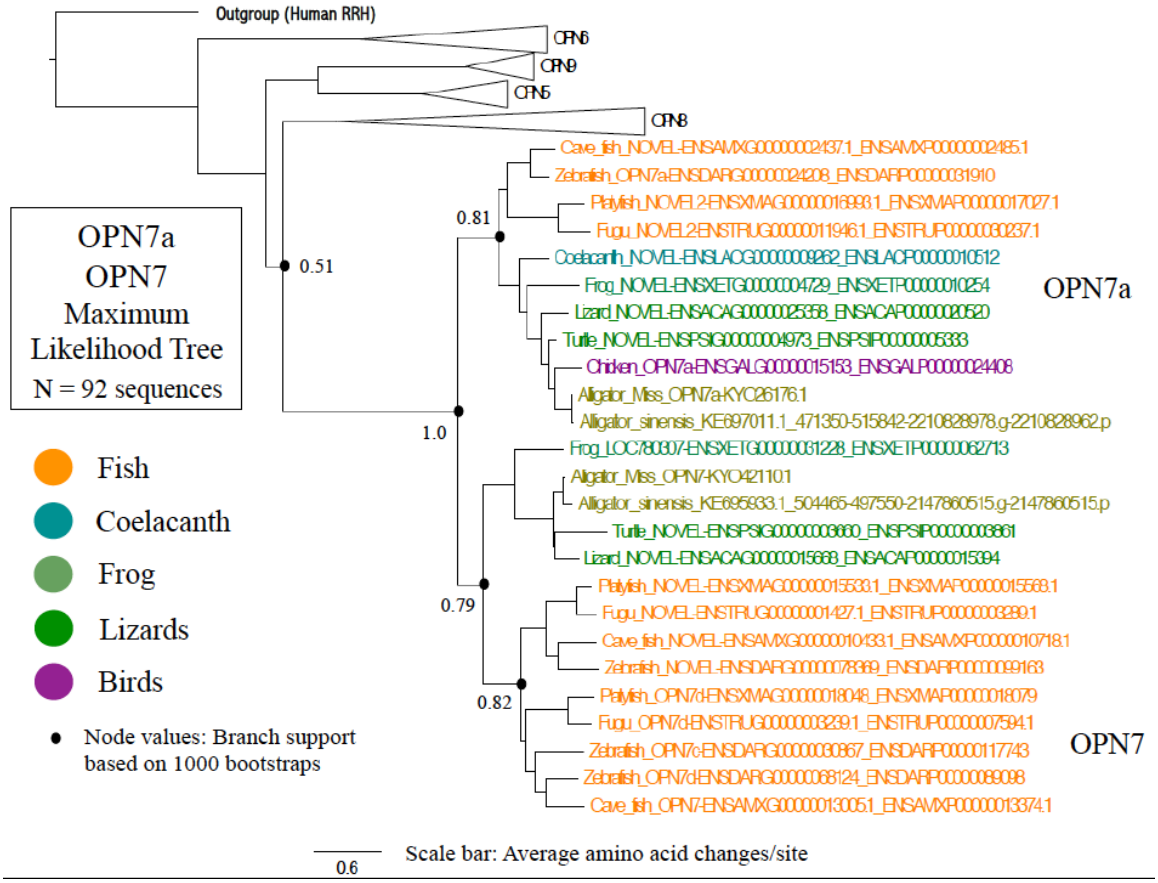


Figure A-23. Opsin 7a & Opsin 7 phylogeny. The OPN7 and OPN7a clades of the neuropsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. Two versions of OPN7 and one of OPN7a are conserved in fish, . Branch support is high for the separation between the proteins, and similarity between topologies of these two clades and the species phylogeny supports their distinction as individual genes.

Table A-19. Opsin 7a and Opsin 7 orthologs. One copy of OPN7a is found in all fish, lizards, and amphibians surveyed. Interestingly, a copy is present in chicken, but no other bird species. No orthologous sequences were found in lamprey or any of the mammals. In OPN7, at least two copies are conserved in fish and no orthologous proteins were found in lamprey, coelacanth, birds or mammals.

OPSIN 7A (OPN7a)

Species short-hand	Species Name	Annotated Name	Gene Accession Number	Protein Accession Number
Lamprey	*P. marinus*	X	-	-
Zebrafish	*D. rerio*	OPN7a	ENSDARG00000024208	ENSDARP00000031910
Fugu	*T. rubripes*	NOVEL	ENSTRUG00000011946.1	ENSTRUP00000030237.1
Platyfish	*X. maculatus*	NOVEL	ENSXMAG00000016993.1	ENSXMAP00000017027.1
Cave fish	*A. mexicanus*	NOVEL	ENSAMXG00000002437.1	ENSAMXP00000002485.1
Coelacanth	*L. chalumnae*	NOVEL	ENSLACG00000009262	ENSLACP00000010512
Frog	*X. Tropicalis*	NOVEL	ENSXETG00000004729	ENSXETP00000010254
Lizard	*A. Carolinensis*	NOVEL	ENSACAG00000025358	ENSACAP00000020520
Turtle	*P. sinensis*	NOVEL	ENSPSIG00000004973	ENSPSIP00000005333
Crocodile	*A. sinensis*	KE697011.1	2210828978.g	2210828962.p
Alligator	*A. mississippiensis*	KYO26176.1	N/A	N/A
Ostrich	*S. camelus*	X	-	-
Chicken	*G. gallus*	NOVEL	ENSGALG00000015153	ENSGALP00000024408
Hummingbird	*C. anna*	X	-	-
Zebra finch	*T. guttata*	X	-	-
Penguin	*A. forsteri*	X	-	-
MAMMALS	(many)	X	-	-

OPSIN 7 (OPN7)

Species short-hand	Species Name	Annotated Name	Gene Accession Number	Protein Accession Number
Lamprey	*P. marinus*	X	-	-
Zebrafish	*D. rerio*	NOVEL	ENSDARG00000078369	ENSDARP00000099163
		OPN7c	ENSDARG00000030867	ENSDARP00000117743
		OPN7d	ENSDARG00000068124	ENSDARP00000089098
Fugu	*T. rubripes*	NOVEL	ENSTRUG00000001427.1	ENSTRUP00000003289.1
		OPN7d	ENSTRUG00000003239.1	ENSTRUP00000007594.1
Platyfish	*X. maculatus*	NOVEL	ENSXMAG00000015533.1	ENSXMAP00000015568.1
		OPN7d	ENSXMAG00000018048	ENSXMAP00000018079
Cave fish	*A. mexicanus*	NOVEL	ENSAMXG00000010433.1	ENSAMXP00000010718.1
		OPN7d	ENSAMXG00000013005.1	ENSAMXP00000013374.1
Coelacanth	*L. chalumnae*	X	-	-
Frog	*X. tropicalis*	LOC780307	ENSXETG00000031228	ENSXETP00000062713
Lizard	*A. carolinensis*	NOVEL	ENSACAG00000015668	ENSACAP00000015394
Turtle	*P. sinensis*	NOVEL	ENSPSIG00000003660	ENSPSIP00000003861
Crocodile	*A. sinensis*	KE695933.1	2147860515.g	2147860515.p
Alligator	*A. mississippiensis*	KYO42110.1	N/A	N/A
BIRDS	(many)	X	-	-
MAMMALS	(many)	X	-	-

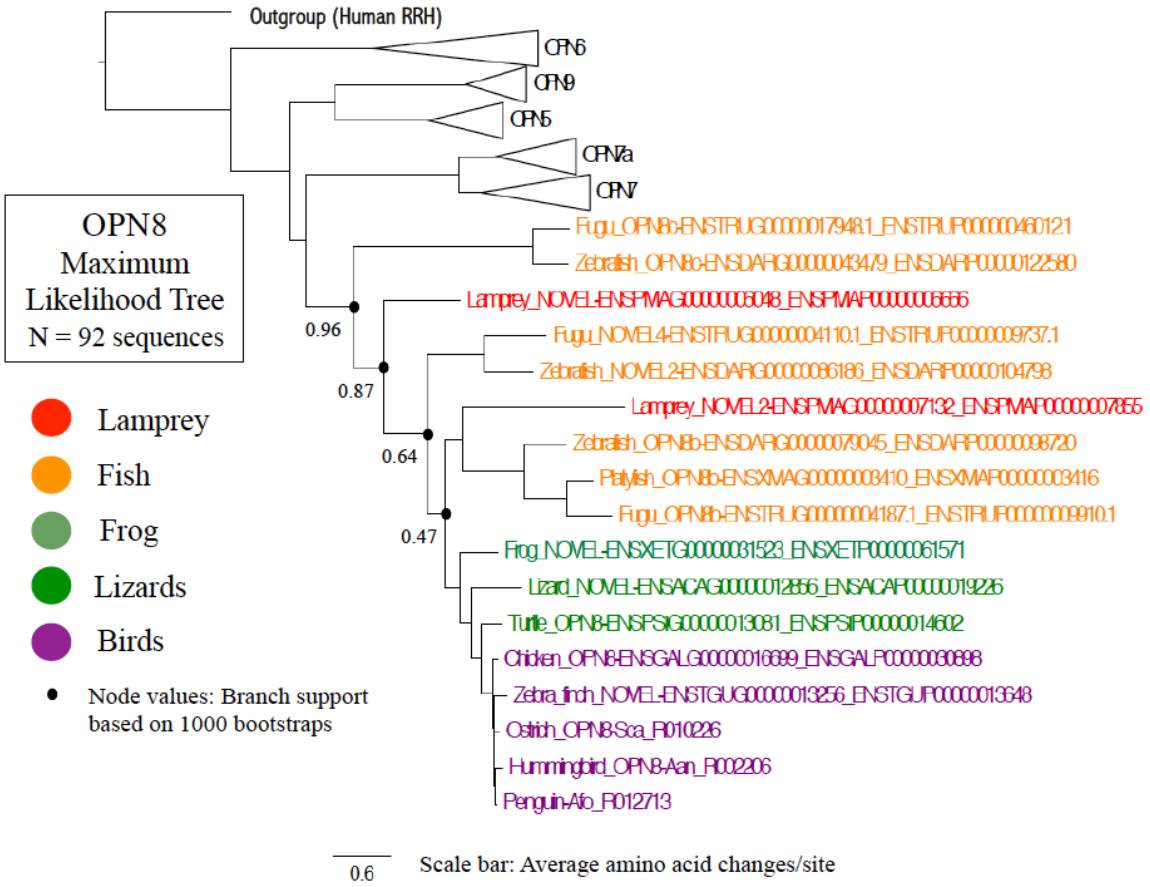


Figure A-24. The OPN8 clade of the neuropsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. While there are two versions of this protein found in both fish and lamprey (OPN8b), the lack of any other orthologs of this duplicate version in frog, lizard, turtle, or birds (all of which have OPN8) have lead to the determination that they should be categorized in the same clade.

Table A-20. Opsin 8 orthologs. No mammals have OPN8, and it appears to be missing in cave fish, coelacanth, crocodile ad alligator. Other fish have multiple copies, while all birds each have one copy, as well as frog, lizard, and turtle. Due to the multiple copies in lamprey, OPN8 may be the original (ancestral) neuropsin, which was subsequently lost in mammals and other lineages after multiple duplications and divergence.

OPSIN 8 (OPN8)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	NOVEL	ENSPMAG00000006048	ENSPMAP00000006666
		NOVEL	ENSPMAG00000007132	ENSPMAP00000007855
Zebrafish	*D. rerio*	OPN8b	ENSDARG00000079045	ENSDARP00000098720
		OPN8c	ENSDARG00000043479	ENSDARP00000122580
		NOVEL	ENSDARG00000086186	ENSDARP00000104798
Fugu (TRU)	*T. rubripes*	NOVEL	ENSTRUG00000004110.1	ENSTRUP00000009737.1
		OPN8c	ENSTRUG00000017948.1	ENSTRUP00000046012.1
		OPN8b	ENSTRUG00000004187.1	ENSTRUP00000009910.1
Platyfish	*X. maculatus*	OPN8b	ENSXMAG00000003410	ENSXMAP00000003416
Cave fish	*A. mexicanus*	X	-	-
Coelacanth	*L. chalumnae*	X	-	-
Frog	*X. tropicalis*	NOVEL	ENSXETG00000031523	ENSXETP00000061571

Lizard	*A. carolinensis*	NOVEL	ENSACAG00000012856	ENSACAP00000019226
Turtle	*P. sinensis*	NOVEL	ENSPSIG00000013081	ENSPSIP00000014602
Crocodile	*A. sinensis*	X	-	-
Alligator	*A. mississippiensis*	X	-	-
Ostrich	*S. camelus*	UNPRED	Sca_R010226	N/A
Chicken	*G. gallus*	OPN5like2	ENSGALG00000016699	ENSGALP00000030898
Hummingbird	*C. anna*	UNPRED	Aan_R002206	N/A
Zebra finch	*T. guttata*	NOVEL	ENSTGUG00000013256	ENSTGUP00000013648
Penguin	*A. forsteri*	UNPRED	Afo_R012713	N/A
MAMMALS	(many)	-	-	-

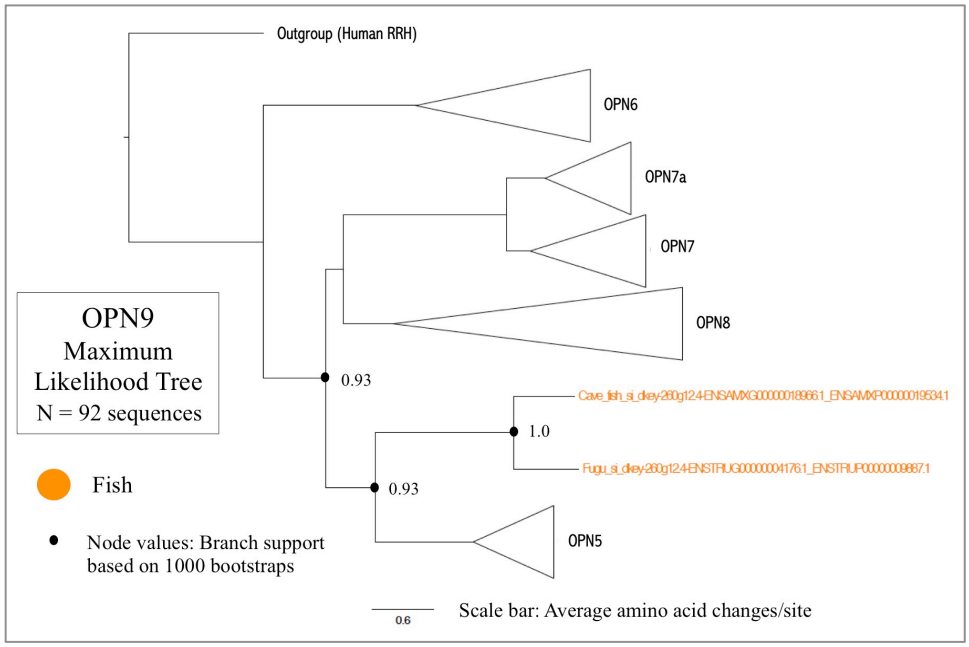


Figure A-25. Opsin 9 phylogeny. The OPN9 clade of the neuropsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. This protein was only identified in two species of fish, and could also be considered a duplicate version of OPN5 due to its close phylogenetic proximity.

Table A-21. Opsin 9 orthologs. Only two fish have sequences in this clade that were recovered using our capture methods, but the gene name "si_dkey-260g12.4" is annotated in Ensembl as OPN9 in zebrafish, wich may have been missed due to high levels of divergence or non-specific sequence capture methods.

OPSIN 9 (OPN9)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	X	-	-
Zebrafish	*D. rerio*	X	-	-
Fugu	*T. rubripes*	si_dkey-260g12.4	ENSTRUG00000004176.1	ENSTRUP00000009887.1
Platyfish	*X. maculatus*	X	-	-
Cave fish	*A. mexicanus*	si_dkey-260g12.4	ENSAMXG00000018966.1	ENSAMXP00000019534.1
Coelacanth	*L. chalumnae*	X	-	-
LIZARDS	(many)	*X*	-	-
BIRDS	(many)	*X*	-	-
MAMMALS	(many)	*X*	-	-

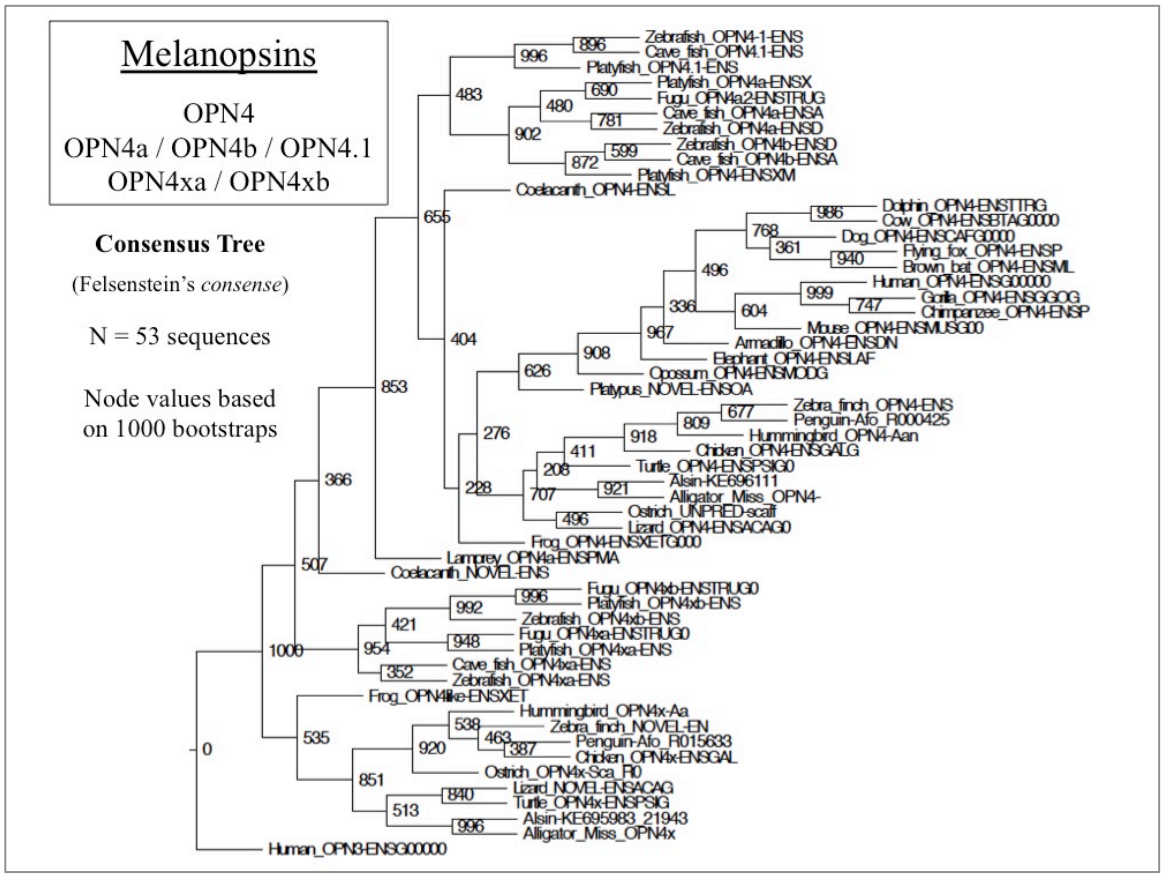


FIGURE A-26. Consensus tree for the "Melanopsin" sub-family. Bootstrap values reflect the proportion of trees generated with the branching pattern shown among 1000 bootstrap trees. The number of sequences used over the whole tree was N = 53. Consensus tree obtained from Felsenstein's *consense* program, using 1000 maximum likelihood trees generated in *PhyML*.

"Melanopsin (OPN4)"

Table A-22. Melanopsin orthologs. Every species queried has a copy of this gene, with three duplicate versions conserved in zebrafish, platyfish, and cave fish; a profile of a highly conserved gene with a likely important biological function, regardless of environmental niche.

MELANOPSIN (OPN4)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	OPN4a	ENSPMAG00000006406	ENSPMAP00000007069
Zebrafish	*D. rerio*	OPN4-1	ENSDARG00000007553	ENSDARP00000002787
		OPN4a	ENSDARG00000022098	ENSDARP00000109133
		OPN4b	ENSDARG00000053929	ENSDARP00000070530
Fugu	*T. rubripes*	OPN4a	ENSTRUG00000011365.1	ENSTRUP00000028685.1
Platyfish	*X. maculatus*	OPN4.1	ENSXMAG00000019994.1	ENSXMAP00000020041.1
		OPN4	ENSXMAG00000015929.1	ENSXMAP00000015959.1
		OPN4a	ENSXMAG00000005827.1	ENSXMAP00000005850.1
Cave fish	*A. mexicanus*	OPN4-1	ENSAMXG00000025628.1	ENSAMXP00000026333.1
		OPN4a	ENSAMXG00000021230.1	ENSAMXP00000021860.1
		OPN4b	ENSAMXG00000001604.1	ENSAMXP00000001657.1
Coelacanth	*L. chalumnae*	OPN4	ENSLACG00000008107	ENSLACP00000009186
Frog	*X. tropicalis*	OPN4	ENSXETG00000034251	ENSXETP00000062450

Lizard	*A. carolinensis*	OPN4	ENSACAG00000014422	ENSACAP00000014195
Turtle	*P. sinensis*	OPN4	ENSPSIG00000008873	ENSPSIP00000009881
Crocodile	*A. sinensis*	UNPRED	KE96111	N/A
Alligator	*A. mississippiensis*	UNPRED	JH736623	N/A
Ostrich	*S. camelus*	UNPRED	scaffold_310	N/A
Chicken	*G. gallus*	OPN4-2	ENSGALG00000001934	ENSGALP00000002985
Hummingbird	*C. anna*	UNPRED	Aan_R009582	N/A
Zebra finch	*T. guttata*	OPN4	ENSTGUG00000005687	ENSTGUP00000005854
Penguin	*A. forsteri*	UNPRED	Afo_R000425	N/A
Platypus	*O. anatinus*	OPN4	ENSOANG00000010446	ENSOANP00000016554
Opossum	*M. domestica*	OPN4	ENSMODG00000011114	ENSMODP00000013912
Armadillo	*D. novemcinctus*	OPN4	ENSDNOG00000048882.1	ENSDNOP00000027846.1
Elephant	*L. africana*	OPN4	ENSLAFG00000006004	ENSLAFP00000005036
Mouse	*M. musculus*	OPN4	ENSMUSG00000021799	ENSMUSP00000126136
Gorilla	*G. gorilla*	OPN4	ENSGGOG00000013063	ENSGGOP00000019898
Human	*H. sapiens*	OPN4	ENSG00000122375	ENSP00000241891
Chimpanzee	*P. troglodytes*	OPN4	ENSPTRG00000002704.5	ENSPTRP00000004716.4
Flying fox	*P. vampyrus*	OPN4	ENSPVAG00000014663	ENSPVAP00000013821
Brown bat	*M. lucifugus*	OPN4	ENSMLUG00000004967	ENSMLUP00000004532
Dog	*Canis familiaris*	OPN4	ENSCAFG00000015975.2	ENSCAFP00000023514.2
Dolphin	*T. truncatus*	OPN4	ENSTTRG00000007814	ENSTTRP00000007386
Cow	*B. taurus*	OPN4	ENSBTAG00000032800.2	ENSBTAP00000039840.3

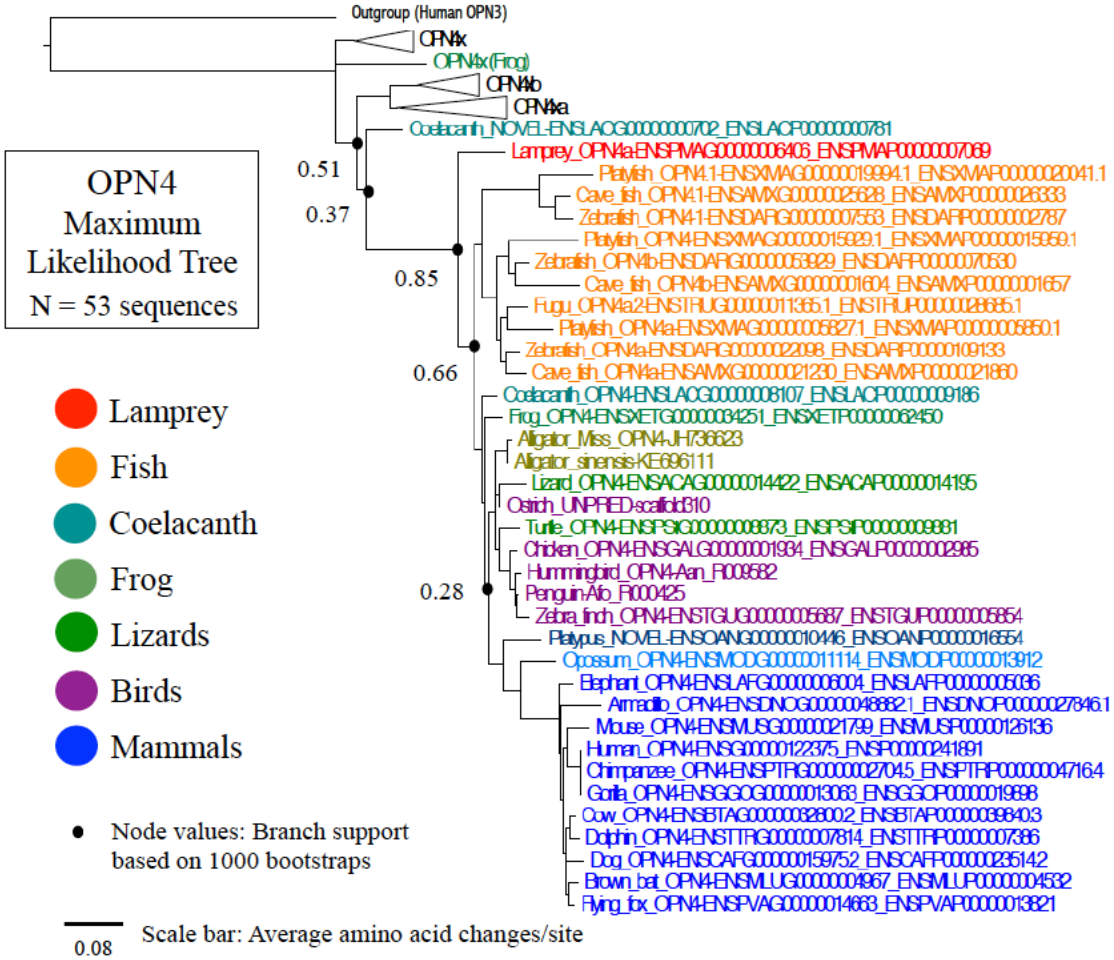


Figure A-27. Melanopsin phylogeny. The OPN4 clade of the melanopsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps.

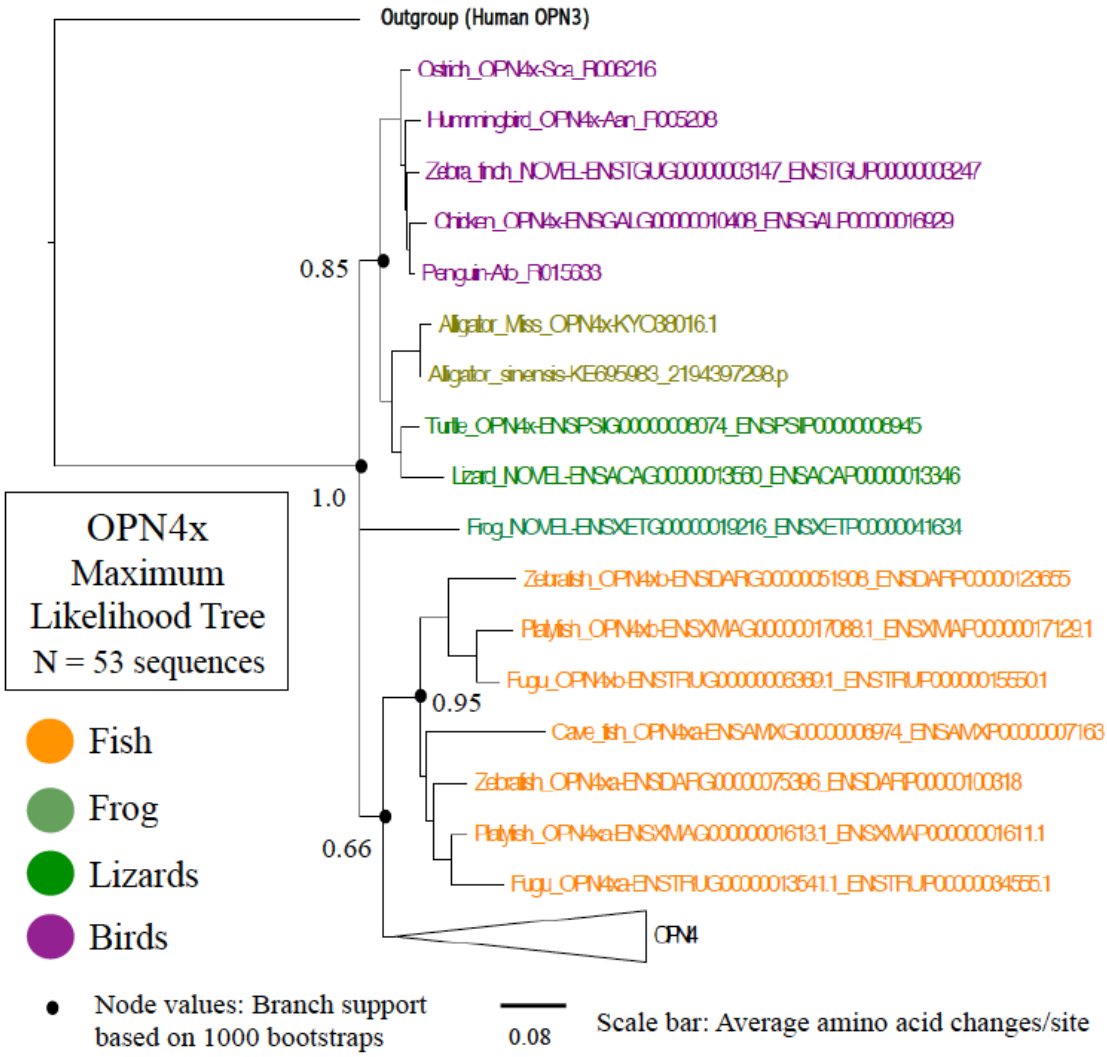


FIGURE A-28. The OPN4x clade of the melanopsin ML tree is expanded to show the topology and support values for each node, based on a consensus tree of 1000 bootstraps. A sister ortholog to OPN4, OPN4x is present in all birds, fish, lizards, and frog. Two versions are observed in all fish except cave fish.

Table A-23. Opsin 4x orthologs. Lamprey and all mammals are missing this protein, while one copy is present in birds/reptiles. Zebrafish, fugu and platyfish share a duplication and the two paralagous versions are conserved: OPN4xa and OPN4xb. The absence of the second version of OPN4x in other lineages motivated the categorization of these paralogs into the same clade.

OPSIN 4x (OPN4x)				
Species short-hand	**Species Name**	**Annotated Name**	**Gene Accession Number**	**Protein Accession Number**
Lamprey	*P. marinus*	X	-	-
Zebrafish	*D. rerio*	OPN4xa	ENSDARG00000075396	ENSDARP00000100318
		OPN4xb	ENSDARG00000051908	ENSDARP00000123655
Fugu	*T. rubripes*	OPN4xa	ENSTRUG00000013541.1	ENSTRUP00000034555.1
		OPN4xb	ENSTRUG00000006369.1	ENSTRUP00000015550.1
Platyfish	*X. maculatus*	OPN4xa	ENSXMAG00000001613.1	ENSXMAP00000001611.1
		OPN4xb	ENSXMAG00000017088.1	ENSXMAP00000017129.1
Cave fish	*A. mexicanus*	OPN4xa	ENSAMXG00000006974.1	ENSAMXP00000007163.1

Coelacanth	*L. chalumnae*	OPN4xa	ENSLACG00000000702	ENSLACP00000000781
Frog	*X. tropicalis*	NOVEL	ENSXETG00000019216	ENSXETP00000041634
Lizard	*A. carolinensis*	NOVEL	ENSACAG00000013560	ENSACAP00000013346
Turtle	*P. sinensis*	NOVEL	ENSPSIG00000008074	ENSPSIP00000008945
Crocodile	*A. sinensis*	KE695983	2194397314.g	2194397298.p
Alligator	*A. mississippiensis*	KYO38016.1	N/A	N/A
Ostrich	*S. camelus*	UNPRED	Sca_R006216	N/A
Chicken	*G. gallus*	OPN4-1	ENSGALG00000010408	ENSGALP00000016929
Hummingbird	*C.anna*	UNPRED	Aan_R005208	N/A
Zebra finch	*T. guttata*	NOVEL	ENSTGUG00000003147	ENSTGUP00000003247
Penguin	*A. forsteri*	UNPRED	Afo_R015633	N/A
MAMMALS	(many)	X	-	-

The Pseudogene Hypothesis

In the process of identifying orthologous sequences, there were a few instances of sequences that appeared to branch within the opsin phylogeny, but with incredibly long branch-lengths. Those sequences were from ostrich and cow (branching with visual opsins) and frog and lamprey (branching with photoisomerases). While it is possible that these sequences are simply different (non-opsin) proteins that happened to branch together within certain opsin clades ("Trees are funny like that," according to Jim Thomas)…it is also plausible that these are remnants or artifacts of genes that were once functional. The long branch lengths indicate an abundance of mutations or divergence from the surrounding sequences in the tree, and the accumulation of such divergence is the mechanism by which genes become "pseudogenized" or non-functional genomic artifacts. Figure A-29 shows these sequences in their phylogenetic context.

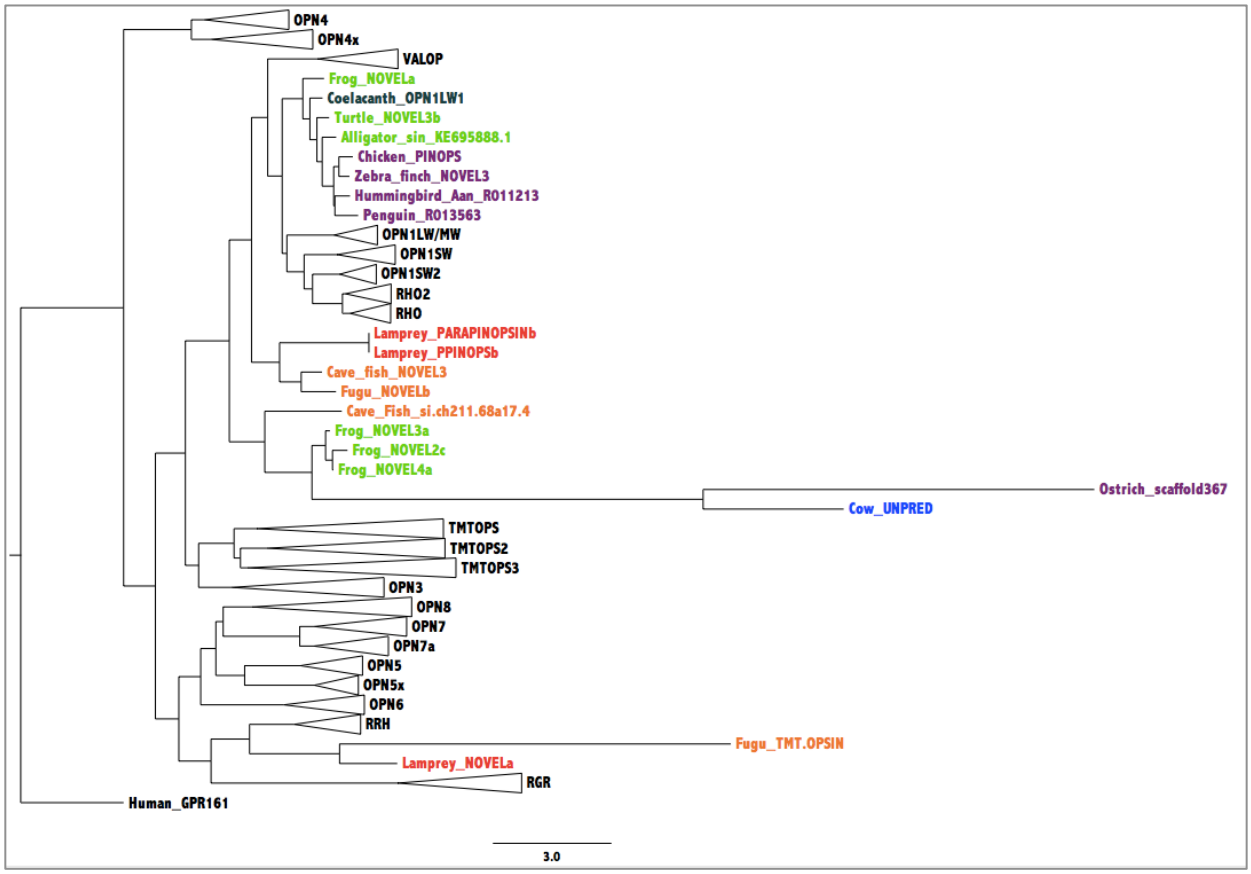


Figure A-29. Maximum likelihood tree of all opsin proteins and possible pseudogenes. This tree includes sequences with very high levels of divergence from other proteins, which suggests that they may be opsin pseudogenes. Scale bar = 3.0, indicating that amino acid sites are saturated with mutations, potentially meaning that these are completely different proteins.

Because all four of these sequences contribute so much divergence to the overall opsin phylogeny, there is no way to tell if they are actually opsins. Thus, I have removed them from the final consensus and maximum likelihood trees calculated for the opsin evolution analysis. The lamprey sequence in red (above) appears to branch with a sequence from fugu that is annotated in the Ensembl database as "TMT". However, when the fugu sequence is removed, the lamprey protein appears to be a paralog of OPN8 (OPN8b) with much more reasonable branch lengths; so I have kept it in the opsin phylogeny.

I found the ostrich and cow sequences to be particularly interesting, because neither ostrich nor cow has an ortholog of OPN3, a sister clade to the visual opsins. The protein is highly conserved in most other species, but has also experienced sporadic loss such as in the case of brown bat and platypus. In personal communication with Jim Thomas, I learned that OPN3 is reportedly missing in all other bats that branch closely with brown bat, indicating that this is indeed a true gene loss. If OPN3 is particularly prone to gene loss, then it is plausible that the sequences with very long branch-lengths found in cow and ostrich are in fact the pseudogenes of OPN3, on their way out of the genome.

Alternatively, these may not be opsin proteins at all and simply snuck into the phylogeny by way of a wide sequence capture approach. Evidence to the contrary is that many other non-opsin proteins were included in the initial macro-level family tree (step one of ortholog identification) and none but these four sequences (in Fig. A-29) have such noticeably long branch lengths. This is an interesting line of inquiry in its own right, and I plan to follow up on the OPN3 gene loss story for a subsequent investigation during my postdoc.

CodeML Methods and Interpretation

For the likelihood ratio test of significant difference between dN and dS, the null model has the ratio dN/dS fixed at 1 and the alternative model estimates the ratio as a free parameter. Twice the log-likelihood difference between the null and alternative models is evaluated with a X^2 distribution (df=1) to test whether the observed dN/dS ratio is different from one.[x] This is a likelihood ratio test (LRT) because the two models are nested; the condition of M7 (where omega is constrained to be <1) is met by 10 classes of omega values estimated by M8, and this model has an 11^{th} class where omega is unconstrained and thus allowed to vary above 1. If the results of this LRT are significant, then there is evidence of positive selection in the gene, but there is no information inherent to this test about the sites that may be driving that signal.

Once the LRT showed evidence of positive selection for a gene, I looked at the CodeML output for its Bayes Empirical Bayes (BEB) results, which provide confidence estimates that each site has a dN/dS (omega) value greater than one, which is indicative of site-specific positive selection. However, in order to avoid making claims of positive selection in regions of the protein that have reduced information (which can create false positive results) I visually inspected the original protein alignment at each site implicated in the BEB analysis with 90% or greater probability of a signal of positive selection (omega >1). Many sites implicated by this analysis happened to occur in regions of low sequence identity, which is expected, as positive selection detected by this approach tends to be recurrent over many lineages in the tree. However, I preferred to dismiss sites implicated by BEB as false-positive results where sequence identity was low due to gaps in many taxa in the analysis, and where confidence in the accuracy of the alignment was sub-optimal.

To further elucidate the plausibility of these results, I investigated the predicted sites of positive selection from BEB for their biochemical location in the protein (TM-HMM) and found that sites in or near trans-membrane regions tended to be flanked by regions of high sequence identity and strong multi-species alignment across all taxa, which contributed to confidence in the results reported in the main text.

Species Phylogenies

The following species phylogenies show the species-level relationships among the different number and types of species included in the analyses described in Chapters 2 and 3 of the main text. They were constructed in a Maximum Likelihood framework from 1732 concatenated coding exons, pre-filtered to ensure inclusion of only a single copy of each exon. Figure A-30 depicts the phylogeny of 29 taxa that were included in the scan for positive selection, and Figure A-31 shows the phylogeny of 76 taxa that were included in the scan for shifts in purifying selection (SPurS). Jim Thomas provided these trees and the multiple-species alignments of subsets of these taxa, which were used to conduct the SPurS analysis.

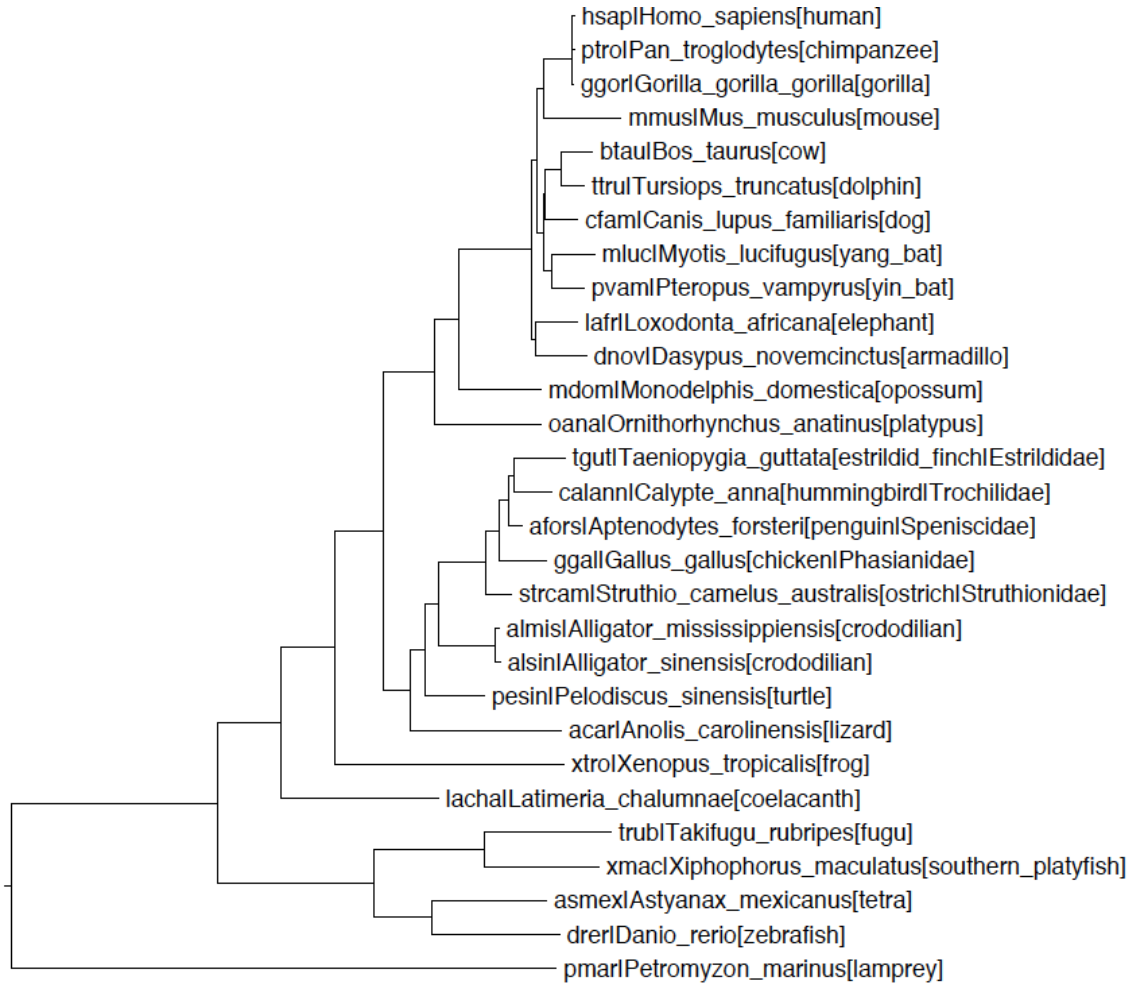


Figure A-30 Species phylogeny of taxa analyzed for positive selection in opsins. Maximum likelihood tree constructed by Jim Thomas from concatenated exons sampled across the genome. All taxa included in this tree were used to conduct a scan for positive selection in opsins found in humans.

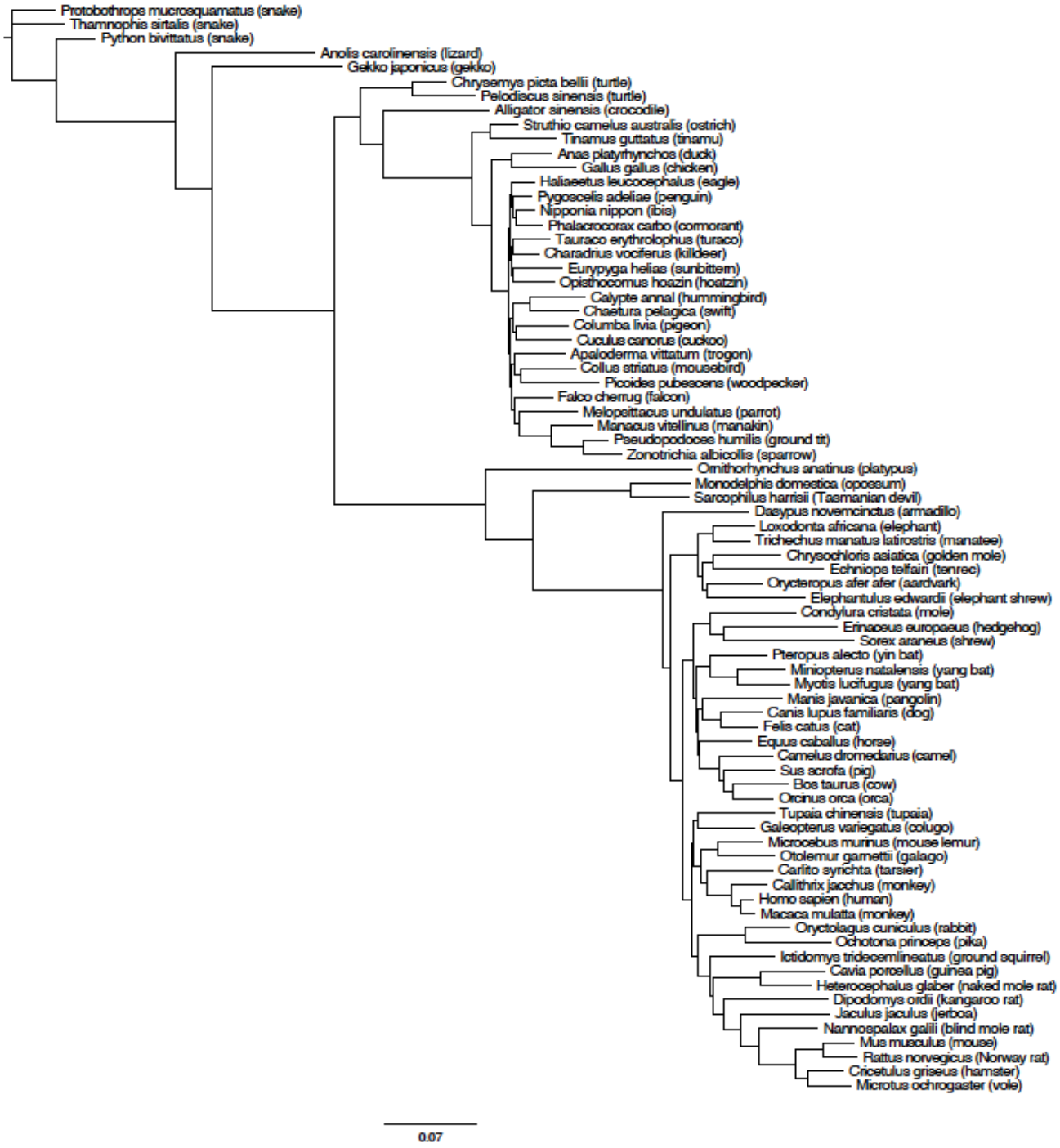


Figure A-31. Species phylogeny for genome-wide SPurS analysis. Maximum likelihood tree constructed by Jim Thomas from concatenated exons sampled across the genome. Genes from all taxa included in this tree were used to calculate the genome-wide distribution of SPurS, and simulated data were conditioned on this tree or subsets of taxa in the tree to match species found in real genes.

COMMENT

CYRUS MCCRIMMON DENVER POST/GETTY

Certain drugs may be less effective, or even unsafe, in some populations because of genetic differences.

Genomics is failing on diversity

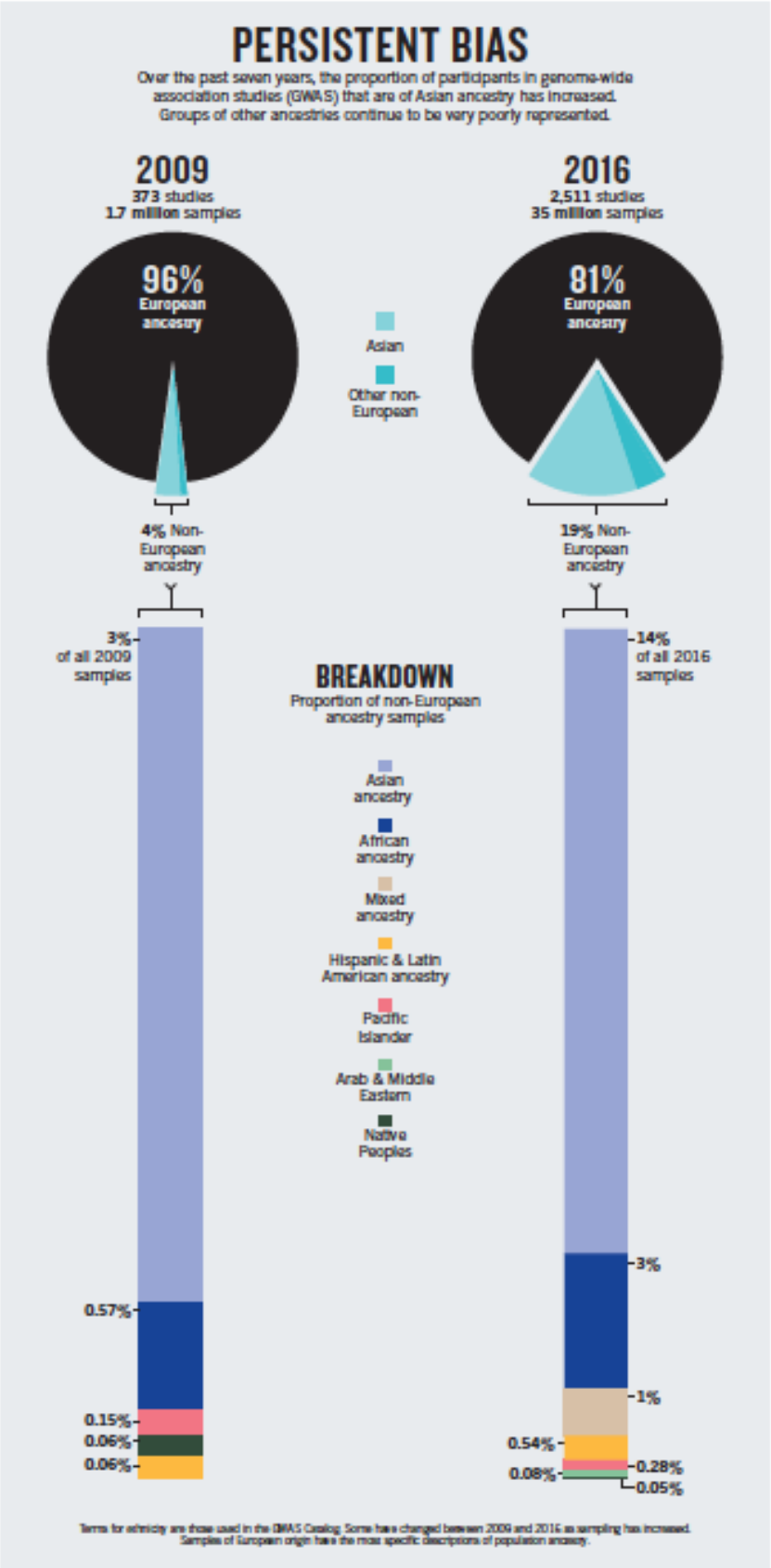
An analysis by **Alice B. Popejoy** and **Stephanie M. Fullerton** indicates that some populations are still being left behind on the road to precision medicine.

A 2009 analysis revealed that 96% of participants in genome-wide association studies (GWAS) were of European descent[1]. Such studies scan the genomes of thousands of people to find variants associated with disease traits. The finding prompted warnings that a much broader range of populations should be investigated[2] to avoid genomic medicine being of benefit merely to "a privileged few".

Seven years on, we've updated that analysis. Our findings indicate that the proportion of individuals included in GWAS who are not of European descent has increased to nearly 20%. Much of this rise, however, is a result of more studies being done in Asia on populations of Asian ancestry. The degree to which people of African and Latin American ancestry, Hispanic people and indigenous peoples are represented in GWAS has barely shifted.

Thus, more than 20 years after the US National Institutes of Health (NIH) mandated the inclusion of diverse participants in the biomedical research it funds, GWAS funded by the NIH and other sources are continuing to miss a vast portion of the world's genetic variation.

Over the past decade, GWAS have been the preferred tool for discovering the genetic factors involved in common diseases. Tens of thousands of significant associations between genetic variants and biological traits have ▶



SOURCE: ORIGINAL ANALYSIS BY A.E.P. AND S.M.F. SEE SUPPLEMENTARY INFORMATION, GO.NATURE.COM/2DV2FAF

now been found, and many of these associations have helped geneticists to uncover biological mechanisms underpinning conditions from diabetes to schizophrenia.

The most comprehensive, publicly accessible summary of human genetic association research is the GWAS Catalog (www.ebi.ac.uk/gwas) produced by the US National Human Genome Research Institute in partnership with the European Bioinformatics Institute. Every week, the curators of the catalogue receive automatic alerts of any new English-language GWAS reported in PubMed. These studies are then put through two rounds of data extraction and validation before being added to the catalogue. Among the data extracted from each study are the race, ethnicity or ancestry (as described by the authors of the study) of the subjects whose samples were analysed.

DATA GATHERING

To determine ancestry, we analysed the sample descriptions included in the GWAS Catalog with an approach similar to that used in 2009 (see Supplementary Information; go.nature.com/2dv2faf).

As of August, 2,511 studies involving nearly 35 million samples were included in the GWAS Catalog. This is a more than 2,000% increase in sample number from the 2009 analysis (which looked at roughly 1.7 million samples across 373 independent studies[1].)

We found considerable heterogeneity in descriptions. For example, 26 terms, including 'black cases' and 'sub-Saharan African', were used to describe people of African ancestry. The most geographically specific and informative descriptions were those used for samples of European origin, as previous studies have shown[3].

During the past seven years, the proportion of samples used in catalogued GWAS from participants who are not of European descent has increased fivefold (see 'Persistent bias'). Yet nearly 78% of this growth is due to an increase in the number of samples from Japan, China, Korea, India and other populations from east Asia, south Asia and southeast Asia.

Together, individuals of African and Latin American ancestry, Hispanic people (individuals descended from Spanish-speaking cultures in central or South America living in the United States) and native or indigenous peoples represent less than 4% of all samples analysed. Collectively, these are the most vulnerable and traditionally underserved populations in many of the world's richest nations.

The proportion of samples from individuals of African ancestry has increased by 2.5%, and the proportion of people of Hispanic or of Latin American ancestry by around 0.5%. In the case of indigenous peoples (including Native Americans, Australian Aboriginals

and Pacific Islanders), representation has decreased slightly since 2009.

By looking up GWAS involving only Asian participants in PubMed (349 studies), we found the institution of the first author of each study. Around 93% of these studies were conducted in Asian countries. That the number of GWAS involving local populations has risen so much in Asia is heartening. But with such a large increase overall in the number of GWAS performed in the past seven years, the lack of growth in representation from other populations is remarkable and deeply disconcerting.

Of course, our analysis does not account for the resampling of data sets across independent studies. Information from some cohorts in publicly available databases has been used multiple times for different GWAS (see Supplementary Information). So the numerous samples of European ancestry used in GWAS could come from a smaller number of actual individuals. Yet if European-ancestry data sets are resampled more often than others, this in itself reflects population-specific differences in research effort.

WHY THE BIAS?

The continuing European bias in GWAS is likely to be the result of logistical, systemic and historical factors.

The more populations that are included in a study, the more variables there are to control for. In trying to keep things as simple as possible, geneticists probably favour the use of existing cohorts, such as that of the Framingham Heart Study, or other large data sets generated by well-established medical centres.

Such organizations collect samples and information from people in the same geographic location, who are presumed to be exposed to shared environmental factors, using uniform collection practices. But for various reasons, some populations are easily bypassed. People may have limited access to certain medical centres, for example, or, for cultural or historical reasons, elect not to contribute their samples to research.

Genotype and phenotype information from diverse populations is available. Researchers using NIH funding are required to submit any such information they have collected to dbGaP, a public database of genotypes and phenotypes. Analogous recommendations are made by other major biomedical funders outside the United States. In Europe, geneticists are encouraged to share similar data through the European Genome-phenome Archive (EGA). Yet for various reasons (such as the difficulties of getting certain kinds of studies funded, a preference for larger sample sizes, a perception that the analysis will be simplified by using data from one ancestry group or a lack of awareness of the diversity of data sets available) geneticists seem to be preferentially using cohorts of European ancestry.

A study of Greenlandic Inuits revealed a previously missed genetic variant associated with height.

Repeated sampling is almost certainly exacerbating the problem. Indeed, to some degree, the over-representation of people of European ancestry in GWAS may be a legacy of earlier biases.

WHAT'S MISSED

Irrespective of what's driving it, the continued under-representation of populations of mixed ancestry or of people whose ancestry is not European is a problem.

Until they are able to conduct amply powered GWAS on each major ancestral population across the world, geneticists will continue to miss important information about disease biology. They won't know how many of the thousands of associations between variants and diseases, and between variants and responses to drugs, observed in populations of European ancestry replicate in other groups. And opportunities will be missed to discover new associations with disease traits in other populations.

"European ancestry in GWAS may be a legacy of earlier biases."

For example, for 25% of the variants in European Americans that GWAS have identified as being associated with body mass index, type 2 diabetes and lipid levels, the strength of the association differs in at least one out of five populations of non-European ancestry[4]. This means that a variant that is associated with diabetes may confer a different risk of disease in someone of European ancestry than in, say, an individual of African ancestry.

Likewise, population-specific differences in the frequencies of variants associated with drug metabolism may mean that certain drugs will be safer and more effective in some populations than in others. The *CYP2D6* gene, for instance, is involved in the metabolism of many commonly prescribed drugs, including tamoxifen, which is used to treat breast cancer. More than 100 different variants of this gene (alleles) — many of which affect an individual's ability to safely digest and use a drug[5] — occur at different frequencies in different populations.

Several associations between drug responses and clinically relevant genetic variants have already been identified with GWAS. In some cases in which the effect sizes are large, significant results have been found with as few as 51 cases and 282 controls[6]. (In this case, patients had different reactions to the lipid-lowering drug simvastatin.) Although physicians must weigh the costs and benefits of using pharmacogenetic testing to guide prescription and dosage decisions for individual patients, these findings suggest that the small samples that have already been collected from under-represented populations could yield leads that have not been identified in populations of European ancestry.

Conducting analyses in other populations is also crucial for assessing the accuracy and broader relevance of a finding. It is possible,

for example, that associations between certain disease traits and variants found in European populations that cannot be replicated in other populations are actually false positives. In fact, the analysis of a broader representation of populations can reveal insights that would have otherwise been missed.

A genome-wide scan in a Greenlandic Inuit population, for example, found last year that a single-nucleotide polymorphism (SNP) in a fatty-acid enzyme affects height in both this population and Europeans[7]. The authors suggest that previous GWAS may have missed this variant because of its low frequency in Europeans (0.017 compared to 0.98 in the Greenlandic Inuit population) — even though it has a much greater effect on height than others previously identified through GWAS.

NEW DIRECTIONS

Increasingly, the sequencing of whole genomes and whole exomes (that is, the complete set of protein-coding genes) are beginning to be used more widely for discovery as costs fall. These may prove more fruitful than GWAS for individual-level diagnosis and treatment. Certainly, they are better suited to revealing rare variations that are clinically informative. (GWAS identify known genetic markers associated with a trait, but not necessarily the mutations that cause the disease.)

Studies that use these new approaches have been slightly more successful than GWAS at recruiting a greater diversity of populations. For example, the International Exome Aggregation Consortium hosts data on genetic variants from more than 60,000 samples, of which 8.6% are from people of African ancestry, 9.5% are from people of Latin American ancestry, and 60.4% are from people of European ancestry[8] (see page 154). The remaining samples (21.5%) are from south Asia, east Asia and the Middle East. Similarly, the TransOmics for Precision Medicine whole-genome sequencing project of the US National Heart, Lung and Blood Institute is growing and currently holds 62,000 samples, of which 50% are from European Americans, 30% are from African Americans, 10% are from Hispanics or Latin Americans, and 8% are from Asians.

Often, large sample sizes are needed to uncover rare genetic variants associated with disease traits. In fact, this realization — from the first generation of exome discovery studies — is driving new interest in ultracheap genotyping arrays (collections of targeted fragments of DNA). Using such arrays, geneticists can speed up the sequencing process and analyse many targeted samples in one go. Exome sequencing combined with the use of genotyping arrays is likely to be the favoured approach over the next decade. Nonetheless, GWAS remains a useful precursor to such studies, as well as to those involving whole-genome sequencing.

And emerging data indicate that inequalities in health care are being exacerbated by findings from whole-exome and -genome sequencing, despite their greater sample diversity compared with GWAS. Patients of African and Asian ancestry are currently more likely than those of European ancestry to receive ambiguous genetic test results after exome sequencing, or be told that they have variants of unknown significance[9]. Furthermore, patients of African ancestry are more likely than those of European ancestry to be wrongly told that a mutation they carry increases their risk of developing a life-threatening heart condition known as hypertrophic cardiomyopathy[10]. Had more ethnically diverse controls been included in the candidate-gene studies that identified these associations, population-specific differences in the frequency of presumed disease-causing variants would have revealed a false positive at the outset.

"Historical, cultural, scientific and logistical factors are sustaining an embarrassing bias in genomics."

WHAT NOW?

The message being broadcast by the scientific and medical genomics community to the rest of the world is currently a harmful and misleading one: the genomes of European descendants matter the most.

Certain efforts, combined with newer data-gathering initiatives, can help to move the needle in the right direction. Some investigators in genomics focus exclusively on diverse populations. For instance, landmark trans-ethnic studies have identified genes associated with traits such as diabetes, levels of lipids and other metabolites, prostate cancer and gene expression[11]. Also, various ventures aim to boost genomics studies in under-represented populations worldwide. The Human Heredity and Health in Africa Consortium, for example, was established by the NIH and the Wellcome Trust in London in 2012 to help build infrastructure and genomics expertise across Africa.

In our view, more fundamental changes are needed — both top-down and bottom-up. Funding agencies should develop financial incentives for the creation of diverse cohorts of study participants. One way for them to do this would be to prioritize grant requests that propose investigations in populations of non-European (and especially of African) ancestry. Given limited budgets, this may need to happen hand in hand with a reduction in the funding of research on existing cohorts of European ancestry for traits and diseases that have been relatively well characterized. (Around 850 genetic associations with height have now been reported by roughly 30 independent GWAS — the vast majority of which have been conducted using individuals of European ancestry.)

Further, all genomics researchers need to recognize the importance of studying under-represented populations to ensure that the benefits of research are distributed fairly and to maximize the potential for discovery. On a practical level, training programmes and new infrastructure, such as good health-care clinics that provide genetic testing in predominantly black or Hispanic neighbourhoods, could enhance trust and allow people to engage in projects as stakeholders rather than as study participants.

A culture shift is required at every level. Efforts to recruit participants for biomedical research in under-represented communities have been most successful when conducted by investigators of concordant racial or ethnic background, and in partnership with institutions trusted by those communities[12] — such as historically black colleges and universities in the United States.

Indeed, to a large extent, the persistent bias in sampling in genomics mirrors the employment trends evident in biomedical institutions worldwide. In the United States in 2012, less than 4% of the tenured and tenure-track faculty members in research-intensive biomedical departments were African American, Hispanic or Native American[13].

A complex web of historical, cultural, scientific and logistical factors is sustaining an embarrassing bias in genomics. Before precision medicine takes hold in clinical practice, we must correct its course. ■

Alice B. Popejoy *is a PhD candidate at the Institute for Public Health Genetics (IPHG) at the University of Washington, Seattle, USA.* **Stephanie M. Fullerton** *is associate professor of bioethics and humanities at the University of Washington, Seattle, USA.*
e-mails: popejoy@uw.edu; smfllrtn@uw.edu

1. Need, A. C. & Goldstein, D. B. *Trends Genet.* **25**, 489–494 (2009).
2. Bustamante, C. D., De La Vega, F. M. & Burchard, E. G. *Nature* **475**, 163–165 (2011).
3. Fullerton, S. M., Yu, J.-H., Crouch, J., Fryer-Edwards, K. & Burke, W. *Hum. Genet.* **127**, 563–572 (2010).
4. Carlson, C. S. *et al. PLoS Biol.* **11**, e1001661 (2013).
5. Desta, Z., Ward, B. A., Soukhova, N. V. & Flockhart, D. A. *J. Pharmacol. Exp. Ther.* **310**, 1062–1075 (2004).
6. Daly, A. K. *Nature Rev. Genet.* **11**, 241–246 (2010).
7. Fumagalli, M. *et al. Science* **349**, 1343–1347 (2015).
8. Lek, M. *et al. Nature* **536**, 285–291 (2016).
9. Petrovski, S. & Goldstein, D. B. *Genome Biol.* **17**, 157 (2016).
10. Manrai, A. K. *et al. N. Engl. J. Med.* **375**, 655–665 (2016).
11. Li, Y. R. & Keating, B. J. *Genome Med.* **6**, 91 (2014).
12. Yancey, A. K., Ortega, A. N. & Kumanyika, S. K. *Annu. Rev. Public Health* **27**, 1–28 (2006).
13. Leboy, P. S. & Madden, J. F. *DNA Cell Biol.* **31**, 1365–1371 (2012).

Sample Descriptions and Ancestry Categories

"African ancestry"

African American
African American ancestry
African ancestry
African cases
African Caribbean ancestry
Afro-Caribbean
Afro-Caribbean cases
Afro-Caribbean controls
Afro-Caribbean individuals
Black cases
Black child cases
Black controls
Black individuals
Gambian ancestry
Malawian ancestry
Moroccan ancestry
Nigerian ancestry
North African ancestry
Seychelles female individuals
Seychelles male individuals
Seychellois ancestry
Sub-Saharan African
Tanzanian ancestry
Tunisian ancestry
Middle Eastern ancestry
Middle Eastern Arab ancestry
Pakistani ancestry
Saudi Arabian ancestry
Turkish ancestry cases
Turkish ancestry
Turkish cases
Turkish controls
Turkish uveitis cases

"Asian ancestry"

Asian ancestry
Bangladeshi ancestry
Chinese ancestry
Dai Chinese ancestry
Dravidian ancestry
East Asian ancestry
East Asian cases
Han Chinese ancestry
Han Chinese cases
Han Chinese controls
Han Chinese individuals
Hong Kong Chinese ancestry
Hui Chinese ancestry
Indian ancestry
Indian Asian ancestry
Japanese ancestry cases
Japanese ancestry
Japanese controls
Japanese ancestry
Jingpo Chinese ancestry
Korean ancestry
West African ancestry
Yoruban ancestry

"Ashkenazi/Jewish"

Ashkenazi Jewish cases
Ashkenazi Jewish controls
Ashkenazi
Jewish cases
Jewish controls
Jewish Israeli cases
Jewish Israeli controls
Jewish-Israeli ancestry

"Arab/Middle Eastern"

Afghanistan ancestry
Arab ancestry
Arab-Israeli ancestry
Arab-Israeli founder
Iranian ancestry
Israeli/Arab controls
Israeli/Arab cases
Lebanese ancestry
Malay ancestry
Malaysian ancestry
Malaysian Chinese ancestry
Mongolian ancestry
Nepalese ancestry
North Indian ancestry
Oriental ancestry
Punjabi Sikh ancestry
She Chinese ancestry
Silk Road individuals
Singaporean ancestry
Singaporean Chinese ancestry
South Asian ancestry
South East Asian ancestry
South Indian
Southern Indian ancestry
Sri Lankan Sinhalese ancestry
Taiwanese ancestry
Taiwanese
Thai ancestry
Thai-Chinese ancestry
Tibetan ancestry
Uighur cases
Uighur controls
Uygur Chinese ancestry
Uygur-Kazakh Chinese
Vietnamese ancestry
Vietnamese-Korean ancestry

"European ancestry"

Amish cases
Amish controls

Amish individuals
Bulgarian ancestry
Carlantino individuals
Carlantino female individuals
Caucasian Eastern Mediterranean ancestry
Cilento individuals
Erasmus Rucphen Family individuals
Erasmus Ruchpen individuals
European
European American cases
European ancestry
European ancestry individuals
European ancestry cases
European ancestry controls
European cases
European child controls
European controls
European individuals
Finland founder cases
Finland founder controls
Finnish Saami individuals
French Canadian individuals
Friuli Venezia Giulia individuals
Hutterite adult individuals
Hutterite individuals
Italian isolated population individuals
Korculan individuals
Korkula individuals
Korkulan individuals
Northern Finnish founder individuals
Old Order Amish individuals
Orcadian female individuals
Orcadian individuals
Romanian founder cases
Russian ancestry
Sardinian cases
Sardinian controls
Sardinian individuals
Sorbian individuals
South African Afrikaner ancestry
Southern European ancestry
Talana adult individuals
Talana individuals
Tatar ancestry
Val Borbera individuals
Vis individuals
Western European ancestry

"Hispanic and Latin American"

Brazilian ancestry
Brazilian individuals
Caribbean Hispanic cases
Caribbean Hispanic controls
Costa Rican ancestry
Dominican Republic ancestry
Hispanic ancestry
Hispanic and unknown ancestry
Hispanic asthmatic individuals
Hispanic cases
Hispanic child cases
Hispanic child controls
Hispanic controls
Hispanic female individuals
Hispanic incident cases
Hispanic individuals
Hispanic male individuals
Hispanic newborn cases
Hispanic newborn controls
Hispanic prevalent cases
Latin American cases
Latin American controls
Latin American individuals
Latino cases
Latino child cases
Latino controls
Latino current smoker
Latino female cases
Latino female controls
Latino individuals
Latino male cases
Latino male controls
Mexican American
Mexican ancestry
Surinamese ancestry

"Native peoples"

American Indian ancestry
Bashkir ancestry
Cape Verdian cases
Martu Australian Aboriginal ancestry
Native Hawaiian ancestry
Native American ancestry
Pima Indian ancestry
Plains American Indian ancestry

"South Pacific Islander"

Filipino ancestry
Filipino female individuals
Filipino male individuals
Kosraen individuals
Micronesian ancestry
Oceania ancestry
Papua New Guinean ancestry
Solomon Islander ancestry

References Cited

[i] Andrew Yates, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil1 Carlos Garcín Girón, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H. Janacek, Nathan Johnson, Thomas Juettemann, Stephen Keenan, Ilias Lavidas, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Michael Nuhn, Anne Parker, Mateus Patricio, Miguel Pignatelli, Matthew Rahtz, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P. Wilder, Amonida Zadissa, Ewan Birney, Jennifer Harrow, Matthieu Muffato, Emily Perry, Magali Ruffier, Giulietta Spudich, Stephen J. Trevanion, Fiona Cunningham, Bronwen L. Aken, Daniel R. Zerbino, Paul Flicek. Ensembl 2016. *Nucleic Acids Res.* 2016 44 Database issue:D710-6. PubMed PMID: 26687719; PubMed CentralPMCID: PMC4702834. doi:10.1093/nar/gkv1157

[ii] Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., & Madden T.L. (2008) "BLAST+: architecture and applications." BMC Bioinformatics 10:421

[iii] G. Zhang. 2015. Genomics: Bird sequencing project takes off. Nature 522. DOI:10.1038/522034d

[iv] Pruitt, K.D., Tatusova, T., Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* (2007) 35(Database):D61–5.

[v] Goodstadt, L., Pontling, C. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput. Biol.* (2006) Vol. 2: e133.

[vi] Ranwez, V., Delsuc, F., Ranwez, S., Belkhir, K., Tilak, M., Douzery, E. OrthoMaM: A database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol. Biol.* (2007) Vol. 7: 241.

[vii] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* (2004) 32(5):1792-1797. doi:10.1093/nar/gkh340.

[viii] "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0." Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. Systematic Biology, 59(3):307-21, 2010.

[ix] Felsenstein, J. 2009. PHYLIP (Phylogeny Inference Package) version 3.7a. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

[x] Yang, Z., and Bielawski, J. Statistical methods for detecting molecular adaptation. *Tree* (2000) Vol. 15, No. 12: 496—503.

VITA

Alice B. Popejoy

396 62nd Street
Oakland, CA 94618

E-mail: alice.popejoy@gmail.com
Mobile phone: (916) 837-2951

EDUCATION

Doctor of Philosophy, Institute for Public Health Genetics, University of Washington, 2017
Certificate in Statistical Genetics, Depts. Biostat/Stat, University of Washington, 2015
B.A. in Biology/French, Hamilton College, Clinton, NY, 2009

PROFESSIONAL EMPLOYMENT

2010—2012 Project Manager, National Science Foundation (NSF) Advancing Ways of Awarding Recognition in Disciplinary Societies (AWARDS Project), Association for Women in Science (AWIS), Alexandria, VA

PUBLICATIONS

Peer-Reviewed Journal Articles

2012 Popejoy AB and PS Leboy. "Is Math Still Just a Man's World?" *Journal of Mathematics and System Sciences* 2:292—298.

2016 Blue EM, Brown LA, Conomos MP, Kirk JL, Nato AQ, Popejoy AB, Raffa J, Ranola J, Wijsman EM, and TA Thornton. Estimating relationships between phenotypes and subjects drawn from admixed families. *BMC Proceedings* 10:357.

2016 Popejoy AB and SM Fullerton. Genomics is failing on diversity. *Nature* 538:161-64.

Book Chapters

2014 Cadwalader EL, Herbers JM and AB Popejoy. Disproportionate Awards for Women in Disciplinary Societies, in Demos V, Berheide CW, MT Segal (ed.), *Gender Transformation in the Academy (Advances in Gender Research, Volume 19)* Emerald Group Publishing Limited, pp.243—263.

2014 Cadwalader EL, Herbers JM and AB Popejoy. Professional societies and gender equity in STEM, in Bilimoria D and L Lord (ed.), *Women in STEM Careers: International Perspectives on Increasing Workforce Participation, Advancement and Leadership* Edward Elgar Publishing Limited, pp.166—182.

Manuscripts in Preparation (Dissertation Research)

2017 Popejoy, AB and JH Thomas. Shifts in Purifying Selection (SPurS) identified by novel statistic to detect divergence-linked changes in amino acid sequences.

2017 Popejoy, AB and J Felsenstein. Conservative Chi-Square Collapse (C3) identifies structure, heterogeneity and outliers in high-dimensional genomics data.

Web-Based Publications

2010 Popejoy AB and C Mooney. Genetic Discrimination: The Best Reason for Universal Health Care You've Never Heard Of. *Intersection* @ Discover Magazine Blog.

Other Publications

2011 Popejoy AB, Leboy PS, Crowley J, and Cook P. Investigating the Gender Gap in SIAM Prizes. *Society for Industrial and Applied Mathematics (SIAM) News* 44:10.
2011 Popejoy AB and PS Leboy. AAS and the Under-recognition of Women for Awards and Prizes. *American Astronomical Society (AAS) Newsletter* (2011) 159:18—19.

AWARDS, GRANTS, AND FELLOWSHIPS

2012 National Science Foundation (NSF) Graduate Research Fellowship (#2012143990)
2014 National Institutes of Health (NIH) Statistical Genetics Training Grant (GM081062) Department of Biostatistics, University of Washington, Seattle, WA
2015 University of Washington Husky 100 Award for Extracurricular Engagement
2016 National Institutes of Health (NIH) Statistical Genetics Training Grant (GM081062) Department of Biostatistics, University of Washington, Seattle, WA
2016 NSF Graduate Research Opportunities Worldwide (GROW) Fellowship, Norwegian Centre for Mental Disorders Research (NORMENT), University of Oslo (UiO) and the Research Council of Norway (Forskningrådet No. 263348)

RESEARCH EXPERIENCE

2010—2012 Research Associate and Project Manager, NSF-ADVANCE Grant, Association for Women in Science (AWIS), Alexandria, VA (Research Mentor: Dr. Phoebe S. Leboy, Department of Biochemistry, UPenn)
2013—2016 Phylogenetics and Bioinformatics, Department of Genome Sciences, University of Washington (Research Mentor: Dr. James H. Thomas)
2014 Admixture Mapping Working Group, Genetic Analysis Workshop 19, Department of Biostatistics, University of Washington (Research Mentors: Dr. Timothy Thornton, Dr. Ellen Wijsman)
2015—2016 Statistical and Population Genetics, Department of Genome Sciences, University of Washington (Research Mentors: Dr. Joseph Felsenstein, Dr. Elizabeth A. Thompson, Dr. Bruce Weir)
2016 Statistical Genetics Intern, Axio Research, LLC, Seattle, WA (Supervisor: Dr. Sangsoon Woo, Dr. David Henderson)

INVITED TALKS

2011 American Astronomical Society (AAS) Executive Board Meeting, Boston, MA
2011 STEM Education Panel, American Enterprise Institute (AEI), Washington, DC
2012 Public Policy for Scientists, Fred Hutchison Cancer Research Center, Seattle, WA
2017 Whose Genomes Matter? University of Southern California, Los Angeles, CA

CONFERENCE ACTIVITY

2009 "Genetic Information and Discrimination: Perspectives from the United States," Ninth Asian Bioethics Association Meeting, Yogyakarta, Indonesia
2010 "Scholarly Recognition for Women in Scientific Disciplinary Societies," National Institutes of Health (NIH) Annual Meeting, Bethesda, MD
2011 Joint Mathematics Annual Meeting, Science Education Session, Boston, MA
2012 "Advancing Ways of Awarding Recognition in Disciplinary Societies (AWARDS)," National Science Foundation (NSF) ADVANCE PI Meeting, Arlington, VA
2012 American Association for the Advancement of Science (AAAS), Washington, DC
2014 "Non-visual Human Opsin Evolution and Implications for Human Health" American Society of Human Genetics (ASHG) Meeting, San Diego, CA
2014 Summer Institutes in Statistical Genetics (SISG), UW Dept. of Biostatistics
2015 "Illuminating Human Photoreceptors: An Evolutionary Investigation of Non-Visual Opsins," Public Health Genomics Symposium, Seattle, WA
2015 Summer Institutes in Statistical Genetics (SISG), UW Dept. of Biostatistics
2016 American Society of Human Genetics (ASHG) Annual Meeting, Vancouver, BC
2016 International Genetic Epidemiology Society (IGES) Meeting, Toronto, CA

DEPARTMENTAL TALKS

2009 "Investigating and Informing Public Knowledge and Attitudes on DNA, Genetic Discrimination, and the Law," Department of Biology, Hamilton College, NY.
2014 "Biological Light Receptors: PheWAS, Statistics, & ELSI," Statistical Genetics, University of Washington, Seattle.
2015 "Capstone Project in Statistical Genetics: Illuminating Human Photoreceptors," Population Genetics, University of Washington, Seattle.
2016 "Illuminating Human Photoreceptors: signals of positive selection in melanopsin," Statistical Genetics, University of Washington, Seattle.
2016 "Birth-and-Death Evolution of Vertebrate Photoreceptors." Population Genetics.
2016 "Whose Genomes Matter?" Department of Epidemiology, U. of Washington.

SEMINAR TALKS

2013 "Genomics and Genetics of Human and Primate Y Chromosomes" (Hughes JF and S Rosen, 2012), Statistical Genetics Seminar.
2014 "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants" (Kircher M, Witten D, Jain P, O'Roak B, Cooper G, Shendure J. *Nature Genetics* 2014), Molecular Population Genetics and Evolution.
2014 "Haplotypes vs. Single Marker Linkage Disequilibrium Tests: What do we gain?" (Akey J, Jin L, and M Xiong, 2001), Statistical Genetics Seminar.
2015 "Low Template DNA Profile Analysis as Forensic Evidence: Challenges and Solutions" (Balding DJ and J Buckleton, 2009), Statistical Genetics Seminar.
2015 "Exome-Based Mapping and Variant Prioritization for Inherited Mendelian Disorders" (Koboldt DC, Larson DE, Sullivan LS, Bowne SJ, Steinberg KM, Churchill JD, Buhr AC, Nutter N, Pierce EA, Blanton SH, Weinstock GM, Wilson RK, and SP Daiger, 2014), Statistical Genetics Seminar.

2016 "Combined Sequence-based and Genetic Mapping Analysis of Complex Traits in Outbred Rats" (Baud A, Hermsen R, Guryev V, Stridh P, Graham D, McBride MW, Foroud T, Calderari S, Diez M, et al., 2013), Statistical Genetics Seminar.
2016 "Population Structure of UK Biobank and Ancient Eurasians Reveals Adaptation at Genes Influencing Blood Pressure" (Galinsky K, Loh P, Mallick S, Patterson N, and Price A., 2016), Statistical Genetics Seminar.

TEACHING EXPERIENCE

2009 Teaching Assistant, Department of French, Hamilton College
2015 Study Group Mentor, Population Genetics (Professor: Joseph Felsenstein) Department of Genome Sciences, University of Washington
2016 Tutor, High School Mathematics (Algebra II)
2016 Teaching Assistant, Implications of Public Health Genomics for Society, Institute for Public Health Genetics (IPHG), University of Washington
2017 Teaching Assistant, Genetic Epidemiology, Institute for Public Health Genetics, University of Washington

UNIVERSITY OF WASHINGTON SERVICE AND LEADERSHIP

Science Policy Committee Chair, Graduate and Professional Student Senate, 2013—2014
Graduate and Professional Student Senate (GPSS) President, 2014—2015
University of Washington Board of Regents, Ex-Officio Member, 2014—2015
University of Washington Alumni Association Board of Trustees, 2014—2015
Faculty Senate Executive Committee, Ex-Officio Member, 2014—2015
Graduate School Council, Ex-Officio Member, 2014—2015
Provost Advisory Committee for Students (PACS) Chair, 2015—2016
Dean's Advisory Committee for Students (DACS) Co-Chair, 2016—2017

MEDIA COVERAGE

"Further bias in personal genomics" (Editor's note, *Frontline Genomics*)
"Genomics Medicine Ireland raises $40 million for large-scale genomics research" (Juliet Preston, *Med City News* 2016)
"Why racial diversity in genetics studies matters in patient care" (Elizabeth Newbern, *Genetic Literacy Project* 2016)
"Genomics continues to fail on diversity" (Timothy Kenney, *The Daily* 2016)
"Among the Multitude: A Look at the Complexity of Diversity in Genomics" (Benjamin Ross, *Bio-IT World* 2016)
"Solving the Lack of Diversity in Genomic Research" (Emily Mullin, *MIT Technology Review* 2016)
"Personal Genomics Open Access Datasets Even More European-Biased Than Scientific Literature?" (Genome Diary, *Genomes, Web 2.0 and Bioethics* 2016)

RELATED PROFESSIONAL SKILLS

Summer Institutes for Statistical Genetics (SISG)
Elements of R for Genetics and Bioinformatics, Population Genetic Data Analysis

Summer Institutes for Big Data (SIBD)
Supervised Methods for Statistical Machine Learning
High-Dimensional Omics Data Analysis

TEACHING AREAS

Public Health and Genome Sciences
Public Health Genomics
Evolutionary Genomics
Genetic Epidemiology
Bioinformatics and Computational Biology

Social Sciences
Science Communication
Ethical, Legal, and Social Implications of Genomics Research

LANGUAGES

Computer Programming Languages
Python (proficient)
R / R Studio (intermediate)
STATA (intermediate)

Spoken Languages
English (native)
French (fluent)
Norwegian (conversational)
Spanish (novice)
Japanese (novice)

REFERENCES

Joseph Felsenstein
Professor, Department of Genome Sciences
Box 355065, University of Washington, Seattle, WA 98195
(206) 543-0150 / joe@gs.washington.edu

James H. Thomas
Professor, Department of Genome Sciences
Box 355065, University of Washington, Seattle, WA 98195
(206) 543-0754 / jht@u.washington.edu

Stephanie (Malia) Fullerton
Associate Professor
Department of Bioethics & Humanities
Office A-204F, University of Washington, Seattle, WA 98195
(206) 616-1864 / smfllrtn@u.washington.edu

ADDITIONAL REFERENCES

Elizabeth A. Thompson
Professor, Department of Statistics
Box 354322, University of Washington, Seattle, WA 98195
(206) 543-7237 / eathomp@u.washington.edu

Ellen Wijsman
Professor, Department of Biostatistics
Box 359460, University of Washington, Seattle, WA 98195
(206) 543-8987 / wijsman@u.washington.edu

Bruce Weir
Professor, Department of Biostatistics
Box 357232, University of Washington, Seattle, WA 98195
(206) 221-7947 / bsweir@u.washington.edu