Evolutionary-based methods for predicting genotype-phenotype associations in the mammalian genome

by

Raghavendran Partha

Bachelors and Masters, Indian Institute of Technology Madras, 2014

Submitted to the Graduate Faculty of
the School of Medicine in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Raghavendran Partha

It was defended on

July 23rd 2019

and approved by

Maria Chikina, Assistant Professor, Department of Computational and Systems Biology, University of Pittsburgh

Chakra Chennubhotla, Associate Professor, Department of Computational and Systems Biology, University of Pittsburgh

Miler Lee, Assistant Professor, Department of Biological Sciences, University of Pittsburgh

Andreas Pfenning, Assistant Professor, Computational Biology Department, Carnegie Mellon University

Mark Rebeiz, Associate Professor, Department of Biological Sciences, University of Pittsburgh

Dissertation Director: Nathan Clark, Associate Professor, Department of Computational and Systems Biology, University of Pittsburgh

Evolutionary-based methods for predicting genotype-phenotype associations in the mammalian genome

Raghavendran Partha, PhD

University of Pittsburgh, 2019

Phenotypic and genotypic variation between species are the result of millions of experiments performed by nature. Understanding why and how phenotypic complexity arises is a central goal of evolutionary biology. Technological advancements enabling whole genome sequencing have laid the foundation for developing comparative genomics-based tools for inferring genetic elements underlying phenotypic adaptations. The work covered as part of this thesis will develop these tools drawing from principles of convergent evolution, aimed at generating specific functional hypotheses that can help focus experimental efforts. These tools will be relevant for characterizing context-specific functions of cis-regulatory elements as well as protein-coding genes, where a large number lack functional annotation beyond domain homology. Expanding from one-dimensional approaches studying proteins in isolation, we propose to build an integrated co-evolutionary framework that will serve as a powerful tool for protein interaction prediction. In this dissertation, we discuss these ideas through the following three projects.

In chapter 1, we perform a genome-wide scan for genes showing convergent rate changes in four subterranean mammals, and study the underlying changes in selective pressure causing these convergent shifts in rate. Using a new variant of our rates-based method, we demonstrate that eye-specific regulatory regions show strong rate accelerations in the subterranean mammals. This study demonstrates the potential of convergent evolution-based tools in the functional annotation of eye-specific genetic elements.

In chapter 2, we build a robust method to infer shifts in rate associated with a wide range of evolutionary scenarios. We investigate the statistical underpinnings of our rates-based framework and identify the best performing variant of our method across real and simulated phylogenetic datasets. We distribute these tools to the research community, enabling large scale generation of specific functional hypotheses for regulatory regions.

In chapter 3, we propose to construct a powerful framework for protein interaction prediction using integration of proteome-wide co-evolutionary signatures. We systematically benchmark the predictions of our coevolutionary framework using known functional interactions among proteins across various scales. We make the predictions of the framework publicly available, useful for functional annotation of less well-characterized genes.

Table of Contents

List of Tables

List of Figures

Preface

Firstly, I would like to express my gratitude to Dr Nathan Clark for being a better mentor and advisor than I could have hoped for. Amidst all the difficulties I have faced in the past five years in my personal life, he has been a constant source of support and guidance without which I could not have continued in this PhD program. I will definitely miss my weekly interactions with him; they have been instrumental in my development as a scientist and also an individual. I would like to thank Dr Maria Chikina who has, in principle, been my co-advisor. Critical aspects of my dissertation would not have been possible if not for her guidance. I consider myself extremely lucky to have had the opportunity to closely work with and learn from two incredible scientists during my PhD.

Being a part of Clark/Chikina labs has been a wonderful experience, and I would like to thank all past and present members. Specifically, I learnt a great deal through my interactions with Dr Wynn Meyer, Amanda Kowalczyk, Wayne Mao, Melissa Plakke, Elysia Saputra, Dr Camille Meslin, Jiaxuan Yang, and Luisa Cusick. Members of the lab embody ideal qualities in scientists - ambitious and smart, yet supportive and kind. I would also like to thank all members of the CPCB program including faculty, students, and most importantly administrative staff. They are a fantastic group of individuals that have created a highly supportive and engaging environment for budding scientists.

I would like to acknowledge members of my thesis committee - Dr Chakra Chennubhotla, Dr Miler Lee, Dr Andreas Pfenning, and Dr Mark Rebeiz, for their inputs on improving my dissertation.

I would like to thank all members of my family in the US who have been an amazing support system over the years. I am deeply grateful for my friends here in Pittsburgh who have been nothing short of family outside of my real family.

The single most important factor behind the successful completion of my PhD is my mother Ramani. She has had to assume the roles of both parents after the passing away of my father midway through my PhD. I dedicate this PhD to her unconditional love and support. I will continue to remain grateful to my father who always strived to create the best possible environment for me to give my best shot at my goals in life. I only wish that he was here to be a part of this joyous moment, of me becoming the next Dr Partha.

The work presented in Chapters 2 and 3 of this thesis have been published as journal articles - in Partha *et al.* (2017) and Partha *et al.* (2019) respectively. Chapter 4 is based on unpublished work that is set to be submitted for publication.

1.0 Introduction

The natural world is characterized by diversity across many scales. Diversity in the organisms that inhabit this world ranging from prokaryotic microorganisms to modern mammals. Focusing on the mammalian species, we observe some common characteristics including possession of a neocortex, some hair, three middle ear bones among other traits. However, mammals are characterized by diversity at many levels. Some mammals are egg-laying while a majority of them give birth to live young. In terms of size, there is roughly a million fold difference in the weight of small versus large mammals. Curiously, many of these large mammals also tend to be longer lived. Mammals have also evolved unique traits or phenotypes that allow them to adapt to their respective environments and overcome the associated challenges. Understanding the genetic basis of this phenotypic diversity is a problem of outstanding interest in biology. The overarching goal of mapping the phenotypes or traits to the genotypes or the genetic makeup of a species has been an important topic of research. In the recent past, improvements in sequencing technologies have resulted in whole genome sequencing efforts for many mammals. Beginning with the original human genome project, we now have the genome sequences of several mammals and vertebrates among other species (Lindblad-Toh *et al.*, 2011). Newer efforts to sequence the genomes of additional species are also being pursued including initiatives such as the 200 mammal sequencing project (`http://grantome.com/grant/NIH/R01-HG008742-01A1`). One of the goals of these projects is to sequence and align the genome sequences of these species to the human sequence so as to enable comparative genomic analyses. Concurrent to the sequencing of these genomes, bioinformatics tools have been developed to align these large genomic assemblies. Such genome-scale alignments allow us to compare these genomes at varying scales. We can compare how the structures of genomes evolve between closely related species, or more locally, how sequences of various genes have changed across these species. These local alignments in turn allow us to reconstruct the evolutionary relationships between these species in the form of phylogenetic trees.

Through the comparisons of genomes, we can readily observe one property – sequences of the genome that are similar or conserved versus dissimilar. Sequence conservation is a useful property to study because it serves as a proxy for function. This is because if the sequence is conserved it implies that there is a force of selection against mutations in this region. In comparison, regions that are not under such constraints are usually neutrally evolving, meaning mutations in such regions do not have a negative consequence to the organism. In many cases, we see that exons of protein-coding genes show high levels of sequence conservation as it is codes for the structure of the protein and subsequently its function. Contrasting to this, we typically observe a much smaller fraction of intergenic regions showing high sequence conservation. Several studies have looked at the conservation across the sequenced mammals to identify these functional regions. The findings show that a remarkable 80% of the exons in protein-coding genes in humans have detectable levels of conservation with the Opossum, a distantly related marsupial mammal (Hardison, 2010). Perhaps as one would have expected, mammals that are more closely related have higher levels of conservation. Even though the fraction of the total genome that is conserved declines rapidly with distance, we find that there are many local regions that are conserved to a very high extent. And surprisingly, many non-coding regions are conserved across mammals. These include important regulatory regions such as enhancers, promoters but a vast majority of these conserved regions have functional annotations that are yet to be discovered (Siepel *et al.*, 2005).

Conservation is indeed a useful proxy for function, but it falls short there - it does not necessarily reveal what that function is. In order to identify the biological function or the associated phenotype of such conserved elements alternative strategies are necessary. Genome-wide experimental initiatives such as the ENCODE and the RoadMap epigenomics projects have been undertaken with the aim of elucidating components regulating cellular functions and mechanisms (Bernstein *et al.*, 2010; Andersson *et al.*, 2014). These efforts have identified a wealth of biochemically active elements across cell types as well as developmental time points in the human and mouse genomes. As a result of this welcome explosion of newly acquired data, the research community faces a more challenging problem - developing tools that can successfully unlock the association between variation at the phenotypic level to that at the level of the genotype. In other words, approaches that provide effective solutions to the problem of assigning specific biological functions to these regulatory elements have not kept pace with the rate at which these elements are identified. To develop effective solutions for a problem of this scale, there is a need for inter-disciplinary efforts that combine insights from complementary fields of research.

In this thesis, we seek to tackle this problem from the perspectives of convergent evolution and coevolution. Nature has provided countless examples of multiple unrelated lineages showing phenotypic adaptation to similar environmental challenges. Examples of convergent evolution include the evolution of winged flight, evolution of structures that enabled three different clades of mammals adapt to the marine environment, dietary adaptations and so on. Another form of convergent evolution is in the convergent loss of phenotypes or traits, such as the loss of vision in subterranean mammals. The availability of diverse patterns of phenotypic evolution in nature opens up the opportunity to develop evolutionary-based approaches aimed at inferring the changes at the genetic level underlying said instances of phenotypic convergence. Alternatively, genetic elements showing convergent changes in species characterized by phenotypic convergence are thus strong candidates for a functional role in the phenotype. Evolutionary methods inferring phenotypic associations for genetic elements based on patterns of convergence can be particularly powerful in the context of inferring candidate genetic elements amidst the background of the entire genome (Hiller *et al.*, 2012). This power is borne out of the presence of multiple independent instances of the phenotypic change which enables the identification of relevant regions of interest. Another force of evolution that can be leveraged to reveal functional associations in the cell is coevolution. As components of cellular systems, proteins do not act in isolation, and coordinate with other proteins resulting in functional pathways (Clark *et al.*, 2012b). Therefore, there is a shared pressure to maintain the functionality of a pathway on the participating proteins. Effective methods to identify coevolving pairs of genes therefore possess the power to predict uncharacterized components of genetic networks underlying phenotypic adaptations and biological functions.

2.0 Evolution of vision-related genetic elements across subterranean mammals

2.1 Introduction

The subterranean habitat has been colonized by numerous animal species for its shelter and unique sources of food (Andersen, 2004; Nevo, 1979). Obligate fossorial species in particular have adopted the underground as a dedicated home, yet the intense demands on life underground often require unique specializations. For one, the air in tunnels is often low in oxygen (hypoxic) and high in carbon dioxide (hypercapnic) (Nevo, 1979). This dark environment also requires enhanced senses to compensate for loss of vision. These and other subterranean specializations have been reported in many independent evolutionary lineages of insects, amphibians, reptiles, and mammals (Leys *et al.*, 2003; Lacey *et al.*, 2011; Albert *et al.*, 2007; Wilkinson, 2012). Within mammals alone, there are several unrelated subterranean species, including the true moles (family Talpidae), the African golden moles (Chrysochloridae), and the marsupial moles (Notoryctidae). There are also at least three unrelated lineages of subterranean rodents: the naked mole-rat (Heterocephalus glaber), blind mole-rats (Spalacidae), and the pocket gophers (Geomyidae).

Vision in many subterranean mammals is limited, and the degree of limitation in each species is related to its extent of underground habitation (Němec *et al.*, 2008; Quilliam, 1966; Sanyal *et al.*, 1990). For example, star-nosed moles (Condylura cristata) that share their time above ground and underground possess diminutive eyes with thick eyelids (Catania, 1999), while the naked mole-rat (Heterocephalus glaber), which spends almost all its time underground, shows tiny eyes that are rarely opened (Hetling *et al.*, 2005). Even more extreme are the completely subcutaneous eyes of the cape golden mole (Chrysochloris asiatica) and the blind mole-rat (genus Nannospalax), which is thought to reflect their strictly subterranean lifestyle (Sweet, 1909; Sanyal *et al.*, 1990). While some degree of visual regression is shared between subterranean mammals, not all visual structures and genetic pathways have regressed to the same degree. For instance, the eyes of true moles and mole-rats show anatomical regression but retain ocular architecture, suggesting that basic eye developmental programs must be largely intact (Carmona *et al.*, 2008, 2010; Quilliam, 1966). Embryonic eye development happens normally in these mammals with disruption occurring later on. Furthermore, analyses on their eye phenotypes display common characteristics. They always exhibit a small eye, with some lacking vitreous. They more often have defects in their lens, where vacuoles manifest in the fiber layer and thin epithelia unusually surrounds the entire structure. The cornea, the transparent layer forming the front of the eye, in these species is thin and shows reduction in corneal stroma. A surprising finding is the anomalous overgrowth of the iris and the proximal ciliary body. Retinas are also greatly remodeled in that they are dominated by rod cells and contain more short-wavelength cones (S-cones) than medium/long wavelength M/L-cones. However, the retina and optic nerve show a severe lack of retinal ganglion cells, therefore suggesting poor vision or lack of it. Thus, the subterranean mammals already exhibit a contrasting reduction of some eye structures alongside the overgrowth of others. The convergent loss of vision and visual structures in subterranean mammals allows us to ask which genetic regions genic or non-genic contributed to regression in these species and which were conserved.

The genetic causes of these malformations have been probed through studies of blind cavefish and evolutionary analysis of retinal genes in subterranean mammals (Jeffery, 2009; Emerling and Springer, 2014). Pioneering work by Hendriks *et al.* (2006) found the evolutionary rate of the lens and retina protein αA-crystallin to be markedly accelerated in the Middle Eastern blind mole-rat (Spalax ehrenbergi), as would be expected under relaxed constraint (Hendriks *et al.*, 2006). Furthermore, Emerling and Springer revealed that regressive genetic changes in retinal proteins are unevenly distributed across different visual pathways and eye tissues (Emerling and Springer, 2014). For one, cones, which are responsible for vision in bright light, preferentially contain genetic lesions such as stop codons, as opposed to rod genes. Previous studies have placed more emphasis on retinal components of vision and connections to the visual cortex because it is these components that sense light and transmit images to the brain for vision (Cooper *et al.*, 1993; Emerling and Springer, 2014). However, less emphasis has been placed on the genes contributing to other eye tissues, such as the cornea. To gain a more comprehensive understanding of regressive evolution, we should endeavor to determine how degeneration has proceeded differently across eye tissues and developmental stages, and more generally how we can identify new and perhaps unsuspected genetic elements that have responded to the subterranean environment.

The genomes of four subterranean mammals have been sequenced and studied for changes in response to their unique environment. The naked mole-rat genome revealed genetic changes in key genes involved in thermogenesis and circadian rhythm, as well as gene loss and deactivating mutations in core visual perception genes (Kim *et al.*, 2011). The blind mole-rat genome (Nannospalax gailili) also yielded diverse insights into its subterranean adaptations, such as an impactful change to the P53 protein allowing cells to escape hypoxia-induced apoptosis, as well as up-regulation of specific pathways in response to hypoxia and hypercapnia (Fang *et al.*, 2014). Additionally, the blind mole-rat genome yielded evidence of convergent evolution in circadian rhythm and hemoglobin genes, since some of their amino acid changes were mirrored in the naked mole-rat. Similar parallel evolution was seen in the deactivation of visual perception genes in the blind mole-rat and naked mole-rat. These instances of convergent change highlight a potential strategy to discover additional genetic regions that repeatedly respond to the subterranean environment. When evolutionary changes are observed in multiple, independent subterranean lineages, their convergence provides some evidence that the changes are in response to the environment, rather than unrelated species-specific conditions or even neutral processes (Losos, 2011; Stern, 2013; Rosenblum *et al.*, 2014).

There is much interest in using convergent evolution to reveal genetic changes related to environmental shifts without a priori expectations of which regions might respond. One strategy has been to search for convergent amino acid substitutions at specific protein sites (Foote *et al.*, 2015; Liu *et al.*, 2010; Dobler *et al.*, 2012). A complementary strategy is to search for convergent changes in selective pressure on larger functional regions, such as genes or regulatory sequences, because evolution at different nucleotides within a gene could nevertheless lead to convergent phenotypic effects. In practice, convergent changes in selective pressure are inferred by studying evolutionary rates, because selective constraint slows evolution, while lack of constraint and adaptation speed it. Computational methods employing this strategy search for functional elements whose evolutionary rates changed on those branches exhibiting the convergent environmental change (Marcovitz *et al.*, 2016; Hiller *et al.*, 2012;

Chikina *et al.*, 2016; Lartillot and Poujol, 2011). One demonstration of this approach by our group identified genes that convergently responded when mammalian lineages shifted from a terrestrial to a marine environment (Chikina *et al.*, 2016). Another recent study by Prudent *et al.* (2016) demonstrated that regions showing convergent rate acceleration in the subterranean environment were enriched in visual perception genes and also contained circadian rhythm genes (Prudent *et al.*, 2016). Together, these studies show the promise of convergent rates to reveal genes underlying major changes in morphology and physiology related to drastic environmental shifts. To determine the demands placed upon subterranean species by their extreme environment, we searched for genes exhibiting convergent rate changes in four subterranean mammals. We report a large set of genes showing marked relaxation of constraint in subterranean species and which were highly enriched for visual functions. This set also contained many genes of undetermined function, which could be unrecognized causative genes in eye-related diseases. Finally, we pinpointed the eye-specific transcriptional enhancers in the Pax6 gene region using a new variant of our method, demonstrating the potential to detect new eye-specific enhancers at key developmental genes.

2.2 Materials and Methods

2.2.1 Adding Nannospalax galili orthologs to alignment

Given the absence of Nannospalax galili (blind mole-rat or BMR) in the 100-species alignments made available by the UCSC genome browser, we employed a custom approach to add the correct BMR orthologous sequence based on its closest relative on the mammalian species phylogeny, mouse. Using the publicly available BMR gene models in NCBI, we first perform pairwise reciprocal nucleotide blast of all BMR gene cDNA sequences and the corresponding cDNA sequences of all genes in the mouse mm9 genome. For every mm9 gene sequence, we subsequently identify the correct BMR ortholog using the InParanoid program as follows - the program clusters pairs of sequences from the two queried genomes into groups of orthologs, and the BMR sequence forming the main ortholog pair (pairs with mutually best hit) in every group was identified as the correct ortholog (Remm *et al.*, 2001). We then perform a profile alignment using the openly available muscle program to add the identified BMR ortholog to the genes multi-species alignment (Edgar, 2004). For the study of non-genic regions in the Pax6 window, we utilized a simpler approach to identify the BMR orthologous region. For each non-genic region of interest, we performed blastn with the mm9 orthologous sequence as the query against the BMR assembled genome with the default Expect (E) value of 10 (NCBI Resource Coordinators 2016). The resulting best scoring blastn hit in the BMR genome, if any, was added to the non-genic regions multi-species alignment (obtained from the UCSC genome browser) using the profile alignment utility of the muscle program (Edgar, 2004; Haeussler *et al.*, 2019).

2.2.2 Calculating gene correlations with subterranean environment

Using the 100-species amino acid alignments from the multiz alignment available at the UCSC genome browser (Blanchette *et al.*, 2004; Harris, 2007; Haeussler *et al.*, 2019), those alignments with a minimum of 10 species were selected for study. We pruned each alignment to include only the species of interest represented in the proteome-wide average tree (Figure 2.3.1B) after adding the BMR ortholog of the corresponding gene sequence to this alignment as described in the previous section. For each resulting amino acid alignment we estimated branch lengths using the aaml program from the phylogenetic analysis using maximum likelihood (PAML) package (Yang, 2007). Branch lengths were estimated under an empirical model of amino acid substitution rates with rate variability between sites modeled as a gamma distribution approximated with four discrete classes (for computational efficiency) and an additional class for invariable sites (aaml model Empirical+F) (Whelan and Goldman, 2001; Yang, 1996). Branch lengths were estimated on a published mammalian species tree topology (Murphy *et al.*, 2004), modified to include the Nannospalax galili (blind mole-rat) whose position in the tree inferred based on existing literature on its ancestry (Fang *et al.*, 2014). For the analyses involving conserved non-genic regions near Pax6, we first identified the regions of interest based on the human phastCons track generated from the 100-way vertebrate multiz alignment, eliminating any region of overlap with the human mRNAs track. For each such non-coding element we obtained the 100-way multiz alignment, further selecting only for the species that are present in our species set of interest after adding the BMR ortholog of the corresponding non-genic sequence to this alignment as described in the previous section. We estimate the branch lengths using the baseml program of the PAML package under the general reversible process (REV) model for nucleotide substitution rates, with rate variability between sites modeled as a gamma distribution approximated with four discrete classes and an additional class for invariable sites (Rodriguez *et al.*, 1990; Blanchette *et al.*, 2004).

Raw branch lengths were transformed into relative rates using a projection operator method (Sato *et al.*, 2005). These branch-specific relative rates were then used to perform a Mann-Whitney U test and correlation analysis over the binary variable of subterranean or aboveground (i.e., not subterranean) branches (Figure 2.3.1A). Subterranean branches are those leading to the star-nosed mole (Condylura cristata), cape golden mole (Chrysochloris asiatica), naked mole-rat (Heterocephalus glaber) and blind mole-rat (Nannospalax galili).

2.2.3 Functional enrichment analysis

We performed functional enrichment analysis using the GOrilla tool by searching for enriched GO terms in the 500 most convergent genes compared to the full background set of 18,980 (Eden *et al.*, 2009). In addition to this, functional information for subterranean-associated genes were mined from the Uniprot and RefSeq databases, and from literature cited directly (Consortium, 2007; Pruitt *et al.*, 2007). Enrichment analysis was performed using the hypergeometric test with the background set of genes restricted to genes that were tested for mole convergence and had at least one annotation in the corresponding annotation file. Correction for multiple testing was performed using false discovery rate q-values (Storey, 2002). We used two sources of annotations, the ”canonical pathways” from MSigDB (Liberzon *et al.*, 2011) and mammalian phenotypes from MGI (Smith and Eppig, 2009). The mammalian phenotype annotations were compiled by associating gene symbols listed in the genotype name to the reported phenotypes and all their ancestors in the mammalian phenotype ontology here.

2.2.4 Tissue-specific gene analysis

In order to determine how specific eye tissues have evolved across subterranean species we first identified tissue-specific gene sets using microarray expression data from 91 mouse tissues (Su *et al.*, 2004). We isolated tissue-specific genes for cornea, iris, lens and retina (including retinal pigmented epithelium). These sets were defined as those with significant differential expression only in the tissue of interest compared to all other tissues at an alpha of 0.05 (T-test).

2.2.5 Phylogenetic models of selective pressure

The top 200 subterranean-accelerated genes were subjected to phylogenetic models of codon evolution to test for significant evidence of relaxation of constraint or positive selection over the subterranean mammal branches. Using PAML, we ran codeml using 5 different models: the branch-site neutral model (BS Neutral), the branch-site selection model (BS Alt Mod), sites neutral model (M1), positive selection model (M8) and its null model (M8A) (Yang, 2007). To assess significance of relaxation of constraint on subterranean mammal branches we performed likelihood ratio tests (LRT) between BS Neutral and its nested null model M1. LRTs between BS Alt Mod and its null BS Neutral were used to infer positive selection on subterranean mammal branches. Probabilities were assigned for each of these 2 LRTs using the chi-square distribution with 1 degree of freedom. Mammal-wide positive selection was inferred using the M8 vs M8A models and their respective LRT, using 1 degree of freedom chi square distribution to assess LRT significance. For calculating the correlation between mole-acceleration and degree of tissue-specificity of genes, we estimate mole-acceleration of each gene as follows: using a branch-site selection model (BS Alt Mod) we estimate two different values of $\omega(dN/dS)$ one for the four subterranean branches and one for the rest of the branches on the tree. Mole-acceleration was calculated as the difference in the two ω values that were estimated.

2.3 Results

2.3.1 Many genes have altered evolutionary rates in subterranean mammals

We first sought to identify the genes that responded to conditions in the subterranean environment. Accordingly, we used relative evolutionary rate (RER) methods to identify protein-coding genes that evolved at a more rapid rate specifically on subterranean branches of the mammalian phylogenetic tree. Subterranean branches consisted of those leading to the star-nosed mole (Condylura cristata), the cape golden mole (Chrysochloris asiatica), the naked mole-rat (Heterocephalus glaber) and the blind mole-rat (Nannospalax galili). Each of these species represents a lineage that independently colonized the subterranean habitat, as each is more closely related to aboveground mammals than they are to each other (Figure 2.3.1A). Hence, similar phenotypic changes within these species are regarded as convergent traits. To demonstrate our RER methods, we first present the case of the eye-specific gene LIM2, which encodes Lens intrinsic membrane protein 2. First, the amount of amino acid divergence in LIM2 on each mammalian branch was quantified using sequences from 39 species and standard evolutionary models (Figure 2.3.1B) (see Materials and methods). The resulting LIM2 tree is markedly different from the genome-wide average tree in Figure 2.3.1aim1A, and reveals distinctly high amounts of divergence in LIM2 for the four subterranean species. This rapid divergence probably resulted from loss of selective constraint in the dark subterranean environment. To quantify this rate acceleration in the LIM2 tree, we normalized all branch lengths for the expected amount of change as defined by the genome-wide average divergence for each branch. This average, after scaling (see Materials and methods), should reflect both the underlying speciation times in the mammalian phylogeny as well as changes in demographic factors affecting substitution rates. The resulting RER values for each branch are plotted in Figure 2.3.1C. An RER of zero indicates that LIM2 evolved at exactly the expected rate on that branch, while positive and negative values reflect faster and slower rates, respectively. By examining RERs it becomes clear that LIM2 changed at abnormally rapid rates in the four subterranean mammals; the rates for all four subterranean species are more rapid than all aboveground species, and this difference is supported statistically (p=0.00084, MannWhitney U test). Thus, extending the RER calculations to all other genes, we can distinguish the functions that responded during adaptation to the subterranean environment. Importantly, the convergence of these species allows us to confidently infer genes that responded specifically to subterranean life, because faster rates in all four species are not likely to be due to random fluctuations, as reflected by the low P-value for LIM2.

We performed the same RER analysis on 18,980 protein-coding genes to determine which shifted to faster or slower evolutionary rates specifically in subterranean species. We will hereafter refer to such genes as mole-accelerated and mole-decelerated, respectively (see Materials and methods). At a false discovery rate (FDR) of 15%, we identified 55 mole-accelerated genes. We expect mole-accelerated genes to result from either selection for amino acid changes (i.e., positive Darwinian selection) or, alternatively, from a reduction in purifying selection, as suggested for the LIM2 protein. At the other extreme, we identified 1306 mole-decelerated genes at the same FDR. We expect genes to show rate deceleration if there is stronger purifying selection on that genes function in the subterranean environment, perhaps as the result of increased importance for fitness.

2.3.2 Vision-related functions are enriched among mole-accelerated genes

Genes with the strongest evidence of mole-acceleration were consistently associated with function in two organs, eye and skin. To illustrate, 17 of the top 30 mole-accelerated genes are expressed solely in eye tissues or are associated with eye-related disorders, whereas three accelerated genes are associated with skin, hair, and nails (Table 1). Among the genes showing very strong signals of mole-acceleration, we find proteins tha are specifically expressed in tissues of the eye such as the retina-specific proteins ROM1 and GNAT1 (Figure 2.3.2). The complete list of the 55 mole-accelerated genes similarly contains a large proportion that are related to vision and external tissues (Table S1 in Supplementary file 1), and they were highly enriched for functional annotations including eye morphology, photoreceptors, visual signal transduction, and eye-related mutant phenotypes (Table S2 in Supplementary file 1). The strength of this enrichment is clearly illustrated by examining all genes annotated to

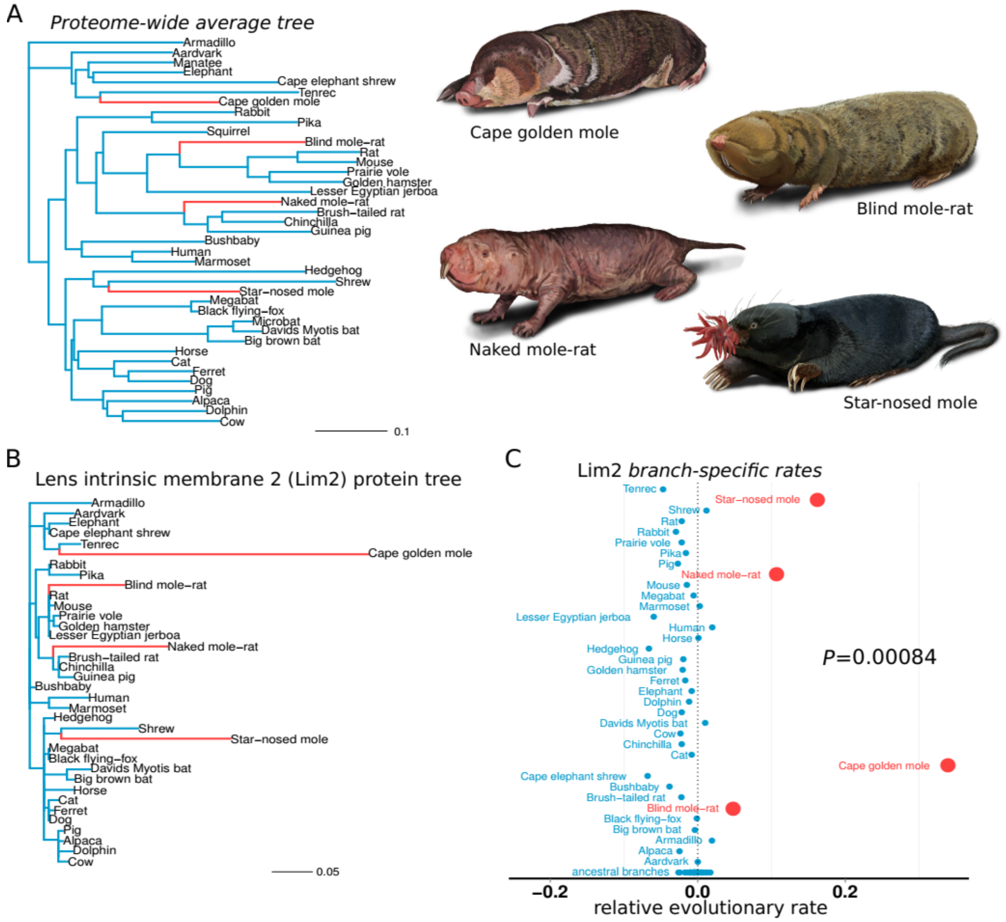


Figure 2.3.1: Lens intrinsic membrane protein 2 (LIM2) evolutionary rates across species. (A) Mammalian transitions to a subterranean environment occurred in four lineages shown in red. (B) LIM2 protein-coding sequence shows accelerated rates of evolution on subterranean branches. (C) Relative evolutionary rates of LIM2 showed the strongest acceleration on the subterranean branches amongst all of the genes studied. Illustrations by Michelle Leveille (Artifact Graphics)

the Gene Ontology (GO) term visual perception, because a large fraction of genes that have this annotation are ranked very highly in the list of mole-accelerated genes (Figure 2.3.3A subterranean). Furthermore, if we were to employ mole-acceleration as a sole predictor of visual function, a search would correctly identify many known visual perception genes with high accuracy, even when searching the entire genome (Figure 2.3.3B). This strong enrichment allows us to pose specific hypotheses in subsequent sections about which tissues and genetic pathways were altered during the regressive evolution of the eye.

We performed a control analysis to demonstrate that these functional enrichments are unique to subterranean species. We chose four aboveground species (Control species) for which there is no reason to expect phenotypic convergence and whose branch lengths are similar to the moles pika, guinea pig, squirrel and cow. Whereas mole-accelerated genes were enriched in 15 GO categories at a FDR of 15%,control-accelerated genes had no enriched categories at the same FDR (Table S2 in Supplementary file 1). Furthermore, these control species showed no enrichment of visual perception genes specifically (Figure 2.3.3). There were also 1306 mole-decelerated genes that evolved at significantly slower rates in subterranean species than in other mammals (Table S3 in Supplementary file 1). Although mole-decelerated genes are individually significant, only one GO category showed significant functional enrichment GO Biological Process: Nucleic acid binding transcription factor activity at an FDR of 15% (Table S4 in Supplementary file 1). A similar control analysis showed 626 genes as being significantly decelerated at an FDR of 15%, and these control-decelerated genes were enriched in 24 GO categories. Therefore, despite there being vastly more mole-decelerated genes than mole-accelerated genes, mole-decelerated genes as a group do not show strong functional enrichment. This result stands in stark contrast to the strong enrichment seen in the mole-accelerated genes.

2.3.3 Most mole-accelerated genes are under relaxed constraint

Accelerated rates could have resulted from adaptive evolution or, alternatively, from relaxation of constraint. We distinguished between these scenarios using codon-based evolutionary models to detect signatures of adaptive evolution. We tested whether the nonsynonymous to synonymous rate ratio (dN/dS) was significantly greater than 1 the expectation for positive selection for any portion of the gene specifically on the subterranean species branches, and also more generally across the entire mammalian phylogeny (Yang, 2007). Of the top 55 mole-accelerated genes, only one gene rejected a neutral model not allowing

Gene	P-value	Tissues	Description
LIM2*	0.00084	Lens	Lens intrinsic membrane protein 2
CRYBB3*	0.00087	Lens	Lens-specific crystallin, beta B3
R0M1*	0.00096	Retina	Retinal outer segment membrane protein 1
CRYBA1*	0.00098	Lens	Lens-specific crystallin, beta Al
CRYGC*	0.00119	Lens	Lens-specific crystallin, gamma C
CRYBB2*	0.00128	Lens	Lens-specific crystallin, beta B2
GPR89B	0.0013	Ubiquitous	G-protein-coupled receptor 89B, pH mediator in Golgi
GNAT1*	0.00133	Retina	Rod cell-specific G-protein, subunit alpha
GPRS9A	0.00134	Ubiquitous	G-protein-coupled receptor 89A, pH mediator in Golgi
NRL*	0.00138	Retina	Neural retina leucine zipper responsible for expression of rhodopsin
CRYGS*	0.00146	Lens	Lens-specific crystallin, gamma S
GRM6*	0.0015	Retina	Metabotropic glutamate receptor 6, required for normal vision
GBX2	0.00165	Embryo	Gastrulation brain homeobox 2, developmental transcription factor
LGSN*	0.00171	Lens	Lengsin, lens protein with glutamine synthetase domain
CRYBB1*	0.00183	Lens	Lens-specific crystallin, beta Bl
KLHDC3	0.00186	Ubiquitous	Kelch-domain-containing 3, high expression in brain
KRT81#	0.00186	Hair	and nails Keratin 81, primarily in hair cortex
WDFY1	0.00192	Ubiquitous	WD repeat and FYVE-domain-containing 1, endosomal protein
KRT9#	0.00195	Skin	Keratin 9, specific to palms of hands and soles of feet
POMP#	0.00199	Ubiquitous	Proteasome maturation protein, associated with rare skin disorder
RRH*	0.00201	Retina	Retinal pigment epithelium-derived rhodopsin homolog
DPCD*	0.00201	Ciliated	cells Deleted in primary ciliary dyskinesia; maintenance of ciliated cells
RAD54L	0.00217	Ubiquitous	RAD54-like: DNA double-strand break repair
TATDN1	0.00235	Ubiquitous	TatD DNase-domain-containing 1
ITLN2	0.00244	Small	intestine Intelectin 2, may play a role in defense against pathogens
STX3*	0.00245	Ubiquitous	Syntaxin 3, associated with congenital cataracts and intellectual disability
SKJV2L*	0.00254	Ubiquitous	DEAD box protein, yeast SKI2 homolog, implicated in macular degeneration
DPY19L1	0.00254	Ubiquitous	dpy-19-like 1 (Caenorhabditis elegans), probable C-mannosyltransferase
TFPT	0.00266	Ubiquitous	TCF3 (E2A) fusion partner (in childhood leukemia)
RSI*	0.00275	Retina	Retinoschisin 1, extracellular protein involved in organization of retina

*related to vision. #related to skin and hair.

Table 1: Top 30 subterranean-accelerated genes.

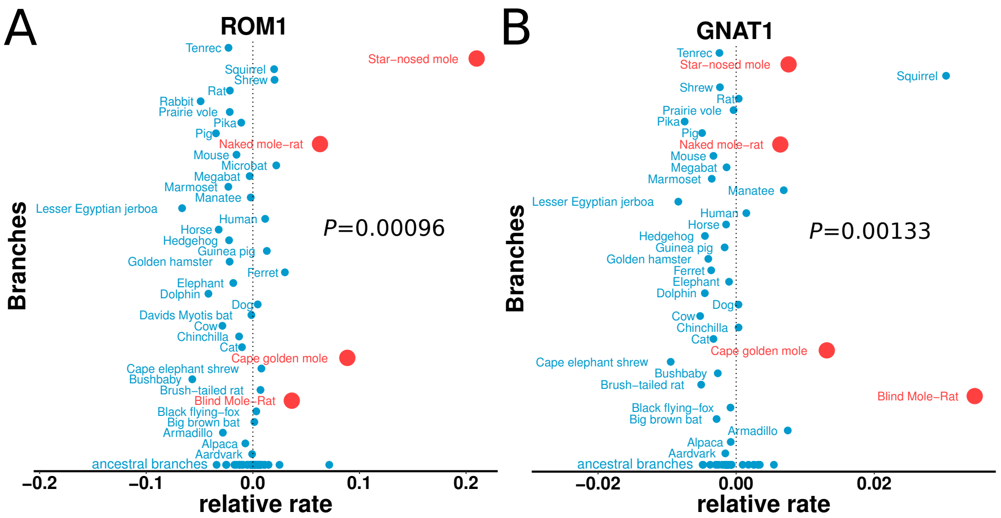


Figure 2.3.2: Relative evolutionary rates of two retinal proteins across species. Relative evolutionary rates of two retinal proteins, (A) Retinal outer segment membrane protein 1 (ROM1) and (B) Rod cell-specific G protein, subunit alpha (GNAT1), show strong acceleration in the subterranean mammals (marked in red).

dN/dS ratios exceeding 1 in favor of a model allowing positive selection ($dN/dS > 1$) on subterranean branches (Table S5 in Supplementary file 1). This gene is involved in connective tissue and hair structure (KRTAP17-1). The other accelerated genes did not show evidence of adaptive evolution and thus are probably under relaxed constraint. Almost all accelerated genes rejected a model requiring them to have identical constraints in all mammals (model M1) in favor of a model that allowed subterranean-specific relaxation of constraint (model BS1) (Table S5 in Supplementary file 1). Some of these genes seem to have lost all functional constraint because they show genetic lesions such as stop codons and frameshifts in some subterranean species (Table S6 in Supplementary file 1). This evidence of relaxed constraint is consistent with the expectation that some vision-related genes have been undergoing regressive evolution.

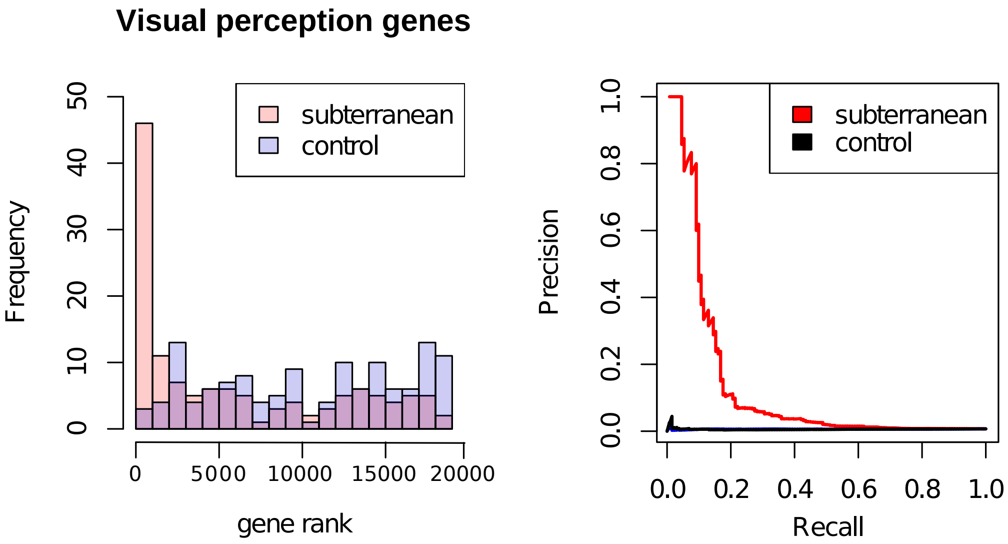


Figure 2.3.3: Enrichment of visual perception genes. (A) Histogram of the rankings of 189 visual perception genes based on their mole-acceleration. (B) Mole-acceleration can equivalently serve as a predictor for function in visual perception. The plot shows the Precision-Recall values at varying p-value thresholds reflecting the fraction of visual perception genes significant at a particular threshold (Precision) and the fraction of visual perception genes retrieved at the same threshold (Recall)

2.3.4 Skin-related genes were accelerated possibly in response to the demands of tunneling

The fossorial lifestyle of subterranean species has selected for traits related to digging and locomotion underground (Nevo, 1979). Perhaps because of this selective pressure, many of the top moleaccelerated genes encode proteins that are structural components of skin, hair and epithelial connective tissues. The reasons for their acceleration are the result of relaxation of constraint on their coding sequence. Genes encoding keratin proteins 9, 12, and 81 (KRT9, KRT12, KRT81) were studied using codon models and the results indicated that they experienced relaxed constraint in subterranean species but not positive selection for amino acid diversification (Table S5 in Supplementary file 1). They contain early stop codons in multiple subterranean species, which is consistent with complete loss of constraint (Table S6 in Supplementary file 1).

The convergent acceleration and pseudogenization of KRT9 is particularly interesting in relation to burrowing (Figure 2.3.4). In mice, KRT9 expression is confined to footpads, and Krt9-/- null mutants develop footpad calluses due to hyperproliferation of skin (Fu *et al.*, 2014). In humans, Keratin 9 is expressed solely on the palms of hands and soles of feet, and mutations lead to a skin disorder characterized by hyperkeratosis (thickening) of the surfaces of palms and soles epidermolytic palmoplantar keratoderma (Hennies *et al.*, 1995). By extension, the loss of KRT9 in subterranean species may also have led to hyperproliferation of footpads, which could carry benefits for tunneling. For example, the star-nosed mole digs with its forepaws, and naked mole-rats collect and remove dirt with their feet (Hamilton, 1931; Jarvis *et al.*, 2014). Such abrasive tasks could place high demands on the footpad surfaces. In addition, mole-acceleration of the POMP gene could similarly have resulted from demands on footpads. A human mutation in POMP is associated with KLICK syndrome, a skin disorder characterized again by hyperproliferation and thickening of palms and footpads (Dahlqvist *et al.*, 2010).

There were also skin- and hair-related genes that were identified outside the mole-accelerated genes discovered at a FDR of 15%, showing increased positive selection in subterranean species, rather than loss of function. One such gene, COL4A4, a gene encoding a subunit of Type IV collagen was strongly accelerated, did not contain genetic lesions, and showed evidence of positive selection in subterranean species (Table S1, S5, S6 in Supplementary file 1). Type IV collagen is the major structural component of the basal lamina in many tissues including skin epithelium, and is composed of 6 subunits, three of which were notably mole-accelerated (COL4A4, COL4A5, COL4A3). On average the 6 subunits were more accelerated than 71% of all other genes, which is a significant difference (P = 0.0342, Mann-Whitney U test). While Type IV Collagen seems to have responded to the subterranean environment, other major components of the basal lamina, the laminin proteins (e.g., LAMA1), were not notably accelerated.

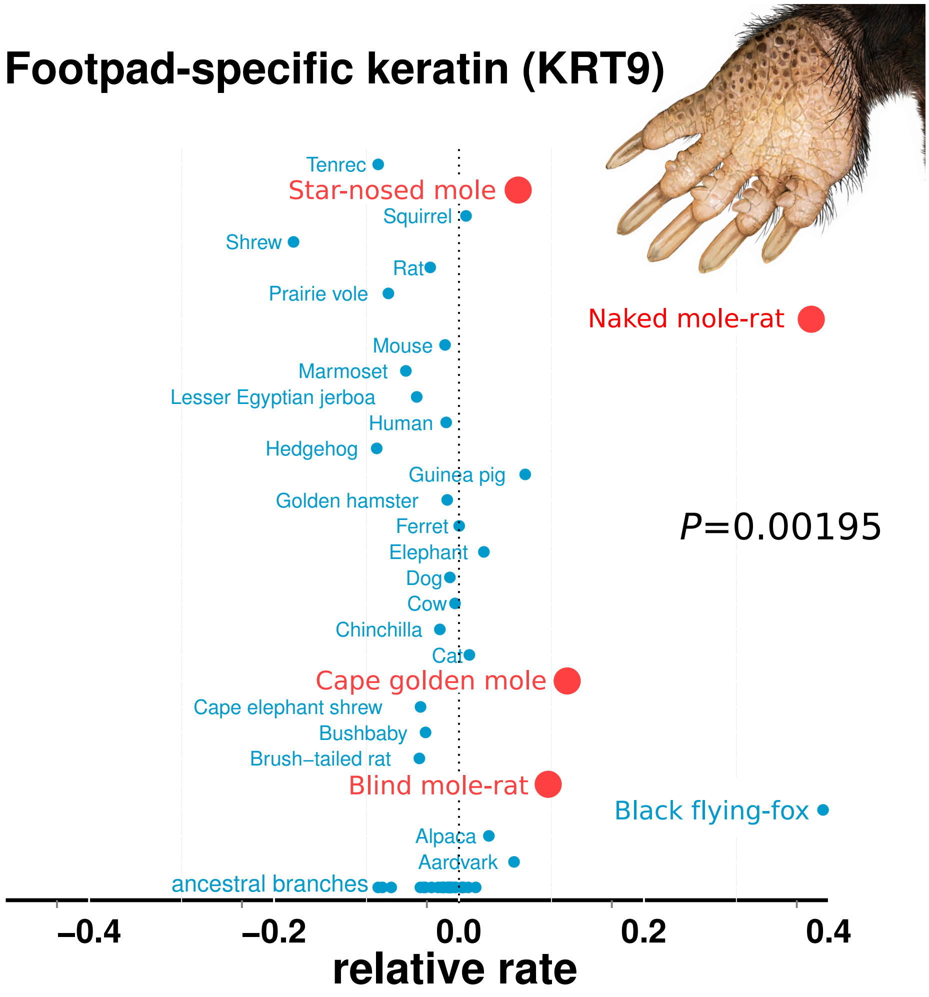


Figure 2.3.4: Relative rates of footpad-specific keratin 9 (KRT9). KRT9 shows strong acceleration on the subterranean branches. The image shown is the footpad of the star-nosed mole, showing characteristic hyperkeratosis. Keratin 9 mutations also lead to hyperkeratosis in mouse models and humans. Illustrations by Michelle Leveille (Artifact Graphics).

2.3.5 Regressive evolution is limited to lens, retina, and eye-specific developmental genes

In order to compare how specific eye tissues have evolved in subterranean species we first compiled tissue-specific gene sets using expression data from 91 mouse tissues (Su *et al.*, 2004). We identified tissue-specific genes for cornea, iris, lens and retina by selecting those genes with significant differential expression in the tissue of interest but not in other tissues. Using literature we also compiled a set of 71 important eye developmental genes (Table S7 in Supplementary file 1). We first asked if there is a relationship between the degree of tissue-specificity and the degree of mole-acceleration measured as the difference in dN/dS between subterranean and aboveground species (Figure 2.3.5A). We found a clear positive correlation between eye tissue-specificity and mole-acceleration, which is consistent with a greater relaxation of constraint on genes with few or no roles outside the eye. Next, we asked which genes with eyetissue specific expression showed acceleration and found that genes specifically expressed in cornea a protective tissue of the outer eye and the iris were not accelerated in subterranean species when compared to a set of randomly chosen genes (background) (Figure 2.3.5B, C). In contrast, many lens- and retina-specific genes are accelerated. On average, lens genes are more accelerated than 84% of background genes, and retina genes are more accelerated than 82% (P = 9.07e-06 and P = 6.10e-10 for lens and retina respectively, Mann-Whitney U test). The contrast between the front and the interior of the eye suggests that the sensory functions of the inner eye, such as phototransduction and the visual cycle, are under relaxed constraint, while the protective function of the cornea is not. Indeed, two of these subterranean species have eyes that are open to the environment, such that the cornea may continue to serve as a barrier to pathogens and debris. Eye developmental genes as a whole were not accelerated compared to background, which may reflect the fact that most of them, such as Sonic Hedgehog (Shh), are important in the development of non-eye tissues. However, five eye-specific developmental genes were notably present at the top of the accelerated list (VAX2, NRL, FOXE3, CRX, ALDH3A1), while no eye-specific genes were found lower in the list (Table S7 in Supplementary file 1). This is consistent with the positive relationship between eye-specificity and relaxation of constraint (Figure 2.3.5A).

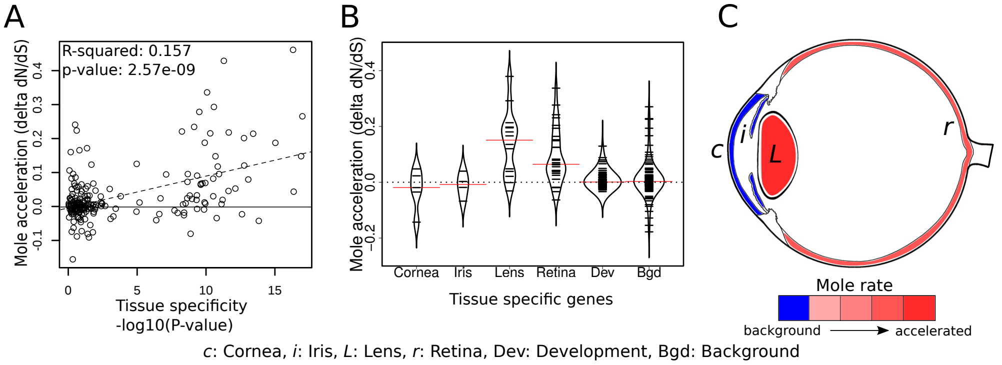


Figure 2.3.5: Tissue-specific retinal and lens genes are highly accelerated in subterranean species. (A) Ocular genes that are more tissue-specific exhibit stronger acceleration in subterranean mole species. (B) Panels of tissue-specific genes were tested for their relative accelerations in the subterranean mammals. Retina- and lens-specific genes show many cases of acceleration in the subterranean environment. (C) Representation of average mole-acceleration for genes specifically expressed in four different tissues of the eye.

2.3.6 Eye-specific enhancers of PAX6 show convergent acceleration in subterranean mammals

Although we observe specific instances wherein eye developmental genes show accelerated rates in subterranean mammals, there is no significant global trend. This is understandable given that a majority of these transcription factors have important roles in the development of non-eye related tissues. For example, PAX6 is important in the development of pancreas and brain, in addition to the eye (Kammandel *et al.*, 1999; Xu *et al.*, 1999; Kleinjan *et al.*, 2006). Hence the protein-coding sequences of these transcription factors experience selective pressure against deleterious mutations. However, regulatory regions controlling the expression of these developmental genes in the eye might be under relaxed constraint in the subterranean mammals given the relaxation on maintaining the functionality of visual pathways. We hypothesize that these eye-specific cis-regulatory elements (CREs) would thus show accelerated rates of evolution in the subterranean mammals.

We tested this hypothesis by applying our evolutionary-based method toward identifying eye-specific regulatory elements controlling the expression of the developmental transcription factor PAX6. We chose the PAX6 system as extensive effort has gone into characterizing the spatiotemporal regulation of its expression (Kammandel *et al.*, 1999; Xu *et al.*, 1999; Dimanlig *et al.*, 2001; Kleinjan, 2001; Kleinjan *et al.*, 2006; Griffin *et al.*, 2002), and there exists comprehensive annotation of cis-regulatory elements (cre) controlling the expression of PAX6 in various tissues including the eye. Based on existing literature on transcriptional regulation of PAX6 expression, we identified a 500kb window containing Pax6 and its neighboring gene Elp4 as our genomic window of interest (Kleinjan *et al.*, 2006). Experiments involving transgenic mice revealed various tissue-specific enhancers in a 200-kb region within this genomic window to be important for PAX6 expression. We subsequently identified 150 highly conserved non-coding elements in this genomic window and estimated their evolutionary rates on each mammalian branch. We subsequently calculated the relative rates of the branches using the same projection operator method as was employed for the protein-coding gene trees. We then employed the Mann-Whitney U hypothesis-testing framework to identify non-coding elements evolving at an accelerated rate specifically on the subterranean branches (Methods).

The results of our analyses show that the 3 regions showing the strongest signals of convergent acceleration in the subterranean mammals highly overlap the regions previously annotated to be enhancers important for regulation in eye-specific tissues (Figure 2.3.6A, B) i. cre149 is a 558 basepair (bp) region containing the 530-bp region annotated as the alpha, intron 4 retinal enhancer (Kammandel *et al.*, 1999). ii. cre21 is a 552-bp region located within the fragment containing HS2 and HS3 of the Distal Regulatory Region, a retina-specific enhancer of PAX6 (Kleinjan, 2001). iii. cre86 is a 429-bp region containing the 341-bp long ectodermal enhancer, which has been shown to be important in driving the expression of PAX6 in developing lens (Dimanlig *et al.*, 2001). Regions overlapping an enhancer element shown to be regulating PAX6 expression in lens, hindbrain and diencephalon (the EI enhancer element) do not show significant rate acceleration in the moles (Kleinjan, 2001). This is in concordance with our expectation that only eye-specific elements show convergent acceleration, and hence the regions overlapping the EI enhancer do not show acceleration given their importance for PAX6 expression in non-eye tissues. Similarly, a 120-bp region overlapping the pancreas enhancer also does not show significant rate acceleration in the moles, as expected (Xu *et al.*, 1999; Kleinjan *et al.*, 2006). In addition to the eye-specific enhancer elements, we observe other regions showing comparable rate acceleration in the moles that are not yet characterized (Table S8 in Supplementary file 1). These regions are thus candidate cis-regulatory elements for PAX6 expression in the eye. This preliminary study of the PAX6 transcriptional regulatory module serves to confirm our hypothesis that eye-specific regulatory elements are under relaxed constraint and thus show accelerated rates of evolution in the subterranean mammals.

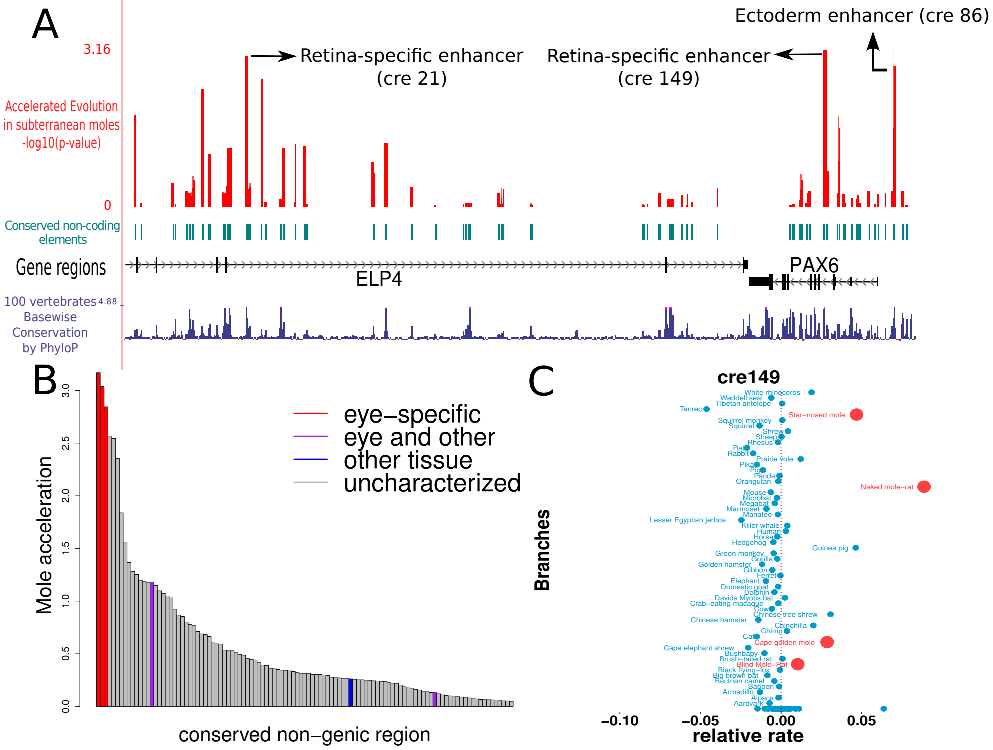


Figure 2.3.6: Mole-acceleration of eye-specific enhancers in the Pax6 gene region. (A) Genomic region spanning Pax6 and its neighbor Elp4. The three most accelerated non-coding regions identified in this analysis are consistent with the eyespecific enhancers regulating Pax6 expression in the eye. (B) The mole-acceleration scores for the three eye-specific enhancers of Pax6 are the highest among 150 regions analyzed. (C) The relative rates in each species for the most accelerated region cre149.

2.3.7 Mole-accelerated non-coding elements are strongly enriched near transcription factors driving eye development

Expanding from our analysis of PAX6, we perform a large-scale scan for convergently accelerated non-coding elements near transcription factors in the mammalian genome. We compiled two sets of transcription factors one comprising 20 genes known to be important in eye development (Eye set), such as PAX6, PAX2, OTX2 etc., and another set consisting of an equal number of tissue-specific transcription factors expressed in other tissues and with no evidence of expression in eye (Other set) such as HOXA9, PAX8, SOX13 (Table S9 in Supplementary file 1). We identified 200 conserved non-coding elements near each gene in both sets totaling to 8,000 elements split equally between the two gene sets (Figure 2.3.7A). We subsequently applied our method and calculated the mole-acceleration of each element. This large-scale scan revealed a total of 17 elements as convergently accelerated at a FDR of 10% (Figure in 2.3.7A). Fourteen of the seventeen elements are found nearby genes belonging to the Eye set, reflecting a significant enrichment of mole accelerated elements near transcription factors driving eye development (Hypergeometric test, p-value = 0.001). We subsequently checked the genomic locations of these mole-accelerated elements to ensure that they are not clustered at the same locus for instance. These seventeen elements are found nearby 14 unique genes, with 11 unique genes belonging to the Eye set, and 3 genes belonging to the Other set, further showcasing the strong enrichment of unique eye developmental transcription factors found near mole-accelerated elements (Hypergeometric test p-value = 0.0016).

2.3.8 FANTOM5 eye enhancers show strong convergent acceleration in subterranean mammals

The FANTOM5 consortium has identified putative enhancer sites in the human and mouse genome based on bidirectional enhancer transcription across tissues as well at multiple developmental time points (Andersson *et al.*, 2014). These putative enhancer sites include genomic regions transcribed in the eyeball of mouse embryo at four developmental time points. Based on this resource, we compiled two sets of FANTOM5 enhancer sites a set consisting of 900 genomic regions with non-zero expression in the eyeball across four developmental time points (Eye enhancers), and another set consisting of 6,000 regions with zero expression across the same samples (Other enhancers). We subsequently calculated the convergent rate acceleration of these genomic elements in the four subterranean mammals and compared the acceleration observed of the Eye enhancers to that of the Other enhancers. Our analysis revealed a strong enrichment of FANTOM5 Eye enhancers showing convergent rate acceleration in the four subterranean species, in comparison to the four control species (Figure 2.3.7B). We observe 62 FANTOM5 enhancers in total showing significant mole acceleration at a FDR of 15% (Table S10 in Supplementary file 1). Fifteen of these correspond to FANTOM5 Eye enhancers set, reflecting a significant enrichment of detecting FANTOM5 eye enhancers using mole-acceleration (Hypergeometric test p-value = 0.006).

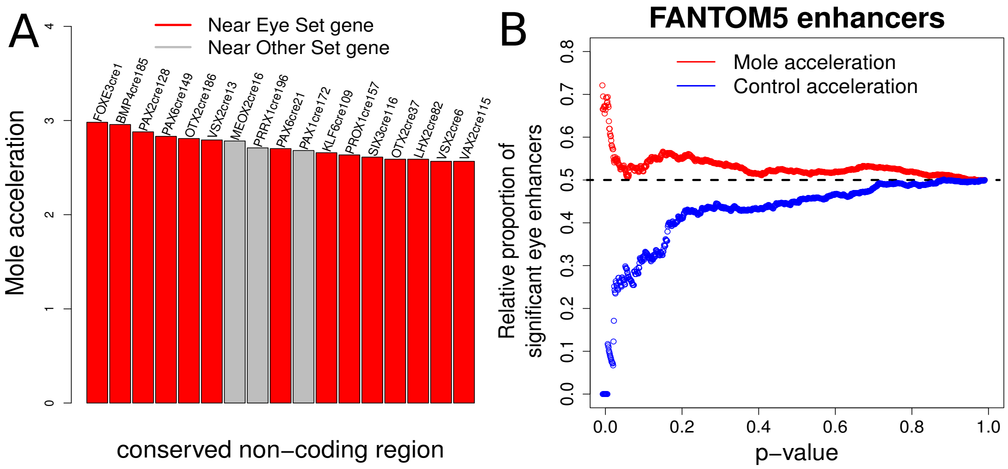


Figure 2.3.7: Evidence of mole-acceleration in candidate eye-specific enhancers. (A) Enrichment of mole-accelerated elements near eye developmental transcription factor genes. The bar plot shows the 17 mole-accelerated conserved non-coding elements identified. (B) FANTOM5 Eye enhancers show strong mole-acceleration. The plot shows the relative proportion of FANTOM5 eye enhancers identified among all enhancers significant at the corresponding p-value threshold.

2.3.9 Some aboveground species exhibit gene acceleration indicative of their altered visual capacities

To systematically understand differences in visual capabilities of mammals, we studied the overall relative rates of evolution of visual genes across all mammals. Our gene set of interest (189 in total) comprised of all genes with Visual perception GO term annotation, excluding developmental transcription factors. For each species, we then calculated the mean relative rate across all the genes (Figure 2.3.8). We observed the four subterranean mammals to be among the accelerated species (with $mean > 0$), as was our expectation. However, we additionally observed aboveground species with overall rate accelerations comparable to the moles, such as the armadillo, thirteen-lined ground squirrel, big brown bat, Davids myotis bat and shrew. Notably all of these aforementioned mammals show varying types of visual regression the armadillo has poor vision characterized by a lack of cone cells in their retina (McDonough and Loughry, 2013), and shrews also have poor vision and diminutive eyes, which in some species are hidden in fur (Nowak, 1999). The nocturnal big brown bat and Davids myotis bat possess reduced eyes and rely on echolocation for navigation (Koay *et al.*, 1998). The thirteen-lined ground squirrel, for which we observe a rod cell-specific acceleration, displays a rare visual trait the central region of its retina is dominated by cone photoreceptors in contrast to most mammals (Brooks *et al.*, 2016). These scenarios could have important implications because the ground squirrel is used as a model for vision research (Li *et al.*, 2010; Chen and Li, 2012).

2.4 Discussion

The independent transitions of four mammals to a subterranean environment is accompanied by convergent phenotypic changes as a result of adaptation to new environmental stresses in the underground ecotope. Here, we report a genome-wide effort encompassing both coding and regulatory regions to identify the changes in genotype accompanying phenotypic adaptation by studying changes in their evolutionary rates. Our study reveals

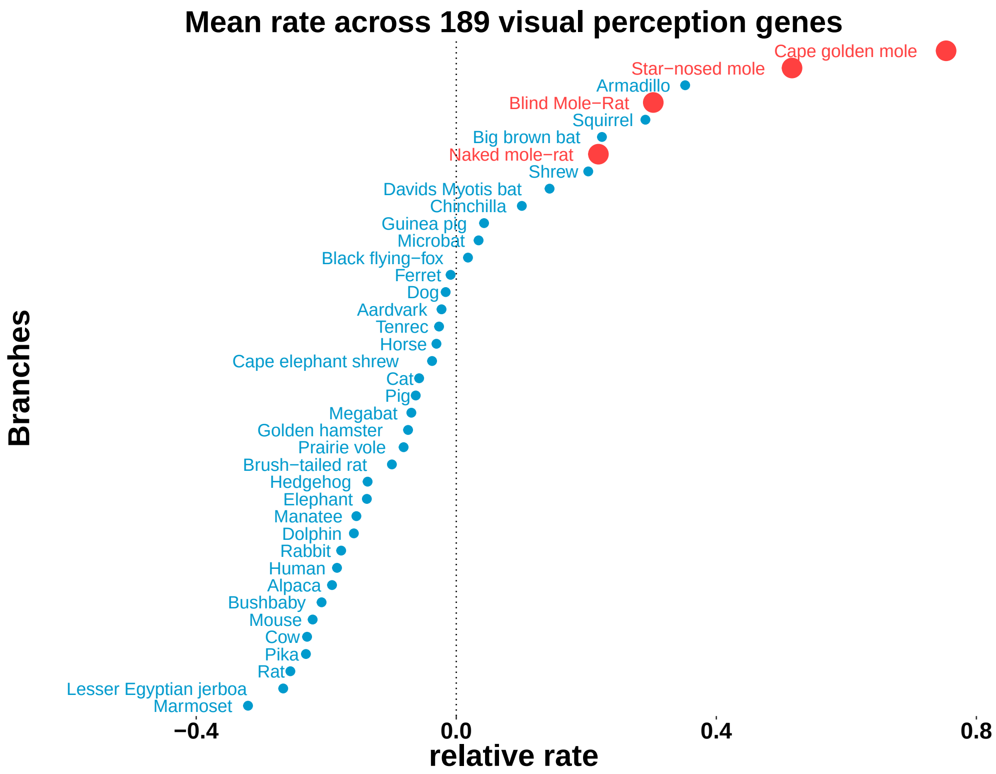


Figure 2.3.8: Some aboveground species show accelerated rates of evolutionary change in visual perception genes. On the basis of the relative evolutionary rates across all species for 189 genes with the GO term annotation visual perception, we calculated the species-wise mean relative rate across of the genes.

that genes showing convergent acceleration in subterranean species are highly enriched for function in visual pathways. The decreased selective pressure on visual pathways in the dim-light subterranean environment leads to a relaxation of constraint on genetic elements involved in various eye-related phenotypes including eye morphology, photoreception, visual transduction etc. In addition to genes in visual pathways, we observe many genes involved in skin-related phenotypes to have an accelerated rate of evolution in the subterranean mammals. While we see acceleration in visual genes primarily as a result of relaxation of constraint, we see that some skin-related genes also show acceleration due to positive selection, perhaps as a result of selection of traits contributing to a fossorial lifestyle. Aside from these two phenotypes, we do not observe a comparably strong enrichment for genes involved in the other environmental challenges associated with a subterranean lifestyle, such as hypoxia, hypercapnia and high infectivity. It is possible that the subterranean mammals may show species-specific adaptations to these stresses, whereas our analysis from a convergent evolutionary perspective reflects changes common to all the species.

Closer examination of the accelerated genes enriched for vision-related pathways reveals that accelerated genes tend to be lens- or retina-specific. On the other hand, genes encoding specifically for the outer ocular structure, the cornea, do not show significant acceleration, indicating preservation of developmental programs important for ocular architecture. In two of the four moles with non-subcutaneous eyes, the cornea can come into direct contact with external environment, perhaps necessitating the proper development of the structure in the highly infective subterranean niche. Lens- and retina-specific genes involved with the processes of photoreception and phototransduction would be under greater relaxed constraint given the dim-light environment, accruing damaging mutations at a much higher rate. Our analyses also reveal genes associated with congenital eye diseases to be accelerated in the four subterranean mammals. For the lens, which is largely made up of crystallins, we find many crystallin genes (CRYBB3, CRYBA1, CRYBB1, CRYGC, CRYGS, etc.) in our accelerated set of genes contributing to various forms of cataracts (Graw, 2009). Similarly, we find multiple genes involved in ciliopathies to be accelerated including deleted in primary ciliary dyskinesia (DPCD), IQCB1 a component of primary cilia, and ciliary neurotrophic factor (CNTF). Further inspection of the accelerated list of genes could potentially reveal new candidate genes important for congenital eye diseases.

Genes involved in the embryonic development of eye do not show significant global acceleration, potentially due to their pleiotropic nature; these developmental transcription factors tend to have important regulatory roles in non-eye related pathways that are not under relaxed constraint. However, we hypothesize that eye-specific regulatory elements of these developmental genes are under relaxed constraint in the moles. We developed a novel variant of our evolutionary rates based approach to study the convergent acceleration at the non-coding level, and successfully proved our hypothesis. Although the strong rate acceleration in the three eye-specific enhancers of PAX6 suggests relaxation of constraint in the subterranean mammals, in the absence of functional tests we cannot be sure that the eye-specific activity is truly lost. Furthermore, we found an enrichment of such convergently accelerated non-coding regions preferentially near eye developmental transcription factors, identifying potential enhancer elements driving the expression of these genes specifically in the eye. As a large-scale validation approach we show that rate acceleration in subterranean mammals strongly overlaps regions identified as eye enhancers by the FANTOM5 consortium. These proof-of-principle analyses serves to illustrate the power of convergent evolution-based tools for the identification of eye-specific regulatory elements. Despite the apparent rapid rate of enhancer evolution across mammals, our methods and those of colleagues showcase the utility of applying evolution-based approaches to conserved non-coding regions in identifying regulatory elements underlying important developmental functions (Villar *et al.*, 2015; Marcovitz *et al.*, 2016). These methods present a unique opportunity to perform genome-wide scans for eye- and other tissue-specific regulatory elements, and potentially serve as complementary approaches to genome-wide assays in the identification of active enhancer elements in the genome. As more genomes are sequenced, we expect these methods to become more powerful in revealing gene regulatory changes underlying convergent phenotypes.

Overall, our results suggest that genes and non-coding regions involved in vision pathways are accumulating deleterious mutations by neutral processes, given the relaxation of constraint on these pathways in the subterranean environment. However, this does not preclude the possibility that the initial inactivating mutations in these pathways were adaptive in nature. The initial shutdown of eye development may have been caused by positively selected changes, followed by continued regression of structural and physiological eye genes through neutral processes. Indeed, there is evidence of such a progression of events during eye regression in blind cavefish (Jeffery, 2005). Adaptive forces for reduced eyes may have been driven by the energetic costs of maintaining functioning eyes and the risk of pathogen entry through the eye (Moran *et al.*, 2015). We note that our rate-based analysis detects signatures of sequence divergence based on what is observed at the end of these processes and does not shed light on the nature of the initial inactivating changes. Additionally, our methods detect convergent changes in the rates of evolution of genes and hence are not designed to detect species-specific changes that might contribute to the subterranean adaptation.

Our results showcasing convergent acceleration in rates of visual genes strongly supports previous reports of visual regression in the subterranean habitat. Emerling and Springer studied the regression of retinal genes in three of these four subterranean species and showed that a decrease in the amount of light entering the retina is associated with higher incidence of inactivating mutations in retinal genes (Emerling and Springer, 2014). They found a significantly higher number of retinal pseudogenes in the moles compared to closely related subaerial species, an observation concordant with our results based on rate acceleration. Genome sequencing efforts of naked mole-rat and blind mole-rat also showed a strong enrichment of pseudogenes in visual pathways associated with degradation of vision in these species (Cooper *et al.*, 1993; Kim *et al.*, 2011). A genome-wide study by Prudent *et al.* (2016) detected significant genomic differences in genes involved in vision-related pathways such as eye development and perception of light in two of these four subterranean mammals, namely cape golden mole and blind mole-rat. Using our rates-based framework we perform a more rigorous investigation of convergently evolving genes in a larger set of four subterranean species, as well as elucidating the tissue-specificity and underlying reasons for their convergent rate changes. More importantly, in a first-of-its-kind demonstration at the non-coding level, we applied our methods to successfully detect eye-specific enhancers.

Visual regression is not limited to these four mole species, and mammals display specific types of regression and other general differences in visual capabilities. Our analysis of visual gene rates across other species revealed interesting patterns and trends, wherein some aboveground species with poor or remodeled visual systems showed mean rate acceleration comparable to subterranean mammals (Figure 2.3.8). This provides an opportunity to further probe specific differences in the development and function of visual systems in terms of the specific pathways that are relaxed or under constraint across species. Additionally, integrating these other species into our rate-based framework can help in fine-tuning the predictive power of the evolutionary-based approaches. Deliberate selection of foreground branches based on specific combinations among these vision-impaired mammals might greatly improve the power of the methods in detecting convergent changes, especially at the non-coding level. In this regard, the availability of rich and diverse phenotypic annotations across mammals further lays the ground for the development of evolutionary-based approaches in functional and phenotypic annotation of non-coding regions (O'Leary and Kaufman, 2011; O'Leary *et al.*, 2013; Marcovitz *et al.*, 2016).

3.0 An improved method for robust estimation of relative evolutionary rates

3.1 Introduction

Understanding the relationship between phenotype and genotype is a fundamental question in biological research. A mechanistic characterization of this relationship hinges on our ability to define how specific genetic elements contribute to biological processes at the molecular, cellular, and organismal level. High-throughput sequencing has enabled new experimental approaches that have uncovered a wealth of genetic elements with putative regulatory roles across tissues (Dunham *et al.*, 2012; Andersson *et al.*, 2014; Romanoski *et al.*, 2015). However, identifying the precise biological functions of these elements remains a challenge. Even beyond non-coding elements, the precise biological roles of many protein-coding genes are still poorly understood, and many genes with statistical disease associations still lack a mechanistic explanation (Pennacchio *et al.*, 2013; Radivojac *et al.*, 2013; Sánchez and Huarte, 2015; Shlyueva *et al.*, 2014). While experimental validation for functional annotation remains challenging, there is considerable interest in developing new tools that can use existing data resources to further elucidate the function of genetic elements. These approaches have the potential to improve the diagnosis of disease susceptibility and the development of therapeutic interventions (Manolio *et al.*, 2009; Esteller, 2011).

Computational approaches learning from patterns of convergent phenotypic evolution across species provide a complementary approach to predict genotype-phenotype associations. The natural world is replete with examples of phenotypic convergence ranging from the independent evolution of flight in birds and mammals to diving in species that transitioned from a terrestrial to marine habitat to loss of complex phenotypes such as eyesight in animals colonizing the subterranean niche. Genome-scale studies aimed at identifying the genetic basis of phenotypic convergence take advantage of the growing availability of whole genome sequences for species across several orders, alongside the development of comparative methods to predict orthologous sequences (Eisen, 1998; Pellegrini *et al.*, 1999; Li *et al.*, 2014). A common approach in such studies is to identify convergence at the molecular level, including substitutions at specific nucleotide or amino acid sites (Zhang and Kumar, 1997; Parker *et al.*, 2013; Stern, 2013; Foote *et al.*, 2015; Thomas and Hahn, 2015; Zou and Zhang, 2015). An alternative strategy to investigate the genetic basis of convergence is to search for convergent changes at the level of larger functional regions rather than specific nucleotide or amino acid sites. Sets of genes associated with a phenotype can respond to convergent changes in the selective pressure on the phenotype through non-identical changes in the same gene, and as such, sites-based methods can fail to detect them. These limitations have encouraged researchers to search for convergent shifts in evolutionary rates of individual protein-coding genes and more recently conserved non-coding elements (Lartillot and Poujol, 2011; Hiller *et al.*, 2012; Chikina *et al.*, 2016; Marcovitz *et al.*, 2016; Prudent *et al.*, 2016). An increased selective constraint can manifest as a slower evolutionary rate, whereas faster evolutionary rates can result from a release of constraint or from adaptation. Thus phenotypic associations for genetic elements can be predicted from correlated changes in their evolutionary rates on phylogenetic branches corresponding to the phenotypic change. Example approaches based on evolutionary rates include the Forward/Reverse Genomics methods that have identified protein-coding and non-coding genetic elements showing convergent regression in subterranean mammals and loss of limb-regulatory elements in snake lineages (Hiller *et al.*, 2012; Marcovitz *et al.*, 2016; Prudent *et al.*, 2016; Roscito *et al.*, 2018).

We previously developed an evolutionary-rates-based method to identify genetic elements showing convergent shifts in evolutionary rates associated with two distinct phenotypic transitions (Chikina *et al.*, 2016; Partha *et al.*, 2017). Our original method calculates gene-specific evolutionary rates using a linear model, and gene-trait associations are inferred using correlations of these rates with the phenotype of interest. A genome-wide scan using this method to find protein-coding genes associated with the transition to the marine environment identified hundreds of genes that showed accelerated evolutionary rates on three marine mammal lineages (Chikina *et al.*, 2016). These accelerated genes were significantly enriched for functional roles in pathways important for the marine adaptation including muscle physiology, sensory systems and lipid metabolism. More recently, using our methods we detected an excess of vision-specific genes as well as enhancers that showed convergent rate acceleration on the branches corresponding to four subterranean mammals (Partha *et al.*, 2017). Genes showing convergent rate shifts associated with these two phenotypic transitions typically follow one of the following modes of change in the selective pressure 1. relaxation of constraint, 2. positive selection. Marine-accelerated and subterranean-accelerated genes identified in earlier scans were further probed using phylogenetic models of selective pressure to identify the underlying evolutionary process. In both cases, we found an excess of genes under relaxed constraint, as well as a smaller number of genes under positive selection. Overall, genome-scale efforts both from our group and others to find genetic elements responding to convergent changes in the selective pressures in their environment are gaining momentum in accurately describing precise genotype-phenotype associations.

Our original evolutionary-rates method has an important statistical limitation, namely strong mean-variance trends in the computed evolutionary rates. The distributions of branch lengths of gene trees in phylogenetic datasets are influenced by the choice of species, divergence from the most recent common ancestor, and species-specific properties, such as generation time, in addition to gene-specific constraints on the sequence evolution. These factors cause large differences in the average lengths as well as the variance of the branch lengths across the branches studied. In this paper, we illustrate how this limitation can adversely impact the confidence with which we infer phenotypic associations for genetic elements, in particular making them sensitive to certain factors in phylogenomic analyses including choice of taxonomic groups and average rates of sequence divergence on phylogenetic branches showing the convergent phenotype. We demonstrate how introducing long branches in phylogenetic trees via the inclusion of distantly related species impacts the reliable estimation of evolutionary rates using gene trees across mammals, as well using a first-of-its-kind model for simulating gene trees. We present key improvements to our methods that address these limitations and overcome them. The next section New Approaches presents a detailed walk-through of our current approach to calculate relative evolutionary rates, the illustration of mean-variance trends (heteroscedasticity) in these rates, and our methodological updates that correct for the problem of heteroscedasticity in the rates. We subsequently demonstrate the improved reliability in relative rate calculations using our updated method, and, more importantly, in the robust detection of convergent rate shifts across a range of evolutionary scenarios in real and simulated phylogenetic datasets.

3.2 Materials and Methods

3.2.1 Protein-coding gene trees across 63 mammalian species

We downloaded the 100-species multiz amino acid alignments available at the UCSC genome browser(Haeussler *et al.*, 2019), and retained only alignments with a minimum of 10 species. We then pruned each alignment down to the species represented in Figure 3.2.1 of the proteome-wide average tree. We added the blind mole rat ortholog of each gene based on the methods described in Partha *et al.* (2017). We estimated the branch lengths for each amino acid alignment using the aaml program from the package PAML (Yang, 2007). We estimated these branch lengths on a tree topology modified from the timetree published in Meredith *et al.* (2011). We attempted to resolve conflicts between the topology inferred in Meredith *et al.* (2011). compared to that in Bininda-Emonds *et al.* (2007). based on a consensus of various studies employing a finer scale phylogenetic inference of the species involved. The differences between our final topology, which we call Meredith+ topology and the Meredith *et al.* (2011) topology include setting the star-nosed mole as an outgroup to the hedgehog and shrew; cow as an outgroup to the Tibetan antelope, sheep and goat; and the ursid clade as an outgroup to mustelid and pinniped clades. For more details about the literature surveyed to resolve these differences, please refer to Meyer *et al.* (2018). The topology of our final Meredith+ tree compared to the UCSC topology tree is reported in Figure 3.2.1. In order to perform analyses benchmarking the method robustness to tree topology, we additionally generated the protein-coding gene trees based on the UCSC tree topology.

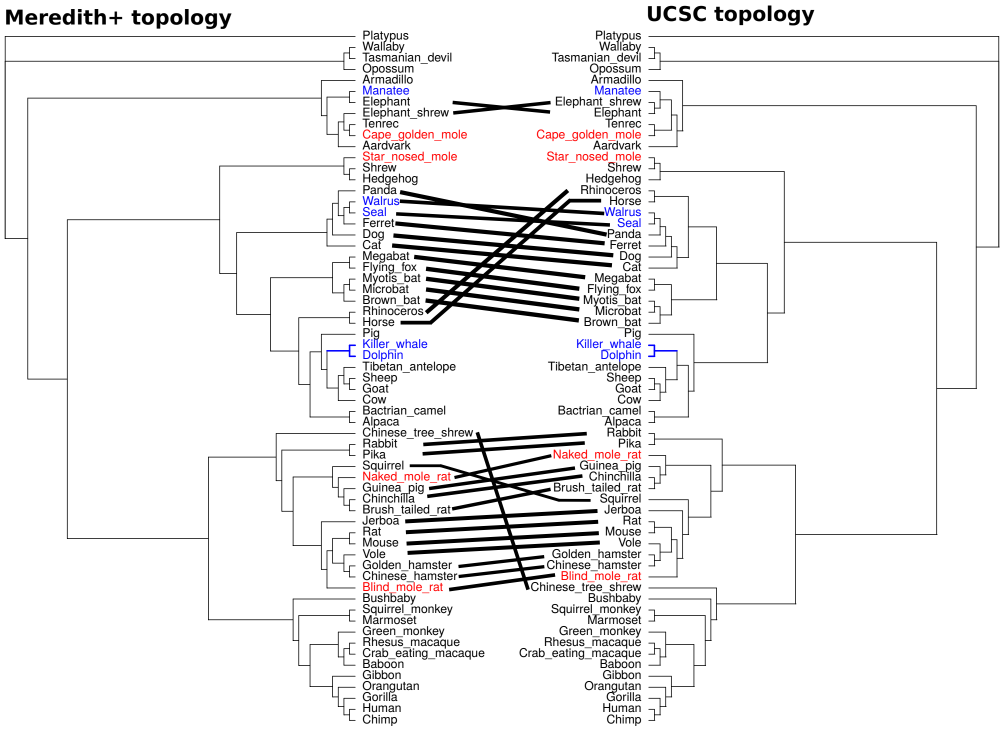


Figure 3.2.1: Cladograms describing relationships between 63 mammalian species used for constructing genome-wide maximum likelihood protein-coding gene trees. Final version of the tree we modified from the topology reported in Meredith et al. (left), and tree reported in UCSC genome browser (right). Key differences between the placement of species are highlighted using black lines. Species corresponding to subterranean and marine mammals are highlighted in red and blue respectively.

3.2.2 Genes showing eye-specific expression

Refer Methods 2.2.4

3.2.3 Calculating concordance in relative rates ranks across datasets with and without non-placental mammals

To estimate the robustness of relative rates calculations to inclusion of non-placental mammals, we calculate the concordance in relative rates ranks across two phylogenetic datasets with and without the non-placental mammals respectively. For each of the 55 eye-specific genes, we rank the extant branches in trees based on the ordering of relative rates independently in the two datasets. We then fit a linear model between the ranks across these two datasets, while forcing a slope coefficient of 1. We subsequently estimate the concordance in the ranks as the mean squared error of the residuals of this linear model. Lower MSE values reflect better concordance in the ranks, and thus superior robustness. We subsequently compare these MSE values for each eye-specific gene obtained using the original and updated methods to calculate relative rates. A positive MSE(original)-MSE(updated) value implies the updated method shows improved concordance in the ranks of relative rates, across datasets with and without the non-placental mammals respectively.

3.2.4 Simulating phylogenetic trees

Phylogenetic branch lengths have units of number of substitutions per site and thus can be thought of as normalized count data. However, we find that a Poisson distribution is unsuitable in this case as the real branch length data shows considerable overdispersion, that is the variance is higher than the mean (Figure 3.2.2). We thus model the branch lengths of the simulated trees using a negative binomial distribution, following ideas from studies simulating expression counts for RNAseq analysis (Robinson *et al.*, 2009; Di *et al.*, 2011; Law *et al.*, 2014; Ritchie *et al.*, 2015). We simulated datasets of phylogenetic trees using the UCSC tree topology and branch lengths from the average proteome-wide tree across 19,149 mammalian protein-coding gene trees across 62 mammals. Figure 3.2.3 describes the tree topology used for the simulations. We simulate the branch lengths (or rates) for every branch (j) on each tree (i) according to the following formula

$$b_{ij} = Poisson(Gamma(\alpha_i \lambda_j, \alpha_i \lambda_j - sqrt(\alpha_i \lambda_j)))$$

where Gamma is parametrized by mean and variance. Here, α_i is a gene-specific scaling term, λ_j is the average rate of the corresponding branch so that $\alpha_i * \lambda_j$ is the expected rate on the ijth branch, and the simulated rate is drawn from a Gamma distribution with that mean. The composite Poisson-Gamma distribution is equivalent to the negative binomial distribution and thus in our simulation the mean variance relationship has a quadratic component, matching what we observe in real data (Figure 3.2.2). We simulate two classes of trees in every dataset based on different input parameters. We simulate control trees, trees where the λ_j are simply the average rate on the branch j. These control trees do not show any explicit convergent rate shift on any of the branches. We additionally simulate positive trees showing convergent rate acceleration on foreground (fgd) branches by sampling at $\lambda_{fgd}^{positive} = m * \lambda_{fgd}^{control}$, only on these branches (m = 1.5, 1.75, or 2). Thus, the foreground branches in positive trees are effectively sampled at an accelerated rate compared to the foreground branches in control trees.

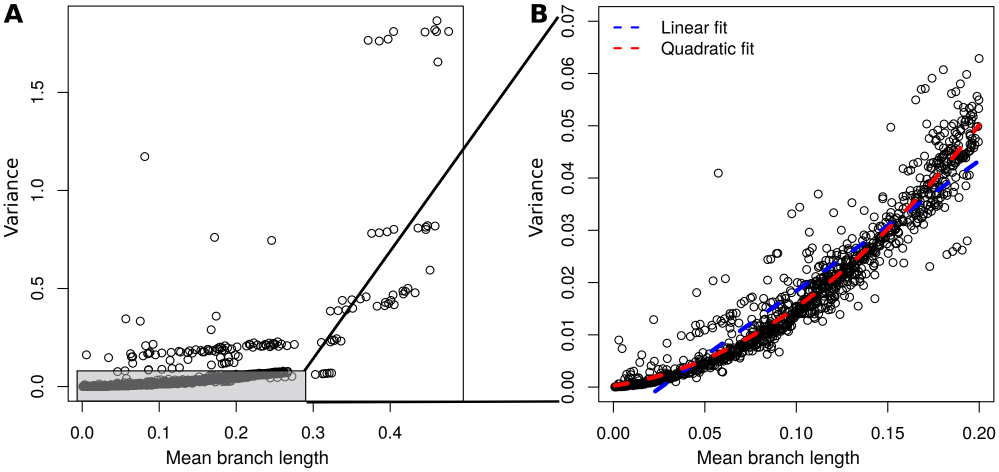


Figure 3.2.2: Mean-Variance trends in branch lengths of 19,149 protein-coding gene trees across 63 mammalian species. Panel A represents the full data, and B shows a zoomed in version containing 80% of the data. The blue dashed line corresponds to the linear model fit between the variance and mean, and the red dashed line represents the fit from a linear model including a quadratic term of the mean in addition to the linear term.

3.2.5 Estimating foreground rate multiplier (m) for genes showing convergent rate acceleration in subterranean mammals

We compared our choices for the foreground rate multiplier (m = 1.5, 1.75, or 2) in simulations to that observed in real data using branch lengths of ten genes showing strong

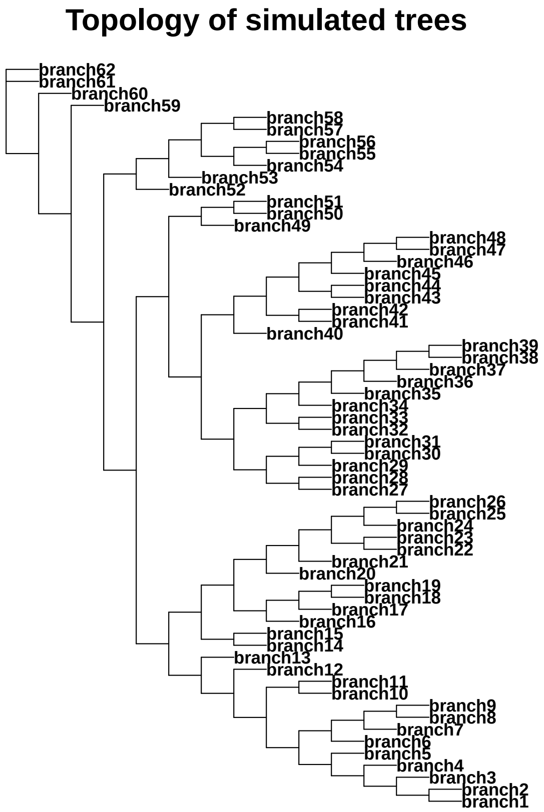


Figure 3.2.3: Topology describing the relationship between branches of simulated trees. The topology is constructed based on the relationships of 62 mammalian species as reported in the UCSC genome browser.

convergent rate acceleration in the four subterranean mammals (moles). Of the 55 genes identified in Partha *et al.* (2017). as showing strongest convergent rate acceleration in the moles, we chose the top eight genes showing relaxation of constraint, and two genes undergoing positive selection on the four mole branches (Table S5 in Supplementary file 2). For each of these ten genes, we estimated the mole foreground rate multiplier as follows: we first fit a linear model between the gene branch lengths and the average branch lengths. Based on the predicted values for the mole branches from this linear model, we calculate the foreground rate multiplier for each mole branch by dividing the real mole branch length by their predicted value. The mole foreground rate multiplier estimate for each gene is subsequently calculated as the mean of the four individual foreground rate multipliers. Table 2, shows the mole foreground rate multiplier estimates for these ten genes.

Gene	Mole foreground multiplier estimate	Evolutionary mode
LIM2	8.63	Relaxed
CRYBB3	5.36	Relaxed
CRYBB2	4.87	Relaxed
CRYGC	4.62	Relaxed
CRYBA1	3.89	Relaxed
GPR89B	3.30	Relaxed
KRTAP17-1	3.22	Positive selection
GNAT1	2.66	Relaxed
ROM1	2.58	Relaxed
COL4A4	1.70	Positive selection

Table 2: Mole foreground multiplier estimates for genes showing strong convergent rate acceleration on mole branches.

3.2.6 Calculating gene-trait correlations

The gene-trait correlations are computed under a Mann-Whitney U testing framework over the binary variable of foreground vs background branches. In the subterranean example, the four subterranean branches (Figure 3.3.1) are designated as foreground. We calculate a foreground acceleration score reflecting the strength of convergent rate acceleration on the foreground branches. The value is calculated as the negative logarithm of the p-value of the Mann-Whitney test multiplied by the direction of the correlation as given by the sign of the rho statistic. A positive rho statistic indicates rate acceleration in the foreground species, and the negative logarithm of p-value reflects the strength of the convergent rate shift. In simulated trees study, we generated trees for three sets of foreground branches with different branch length distributions - short, intermediate, and long as illustrated in Figure 3.3.8 and

Figure 3.3.9.

$$Foreground\,acceleration\,score = Sign(Rho) * [-log_{10}P]$$

where rho and P are the correlation coefficient and statistical significance of the Mann-Whitney test for association between relative rates and binary trait.

3.2.7 Gene Ontology term enrichment analysis

Refer Methods 2.2.3

3.3 Results

3.3.1 Need for new approaches to estimate relative evolutionary rates

3.3.1.1 Original relative-evolutionary-rates method for predicting phenotypic associations of genetic elements Our method infers genetic elements associated with a convergent phenotype of interest based on correlations between that phenotype and the rates of evolution of genetic elements. As input, the phenotype is encoded as a binary trait on a phylogenetic tree, and the evolution of each genetic element is similarly described by phylogenetic trees with the same fixed topology. Figure 3.3.1 provides an illustration of our method capturing the convergent acceleration of the Lens Intrinsic membrane 2 protein Lim2 on four subterranean mammal branches. We use maximum likelihood approaches to estimate the amount of sequence divergence of each genetic element on branches of the phylogenetic tree (Yang, 2007). Using each tree's branch lengths, we calculate the average tree across the individual trees reflecting the expected amount of divergence on each branch. Relative evolutionary rates (RERs) on individual trees are then calculated as the residuals of a linear regression analysis where the dependent variable corresponds to the branch lengths of individual trees, and the independent variable corresponds to branch lengths of the average tree. Thus the relative rates reflect the gene-specific rate of divergence in each branch, factoring out the expected divergence on the branch due to genome-wide effects (such as mutation rate, time since speciation, etc.). The relative rates method works downstream of estimating the trees, and hence considers protein-coding gene trees, non-coding genetic element trees, and simulated gene trees equivalently. For the sake of simplicity, we refer to the relative rates on the branches of each tree as the gene-specific relative rate; the term gene could in principle be referring to a protein-coding gene, non-coding genetic element, or a simulated tree depending on the dataset being studied.

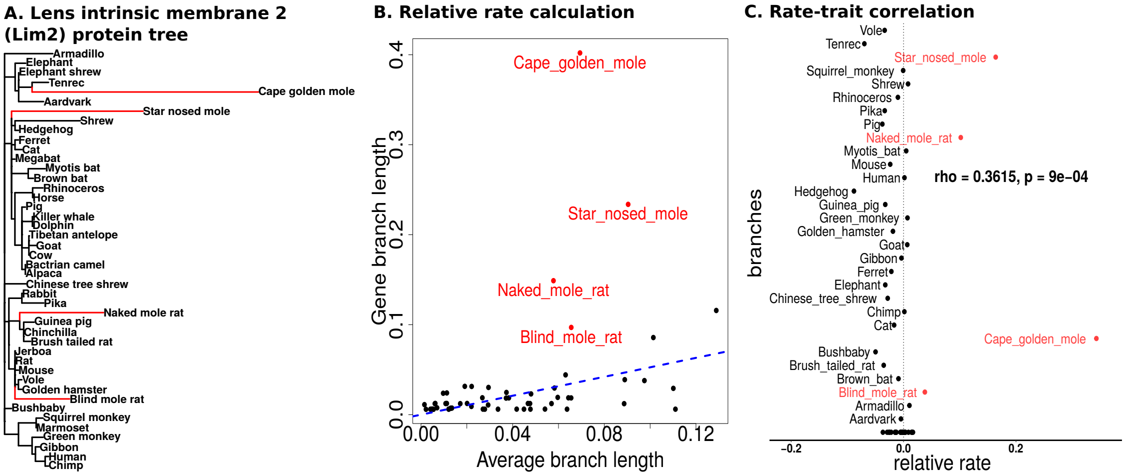


Figure 3.3.1: Predicting gene-trait associations using relative rates method. A. Lens Intrinsic Membrane 2 (Lim2) protein-coding gene tree. Our phylogenetic dataset is comprised of trees constructed from alignments of protein-coding genes in the mammalian genome across 59 species of placental mammals. B. Relative rates on branches of phylogenetic trees are calculated using linear regression. C. Gene-trait associations are identified using correlations of relative rates of the gene with binary trait of interest.

3.3.1.2 Estimating mean-variance trends in relative rates Our original method calculates the gene-specific rates by correcting for the genome-wide effects on branch lengths using linear regression. Consequently, the variance of the relative rates on individual branches strongly depends on the average length of the branch, illustrated here using an example protein-coding gene tree for MFNG, Manic Fringe Homolog Drosophila (Figure 3.3.2A). We see that longer branches have relative rates showing a higher variance, as can be inferred from the increasing spread of the relative rates. This pattern becomes clearer when we plot the genome-wide variance in relative rates for branches of different average lengths (Figure 3.3.2B). In statistical terms, the relative rates are heteroscedastic, meaning they show unequal variance across the range of values of the dependent variable, here the average branch length. The presence of a non-constant mean-variance trend in the residuals stands in violation of one of the assumptions underlying linear regression, namely homoscedasticity, or constant variance of residuals with respect to the dependent variable. More importantly, we suspect that this heteroscedasticity of the relative rates adversely affects the confidence with which we can infer rate shifts on specific branches. For example, the presence of a mean-variance trend can increase the likelihood of observing higher relative rates on longer branches by chance, rather than due to gene-specific changes reflecting changes in selective pressure. A potential negative consequence could be a higher proportion of false positives while inferring convergent rate changes on such branches.

3.3.1.3 Updated method to calculate relative rates In this study, we present an approach relying on a combination of data transformation and weighted linear regression to calculate relative evolutionary rates that addresses the statistical limitations resulting from relative rates calculated using naive linear regression. The proposed method updates are based on the ideas presented in Law *et al.* (2014), who developed new linear modeling strategies to handle issues related to mean-variance relationship of log-counts in RNA-seq reads (Law *et al.*, 2014; Ritchie *et al.*, 2015). We represent the branch lengths on individual gene trees as a matrix Y, where rows correspond to individual genes (g), and columns to the branches (b) on these trees. We first transform the branch length data using a square-root transformation (3.1).

$$Y_{gb}^{'} = \sqrt{Y_{gb}} \tag{3.1}$$

Following the transformation, we perform a weighted regression analysis to calculate the relative evolutionary rates as follows: we calculate the average tree and perform a first-pass of linear regression using the transformed branch length matrix (3.2, 3.3, and 3.4).

$$x_b = \bar{Y_b^{'}} \tag{3.2}$$

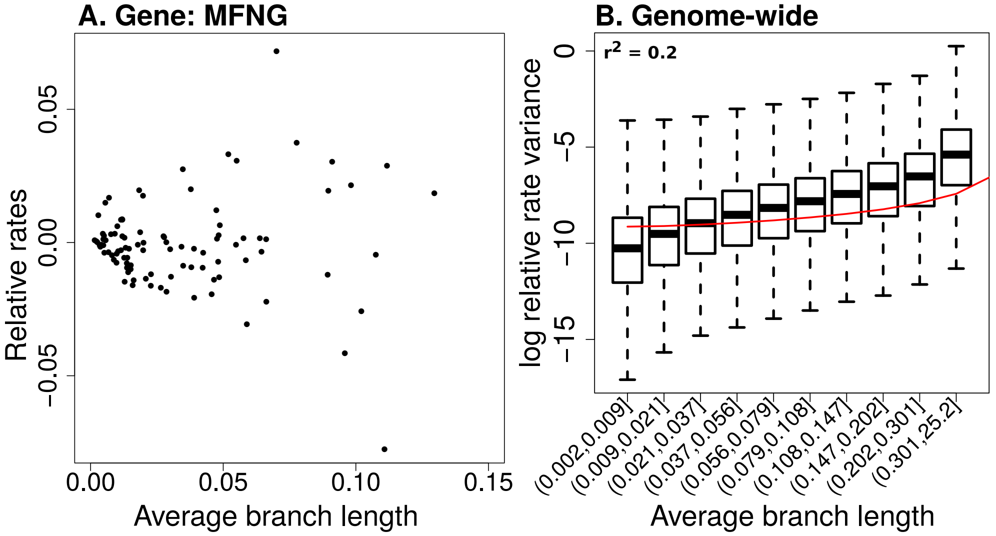


Figure 3.3.2: Heteroscedasticity in the relative rates computed using current method. A. Relative rates on branches of Manic Fringe (MFNG) gene tree, calculated using original method. Heteroscedasticity in the relative rates can be visualized as the increase in the variance of the relative rates with increasing average branch length B. Genome-wide mean-variance trends in relative rates. Higher variance in relative rates are observed with increasing branch lengths.

where x_b is the branch length for branch b in the average tree.

$$\hat{\beta} = (X^T X)^{-1} X^T Y' \tag{3.3}$$

$$R = Y' - X\hat{\beta} \tag{3.4}$$

where $\hat{\beta}$ are the coefficients of linear regression, and R is the residuals matrix. We then estimate the mean-variance trends in the residuals of the linear regression analysis by empirically fitting a locally weighted scatterplot smoothing (LOWESS) function capturing the relationship between the log of variance of the residuals and the branch lengths (3.5).

$$log(R^2) = f(Y') \tag{3.5}$$

Subsequent to estimating this function, we assign each gene x branch observation a weight W based on the predicted value for the branch, obtained from the first pass linear regression (3.6)

$$W = e^{-f(X\hat{\beta})} \tag{3.6}$$

For branches that are shorter on average, the variance in the residuals is smaller, thus resulting in a higher weight, and vice versa. Using the computed weights, we perform a weighted regression analysis between the individual branch length (dependent variable) and the average tree (independent variable). The weighted regression analysis attempts to remove the heteroscedasticity in the residuals by computing the residuals after minimizing the weighted sum of squared errors, as opposed to the raw sum of squared errors (3.7, 3.8).

$$\hat{\beta}_{WLS} = (X^T W X)^{-1} X^T W Y^{'} \tag{3.7}$$

$$R = Y^{'} - X\hat{\beta}_{WLS} \tag{3.8}$$

$$r^{'}_{gb} = \frac{r_{gb}\sqrt{w_{gb}}}{\sigma_b} \tag{3.9}$$

where σ_b is the standard deviation of the weighted residuals in branch b. Subsequent to the weighted regression analysis, the weighted residuals ($r^{'}_{gb}$), are estimated by rescaling the regression residuals (r_{gb}) with the weights, and the weighted residuals are additionally standardized to have unit variance within every branch across all genes (3.9). The weighted residuals ($r^{'}_{gb}$) correspond to the weighted relative rate on branch b for gene g. The differences to the relative rate calculations introduced by the updated method result in changes to the scales of the relative rates computed. However, we note that this scale is arbitrary and the downstream gene-trait correlations for binary traits estimated using a Mann-Whitney test (see Methods) depend only on the ranks of the relative rates of each branch within any single gene tree. Figure 3.3.3 shows the workflow for computing relative evolutionary rates using the original and updated method.

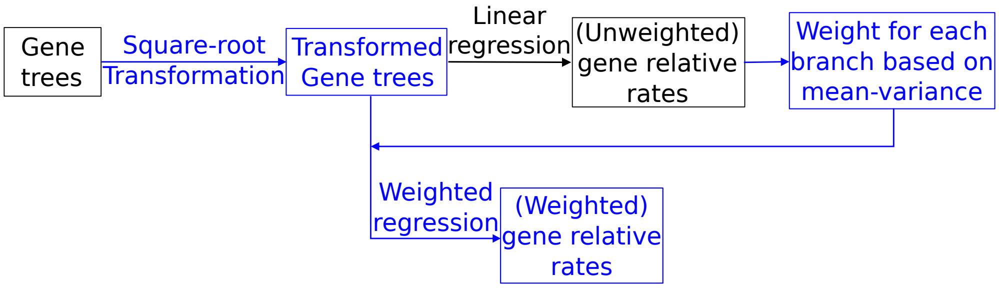

Gene trees
Square-root Transformation
Transformed Gene trees
Linear regression
(Unweighted) gene relative rates
Weight for each branch based on mean-variance
Weighted regression
(Weighted) gene relative rates

Figure 3.3.3: Workflow for calculating relative evolutionary rates using the updated method. Black areas of the workflow represent steps implemented as part of current relative rates method, and blue areas correspond to methodological updates.

3.3.2 Improvements to relative evolutionary rates methods mitigate genome-wide mean-variance relationship

Our updated method to calculate relative rates using data transformation followed by weighted regression produces nearly homoscedastic relative rates that do not show a significant global mean-variance relationship. Figure 3.3.4A shows the relative rates computed for the MFNG protein-coding gene tree using the updated method. In comparison to the original method based on naive linear regression (Figure 3.3.2A), we observe that the up-dated method produces relative rates showing no apparent increase in the variance of relative rates on longer branches of the tree. Plotting the genome-wide mean-variance trends of the relative rates across all branches of all gene trees, we observe that the relative rates cal-culated from transformed-weighted residuals show nearly constant variance across branches of varying lengths (Figure 3.3.4B). We additionally checked the mean-variance relationships from intermediate steps in our method that can estimate relative rates, corresponding to two method variants which do not implement data transformation (linear-weighted regime) or a weighted regression (square-root unweighted regime) (Figure 3.3.5). However, we find that the intermediate regimes, utilizing only one of the method updates (branch length transformation or weighted regression alone) are less effective at eliminating mean-variance trends. A combination of transformation and weighted regression steps works best at producing homoscedastic relative rates.

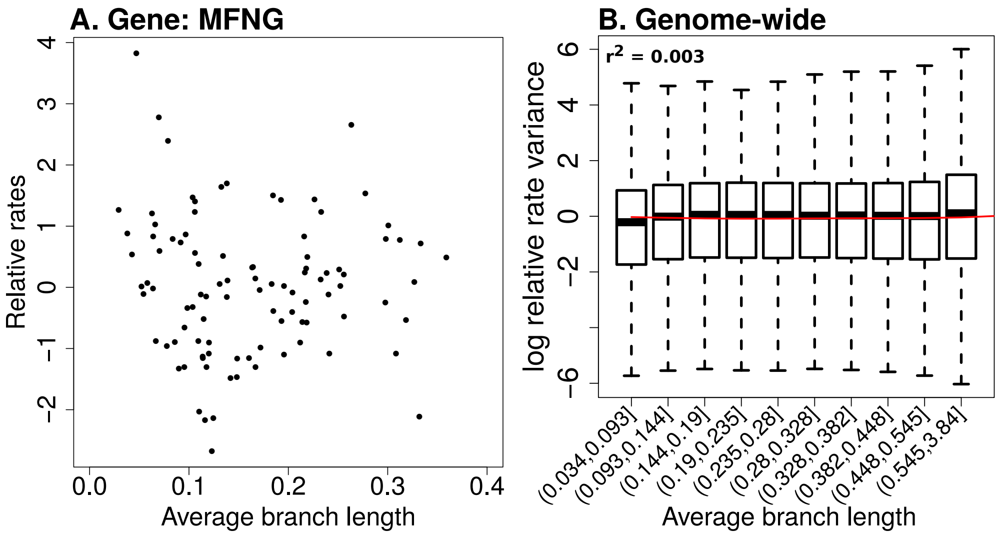


Figure 3.3.4: Updated method to calculate relative rates shows no apparent trends of heteroscedasticity. A. Manic Fringe (MFNG) gene relative rates calculated using the updated method. B. Genome-wide mean-variance trends for relative rates computed using the updated method show constant variance with increasing branch lengths.

3.3.3 Better robustness to inclusion of distantly related species

In earlier applications of our relative rates method to detect genetic elements convergently responding in subterranean mammals and marine mammals respectively, we sampled alignments of placental mammal species to construct phylogenetic trees for each genetic element (Chikina *et al.*, 2016; Partha *et al.*, 2017). These alignments were derived from the placental mammal subset of the 100-way vertebrate alignments made publicly available by the UCSC genome browser (Haeussler *et al.*, 2019). In addition to these placental mammals, the 100-way alignments include four other species of mammals, three marsupials Opossum (monDom5), Wallaby (macEug2), Tasmanian Devil (sarHar1), and one monotreme Platypus (ornAna1). Despite deep conservation of many genetic elements in these non-placental mammals, human-and-mouse centered phylogenomic studies tend to exclude these

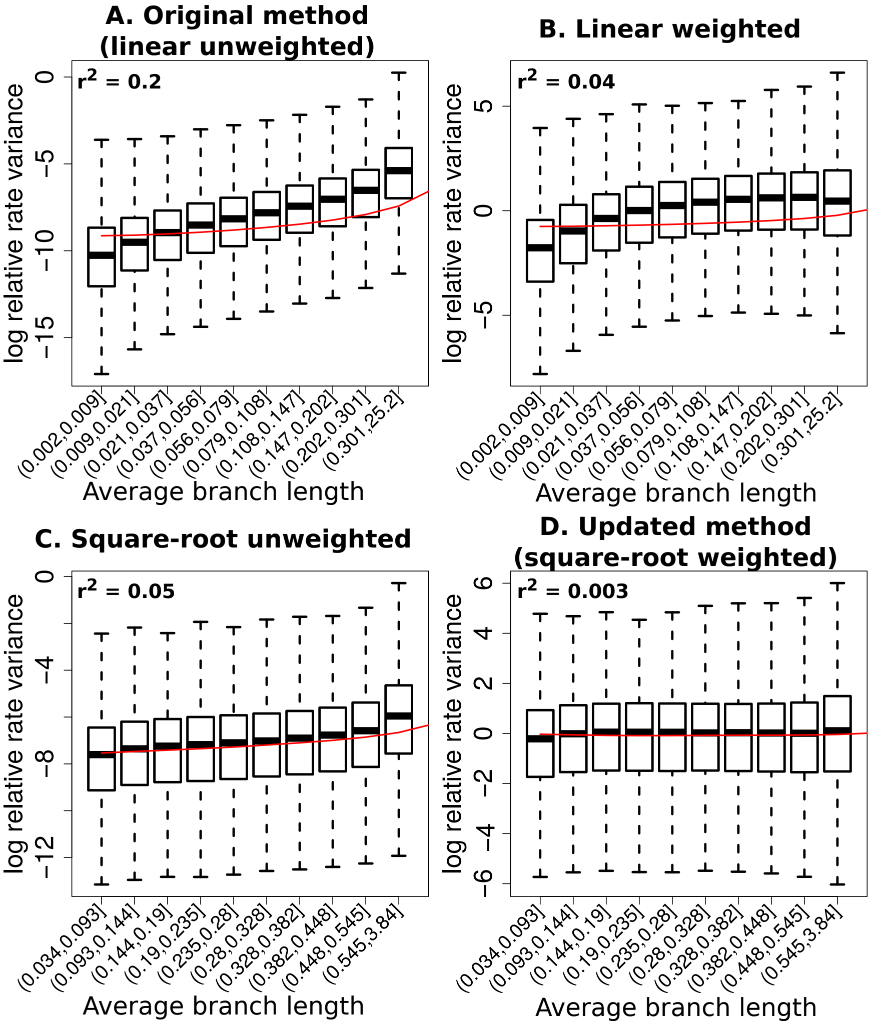


Figure 3.3.5: Comparison of mean-variance trends in relative rates computed using original, updated and intermediate methods. A corresponds to original method, D the updated method. Panels B and C reflect methods that are intermediate to the updated method, with no transformation (B), and no weighted regression (C).

species due to the introduction of long branches in the phylogenetic trees (Parker *et al.*, 2013; Marcovitz *et al.*, 2016; Prudent *et al.*, 2016). For instance, in previous applications of our relative rates method we deliberately excluded these non-placental mammals since they produce wide variations in relative rates due to the introduction of long branches, which would adversely affect the confidence with which we make inferences of convergent rate acceleration in species exhibiting a convergent phenotype (Chikina *et al.*, 2016; Partha *et al.*, 2017). However, scanning for rate-trait associations across tree datasets with higher numbers of species would allow for more statistical power, and hence a relative rates method that can reliably include such distantly related species offers a clear advantage. To this end, we tested the robustness of our updated method to the inclusion of distantly related species at inferring convergent rate shifts. We chose two phylogenetic datasets - 1. Genome-wide protein-coding gene alignments across 59 placental mammal species, and 2. across 63 mammals including four non-placental mammals in addition to the placentals. An example demonstration of how our current method to calculate relative rates is sensitive to the inclusion of non-placental mammals is illustrated in Figure 3.3.6A. Using the Peropsin (RRH) gene for illustrative purposes, we show that the ranks of relative rates computed using the current method considerably vary upon the inclusion of non-placental mammals. These changes in ranks are observed across many branches on the gene tree including one of the four subterranean branches (Cape golden mole). In comparison, the updated method displays a stronger concordance in the ranks of the computed relative rates (Figure 3.3.6A). Consequently, the subterranean acceleration scores for RRH computed using the updated method are more stable with the inclusion of non-placental mammals (Table 3).

Dataset\Method	Original	Updated
With non-placentals	2.70	2.1
Placentals only	1.38	2.0

Table 3: Subterranean acceleration scores for Peropsin (RRH) computed using two methods, and across two datasets. In comparison to the original method, the updated method shows stronger consistency in the scores across the two tree datasets with and without the non-placental mammals. The subterranean acceleration scores reflect the significance of convergent rate acceleration on the four subterranean branches.

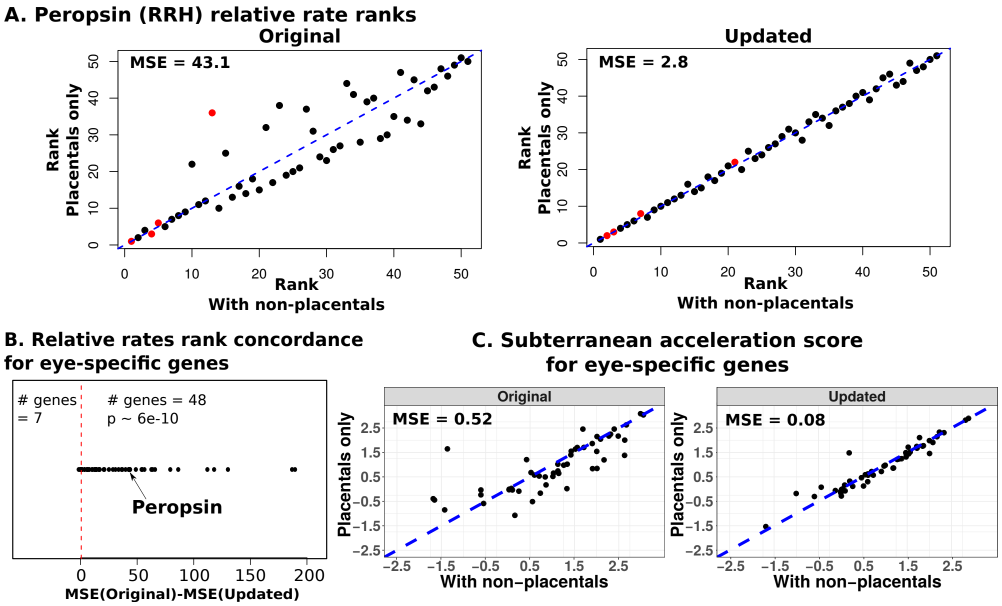


Figure 3.3.6: Comparison of robustness of methods to inclusion of non-placental mammals. A. The relative rate ranks of terminal lineage branches within the RRH tree are plotted with respect to the inclusion of non-placental mammals. Red points denote subterranean branches. B. Updated method shows improved concordance in ranks of relative rates across trees with and without non-placental mammals. C. Updated method shows improved robustness to inclusion of non-placental mammals at detecting subterranean acceleration of eye-specific genes.

We also performed a larger-scale benchmarking of the robustness of our methods to the inclusion of non-placental mammals across 55 genes showing eye-specific expression. These genes were identified based on mouse microarray expression data across 91 tissues (see Methods). We first compared the estimated concordance in ranks of relative rates computed using the original and updated method in trees including and excluding the non-placental mammals. For each gene, we calculated concordance in ranks using the mean squared error of residuals of a linear model (see Methods), where lower MSE values reflect better robustness. We observed that for 48 (out of 55) eye-specific genes, the updated method shows improved concordance in the ranks of relative rates across the two sets of gene trees (Figure 3.3.6B). Using a pairwise Wilcoxon test, we compared the MSE values obtained using the original versus updated method, revealing a statistically significant (P = 6e-10) decrease in MSE values obtained using the updated method.

For each of these eye-specific genes, we also calculated subterranean acceleration scores (see Methods) reflecting the convergent rate acceleration on the four subterranean branches independently in gene trees including and excluding the non-placental mammals. Based on the relative rates calculated using each method, we compared the concordance of the subterranean acceleration scores across the two tree datasets. Ideally, we expect the scores produced by the methods to be highly consistent across the two datasets since the four non-placental mammals are not subterranean, with only minor differences arising due to the inclusion of four additional background species. The results of the analysis revealed that the updated method produces superior concordance in the scores across the two tree datasets, reflecting its improved ability to handle the long branches introduced by the non-placental mammals (Figure 3.3.6C).

3.3.4 Improved power to detect convergent rate shifts in simulated trees

In order to compare the power of our methods to detect convergent rate shifts in branches across a range of evolutionary scenarios, we developed a model to simulate individual gene trees. Such a model allows us to rigorously examine method performance in relation to various parameters in phylogenetic datasets including number of foreground branches, length distribution of foreground branches etc., where foreground branches describe branches showing a convergent phenotype, while background branches do not. The limited availability of ground truth examples of convergently evolving genetic elements calls for the development of biologically realistic simulations of sequence evolution. Using our model to simulate trees (see Methods), we compared the power to detect rate shifts in relation to two factors: 1. Average lengths of foreground branches, in particular extreme foreground branches that are very short or very long on average. 2. Number of foreground branches. We investigated the performance of the updated method in detecting rate shifts in such extreme branches, assessing the power advantage resulting from calculating relative rates that do not suffer from a biased mean-variance relationship. Our model to simulate phylogenetic trees allows for explicit control over choosing foreground branches showing convergent rate acceleration. We simulate control trees, where all branches are modeled to evolve at their respective average rates, and positive trees, where the chosen foreground branches are modeled to evolve at an accelerated rate. Initially, we chose a foreground rate multiplier value of 2, which corresponds to foreground branches in positive trees being sampled at twice their average rates (see Methods). We first compared the heteroscedasticity in the relative rates on the branches of the control trees calculated using the original and updated methods. Similar to the trends observed in mammalian gene trees (Figure 3.3.5), we observed that the updated method outperformed the original method at producing homoscedastic relative rates (Figure 3.3.7). We then calculated a foreground acceleration score for individual simulated trees, both control and positive. A more positive value of this score, calculated as a signed negative logarithm of the p-value, reflects stronger convergent rate acceleration on the foreground branches (see Methods). Subsequent to estimating these scores, we evaluated the performance of the two methods, based on the power to distinguish the positive trees from control trees. In two independent simulation settings with foreground branches of long and short average lengths, we observed that the updated method offers more power to detect positive trees (Figure 3.3.8, see Figure 3.3.9 for precision-recall curves).

We repeated the analyses with more conservative choices for modeling foreground acceleration using foreground rate multiplier values of 1.5 and 1.75 to ensure the improved power was robust to the choice of foreground rate multiplier (m). Consistent with the original analysis, the updated method was more powerful at precise detection of positive trees for all values of m (Figure 3.3.10). We also observed that with increasing values of m, it becomes easier to detect positive trees (Figure 3.3.8B, Figure 3.3.10) which is expected since the foreground branches will be longer for larger values of m. Our choices of foreground rate multiplier values in simulations (m = 1.5, 1.75 and 2) represent challenging scenarios for our method in comparison to foreground rate multiplier estimates observed in real data. For instance, our simulation choices are lower than the foreground rate multiplier estimates for genes showing strong relaxation of constraint in subterranean mammals, and more comparable to the estimates for genes under positive selection (see Methods, Table 2). This proves the utility of our method at detecting genes showing rate acceleration due to positive

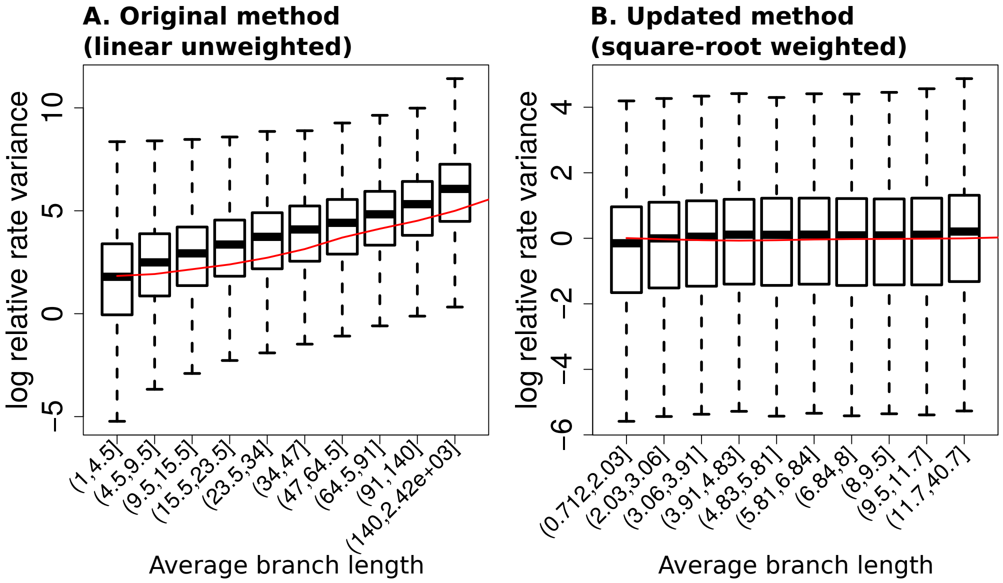


Figure 3.3.7: Mean-variance trends in relative rates on branches of simulated phylogenetic trees computed using the two methods. The original method (A) produces heteroscedastic relative rates that show a strong mean-variance trend, whereas relative rates calculated using the updated method (B) show constant variance across branches of different lengths.

selection, in addition to relaxation of constraint.

We also performed a control analysis using foreground acceleration scores computed using four length-matched control foreground branches that were not the true foreground, proving that the positive trees were not detected due to random chance (Figure 3.3.11). Finally, in addition to the positive trees with foreground branches that were long or short, we compared the power to detect rate acceleration on foreground branches of intermediate length. Consistent with the findings in short/long foregrounds, we find a modest yet significant improvement offered by the updated method (Figure 3.3.12). Overall, we find that our updated method to compute relative rates offers a significantly improved power to detect convergent rate shifts in simulated trees.

We then compared the power to detect rate shifts across varying numbers of foreground branches by simulating positive trees with seven foreground branches of long average lengths

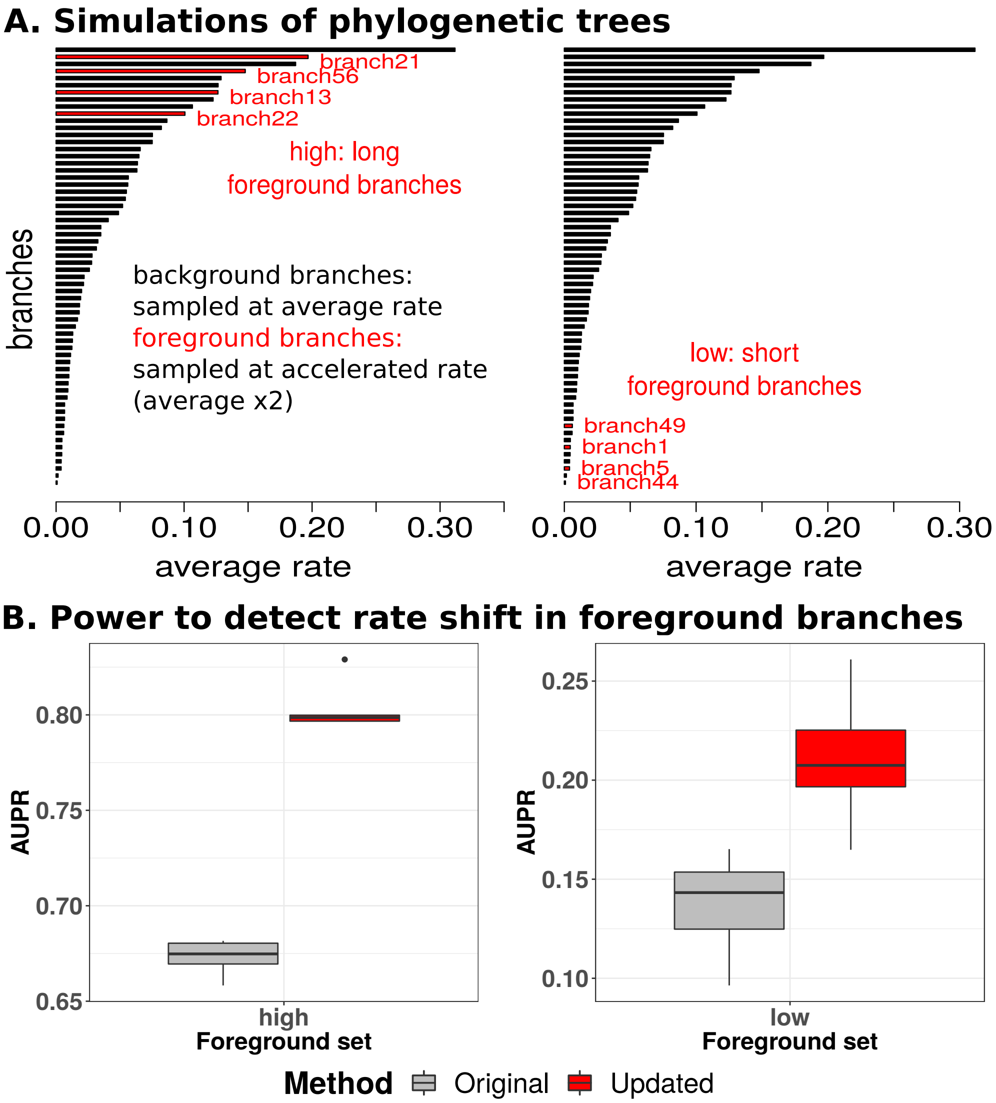


Figure 3.3.8: Comparison of method performance across simulated phylogenetic trees. A. Branch length distributions for simulating phylogenetic trees with foreground branches highlighted in red. B. Power to detect rate shift in foreground branches of simulated trees.

(Figure 3.3.13). We subsequently generated positive trees with subsets of n branches (n ranging from 4 to 7) among these seven foreground branches (Figure 3.3.13). Within each of these datasets, we calculated foreground acceleration scores for control and positive trees using each method independently. We observed that the updated method to calculate relative

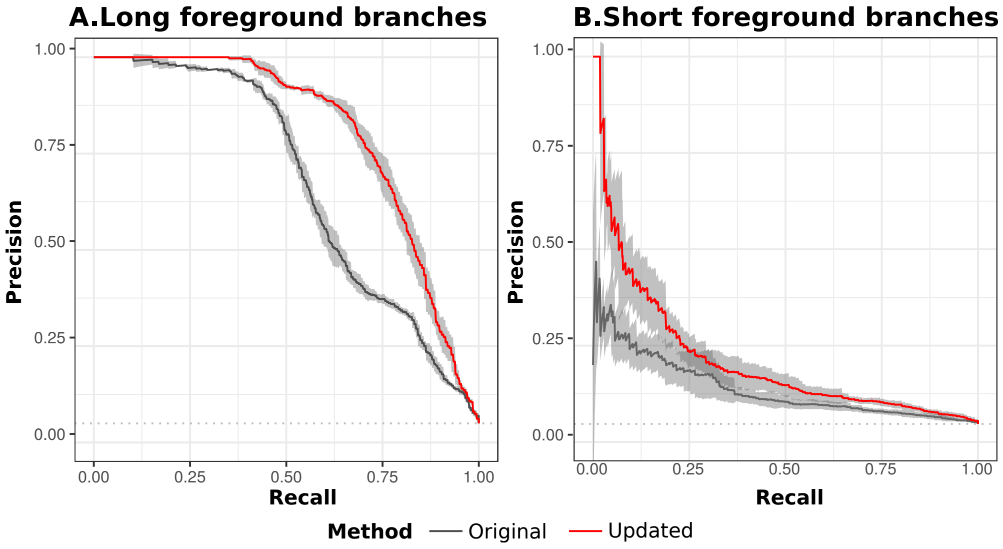


Figure 3.3.9: Precision-recall curves for detecting positive trees in simulations. Updated method outperforms the original method across varying configurations of foreground branches, including long (A) and short (B) foreground branches respectively. Shaded areas reflect 95% confidence intervals.

rates is consistently more powerful than the original method at precise detection of positive trees (Figure 3.3.14A). We repeated the analysis choosing seven foreground branches that were short on average rather than long (Figure 3.3.13) and observed consistent gains in power using updated method to calculate relative rates (Figure 3.3.14B). Applying our method to simulations with varying configurations of foreground branches also revealed that the power to detect foreground acceleration is higher for longer foreground branches. In other words, it is easier to detect rate acceleration on longer foreground branches compared to shorter ones (Figure 3.3.14A vs B). In terms of sequence divergence, longer branches represent instances of higher sequence divergence or more changes, which are easier to detect as the method ranks the rates on branches relative to one another. The increased power to detect rate acceleration therefore becomes especially useful in convergent phenotypes involving short foreground branches, where the improvements are nearly two-fold (Figure 3.3.14B).

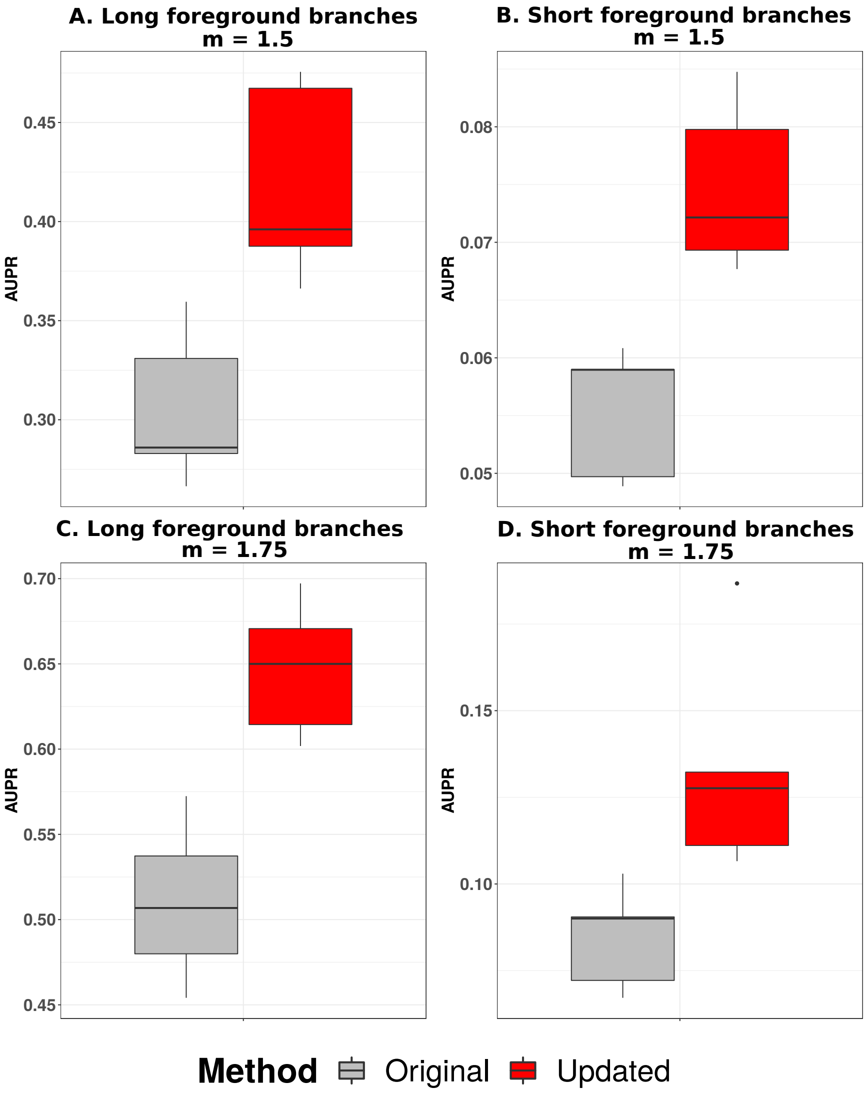


Figure 3.3.10: Power to detect rate shifts in foreground branches sampled at varying accelerated rates. Updated method offers consistently higher power in independent simulations of long (A,C) and short foreground branches (B,D) where the foreground branches are sampled using varying foreground rate multipliers (m = 1.5, A and B; m = 1.75, C and D).

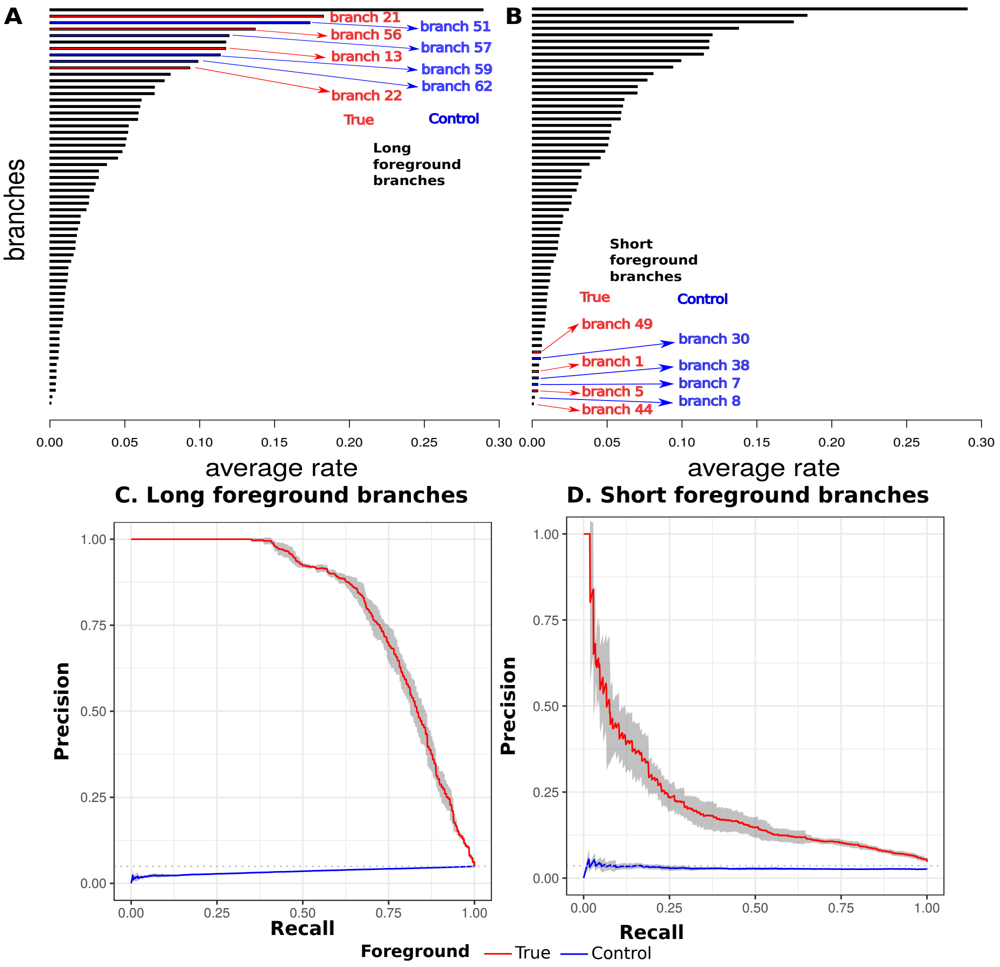


Figure 3.3.11: Comparison of power to detect positive trees using true foreground branches vs control foreground branches. Panels A and B correspond to the true and control foreground branches used in independent simulations with long and short foreground branches respectively. Shaded areas reflect 95% confidence intervals

3.3.5 Relative rates-based inference is robust to minor uncertainties in species tree topology

Our method relies on estimating sequence divergence on branches of phylogenetic trees with a fixed topology. Efforts to better resolve the phylogeny of extant mammals have resulted in continuous updates to the consensus species tree topology (Murphy *et al.*, 2001,

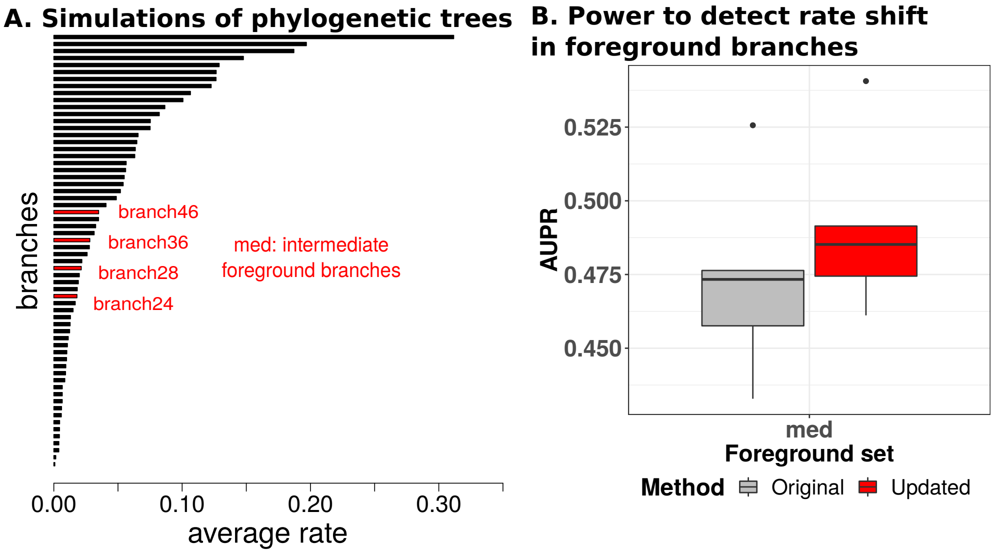
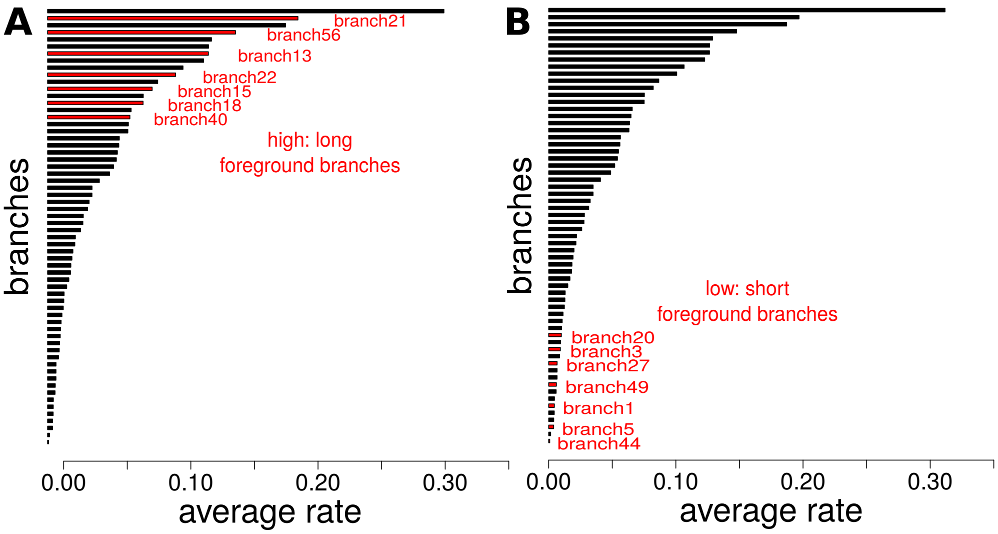


Figure 3.3.12: Simulations of phylogenetic trees with foreground branches of intermediate length A. Foreground branches (in red), used for simulating convergent acceleration on intermediately long branches. B. Comparison of power between the two methods to detect convergent rate acceleration on intermediately long foreground branches.

2007). Topology trees commonly used in phylogenomic analyses of extant mammals include the UCSC genome browsers 100-way tree, as well as the timetrees reported in the Meredith et al and Bininda-Emonds et al (Bininda-Emonds *et al.*, 2007; Meredith *et al.*, 2011; Casper *et al.*, 2018). Differences between these species tree topologies often involve entire clades, and the decision to choose a particular topology tree can potentially strongly influence the outcomes of phylogenetic analyses. Here, we benchmarked the robustness of our relative rates method to the choice of topology tree. We constructed protein-coding gene trees based on two different species tree topologies, namely the UCSC 100-way tree and our modified Meredith et al. (Meredith+) topology tree (see Methods). The Robinson-Foulds metric (calculated using the function RF.dist in the R package phangorn) between these two phylogenies is 22, reflecting differences in 22 partitions of species (Robinson and Foulds, 1981; Schliep, 2011). We observed that both the updated and original methods to calculate relative rates show robust signatures of subterranean rate acceleration for eye-specific genes with respect to the



C

Number of foreground branches	Long foreground branches set	Short foreground branches set
4	21, 56, 13, 22	44, 5, 1, 49
5	21, 56, 13, 22, 15	44, 5, 1, 49, 27
6	21, 56, 13, 22, 15, 18	44, 5, 1, 49, 27, 3
7	21, 56, 13, 22, 15, 18, 40	44, 5, 1, 49, 27, 3, 20

Figure 3.3.13: Simulations of phylogenetic trees with varying numbers of foreground branches. Red branches correspond to the seven foreground branches chosen for simulating trees showing convergent rate acceleration on long (A) and short (B) foreground branches. The foreground branch sets of simulated trees used for comparing the power to detect foreground acceleration across different numbers of long and short branches are given in (C).

species tree topology used (Figure 3.3.15).

3.3.6 Comparison of power to detect enriched pathways associated with two independent convergent phenotypes

Beyond examining individual genes, we further assessed our new methods ability to detect pathway enrichments for genes under relaxation of constraint in subterranean mammals and

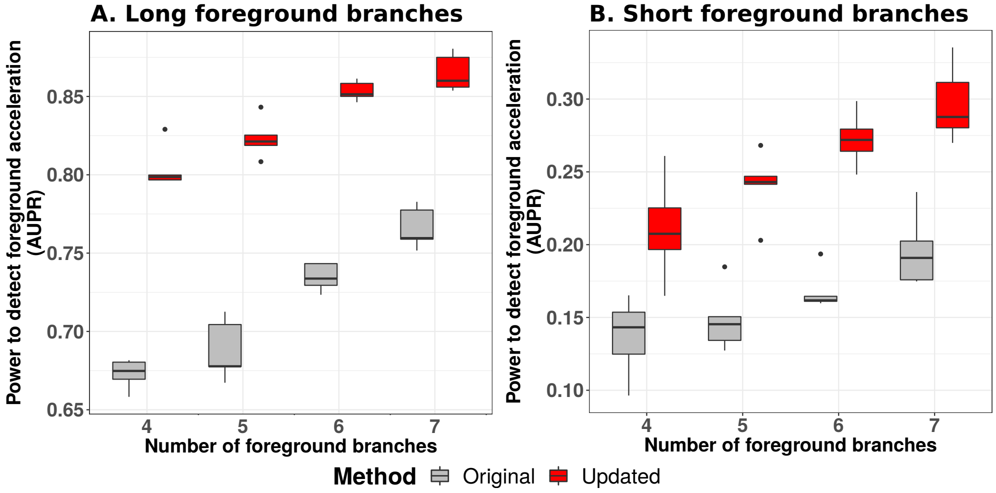


Figure 3.3.14: Improved power to detect foreground rate shifts using the updated method across different numbers of foreground branches. These simulations were performed across two scenarios with different foreground branch sets consisting of short (A) and long branches (B) respectively.

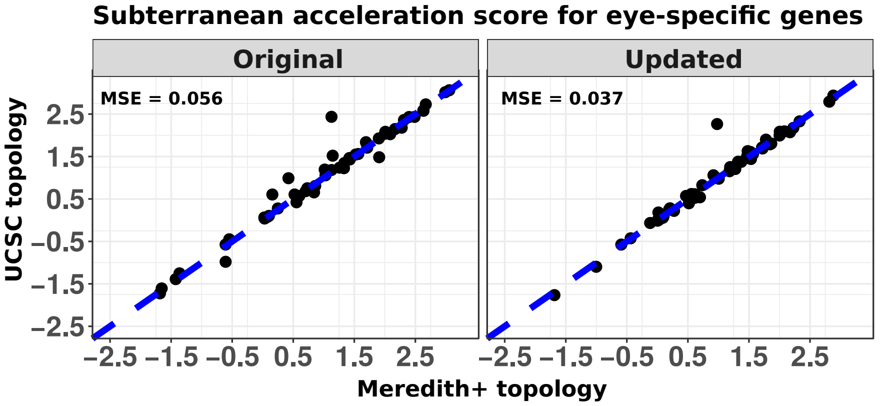


Figure 3.3.15: Comparison of robustness of methods to species tree topology. Both the original and updated relative rates methods are robust to choice of species tree topology used to construct individual gene trees. Points represent the strength of convergent subterranean acceleration for eye-specific genes.

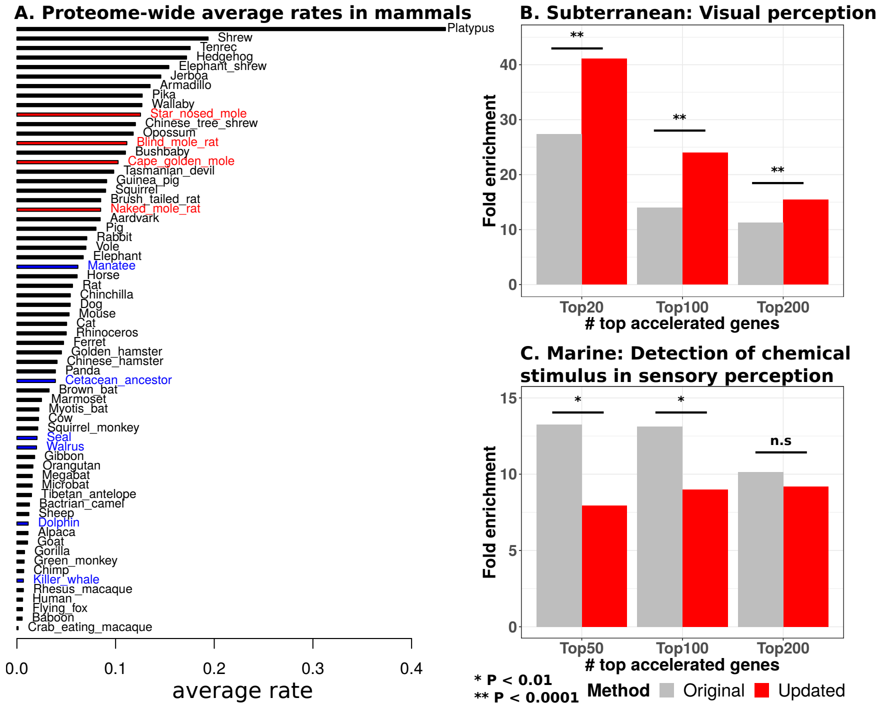


Figure 3.3.16: Comparison of fold enrichments of top enriched terms associated with two convergent phenotypes. A. Branch length distributions representing average rates in protein-coding gene trees across mammals. Foreground branches corresponding to subterranean, and marine mammals, are highlighted in red and blue respectively. B. The barplot compares the fold enrichment for the visual perception GO term across top subterranean accelerated genes. C. The same analysis was repeated with the top enriched term in marine-accelerated genes, namely Detection of chemical stimulus in sensory perception.

marine mammals (see Figure 3.2.1 for respective foreground branches, and Figure 3.3.16 for average rates). Compared to our original method, the updated method detected more enriched Gene Ontology (GO) terms with accelerated evolutionary rates in subterranean mammals (Table 4). Additionally, the fold enrichment for detected terms was significantly stronger with the updated method (Figure 3.3.16, Table S1-S6 in Supplementary file 2). On the other hand, the marine system showed mixed results. Both the updated and the original methods showed approximately equal power to detect enriched GO terms if we only consider the number of terms detected (Table 5, Table S7-S12 in Supplementary file 2). However, when comparing the fold enrichment for detected terms, the original method was significantly better than the updated method (Figure 3.3.16). These contrasting results from the subterranean dataset versus the marine dataset indicate the importance of tailoring the corrections we have developed to the dataset of interest, as well as the importance of taking advantage of simulation-based power and robustness assessments to develop methods that are broadly applicable to many convergent phenotypes.

topN:	subterranean-accelerated GO terms (FDR <0.05)	
number of top accelerated genes	Original method	Updated Method
20	2	9
100	11	28
200	16	32

Table 4: Comparison of numbers of vision-related Gene Ontology terms enriched in top mole-accelerated genes discovered by the original and updated methods.

topN:	Marine-accelerated GO terms (FDR <0.05)	
number of top accelerated genes	Original method	Updated Method
50	16	10
100	27	31
200	59	59

Table 5: Comparison of numbers of Gene Ontology terms enriched in top genes showing marine-acceleration discovered by the original and updated methods.

3.4 Discussion

Our original evolutionary-rates-based method to detect genomic elements underlying convergent phenotypes has already proved to be a valuable technique to detect genes and enhancers associated with transitions to marine and subterranean habitats (Chikina *et al.*, 2016; Partha *et al.*, 2017). However, the original method suffered from reduced power to detect such genomic elements due to a heteroscedastic relationship between the mean and variance of branch lengths for a given branch across all gene trees, i.e. branches that are longer on average have higher variance than branches that are shorter on average.

Here, we developed a method using a square-root transformation and a weighted regression based on the observed mean-variance relationship to correct for the heteroscedasticity. While our objective was to develop a method that robustly handles mean-variance trends in phylogenetic trees, we do not systematically investigate factors underlying this property. Previous genome-scale analyses in modern birds have showed evidence for base composition heterogeneity affecting variance of branch lengths in exon trees (Jarvis *et al.*, 2014). However, in our phylogenetic dataset of mammalian protein-coding genes we found no evidence for base composition heterogeneity influencing sequence divergence at the gene level - we failed to detect any significant global trends between GC-content of our sequences and their raw branch lengths, relative rates computed using our original method, or from our new method (Figure 3.4.1). Further comparative genomics analysis is required to better understand factors influencing branch length distribution patterns in phylogenetic trees. We tested our new method on real and simulated phylogenies and observed improved robustness to wider ranges of branch lengths and increased ability to detect convergent evolutionary rate shifts. Our new method offers increased robustness to the inclusion of distantly-related species with long branch lengths in our phylogeny, namely non-placental mammals. When we compared results from an analysis using only placental mammals and an analysis that included non-placental mammals using both our original and our updated methods, we found that our new method, unlike our original method, is unimpaired by the inclusion of non-placental mammals. By improving our methods robustness to inclusion of long branches, we increased the methods applicability to a broader range of species and hence a broader range

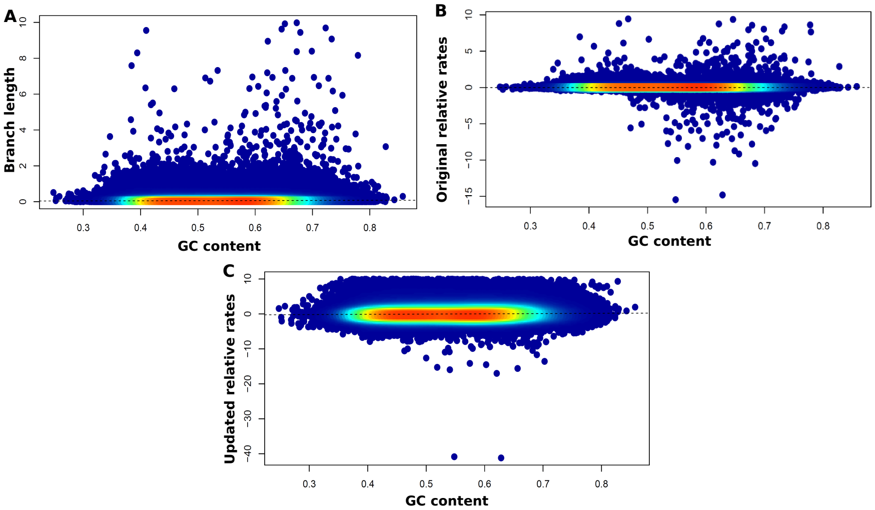


Figure 3.4.1: Scatter plot of trends between GC content of sequences and measures of sequence divergence in phylogenetic trees. GC content of extant orthologs from alignments of 19,149 protein-coding genes, are plotted them against the branch lengths (A), relative rates of the corresponding branches computed using original (B), and updated methods (C) respectively.

of convergent phenotypes. Additionally, our new methods increased power could enable us to discover more convergently evolving genomic elements. One particular incentivizing example for these improvements is the recent efforts to sequence the northern marsupial mole, a completely blind mammal (Archer *et al.*, 2011). When considering using subterranean species to find genes and enhancers associated with vision, the ability to include the non-placental marsupial mole along with the other non-placental mammals in our dataset will allow for more power in a scan for vision-specific genetic elements showing convergent regression in the five blind mammals.

In addition to testing our method on real data, we also developed a simulation-based strategy to represent a true positive case of convergent evolution. Our simulations follow a similar approach to simulating RNA-seq counts where simulated rates are essentially capturing the number of substitutions that occur along a branch (Di *et al.*, 2011). We showed that our new method demonstrates improved detection of rate shifts both when foreground species occupy long, high-variance branches and when foreground species occupy short, low-variance branches. This allows the method to detect convergent rate shifts given a variety of potential configurations of convergently-evolving species. The types of simulations we developed are essential because relatively few concrete instances of sequence-level evolutionary convergence exist, so biologically accurate simulations of such evolution are essential to rigorously test methods that detect shifts in evolutionary rates. One simplification of our simulation method is that all species are present in all simulated trees, which is not the case in real genomic data because of genomic element gain and loss across species. However, maintaining constant species composition in our simulated trees should have little impact on our ability to compare our methods because we expect both to be equally impacted by species presence and absence. A second simplification is that we assume all convergently-evolving species have the same phylogenetic relatedness, i.e. each foreground branch is an independent instance of convergent evolutionary rates. We would like to be able to answer questions about our methods power given more complex phylogenetic configurations. Developing methods to answer those types of questions will require a much higher degree of complexity in our simulations, but it will also allow us to determine which species to add to our genomic datasets to increase our power to find convergently evolving genomic features.

Our improved method has proved valuable for detecting genomic elements associated with two binary traits - subterranean-dwelling or not, and marine-dwelling or not - and we will extend our method for use in convergent continuous traits and non-binary discrete traits. We will also assemble complementary analyses to assess the robustness and power of each method. By extending the scope of our method to non-binary traits, we will expand the potential search-space of our method to a plethora of new convergent phenotypes. Our overarching goal is to develop an entire suite of methods that can utilize any conceivable phenotypes as inputs to accurately and robustly identify convergently evolving genomic elements.

4.0 An integrated map of protein co-evolution across five eukaryotic lineages

4.1 Introduction

Functional interactions between proteins underlie most biological processes at the cellular level. These interactions can occur at various scales ranging from physical interactions between pairs of proteins or binary PPIs, to co-complex interactions among participating subunits of macromolecular complexes (Jones and Thornton, 1996; Fields and Song, 1989; De Las Rivas and Fontanillo, 2010). Beyond these direct and in-direct physical interactions, functional links exist between proteins that are components of the same biological pathways, as a shared regulation of their function is necessary for successfully mediating cellular processes. A major focus of the post-genomic era has been the development of large-scale experimental and computational approaches to map the interactome - to characterize genome-scale PPI networks across multiple species (Havugimana *et al.*, 2012; Wan *et al.*, 2015; Marcotte *et al.*, 1999a,b; Huynen *et al.*, 2003).

The growing availability of genome sequences across closely related species allow for comparative genomic studies of proteins. This permits for co-evolutionary analysis at the molecular level studying the interdependence of evolutionary changes of protein sequences. The premise of co-evolutionary analysis to infer functional interactions between proteins relies on the idea that proteins under shared evolutionary pressures evolve in a codependent manner (Clark *et al.*, 2011; Fraser *et al.*, 2004). An extreme instance of this principle is the basis for methods predicting PPIs by matching phylogenetic profiles — interacting proteins tend to show similar patterns of presence/absence across a set of species (Pellegrini *et al.*, 1999).

Over the past years, substantial progress has been made at developing computational tools to predict PPIs via molecular co-evolution (De Juan *et al.*, 2013; Ramani and Marcotte, 2003). Such methods utilize protein sequences across divergent species, studying relationships between evolutionary rates (Clark *et al.*, 2011; Ochoa and Pazos, 2014). Candidate PPIs are predicted using pairs of protein sequences showing correlated rates of evolution on branches of phylogenetic trees (Pazos and Valencia, 2002; Clark *et al.*, 2011). Early applications using methods such as MirrorTree and ContextMirror, illustrated the utility to detect PPIs in E.coli using 14 prokaryotic genomes (Pazos and Valencia, 2002; Pazos *et al.*, 2008). More recent methods including MatrixMatchMaker (MMM) and Evolutionary Rate Covariation (ERC), have demonstrated the power of coevolutionary approaches to reveal functional links across diverse model organisms such as yeast, drosophila, as well as in humans (Tillier and Charlebois, 2009; Bezginov *et al.*, 2013; Clark *et al.*, 2012b, 2013, 2012a; Priedigkeit *et al.*, 2015).

Evolutionary Rate Covariation measures the statistical covariation of gene-specific rates of sequence evolution in pairs of genes across a set of species (Clark and Aquadro, 2010; Clark *et al.*, 2012a). Participating genes of a pathway experience shared evolutionary constraint to maintain its functionality, thus responding to any changes in selective pressure through parallel changes in their evolutionary rates (Clark *et al.*, 2012b). Strong signatures of ERC have been reported in biological pathways across diverse model organisms, including proteins involved in meiotic crossing over in yeast, fertilization proteins in abalones, proteins belonging to reproductive pathways in Drosophila (Clark *et al.*, 2009, 2013; Findlay *et al.*, 2014). Application of ERC to mammalian genome sequences revealed strong signatures of coevolution among genes associated with genetic diseases in humans, furthermore, providing a complementary approach to predict candidate genes linked to disease (Priedigkeit *et al.*, 2015).

A major challenge in coevolutionary analysis of protein sequences is correcting for the contribution of non-specific factors influencing evolutionary rates of proteins (Clark *et al.*, 2011; Ochoa and Pazos, 2014). Such factors including local mutation rates, time since divergence etc., can impact estimation of gene-specifc evolutionary rates on branches of phylogenetic trees, and therefore at accurate inference of coevolutionary signatures between pairs of proteins (Sato *et al.*, 2005). Methods such as MirrorTree and MMM employ post-processing methods that correct for the background signal using the average coevolutionary signal across all pairs of protein trees (Pazos and Valencia, 2002; Pazos *et al.*, 2008; Tillier and Charlebois, 2009; Clark *et al.*, 2011). Earlier applications of ERC addressed this issue by calculating gene-specific rates of evolution termed relative evolutionary rates (Clark *et al.*, 2012a). The expected divergence for every branch of a phylogenetic tree is estimated using a genome-wide average. Individual gene-specific rates of evolution are subsequently estimated by correcting this expected divergence using linear regression (Wolfe and Clark, 2015; Chikina *et al.*, 2016).

In this study, we report two key contributions building upon previous applications of Evolutionary Rate Covariation to reveal functional links between proteins. First, we employ an improved and more robust method, RERconverge, to estimate gene-specific rates of evolution which are subsequently used to measure ERC between pairs of proteins Kowalczyk *et al.* (2019). RERconverge utilizes a combination of data transformation and weighted regression analysis to estimate robust gene-specific evolutionary rates Partha *et al.* (2019). Correcting the non-specific background variation in evolutionary rates of genes using this improved method, in turn offers an opportunity to robustly calculate rate covariation across larger collections of distantly related species. The second, and perhaps the more important contribution of this study, is the curation of an integrated map of protein co-evolution across five eukaryotic lineages namely mammals, vertebrates, flies, worms, and yeasts, using ERC. We perform preliminary investigations of using this integrated coevolutionary approach to detect functionally interacting proteins at various scales including pairwise protein associations, co-complex interactions in complexes displaying a range of evolutionary conservation, and finally in revealing interactions among genes contributing to genetic diseases in humans. By way of direct comparisons between the Integrated and Mammal-specific ERC analysis, we demonstrate the improved predictive power offered by the integrated approach to reveal functional links between proteins.

4.2 Materials and Methods

4.2.1 Calculating genome-wide phylogenetic gene trees

4.2.1.1 Mammals and Vertebrates Amino acid alignments for 100 vertebrate species were downloaded from the multiz alignment available at the UCSC genome browser (Haeussler *et al.*, 2019). For mammalian gene trees, we removed species from each alignment that are not represented in the mammalian species tree topology (Figure 4.2.1), after adding the Blind Mole-Rat ortholog of the corresponding gene sequence to this alignment as described in detail in (Partha *et al.*, 2017). For the vertebrate gene trees, we pruned each alignment to retain only the species represented in the vertebrate species tree topology. In each of these alignments, we removed low quality orthologs which contained fewer than 50 non-gap amino acid sites or less than 70% non-gap sites. Finally, we filtered out gene alignments which included fewer than 15 species. For each resulting amino acid alignment, we estimated branch lengths using the aaml program from the phylogenetic analysis using the maximum likelihood (PAML) package (Yang, 2007). Branch lengths were estimated under an empirical model of amino acid substitution rates with rate variability between sites modeled as a gamma distribution approximated with four discrete classes (for computational efficiency) and an additional class for invariable sites (aaml model Empirical + F) (Whelan and Goldman, 2001; Yang, 1996). Branch lengths were estimated on the mammalian species tree topology described in detail in Partha *et al.* (2019) (Meyer *et al.*, 2018; Partha *et al.*, 2019; Meredith *et al.*, 2011). For vertebrate gene trees, the species topology was inferred from the topology described in the UCSC genome browser (Haeussler *et al.*, 2019).

4.2.1.2 Flies and Worms The work described in this section was performed by Dr Jae Young Choi, New York University. For the fly dataset, protein coding sequences for species in the phylogeny represented in Figure 4.2.2B were downloaded from the Flybase website or the NCBI genome annotation website. For the worm dataset, protein coding sequences for species in the phylogeny represented in Figure 4.2.2C were downloaded from the Wormbase website. Presence of internal stop codons were ascertained and the sequence was removed if found. Genes represented by multiple transcripts were matched with the longest transcript.

Within each of these two taxonomic groups, orthologous groups of genes were inferred using the Orthofinder algorithm (Emms and Kelly, 2015). In every orthogroup (sets of genes that are orthologs and/or recent paralogs), paralogous genes were filtered out. Orthogroups represented by a minimum of 10 species were analyzed further. The PRANK aligner was used to align gene members in each orthogroup (Löytynoja and Goldman, 2008).

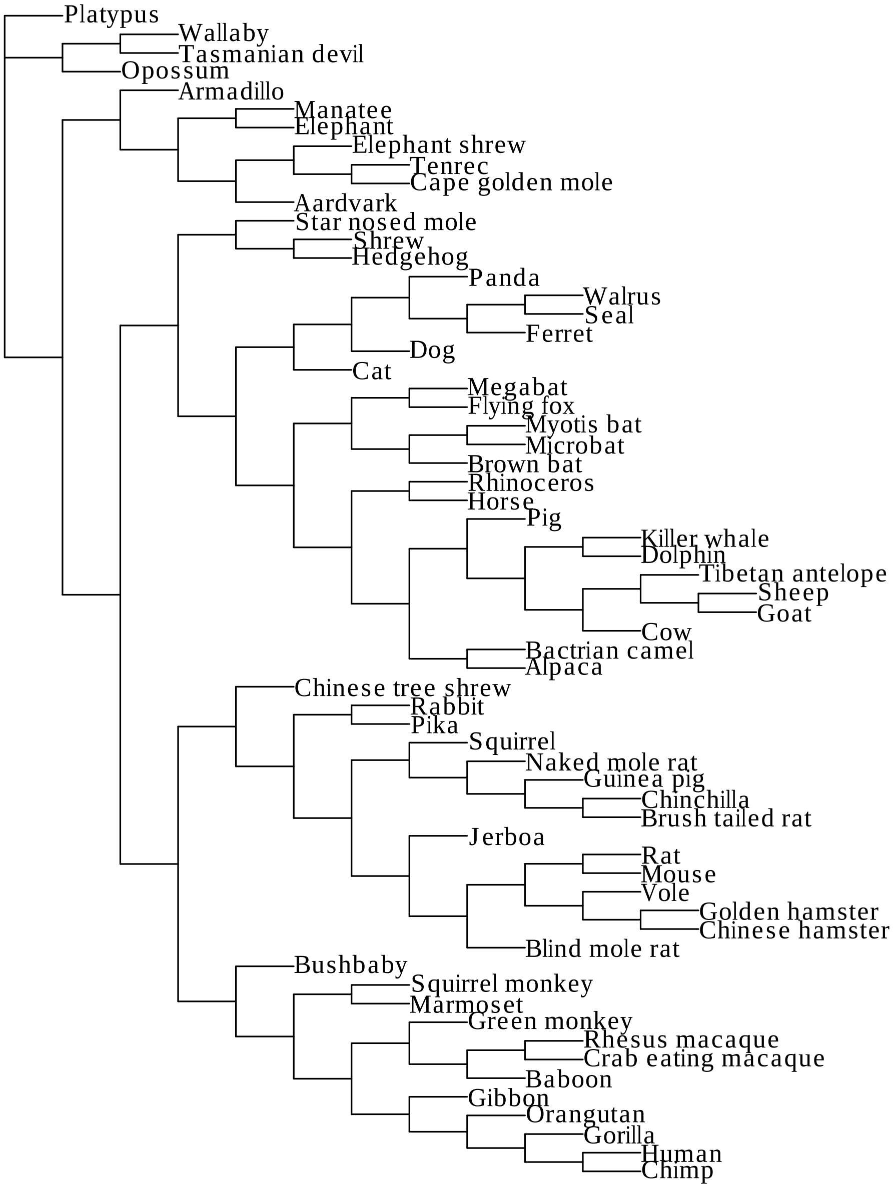


Figure 4.2.1: Species phylogeny across 63 mammals.

Branch lengths on a single fixed species topology (represented in Figure 4.2.2B and C, were estimated through the PAML aaml program (Yang, 2007) using the Whelan and Goldman (WAG) amino acid replacement matrix (Whelan and Goldman, 2001). The final species topology was inferred using a supertree approach combining individual orthogroup topologies that were estimated using RAxML (Stamatakis, 2014). Trees were combined using the matrix representation method implemented in phytools (Revell, 2012).

4.2.1.3 Yeasts Proteome-wide orthologous groups were constructed across 18 species of yeast represented in Figure 4.2.2D. Amino acid sequences across these species were downloaded from the Fungal Genome Research database (http://fungalgenomes.org/) and the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). Orthologous groups for each protein in S.cerevisiae were inferred for all other 17 species using InParanoid (Remm *et al.*, 2001). Only orthologous sequences passing a similarity score cut-off of 50 bits calculated by reciprocal best BLAST were retained. In groups of orthologs containing in-paralogs (orthologs arising due to duplication since speciation), only orthologs with a 100% confidence were considered. Finally, the program MUSCLE was used to align the 4459 orthologous groups of proteins (Edgar, 2004). Branch lengths on a single fixed species topology (represented in Figure 4.2.2D, were estimated through the PAML aaml program (Yang, 2007) using the Whelan and Goldman (WAG) amino acid replacement matrix (Whelan and Goldman, 2001). This species tree topology was inferred using the topology reported in Fitzpatrick *et al.* (2006).

4.2.2 Genome-wide ERC calculations

For each pair of gene trees, we calculate the Evolutionary Rate Covariation (ERC) using correlations of gene-specific relative evolutionary rates (RERs). Gene-specific relative evolutionary rates (RER), reflect sequence divergence on a particular branch after removing effects of non-specific factors affecting divergence including time since speciation and mutation rate. Furthermore, we use a combination of statistical approaches including data transformation and weighted linear regression, to robustly estimate relative evolutionary rates. This procedure, described in detail in Partha *et al.* (2019) and Kowalczyk *et al.* (2019), improves RER robustness to several factors introducing outliers in the dataset, such as the presence of distantly related species in the phylogeny. Using this updated method to estimate gene-specific rates, we observed improved handling mean-variance trends in all five tree datasets (Figures

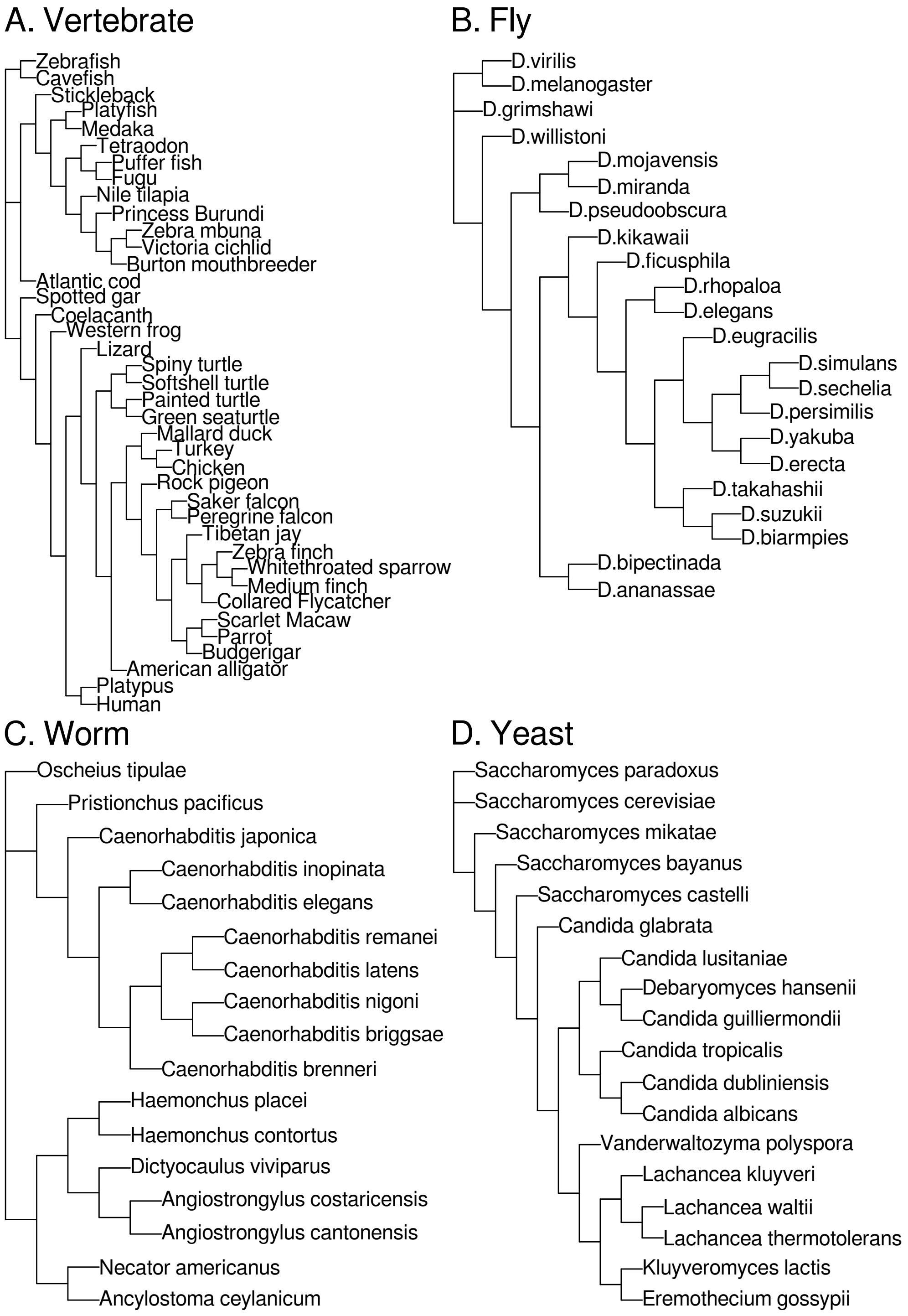


Figure 4.2.2: Species phylogeny across A. Vertebrates B. Flies C. Worms and D. Yeasts

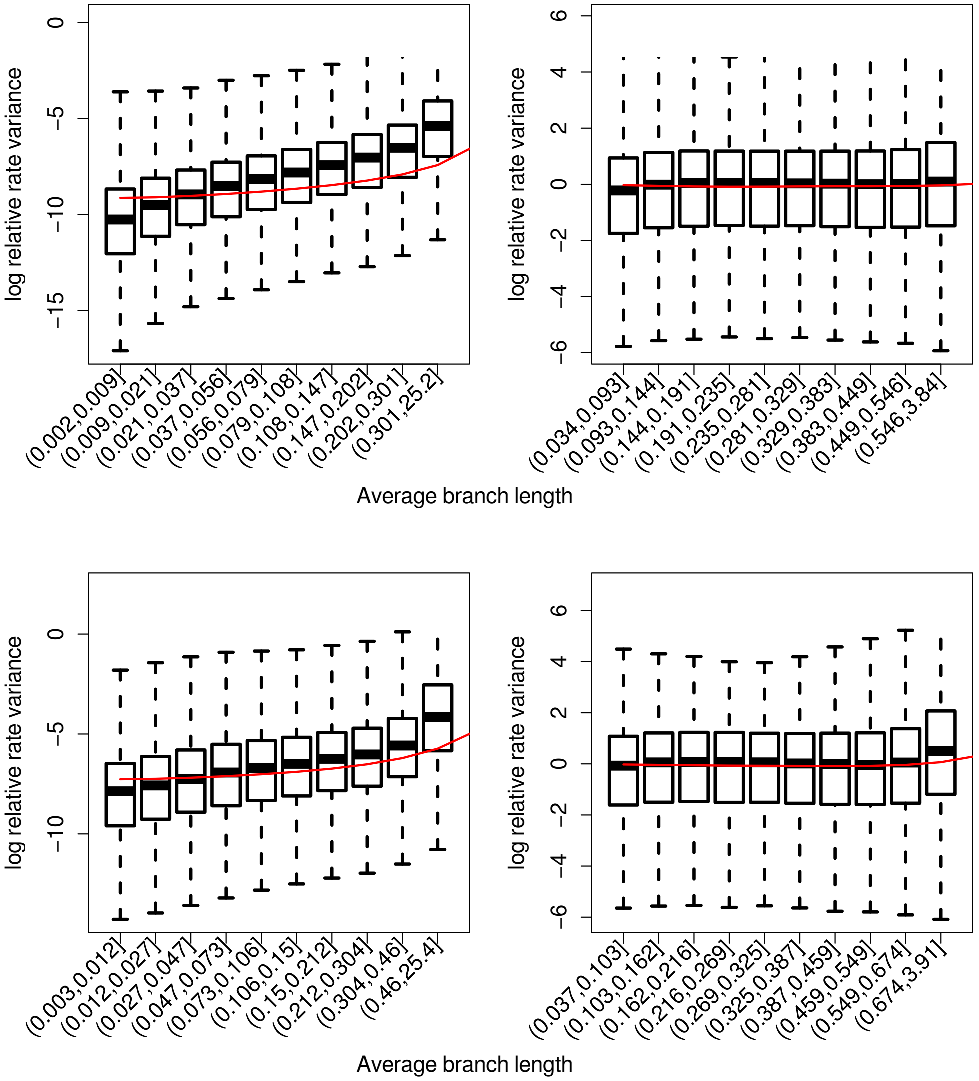


Figure 4.2.3: Mean-variance trends in relative evolutionary rates estimated using original and updated methods offered by RERconverge across Mammal (A, B) and Vertebrate (C, D) trees. Updated method to estimate relative rates shows near constant variance across all branch lengths compared to the original method.

.

4.2.3 and 4.2.4). We finally calculate the ERC as the Winsorized Pearson correlation coefficient between the rates of two genes, for gene pairs with at least 10 branches in common. The winsorization caps the maximum and minimum RER for every gene at their third most

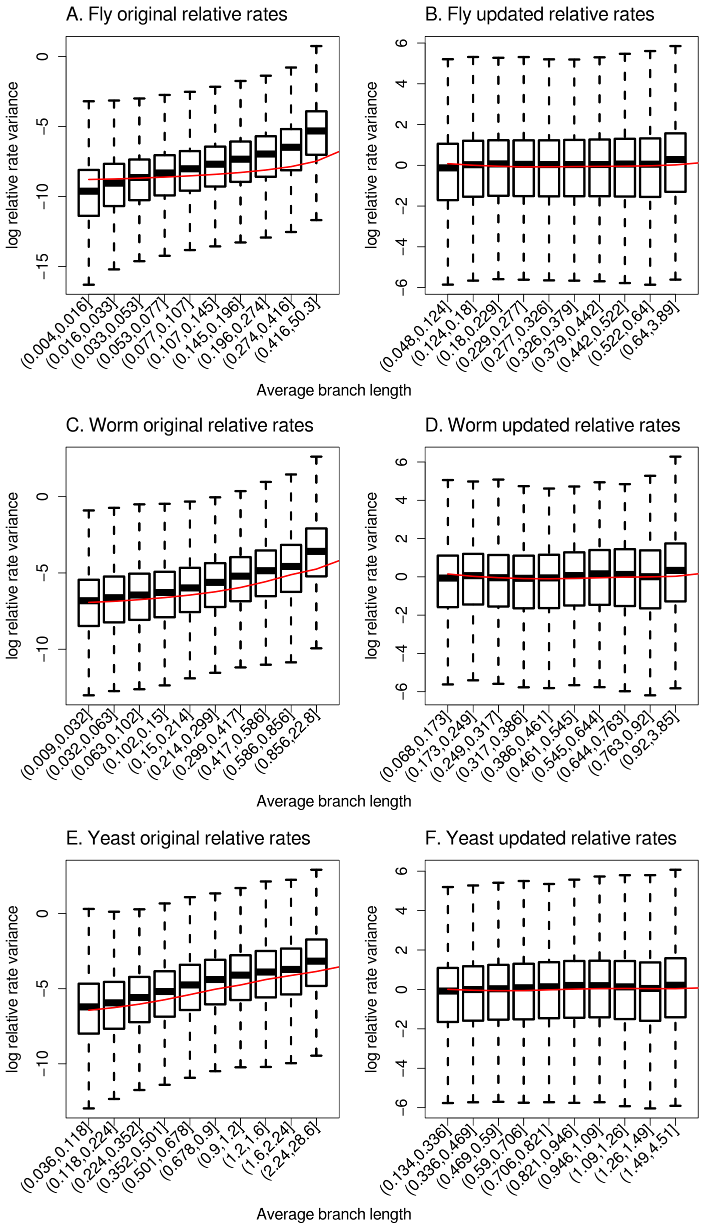


Figure 4.2.4: Mean-variance trends in relative evolutionary rates estimated using original and updated methods offered by RERconverge across Fly (A, B), Worm (C, D), and Yeast (E, F) trees. Updated method to estimate relative rates shows near constant variance across all branch lengths compared to the original method.

.

extreme values. This process precludes spuriously high correlations arising from extreme RERs in outlier branches. Finally, we apply the Fisher z-transformation to each ERC value to account for the differences in variance arising from varying number of common branches used in the ERC calculation (see Results 4.3.1)

4.2.3 Calculating Integrated ERC based on orthology mapping across reference species in each dataset

We calculated genome-wide integrated ERC values based on the five individual ERC datasets from mammals, vertebrates, flies, worms and yeasts. For each human ortholog of the mammalian gene trees, we identified the ortholog corresponding to the reference species in the other four tree datasets. In the case of vertebrates, this process is trivial due to the inclusion of the human orthologs in each gene tree. For the fly, worm, and yeast trees, the reference species were D.melanogaster, C.elegans, and S.cerevisiae respectively. The ortholog assignments for each human (hg19) gene in the respective reference species were identified using the InParanoid ortholog database (Sonnhammer and Östlund, 2015). These orthologs are converted from their UniProt IDs to gene symbols using the appropriate mapping files in the UniProt database for the four reference species (Bateman, 2019). Across these individual maps between human and the respective reference species, only 1:1 ortholog maps were retained. In total, we identified 4764, 1892, and 1878 genes with human orthologs in the flies, worms, and yeasts, respectively. Using these inferred ortholog assignments to map gene identifiers across the tree datasets, we subsequently calculate the integrated ERC for every pair of genes, as the sum of the ERC values across the five datasets.

4.2.4 Filtering protein pairs showing strong sequence similarity and duplicated genes

Ortholog assignments across species typically involve as a first step, computing pairwise similarity scores between gene sequences. Approaches including InParanoid, subsequently identify reciprocal best hits between the genomes of two species to construct the orthologous groups (Sonnhammer and Östlund, 2015). The presence of duplicated genes showing strong sequence similarity in the genome can result in nearly identical sequence alignments, and consequently highly correlated branch lengths in gene trees. Such instances of gene pairs will therefore show spuriously high ERC values. We identified and removed such protein pairs that show strong sequence similarities using gene-by-gene BLAST (Altschul *et al.*, 1990). Using nucleotide sequences of every pair of genes (x, y), we calculated the BLAST Bit scores (using local BLAST) reflecting the extent of sequence similarity with a minimum word size match of 7 and a percent identity of 50%. We then calculated a Relative Blast Bit Score as follows:

$$Relative\,Blast\,BitScore(x,y) = max(\frac{BitScore(x,y)}{BitScore(x,x)}, \frac{BitScore(x,y)}{BitScore(y,y)}) \tag{4.1}$$

where BitScore(x,x) and BitScore(y,y) reflect the Bit scores from the alignments involving gene x with itself, and gene y with itself, respectively. We only retained protein pairs showing a Relative Blast Bit Score lesser than 10%, removing 15,500 gene pairs in total. Our choice of this threshold was informed by the distribution of the Integrated ERC values with respect to the Relative Blast Bit Score (shown in Figure 4.2.5). Additionally, we removed any gene pairs involving genes belonging to two large gene families that have evolved due to repeated gene duplications followed by divergence, namely the Zinc finger genes (genes associated with the GO term "Zinc Finger protein") and the Olfactory receptor genes (genes associated with the GO term "Olfactory receptor") (Emerson and Thomas, 2009; Nei and Rooney, 2005).

4.2.5 Curation of publicly available pairwise protein associations and multiprotein complexes

To investigate coevolutionary signatures across predicted pairwise associations among proteins in humans, we downloaded PPI predictions from the STRING database (`https://bit.ly/2JE37L5`). Interactions among proteins are characterized using multiple attributes including mode of interaction, confidence scores etc (Szklarczyk *et al.*, 2017). Ensembl gene identifiers were mapped to gene symbols using the UCSC genome browser (Haeussler *et al.*, 2019). Interactions in STRING that are directional, for instance activation and inhibition, were summarized using the maximum score reported across either direction. CORUM multiprotein complexes investigated in this study were downloaded from

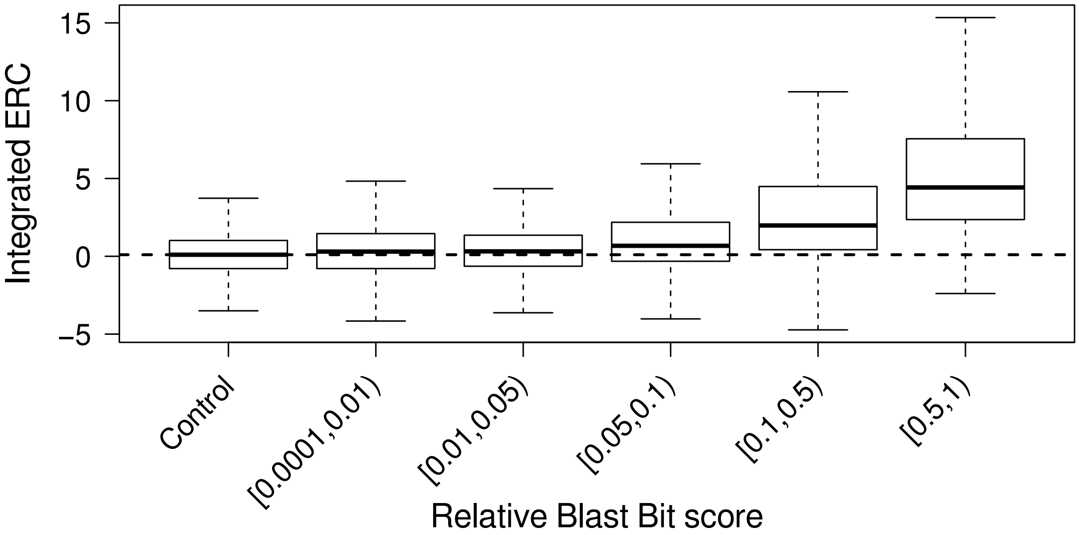


Figure 4.2.5: Distribution of Integrated ERC values with respect to Relative Blast Bit scores. Protein pairs showing strong sequence similarity ($Relative\,Blast\,BitScore >= 0.1$) have an elevated Integrated ERC distribution compared to 1 million Control pairs. Control pairs have no significant local BLAST hits at an E-value threshold of 10

`http://mips.helmholtz-muenchen.de/corum/#download` (Giurgiu *et al.*, 2019). Duplicate complexes (based on the name attribute) were summarized to the complex with higher numbers of participating genes. Ancient metazoan macromolecular complexes used for the study were downloaded from Supplementary Table 4 in Wan *et al.* (2015). In both CORUM and the Wan *et al.* (2015) datasets, only complexes with at least 3 subunits mapped to our list of genes were retained for further analyses. Dataset corresponding to OMIM disease gene groups were downloaded from Supplementary Table 2 in Priedigkeit *et al.* (2015).

4.2.6 Simulating control protein pairs and complexes with matched representation across datasets

For a given list of binary PPIs, we compare the distribution of ERCs (Integrated and/or mammal) with a null distribution generated using control protein pairs randomly sampled under certain constraints. We account for differences in representations of pairs of genes across tree datasets, by generating a control pair for each true pair that has an equal num-ber of tree datasets consisting both genes belonging to the pair. This accounts for any bias in the distribution of ERC values introduced by the differential conservation of genes across the five eukaryotic lineages in the study. To sample control complexes, we similarly applied a constrained random sampling procedure ensuring that the datasets containing the participating genes are matched. For instance, controls for complexes containing genes present only in mammal and vertebrate trees, will contain complex members randomly sampled from the subset of genes present only in mammal and vertebrate trees.

4.3 Results

4.3.1 Fisher z-transformation accounts for differences in ERC distributions arising due to varying numbers of common branches

The genome-wide ERC values in each of the five tree datasets consistently reflect variations in the distribution arising from differences in the number of common branches used to calculate ERC. An example illustration of this problem is shown in the case of ERC values from the mammal dataset (Figure 4.3.1). We see that the ERC values calculated based on fewer common branches show higher variance. In order to correct for this variation, we performed a variance-stabilizing transformation using the Fisher z-transformation. Figure 4.3.1 shows that Fisher z-transformed ERC values have stable variance with respect to the number of common branches. The application of the Fisher z-transformation to the other four ERC datasets yield similar results (Figure 4.3.2). Henceforth, every reference to the term ERC is considered to mean the Fisher z-transformed ERC.

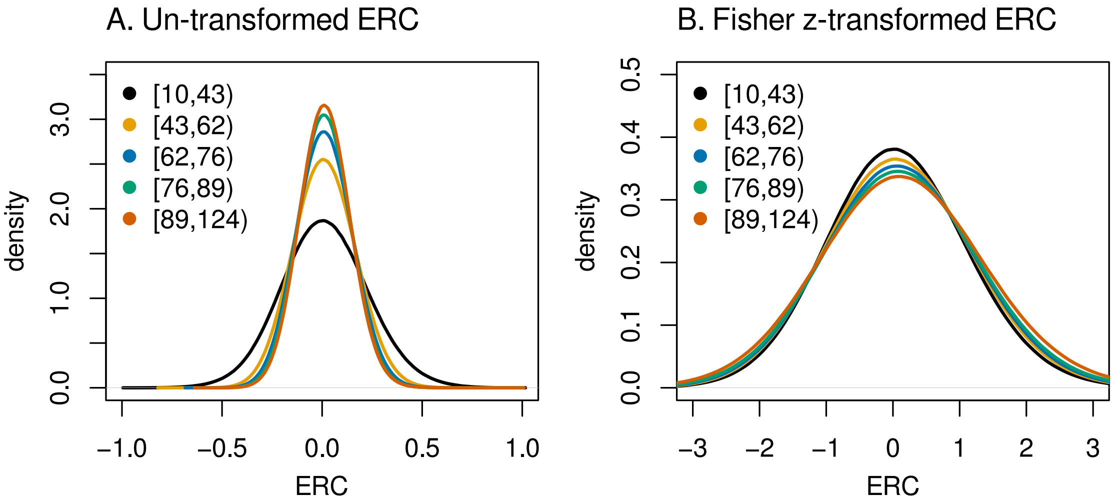


Figure 4.3.1: Fisher z-transformation of ERC values produces stable variance wih respect to the number of common branches. Panels A and B correpsond to distribution of mammal ERC values prior to and after Fisher z-transformation respectively. ERC values are binned according to the number of common branches used to calculate ERC. Bins have equal numbers of protein pairs.

4.3.2 Integrated ERC offers improved power to predict pairwise associations among human proteins

We investigated the power of coevolution to predict protein-protein interactions using our genome-wide Integrated ERCs which combine the signatures of coevolution across the five phylogenetic tree datasets used in the study. Using the STRING database, we downloaded human pairwise protein associations with strong confidence scores computed based on the extent of evidence supporting the interaction (Methods 4.2.5). STRING PPIs are predicted based on multiple sources including experimental evidence, pathway knowledge from manually curated databases, co-expression analysis, text-mining approaches to uncover semantic links etc (Szklarczyk *et al.*, 2017). We calculated the distribution of Integrated ERC values the among these STRING PPIs, contrasting the resulting distributions for high confidence STRING PPIs with randomly sampled representation-matched control pairs. We

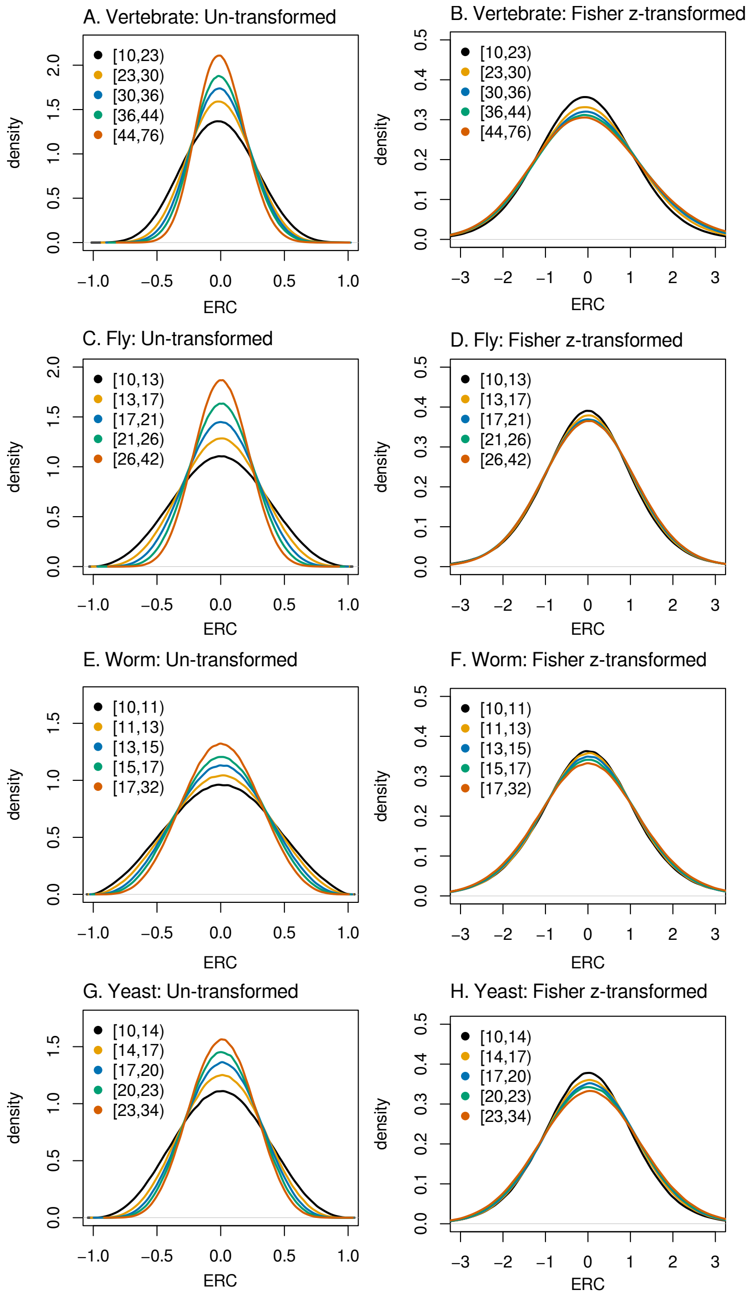


Figure 4.3.2: Fisher z-transformation of ERC values in vertebrate, fly, worm and yeast datasets. ERC values are binned according to the number of common branches. Bins have equal numbers of protein pairs.

observed an elevation of Integrated ERC in STRING protein pairs with the strongest confidence scores (Figure 4.3.3). Additionally, we see that as the confidence score for a given PPI increases, there is a concordant increase in the Integrated ERCs. We further investigated

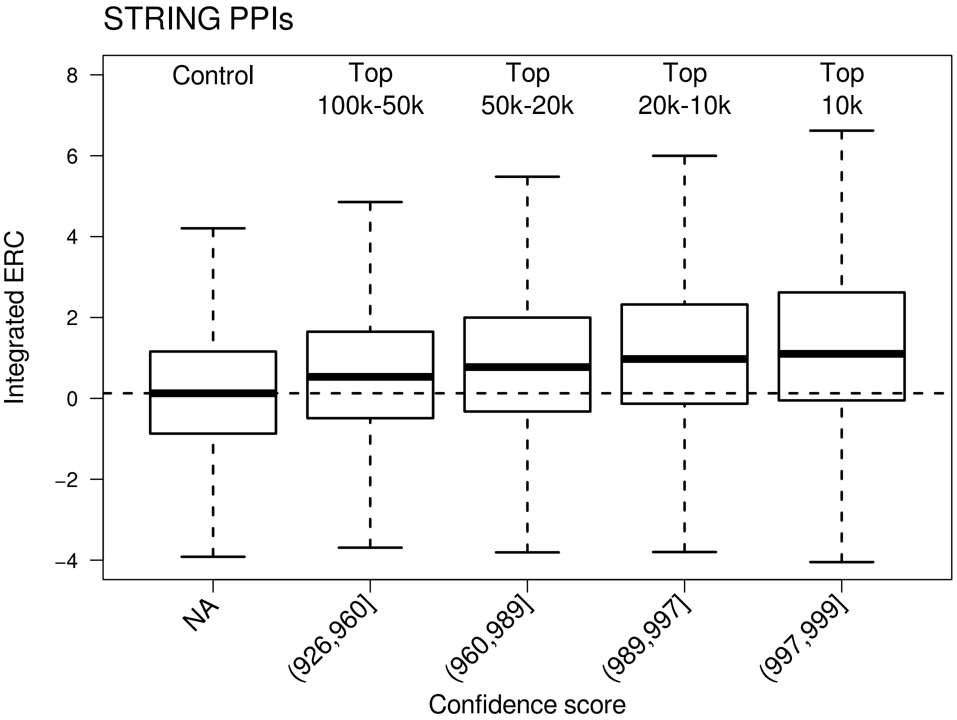


Figure 4.3.3: Integrated ERC values are elevated for STRING binary PPIs with the strongest confidence scores. PPIs are binned into categories of Top 100,000-50,000, Top 50,000-20,000, Top 20,000-10,000, and Top 10,000 pairs based on confidence scores. Control interactions containing 5x pairs of proteins were randomly sampled ensuring matched representation across ERC datasets (Methods 4.2.6). All classes of STRING PPIs shown here are significantly different from Control pairs (Wilcoxon rank sum test, $P < 2.2e - 16$)

the differential enrichments of the Integrated ERCs depending on the mode of interaction of the STRING PPIs. The interactions are classified into one or more of the following six categories activation, binding, expression, inhibition, post-translational modification, and reaction. Among these seven categories, we see proteins pairs characterized as interacting through the following modes – activation, binding, and reaction, typically show stronger elevation in the Integrated ERC values, as reflected by the power to distinguish these PPIs from control PPIs using ERC (Figure 4.3.4).

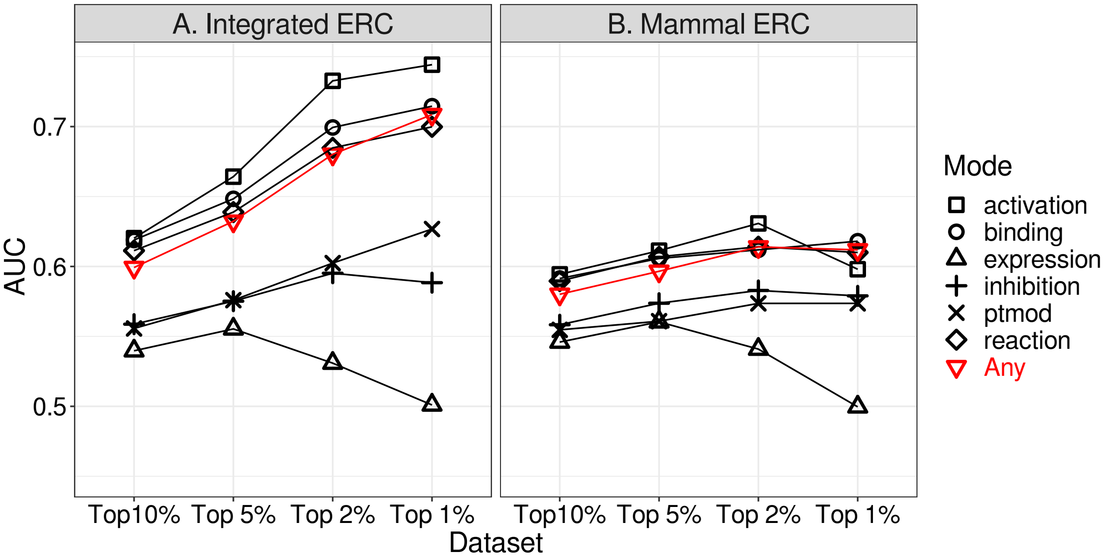


Figure 4.3.4: Distribution of Area under ROC curves comparing the performance of Integrated (Panel A) and Mammal ERC (Panel B) at detecting high-confidence (up to Top 10%) STRING PPIs across seven modes of interaction.

4.3.3 Integrated ERC allows for increased power to detect co-complex interactions

4.3.3.1 CORUM human protein complexes Cellular functions are mediated through interactions among proteins that are a part of protein complexes. The Comprehensive Resource of Mammalian Protein Complexes (CORUM) database provides the largest manually curated set of experimentally characterized protein complexes in mammals (Giurgiu *et al.*, 2019). We benchmarked the power of our coevolutionary method to predict co-complex interactions using the CORUM human protein complexes. We identified a set of 1581 protein complexes with at least 3 genes. For each of these complexes, we calculate the mean Integrated ERC among all pairs of participating genes. We then simulated 100,000 representation-matched random complexes with matched numbers of genes using which we compute the null distribution for the mean Integrated ERCs (see Methods 4.2.6). We calculated an empirical p-value for each complex, which represents the probability of observing an equivalent or more extreme mean Integrated ERC in the simulated complexes. We identified 326 CORUM human complexes that were significant at a FDR of 5%, showing elevated signatures of Integrated ERC among complex subunits. We compared the predictive power of Integrated ERC to detect CORUM co-complex interactions with that of the ERC calculated using just mammalian trees. We observe a larger excess of lower empirical p-values calculated based on mean Integrated ERCs in comparison to mean Mammal ERCs (Figure 4.3.5), as well as significantly more CORUM complexes identified at a FDR of 5% (Integrated ERC: 326 complexes, Mammal ERC: 172 complexes). Integrated ERC serves as a more useful predictor for detecting CORUM complexes among simulated random complexes, offering consistently higher precision over Mammal ERC (see Figure 4.3.5). Table 6 provides a list of the top 20 CORUM complexes showing the strongest signatures of coevolution.

4.3.3.2 Ancient metazoan macromolecular complexes Expanding from the analysis on human protein complexes, we sought to understand the utility of our coevolutionary method in predicting co-complex interactions in complexes that show deeper evolutionary conservation. To this end, we identified a set of 487 ancient metazoan macromolecular complexes curated by Wan *et al.* (2015). These soluble multiprotein complexes were identified using a combination of biochemical fractionation and quantitative mass spectrometry among diverse metazoan model systems. These complexes display a range of evolutionary conservation, ranging from ancient eukaryal modules, to metazoan-specific modules (Wan *et al.*, 2015). We investigate the extent of coevolution among protein pairs within these ancient complexes, undertaking an approach similar in principle to the CORUM complexes. We compared the mean Integrated ERC among complex members for each complex, to its empirical null obtained from 100,000 representation-matched simulated random complexes (see Methods 4.2.6). Using the empirical p-values which reflect how extreme the mean Integrated ERCs are (Figure 4.3.6), we identified 110 complexes significant at a FDR of 5%. Repeating the analysis using Mammal ERCs instead of Integrated ERCs, we discovered 36 complexes significant at a FDR of 5%. Similar to our analysis using the CORUM complexes, we observe

ID	Name	nGenes	Mean Integrated ERC	Mean Mammal ERC	P value Integrated	P value Mammal	Q value Integrated	Q value Mammal
23	Nup 107-160 subcomplex	9	5.209	1.829	1e-06	1e-06	1.86e-05	5.74e-05
25	Anaphase-promoting complex	8	5.099	1.009	1e-06	0.0016	1.86e-05	0.0207
40	CCT complex	8	3.681	1.174	1e-06	0.00051	1.86e-05	0.00922
41	NDC80 kinetochore complex	4	5.229	4.68	1e-06	1e-06	1.86e-05	5.74e-05
48	Condensin I-PARP-1-XRCC1 complex	7	3.5	0.982	1e-06	0.00278	1.86e-05	0.0302
49	Condensin II	5	6.421	2.156	1e-06	2e-05	1.86e-05	0.000795
50	COG complex	8	2.999	1.202	1e-06	0.0003	1.86e-05	0.00604
55	Respiratory chain complex I, mitochondrial	36	1.857	1.176	1e-06	1e-06	1.86e-05	5.74e-05
61	PA28-20S proteasome	16	1.662	0.387	1e-06	0.0271	1.86e-05	0.132
81	Mediator complex	32	1.055	0.429	1e-06	0.00034	1.86e-05	0.00676
115	28S ribosomal subunit, mitochondrial	30	1.31	0.762	1e-06	1e-06	1.86e-05	5.74e-05
116	55S ribosome, mitochondrial	77	1.225	0.59	1e-06	1e-06	1.86e-05	5.74e-05
117	DNA-PK-Ku complex	3	9.558	3.96	1e-06	6.01e-05	1.86e-05	0.00183
118	39S ribosomal subunit, mitochondrial	47	1.195	0.56	1e-06	1e-06	1.86e-05	5.74e-05
123	Spliceosome	136	1.276	0.556	1e-06	1e-06	1.86e-05	5.74e-05
132	ORC complex (origin recognition complex)	6	3.502	1.71	1e-06	0.00013	1.86e-05	0.00331
140	BASC complex	12	2.303	0.687	1e-06	0.00128	1.86e-05	0.0174
154	TFIID complex	10	2.277	1.259	1e-06	3e-05	1.86e-05	0.00108
155	TFIID complex, B-cell specific	11	2.18	1.144	1e-06	2e-05	1.86e-05	0.000795

Table 6: Top 20 CORUM complexes showing elevated Integrated ERC

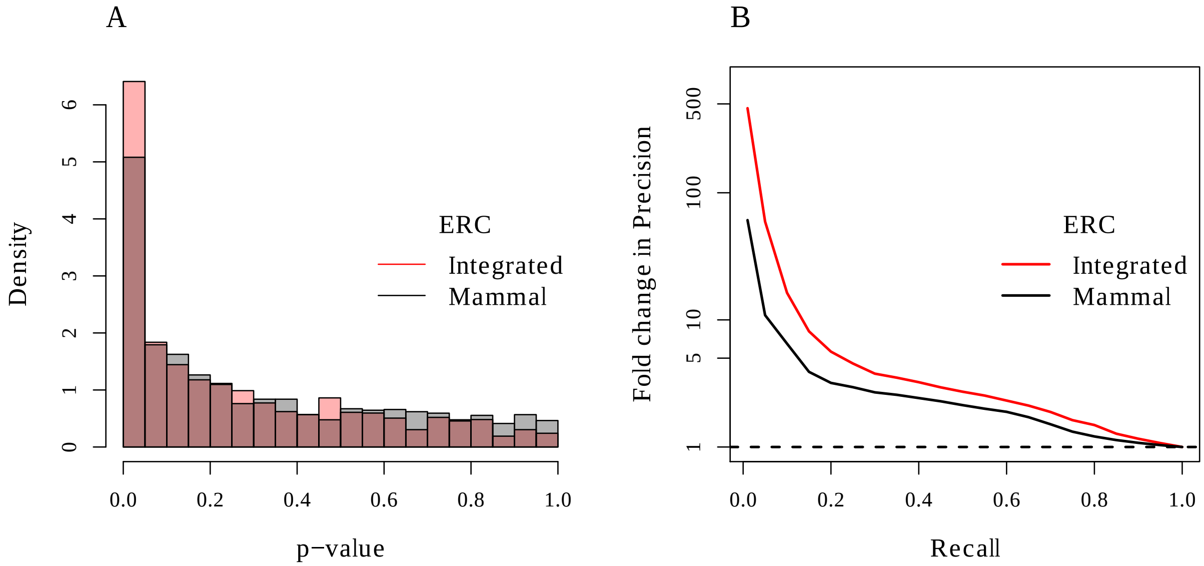


Figure 4.3.5: Strength of coevolution across CORUM human complex subunits. A. Distribution of empirical p-values for CORUM complexes reflecting probability of observing equivalent or higher mean ERC values in 100,000 simulated complexes. There is a larger excess of lower empirical p-values observed using mean Integrated ERCs. B. Fold change in Precision vs Recall curves describing power of Integrated and Mammal ERC values to predict CORUM complexes. Fold change in Precision is calculated as the Precision relative to Precision at 100% Recall. Integrated ERC offers higher power to distinguish CORUM complexes from simulated complexes.

that Integrated ERC is more precise at detecting co-complex interactions in these ancient metazoan complexes (see Figure 4.3.6).

4.3.4 Integrated ERC reflects strong associations among genes in disease gene groups

One previous exploration by Priedigkeit *et al.* (2015) highlighted the presence of elevated signatures of coevolution between genes associated with the same disease. Among genes

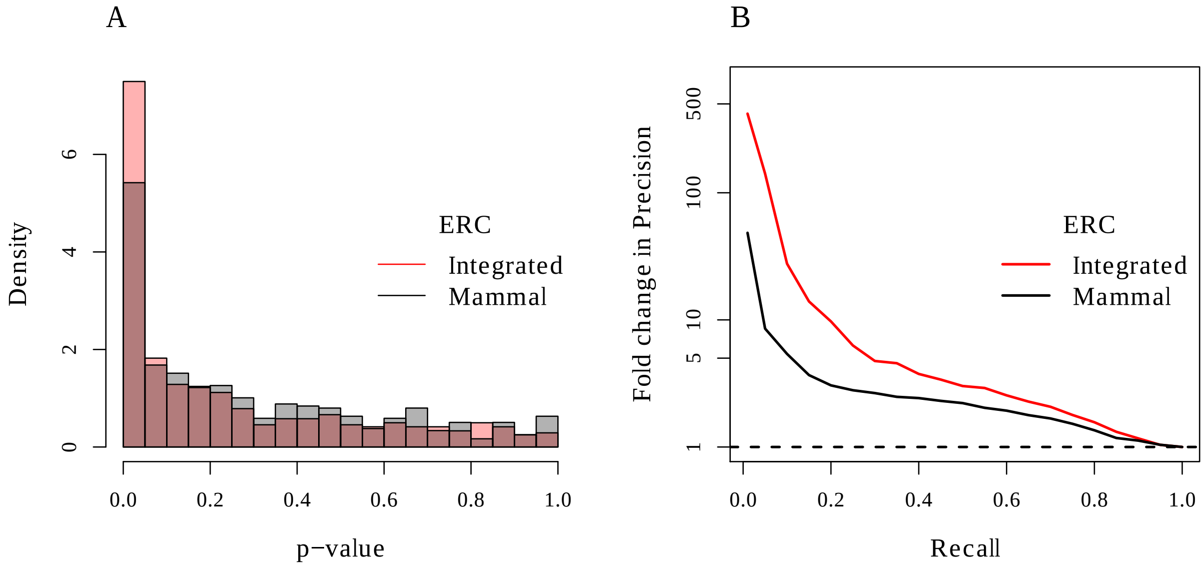


Figure 4.3.6: Strength of coevolution across ancient-metazoan-macromolecular complex subunits. A. Distribution of empirical p-values for these complexes reflecting probability of observing equivalent or higher mean ERC values in 100,000 simulated complexes. There is a larger excess of lower empirical p-values observed using mean Integrated ERCs. B. Fold change in Precision vs Recall curves describing power of Integrated and Mammal ERC values to predict ancient-metazoan-macromolecular complexes. Fold change in Precision is calculated as the Precision relative to Precision at 100% Recall. Integrated ERC offers higher power to distinguish ancient-metazoan-macromolecular complexes from simulated complexes.

involved in 310 distinct diseases curated by OMIM (Hamosh *et al.*, 2000), the study reported 40 disease gene groups (significant at a FDR of 5%) showing elevated ERC signatures between their participating genes. The dataset utilized in this study involved ERC calculations among 17487 genes across 33 mammals (Wolfe and Clark, 2015). Using our newer method to calculate ERCs and larger gene tree datasets, we directly benchmark the improvements in predictive power to detect coevolutionary signatures in OMIM disease gene groups. Using our larger set of gene trees, we identified a list of 320 disease gene groups with at least

ID	Disease group	nGenes	Mean Integrated ERC	Mean Mammal ERC	P value Integrated	P value Mammal	Q value Integrated	Q value Mammal
19	Asphyxiating thoracic dystrophy	4	6.118	2.387	1e-06	0.00032	1.77e-05	0.00328
28	Bardet-Biedl syndrome	14	2.409	0.804	1e-06	0.00015	1.77e-05	0.00217
46	Cataract	26	1.554	1.343	1e-06	1e-06	1.77e-05	2.65e-05
62	Combined oxidative phosphorylation deficiency	13	2.229	0.784	1e-06	0.00029	1.77e-05	0.00318
63	Complement deficiency	17	2.82	2.405	1e-06	1e-06	1.77e-05	2.65e-05
65	Congenital disorder of glycosylation	30	1.152	0.189	1e-06	0.189	1.77e-05	0.354
72	Cranioectodermal dysplasia	4	10.619	5.255	1e-06	1e-06	1.77e-05	2.65e-05
102	Fanconi anemia	16	1.427	1.138	1e-06	1e-06	1.77e-05	2.65e-05
150	Ichthyosis	16	1.489	1.373	1e-06	1e-06	1.77e-05	2.65e-05
164	Leigh syndrome	16	1.505	0.721	1e-06	0.0001	1.77e-05	0.00167
186	Meckel syndrome	10	1.747	0.661	1e-06	0.00693	1.77e-05	0.0501
192	Mental retardation	66	0.645	0.612	1e-06	1e-06	1.77e-05	2.65e-05
202	Mitochondrial complex deficiency	26	1.188	0.531	1e-06	4e-05	1.77e-05	0.000707
226	Night blindness	12	1.881	1.714	1e-06	1e-06	1.77e-05	2.65e-05
244	Peroxisome biogenesis disorder	13	2.116	0.132	1e-06	0.485	1.77e-05	0.614
284	Spherocytosis	5	7.612	7.064	1e-06	1e-06	1.77e-05	2.65e-05
303	Thrombophilia	14	1.966	1.073	1e-06	1e-05	1.77e-05	0.000212
314	Usher syndrome	10	2.343	1.72	1e-06	1e-06	1.77e-05	2.65e-05
56	Ciliary dyskinesia	12	1.257	0.848	1e-05	0.00014	0.000145	0.00212
84	Dysfibrinogenemia	3	6.531	5.494	1e-05	1e-06	0.000145	2.65e-05

Table 7: Top 20 OMIM Disease Gene Groups showing elevated Integrated ERC

3 genes. Following an approach similar to the analyses involving CORUM complexes and the ancient metazoan macromolecular complexes, we calculate empirical p-values for elevated mean Integrated ERCs based on 100,000 representation-matched random disease gene groups (see Methods 4.2.6, Figure 4.3.7 for distribution of empirical p-values). We observed 65 diseases showing elevated mean Integrated ERCs at a FDR of 5%. Performing the analysis using Mammal ERCs between gene pairs instead of Integrated ERCs yielded 43 diseases significant at a FDR of 5%. Comparing the predictive power of Mammal and Integrated ERCs, we observed a slightly higher precision upto 50% recall offered by the integrated coevolutionary approach (Figure 4.3.7). Table 7 provides a list of the top 20 OMIM disease gene groups showing the strongest signatures of coevolution.

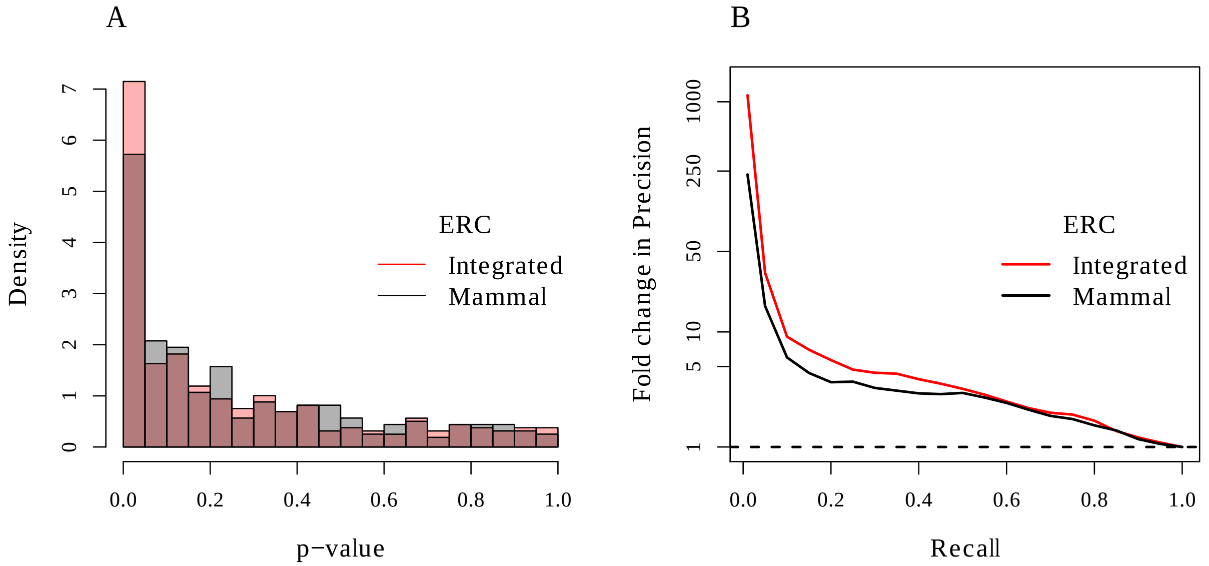


Figure 4.3.7: Strength of coevolution across contributing genes in OMIM disease gene groups (DGGs). A. Distribution of empirical p-values for these DGGs reflecting probability of observing equivalent or higher mean ERC values in 100,000 simulated gene groups. There is a larger excess of lower empirical p-values observed using mean Integrated ERCs. B. Fold change in Precision vs Recall curves describing power of Integrated and Mammal ERC values to predict OMIM disease gene groups (DGGs). Fold change in Precision is calculated as the Precision relative to Precision at 100% Recall. Integrated ERC offers higher power to distinguish OMIM DGGs from simulated complexes

4.3.5 Coevolving protein pairs show strong similarity in their Gene Ontology term associations

Pairs of proteins that are functionally related are likely to show strong similarity in their related functional annotations such as Gene Ontology (GO) term associations (Ashburner *et al.*, 2000; Carbon *et al.*, 2019). We compared the similarity in GO term associations for gene pairs having elevated Integrated ERC values with that of control pairs using the R package GOSemSim (Yu *et al.*, 2010). For each category of GO term annotations, namely, biological pathway, cellular component, and molecular function, GOSemSim quantifies the relatedness of GO terms accounting for their hierarchical structure. We removed any GO term annotation predicted using phylogenetic or computational methods to reduce any inherent bias across our methods and others. Across all three functional categories, we observed significantly higher values of semantic similarity for GO terms associated with protein pairs having higher values of Integrated ERCs (see Figure 4.3.8). Table 8 reports the Wilcoxon-rank sum p-values of the hypothesis test contrasting distributions of semantic similarities for top scoring versus control protein pairs.

Category	Top 1000	Top 10,000	Top 100,000
Biological Pathway	5e-24	3e-69	7e-171
Cellular Component	1e-58	9e-119	4e-140
Molecular Function	6e-12	8e-16	2e-07

Table 8: Wilcoxon-rank-sum p-values contrasting distributions of semantic similarities for top vs control pairs

4.4 Discussion

Methods to infer pairs of coevolving genes have proven valuable in revealing functional relationships among proteins (De Juan *et al.*, 2013). Evolutionary rate covariation which computes the correlation of gene-specific rates of sequence evolution across branches of phylogenetic trees is elevated for gene pairs involved in a variety of biological processes, and across taxonomic groups ranging from prokaryotes to mammals (Clark *et al.*, 2012b, 2013). In this study, we report an integrated map of protein coevolution across five taxonomic groups - mammals, vertebrates, flies, worms, and yeasts. Genome-wide coevolutionary datasets in each of these groups were created using a new and improved method to robustly estimate gene-specific rates of evolution (RERconverge), allowing for a more comprehensive sampling

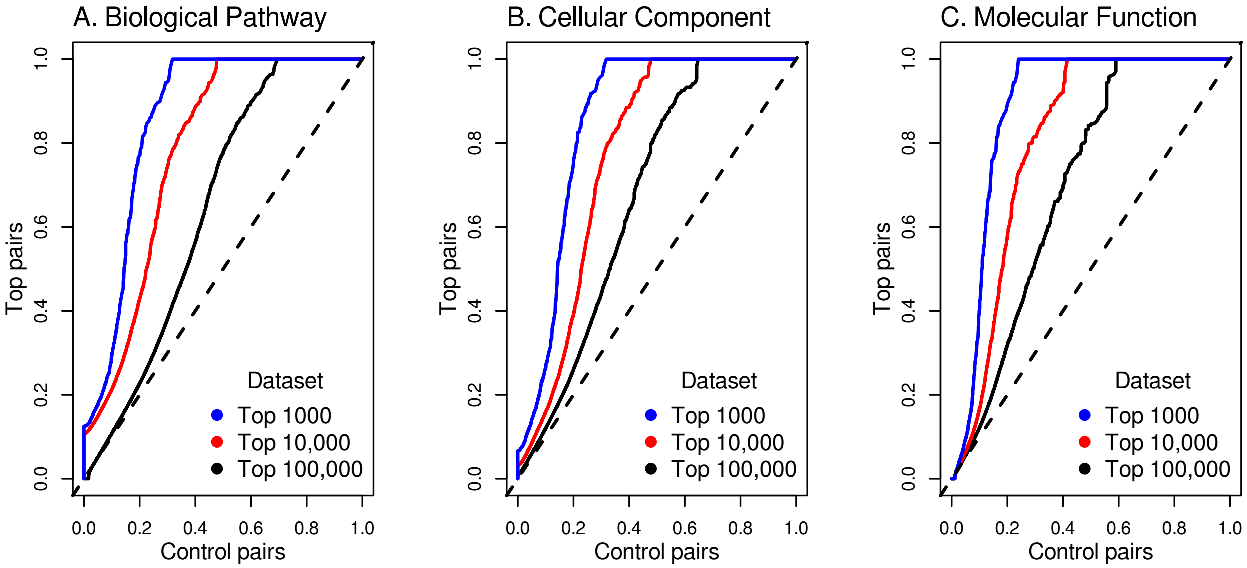


Figure 4.3.8: Protein pairs showing strong signatures of coevolution show significant semantic similarity in their Gene Ontology terms associations. Semantic similarity of GO terms associated with each pair of proteins is compared for pairs with the strongest Integrated ERC values genome-wide, and 500,000 randomly sampled representation-matched control pairs. Panels A, B, and C correspond to Q-Q plots comparing the quantiles of Semantic similarities between top pairs and control pairs, computed using GO categories Biological Pathway, Cellular Component, and Molecular Function respectively. We observed stronger enrichments for higher semantic similarities across all three categories for pairs with higher Integrated ERCs (or fewer top pairs).

of evolutionarily distant species. The mammalian trees were constructed across a sample of 63 species including four distantly related non-placental mammals in addition to the placentals. The evolutionary distance across species in vertebrate, worms, and yeast taxonomic groups are orders of magnitude larger in comparison to the mammal trees. This broad sampling of species within each lineage necessitates estimating evolutionary rates that are robust to statistical properties such as heteroscedasticity, allowed by methodological updates discussed in RERconverge (Kowalczyk *et al.*, 2019; Partha *et al.*, 2019). Using publicly available predictions of orthologs between reference species, we created a genome-wide map of protein coevolution combining the coevolutionary signatures across these five taxonomic groups.

Our integrated method to calculate ERC provides a significant improvement over the mammal-specific method to predict functional links between proteins at various scales. Using pairwise associations among proteins in humans reported in the STRING database, we observed elevated signatures of integrated ERCs specifically for high confidence predictions (Szklarczyk *et al.*, 2017). Further refining interactions based on the mode of action, we observed a preferential enrichment of Integrated ERC for the following interaction modes - activation in signaling pathways, physical binding, catalysis of subsequent reactions in metabolic pathways.

Expanding from pairwise associations among proteins, we tested the power of Integrated ERC to detect co-complex interactions relevant to human, as well as pan-metazoan multiprotein complexes. Using gold-standard complexes curated by the CORUM consortium and independent efforts to characterize evolutionarily conserved metazoan complexes, we demonstrated that an integrated coevolutionary approach is consistently superior at distinguishing multiprotein complexes (Giurgiu *et al.*, 2019; Wan *et al.*, 2015). These complexes have broad functional roles in the cell, such as the conserved Nuclear Pore Complexes (Nup107-160) that regulates cellular traffic between the cytoplasm and the nucleus (Walther *et al.*, 2003), and the BASC complex involved in the repair of aberrant DNA structures (Wang *et al.*, 2000). Genes contributing to genetic diseases in humans show strong signatures of Integrated ERC, reflecting functional links between the genes part of aberrant disease pathways (Priedigkeit *et al.*, 2015; Hamosh *et al.*, 2000). Using Integrated ERC, we observed a modest yet significant improvement in power to detect gene groups involved in diseases with strong genetic basis such as thalassemia as well as in complex diseases such as Coronary artery disease and Hepatitis. Predictions of the Integrated ERC approach can therefore further improve the value demonstrated by earlier applications of ERC at prioritizing candidate disease genes.

These preliminary findings pave the way for further analyses probing the nature of interactions uncovered by coevolutionary methods. One limitation of coevolutionary methods is that some functionally related proteins show well-correlated rates of evolution whereas others do not. To be able to calculate ERC between a pair of proteins, there needs to be sufficient changes at the sequence level such that there is quantifiable variation in their evolutionary rates. Consequently interactions involving proteins whose sequences evolve under strong purifying selection in the sampled species will not show elevated ERCs. In order to fully characterize which groups of functionally related proteins show ERC, characterizing coevolutionary signatures in the context of coexpression, co-localization, stable vs transient physical association, tissue-specific vs ubiquitous interactions etc. is necessary. Such investigations can better inform factors controlling rate variation and covariation, and enrich our current understanding of the model of cellular function from an evolutionary perspective (Lovell and Robertson, 2010; Clark *et al.*, 2012b).

Perhaps the most useful application of our integrated co-evolutionary framework is in the prediction of phenotypic associations for novel or uncharacterized genes. Particularly, conserved signatures of coevolution across evolutionary lineages provides a clear hypothesis for functional roles in lineages lacking phenotypic annotations, based on the annotation in well-characterized species. The predictive capacity of such guilt-by-association analyses can additionally be validated against annotated interactions across lineages. The genome-wide ERC datasets constructed in this study are available for the research community to investigate coevolutionary signatures of candidate groups of genes. These tools will be useful in generating functional hypotheses for the mechanism of action for genes that lack detailed functional annotation.

5.0 Conclusions

Evolutionary-based methods to predict phenotypic associations for genetic elements offer a complementary strategy to characterize components of cellular systems. Advances in sequencing technologies and comparative genomics methods have created an opportunity to infer associations between patterns of sequence evolution of genetic elements and phenotypic evolution. In chapter 1, we illustrate the power of using convergent changes in evolutionary rates to predict vision-related genetic elements. In addition to predicting protein-coding genes involved in visual pathways, we demonstrate the utility of our evolutionary-rates method to predict candidate non-coding elements regulating the expression of genes in a eye-specific manner. Our genome-wide predictions will be highly useful to the biomedical research community investigating genetics of eye disease and development. The results of this project have opened up multiple opportunities for collaborative research including - creating diagnostic panels for genotyping patients suffering from congenital eye disorders, developing effective vectors for precise targeting of gene therapies in the retina and other tissues of the eye. One limitation of our evolutionary-rates approach is that it does not detect sequences that are responsible for lineage-specific adaptations, as well as sequences that have lost or gained function due to evolutionary turnover. However, the utility of our approach in discovering eye-specific genetic elements showing changes preferentially in blind mammals stems from the observation that visual pathways show deep conservation with far more evolutionary distant species such as the zebrafish.

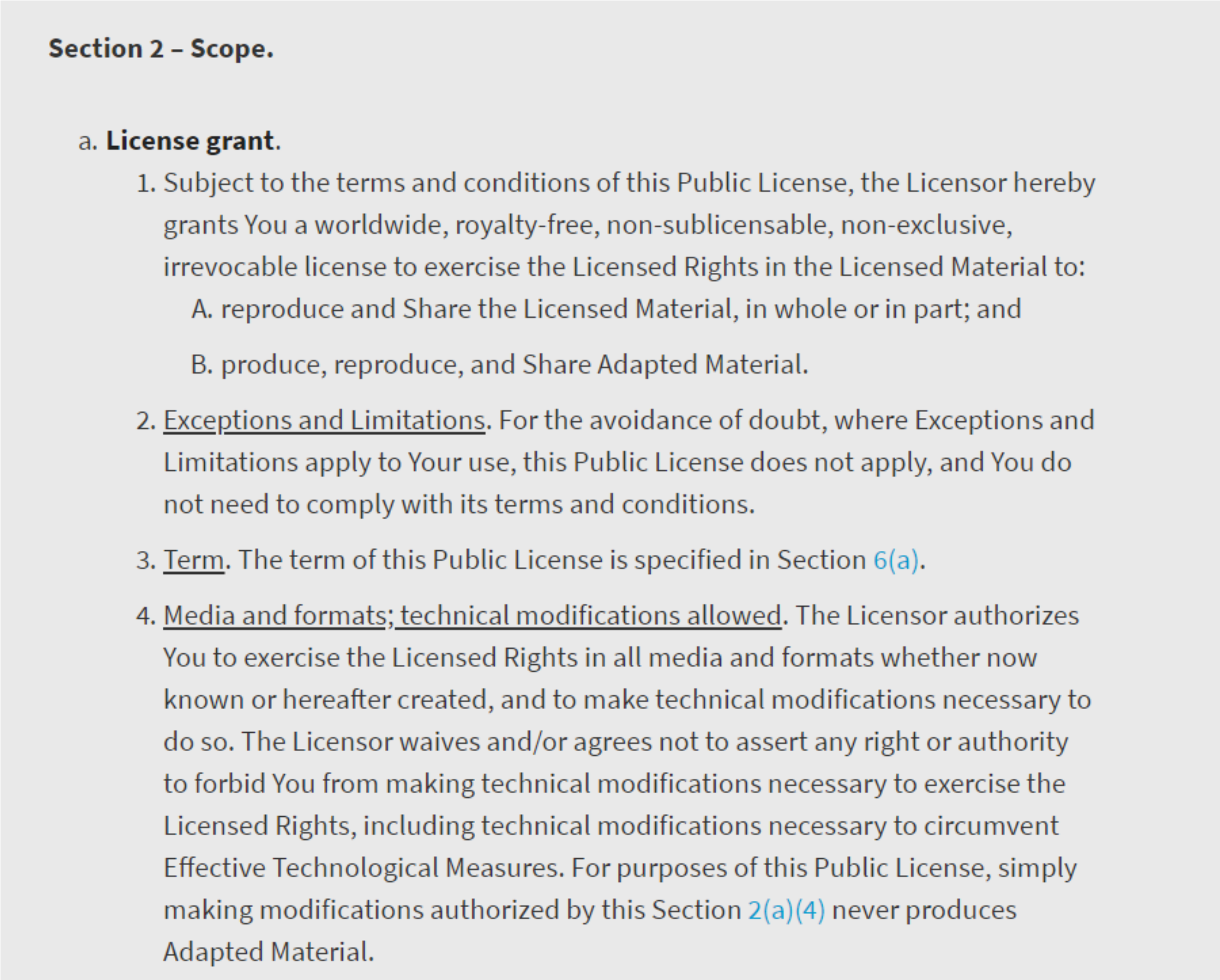
In chapter 2, we present an improved method to estimate gene-specific rates of evolution. We discuss a key statistical issue in our current approach, namely heteroscedasticity and how it affects inferences of shifts in evolutionary rates in phylogenetic datasets. Using a combination of data transformation and weighted regression, we present an improved method that better handles these statistical issues and enables robust inclusion of distantly related species. These advances are important in the context of broadening the applicability of the method to a wide range of evolutionary scenarios. However, there are additional challenges that present opportunities to further refine and improve the method. Some of these include, understanding the power of the method to detect shifts in evolutionary rates in phylogenetically related branches (sister species), expanding the method to include additional covariates related to sequence properties such as GC content or species characteristics such as body mass. With continual efforts to sequence genomes of more species, the power of our evolutionary-based methods can only grow, therefore necessitating method refinements enabling precise identification of convergently evolving genetic elements. Additionally, in the context of subterranean convergence, the method has been powerful at protein-coding gene sequences that predominantly show changes in rate due to relaxed constraint on their sequence evolution rather than positive selection. Improving the power to detect genes undergoing positive selection in species showing phenotypic convergence can therefore uncover novel geneotype-phenotype associations.

In the final chapter of this thesis, we shift gears from binary analyses of gene-trait associations to gene-gene associations. We present an integrated map of protein coevolution across five eukaryotic lineages - mammals, vertebrates, flies, worms, and yeasts. We demonstrate the power of this integrated framework at detecting functional associations between proteins in humans in the context of pairwise associations or PPIs, proteins interacting directly and indirectly in multiprotein complexes, and among genes contributing to genetic diseases. We plan to make this integrated map of protein coevolution publicly available for the research community to investigate functional associations between groups of genes of interest. Such a resource will be useful in generating hypotheses for mechanism of action for uncharacterized genes. This integrated map can also serve as a starting point for systematic investigation of factors controlling evolutionary rate variation and covariation. A better understanding of why certain groups of functionally related genes coevolve while others do not can better inform approaches seeking to utilize the coevolutionary framework for protein interaction prediction.

Appendix A

Permissions to reuse copyrighted content

Permission to reuse content for Chapter 1 from the paper Partha *et al.* (2017) provided by `https://creativecommons.org/licenses/by/4.0/legalcode`:



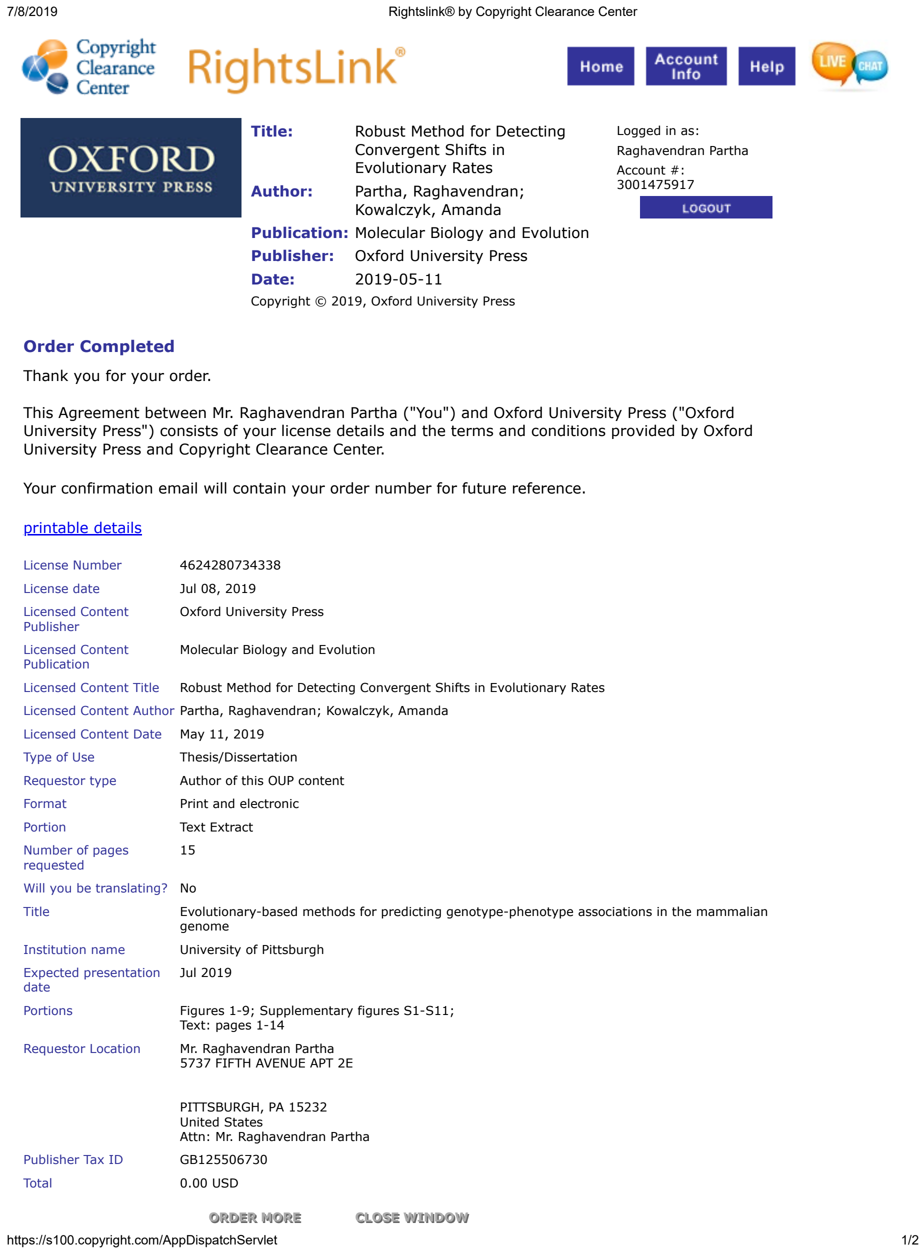
Section 2 – Scope.

a. **License grant.**

1. Subject to the terms and conditions of this Public License, the Licensor hereby grants You a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to:
   A. reproduce and Share the Licensed Material, in whole or in part; and
   B. produce, reproduce, and Share Adapted Material.
2. Exceptions and Limitations. For the avoidance of doubt, where Exceptions and Limitations apply to Your use, this Public License does not apply, and You do not need to comply with its terms and conditions.
3. Term. The term of this Public License is specified in Section 6(a).
4. Media and formats; technical modifications allowed. The Licensor authorizes You to exercise the Licensed Rights in all media and formats whether now known or hereafter created, and to make technical modifications necessary to do so. The Licensor waives and/or agrees not to assert any right or authority to forbid You from making technical modifications necessary to exercise the Licensed Rights, including technical modifications necessary to circumvent Effective Technological Measures. For purposes of this Public License, simply making modifications authorized by this Section 2(a)(4) never produces Adapted Material.

Figure A.0.1: eLIFE license/copyright permissions

Permission to reuse content for Chapter 2 based on the paper Partha *et al.* (2019):



7/8/2019 Rightslink® by Copyright Clearance Center

Title:	Robust Method for Detecting Convergent Shifts in Evolutionary Rates
Author:	Partha, Raghavendran; Kowalczyk, Amanda
Publication:	Molecular Biology and Evolution
Publisher:	Oxford University Press
Date:	2019-05-11

Copyright © 2019, Oxford University Press

Logged in as:
Raghavendran Partha
Account #:
3001475917

Order Completed

Thank you for your order.

This Agreement between Mr. Raghavendran Partha ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

printable details

License Number	4624280734338
License date	Jul 08, 2019
Licensed Content Publisher	Oxford University Press
Licensed Content Publication	Molecular Biology and Evolution
Licensed Content Title	Robust Method for Detecting Convergent Shifts in Evolutionary Rates
Licensed Content Author	Partha, Raghavendran; Kowalczyk, Amanda
Licensed Content Date	May 11, 2019
Type of Use	Thesis/Dissertation
Requestor type	Author of this OUP content
Format	Print and electronic
Portion	Text Extract
Number of pages requested	15
Will you be translating?	No
Title	Evolutionary-based methods for predicting genotype-phenotype associations in the mammalian genome
Institution name	University of Pittsburgh
Expected presentation date	Jul 2019
Portions	Figures 1-9; Supplementary figures S1-S11; Text: pages 1-14
Requestor Location	Mr. Raghavendran Partha 5737 FIFTH AVENUE APT 2E PITTSBURGH, PA 15232 United States Attn: Mr. Raghavendran Partha
Publisher Tax ID	GB125506730
Total	0.00 USD

https://s100.copyright.com/AppDispatchServlet 1/2

Figure A.0.2: Molecular Biology and Evolution license/copyright permissions

Appendix B

List of Supplementary files

Supplementary file 1: link

Supplementary file 2: link

Bibliography

Albert, E.M., Zardoya, R. and García-París, M. (2007). Phylogeographical and speciation patterns in subterranean worm lizards of the genus Blanus (Amphisbaenia: Blanidae). *Molecular Ecology*, **16**(7), 1519–1531.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403–410.

Andersen, D.C. (2004). Below-Ground Herbivory in Natural Communities: A Review Emphasizing Fossorial Animals. *The Quarterly Review of Biology*, **62**(3), 261–286.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T. et al (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, **507**(7493), 455–461.

Archer, M., Beck, R., Gott, M., Hand, S., Godthelp, H. and Black, K. (2011). Australia's first fossil marsupial mole (Notoryctemorphia) resolves controversies about their evolution and palaeoenvironmental origins. *Proceedings of the Royal Society B: Biological Sciences*, **278**(1711), 1498–1506.

Ashburner, M., Blake, J., H, B., JM, C., AP, D., K, D. and Dwigh (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Bateman, A. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, **47**(D1), D506–D515.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R. et al (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*, **28**(10), 1045–8.

Bezginov, A., Clark, G.W., Charlebois, R.L., Dar, V.U.N. and Tillier, E.R. (2013). Coevolution reveals a network of human proteins originating with multicellularity. *Molecular Biology and Evolution*, **30**(2), 332–346.

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L. and Purvis, A. (2007). The delayed rise of present-day mammals. *Nature*, **446**(7135), 507–12.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smith, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D. et al (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, **14**(4), 708–715.

Brooks, M., Allison, W., Li, W., Oel, A., Kim, J.W., Jia, L., Plachetzki, D., Yang, H.J. and Swaroop, A. (2016). Recruitment of Rod Photoreceptors from Short-Wavelength-Sensitive Cones during the Evolution of Nocturnal Vision in Mammals. *Developmental Cell*, **37**(6), 520–532.

Carbon, S., Douglass, E., Dunn, N., Good, B., Harris, N.L., Lewis, S.E., Mungall, C.J., Basu, S., Chisholm, R.L., Dodson, R.J. et al (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, **47**(D1), D330–D338.

Carmona, F.D., Jiménez, R. and Martin, J.M. (2008). The molecular basis of defective lens development in the Iberian mole. *BMC Biology*, **6**, 44.

Carmona, F.D., Ou, J., Jiménez, R. and Collinson, J.M. (2010). Development of the cornea of true moles (Talpidae): Morphogenesis and expression of PAX6 and cytokeratins. *Journal of Anatomy*, **217**(5), 488–500.

Casper, J., Zweig, A.S., Villarreal, C., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Karolchik, D. et al (2018). The UCSC Genome Browser database: 2018 update. *Nucleic Acids Research*, **46**(D1), D762–D769.

Catania, K.C. (1999). A nose that looks like a hand and acts like an eye: The unusual mechanosensory system of the star-nosed mole. *Journal of Comparative Physiology - A Sensory, Neural, and Behavioral Physiology*, **185**(4), 367–372.

Chen, S. and Li, W. (2012). A color-coding amacrine cell may provide a blue-Off signal in a mammalian retina. *Nature Neuroscience*, **15**(7), 954–956.

Chikina, M., Robinson, J.D. and Clark, N.L. (2016). Hundreds of Genes Experienced Convergent Shifts in Selective Pressure in Marine Mammals. *Molecular Biology and Evolution*, **33**(9), 2182–2192.

Clark, G.W., Dar, V.U.N., Bezginov, A., Yang, J.M., Charlebois, R.L. and Tillier, E.R. (2011). Using coevolution to predict protein-protein interactions. *Methods in Molecular Biology*, **781**, 237–256.

Clark, N.L. and Aquadro, C.F. (2010). A novel method to detect proteins evolving at correlated rates: Identifying new functional relationships between coevolving proteins. *Molecular Biology and Evolution*, **27**(5), 1152–1161.

Clark, N.L., Gasper, J., Sekino, M., Springer, S.A., Aquadro, C.F. and Swanson, W.J. (2009). Coevolution of interacting fertilization proteins. *PLoS Genetics*, **5**(7).

Clark, N.L., Alani, E. and Aquadro, C.F. (2012a). Evolutionary Rate Covariation : A bioinformatic method that reveals co-functionality and co-expression of genes. *Genome Research*, **22**(607), 714–720.

Clark, N.L., Alani, E. and Aquadro, C.F. (2012b). Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Research*, **22**(4), 714–720.

Clark, N.L., Alani, E. and Aquadro, C.F. (2013). Evolutionary rate covariation in meiotic proteins results from fluctuating evolutionary pressure in yeasts and mammals. *Genetics*, **193**(2), 529–538.

Consortium, T.U. (2007). The Universal Protein Resource (UniProt). *Nucleic Acids Res*, **35**(Database issue), D193–7.

Cooper, H.M., Herbin, M. and Nevo, E. (1993). Visual system of a naturally microphthalmic mammal: The blind mole rat, Spalax ehrenbergi. *Journal of Comparative Neurology*, **328**(3), 313–350.

Dahlqvist, J., Klar, J., Tiwari, N., Schuster, J., Törmä, H., Badhai, J., Pujol, R., van Steensel, M.A., Brinkhuizen, T., Gijezen, L. et al (2010). A Single-Nucleotide Deletion in the POMP 5 UTR Causes a Transcriptional Switch and Altered Epidermal Proteasome Distribution in KLICK Genodermatosis. *American Journal of Human Genetics*, **86**(4), 596–603.

De Juan, D., Pazos, F. and Valencia, A. (2013). Emerging methods in protein co-evolution. *Nature Reviews Genetics*, **14**(4), 249–261.

De Las Rivas, J. and Fontanillo, C. (2010). ProteinProtein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLoS Computational Biology*, **6**(6), e1000807.

Di, Y., Schafer, D.W., Cumbie, J.S. and Chang, J.H. (2011). The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. *Statistical Applications in Genetics and Molecular Biology*, **10**(1), 1–28.

Dimanlig, P.V., Faber, S.C., Auerbach, W., Makarenkova, H.P. and Lang, R.A. (2001). The upstream ectoderm enhancer in Pax6 has an important role in lens induction. *Development*, **128**(22), 4415–4424.

Dobler, S., Dalla, S., Wagschal, V. and Agrawal, A.A. (2012). Community-wide convergent evolution in insect adaptation to toxic cardenolides by substitutions in the Na,K-ATPase. *Proceedings of the National Academy of Sciences*, **109**(32), 13040–13045.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R. et al (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.

Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics*, **10**(1), 48.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, **32**(5), 1792–7.

Eisen, J.A. (1998). Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research*, **8**(3), 163–167.

Emerling, C.A. and Springer, M.S. (2014). Eyes underground: Regression of visual protein networks in subterranean mammals. *Molecular Phylogenetics and Evolution*, **78**(1), 260–270.

Emerson, R.O. and Thomas, J.H. (2009). Adaptive evolution in zinc finger transcription factors. *PLoS Genetics*, **5**(1).

Emms, D.M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, **16**(1).

Esteller, M. (2011). Non-coding RNAs in human disease. *Nature Reviews Genetics*, **12**(12), 861–874.

Fang, X., Nevo, E., Han, L., Levanon, E.Y., Zhao, J., Avivi, A., Larkin, D., Jiang, X., Feranchuk, S., Zhu, Y. et al (2014). Genome-wide adaptive complexes to underground stresses in blind mole rats Spalax. *Nature Communications*, **5**, 1039–1052.

Fields, S. and Song, O.K. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, **340**(6230), 245–246.

Findlay, G.D., Sitnik, J.L., Wang, W., Aquadro, C.F., Clark, N.L. and Wolfner, M.F. (2014). Evolutionary Rate Covariation Identifies New Members of a Protein Network Required for Drosophila melanogaster Female Post-Mating Responses. *PLoS Genetics*, **10**(1).

Fitzpatrick, D.A., Logue, M.E., Stajich, J.E. and Butler, G. (2006). A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evolutionary Biology*, **6**.

Foote, A.D., Liu, Y., Thomas, G.W., Vina, T., Alföldi, J., Deng, J., Dugan, S., Van Elk, C.E., Hunter, M.E., Joshi, V. et al (2015). Convergent evolution of the genomes of marine mammals. *Nature Genetics*, **47**(3), 272–275.

Fraser, H.B., Hirsh, A.E., Wall, D.P. and Eisen, M.B. (2004). Coevolution of gene expression among interacting proteins. *Proceedings of the National Academy of Sciences*, **101**(24), 9033–9038.

Fu, D.J., Thomson, C., Lunny, D.P., Dopping-Hepenstal, P.J., McGrath, J.A., Smith, F.J., Irwin McLean, W.H. and Pedrioli, D.M. (2014). Keratin 9 is required for the structural integrity and terminal differentiation of the palmoplantar epidermis. *Journal of Investigative Dermatology*, **134**(3), 754–763.

Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C. and Ruepp, A. (2019). CORUM: The comprehensive resource of mammalian protein complexes - 2019. *Nucleic Acids Research*, **47**(D1), D559–D563.

Graw, J. (2009). Genetics of crystallins: Cataract and beyond. *Experimental Eye Research*, **88**(2), 173–189.

Griffin, C., Kleinjan, D.A., Doe, B. and Van Heyningen, V. (2002). New 3 elements control Pax6 expression in the developing pretectum, neural retina and olfactory region. *Mechanisms of Development*, **112**(1-2), 89–100.

Haeussler, M., Zweig, A.S., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Hinrichs, A.S., Gonzalez, J.N. et al (2019). The UCSC Genome Browser database: 2019 update. *Nucleic Acids Research*, **47**(D1), D853–D858.

Hamilton, W.J. (1931). Habits of the Star-Nosed Mole, Condylura cristata. *Journal of Mammalogy*, **12**(4), 345.

Hamosh, A., Scott, A.F., Amberger, J., Valle, D. and McKusick, V.A. (2000). Online Mendelian Inheritance in Man (OMIM). *Human Mutation*, **15**(1), 57–61.

Hardison, R.C. (2010). Chapter 19: Comparative Genomics. *Vogel and Motulsky's Human Genetics: Problems and Approaches (Fourth Edition)*, pages 557–587.

Harris, R.S. (2007). *Improved Pairwise Alignmnet of Genomic DNA*. Ph.D. thesis.

Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S. et al (2012). A census of human soluble protein complexes. *Cell*, **150**(5), 1068–81.

Hendriks, W., Leunissen, J., Nevo, E., Bloemendal, H. and de Jong, W.W. (2006). The lens protein alpha A-crystallin of the blind mole rat, Spalax ehrenbergi: evolutionary change and functional constraints. *Proceedings of the National Academy of Sciences*, **84**(15), 5320–5324.

Hennies, H.C., Küster, W., Mischke, D. and Reis, A. (1995). Localization of a locus for the striated form of palmoplantar keratoderma to chromosome 18q near the desmosomal cadherin gene cluster. *Human Molecular Genetics*, **4**(6), 1015–1020.

Hetling, J.R., Baig-Silva, M.S., Comer, C.M., Pardue, M.T., Samaan, D.Y., Qtaishat, N.M., Pepperberg, D.R. and Park, T.J. (2005). Features of visual function in the naked mole-rat Heterocephalus glaber. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, **191**(4), 317–330.

Hiller, M., Schaar, B.T., Indjeian, V.B., Kingsley, D.M., Hagey, L.R. and Bejerano, G. (2012). A "Forward Genomics" Approach Links Genotype to Phenotype using Independent Phenotypic Losses among Related Species. *Cell Reports*, **2**(4), 817–823.

Huynen, M.A., Snel, B., Von Mering, C. and Bork, P. (2003). Function prediction and protein networks.

Jarvis, E.D., Siavash, M., Ye, C., Liang, S., Yan, Z., Zepeda, M.L., Campos, P.F., Missael, A., Velazquez, V., Samaniego, J.A. et al (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, **346**(6215), 1126–1138.

Jeffery, W.R. (2005). Adaptive evolution of eye degeneration in the Mexican blind cavefish. In *Journal of Heredity*, volume 96, pages 185–196.

Jeffery, W.R. (2009). Regressive evolution in Astyanax cavefish. *Annual review of genetics*, **43**, 25–47.

Jones, S. and Thornton, J.M. (1996). Review Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, **93**, 13–20.

Kammandel, B., Chowdhury, K., Stoykova, A., Aparicio, S., Brenner, S. and Gruss, P. (1999). Distinct cis-essential modules direct the time-space pattern of the Pax6 gene activity. *Developmental Biology*, **205**(1), 79–97.

Kim, E.B., Fang, X., Fushan, A.A., Huang, Z., Lobanov, A.V., Han, L., Marino, S.M., Sun, X., Turanov, A.A., Yang, P. et al (2011). Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature*, **479**(7372), 223–7.

Kleinjan, D.A. (2001). Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. *Human Molecular Genetics*, **10**(19), 2049–2059.

Kleinjan, D.A., Seawright, A., Mella, S., Carr, C.B., Tyas, D.A., Simpson, T.I., Mason, J.O., Price, D.J. and van Heyningen, V. (2006). Long-range downstream enhancers are essential for Pax6 expression. *Developmental Biology*, **299**(2), 563–581.

Koay, G., Kearns, D., Heffner, H.E. and Heffner, R.S. (1998). Passive sound-localization ability of the big brown bat (Eptesicus fuscus). *Hearing Research*, **119**(1-2), 37–48.

Kowalczyk, A., Meyer, W.K., Partha, R., Mao, W., Clark, N.L. and Chikina, M. (2019). RERconverge: an R package for associating evolutionary rates with convergent traits. *Bioinformatics*.

Lacey, E., Patton, J. and Cameron, G. (2011). Life Underground: The Biology Of Subterranean Rodents. *Australian Mammalogy*, **23**(1), 75.

Lartillot, N. and Poujol, R. (2011). A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular Biology and Evolution*, **28**(1), 729–744.

Law, C.W., Chen, Y., Shi, W. and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, **15**(2), R29.

Leys, R., Watts, C.H.S., Cooper, S.J.B. and Humphreys, W.F. (2003). Evolution of subterranean diving beetles (Coleoptera: Dytiscidae hydroporini, Bidessini) in the arid zone of Australia. *Evolution*, **57**(12), 2819–2834.

Li, W., Chen, S. and Devries, S.H. (2010). A fast rod photoreceptor signaling pathway in the mammalian retina. *Nature Neuroscience*, **13**(4), 414–416.

Li, Y., Calvo, S.E., Gutman, R., Liu, J.S. and Mootha, V.K. (2014). Expansion of biological pathways based on evolutionary inference. *Cell*, **158**(1), 213–225.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P. and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**(12), 1739–1740.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E. et al (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**(7370), 476–482.

Liu, Y., Cotton, J.A., Shen, B., Han, X., Rossiter, S.J. and Zhang, S. (2010). Convergent sequence evolution between echolocating bats and dolphins. *Current Biology*, **20**(2).

Losos, J.B. (2011). Convergence, adaptation, and constraint. *Evolution*, **65**(7), 1827–1840.

Lovell, S.C. and Robertson, D.L. (2010). An integrated view of molecular coevolution in protein-protein interactions. *Molecular Biology and Evolution*, **27**(11), 2567–2575.

Löytynoja, A. and Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**(5883), 1632–1635.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. et al (2009). Finding the missing heritability of complex diseases. *Nature*, **461**(7265), 747–53.

Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D. (1999a). A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**(6757), 83–86.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999b). Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**(5428), 751–753.

Marcovitz, A., Jia, R. and Bejerano, G. (2016). "reverse Genomics" Predicts Function of Human Conserved Noncoding Elements. *Molecular Biology and Evolution*, **33**(5), 1358–1369.

McDonough, C.M. and Loughry, W.J. (2013). *The nine-banded armadillo: a natural history*, volume 11. University of Oklahoma Press.

Meredith, R.W., Janečka, J.E., Gatesy, J., Ryder, O.A., Fisher, C.A., Teeling, E.C., Goodbla, A., Eizirik, E., Simão, T.L., Stadler, T. et al (2011). Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science*, **334**(6055), 521–524.

Meyer, W.K., Jamison, J., Richter, R., Woods, S.E., Partha, R., Kowalczyk, A., Kronk, C., Chikina, M., Bonde, R.K., Crocker, D.E. et al (2018). Ancient convergent losses of Paraoxonase 1 yield potential risks for modern marine mammals. *Science*, **361**(6402), 591–594.

Moran, D., Softley, R. and Warrant, E.J. (2015). The energetic cost of vision and the evolution of eyeless Mexican cavefish. *Science Advances*, **1**(8), e1500363.

Murphy, W.J., Eizirik, E., O'Brien, S.J., Madsen, O., Scally, M., Douady, C.J., Teeling, E., Ryder, O.A., Stanhope, M.J., De Jong, W.W. et al (2001). Resolution of the early placental mammal radiation using bayesian phylogenetics. *Science*, **294**(5550), 2348–2351.

Murphy, W.J., Pevzner, P.A. and O'Brien, S.J. (2004). Mammalian phylogenomics comes of age. *Trends in Genetics*, **20**(12), 631–639.

Murphy, W.J., Pringle, T.H., Crider, T.A., Springer, M.S. and Miller, W. (2007). Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Research*, **17**(4), 413–421.

Nei, M. and Rooney, A.P. (2005). Concerted and Birth-and-Death Evolution of Multigene Families. *Annual Review of Genetics*, **39**(1), 121–152.

Němec, P., Cveková, P., Benada, O., Wielkopolska, E., Olkowicz, S., Turlejski, K., Burda, H., Bennett, N.C. and Peichl, L. (2008). The visual system in subterranean African mole-rats (Rodentia, Bathyergidae): Retina, subcortical visual nuclei and primary visual cortex. *Brain Research Bulletin*, **75**(2-4), 356–364.

Nevo, E. (1979). Adaptive Convergence and Divergence of Subterranean Mammals. *Annual Review of Ecology and Systematics1*, **10**, 269–308.

Nowak, R.M. (1999). *Walker's Mammals of the World*. Number v. 1 in Walker's Mammals of the World. Johns Hopkins University Press, fifth edition.

Ochoa, D. and Pazos, F. (2014). Practical aspects of protein co-evolution. *Frontiers in Cell and Developmental Biology*, **2**(April), 1–9.

O'Leary, M.A. and Kaufman, S. (2011). MorphoBank: phylogenomics in the "cloud". *Cladistics*, **27**, 1–9.

O'Leary, M.A., Bloch, J.I., Flynn, J.J., Gaudin, T.J., Giallombardo, A., Giannini, N.P., Goldberg, S.L., Kraatz, B.P., Luo, Z.X., Meng, J. et al (2013). The placental mammal ancestor and the Post-K-Pg radiation of placentals. *Science*, **339**(6120), 662–667.

Parker, J., Tsagkogeorga, G., Cotton, J.A., Liu, Y., Provero, P., Stupka, E. and Rossiter, S.J. (2013). Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*, **502**(7470), 228–231.

Partha, R., Chauhan, B.K., Ferreira, Z., Robinson, J.D., Lathrop, K., Nischal, K.K., Chikina, M. and Clark, N.L. (2017). Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *eLife*, **6**, e25884.

Partha, R., Kowalczyk, A., Clark, N. and Chikina, M. (2019). Robust method for detecting convergent shifts in evolutionary rates. *Molecular Biology and Evolution*.

Pazos, F. and Valencia, A. (2002). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering, Design and Selection*, **14**(9), 609–614.

Pazos, F., Juan, D., Izarzugaza, J.M., Leon, E. and Valencia, A. (2008). Prediction of protein interaction based on similarity of phylogenetic trees. *Methods in Molecular Biology*, **484**, 523–535.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(8), 4285–8.

Pennacchio, L.A., Bickmore, W., Dean, A., Nobrega, M.A. and Bejerano, G. (2013). Enhancers: Five essential questions. *Nature Reviews Genetics*, **14**(4), 288–295.

Priedigkeit, N., Wolfe, N. and Clark, N.L. (2015). Evolutionary Signatures amongst Disease Genes Permit Novel Methods for Gene Prioritization and Construction of Informative Gene-Based Networks. *PLoS Genetics*, **11**(2), 1–17.

Prudent, X., Parra, G., Schwede, P., Roscito, J.G. and Hiller, M. (2016). Controlling for Phylogenetic Relatedness and Evolutionary Rates Improves the Discovery of Associations between Species' Phenotypic and Genomic Differences. *Molecular Biology and Evolution*, **33**(8), 2135–2150.

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

Quilliam, T.A. (1966). The mole's sensory apparatus. *Journal of Zoology*, **149**(1), 76–88.

Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A. et al (2013). A large-scale evaluation of computational protein function prediction. *Nature Methods*, **10**(3), 221–227.

Ramani, A.K. and Marcotte, E.M. (2003). Exploiting the co-evolution of interacting proteins to discover interaction specificity. *Journal of Molecular Biology*, **327**(1), 273–284.

Remm, M., Storm, C.E. and Sonnhammer, E.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, **314**(5), 1041–1052.

Revell, L.J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**(2), 217–223.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, **43**(7), e47.

Robinson, D.F. and Foulds, L.R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**(1-2), 131–147.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

Rodriguez, F., Oliver, J. and Marin, A. (1990). The General Stochastic Model of Nucleotide Substitution The G4H Model. *J. theor. Biol.*, **142**(4), 485–501.

Romanoski, C.E., Glass, C.K., Stunnenberg, H.G., Wilson, L. and Almouzni, G. (2015). Epigenomics: Roadmap for regulation. *Nature*, **518**(7539), 314–316.

Roscito, J.G., Sameith, K., Parra, G., Langer, B.E., Petzold, A., Moebius, C., Bickle, M., Rodrigues, M.T. and Hiller, M. (2018). Phenotype loss is associated with widespread divergence of the gene regulatory landscape in evolution. *Nature Communications*, **9**(1), 1–36.

Rosenblum, E.B., Parent, C.E. and Brandt, E.E. (2014). The Molecular Basis of Phenotypic Convergence. *Annual Review of Ecology, Evolution, and Systematics*, **45**(1), 203–226.

Sánchez, Y. and Huarte, M. (2015). Long Non-Coding RNAs: Challenges for Diagnosis and Therapies. *Nucleic Acid Therapeutics*, **23**(1), 15–20.

Sanyal, S., Jansen, H.G., de Grip, W.J., Nevo, E. and de Jong, W.W. (1990). The Eye of the Blind Mole Rot, Spalax ehrenbergi. *Investigative Ophthalmology & Visual Science*, **31**(7), 1398–1404.

Sato, T., Yamanishi, Y., Kanehisa, M. and Toh, H. (2005). The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, **21**(17), 3482–3489.

Schliep, K.P. (2011). phangorn: Phylogenetic analysis in R. *Bioinformatics*, **27**(4), 592–593.

Shlyueva, D., Stampfel, G. and Stark, A. (2014). Transcriptional enhancers: From properties to genome-wide predictions. *Nature Reviews Genetics*, **15**(4), 272–286.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S. et al (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, **15**(8), 1034–50.

Smith, C.L. and Eppig, J.T. (2009). The mammalian phenotype ontology: Enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, **1**(3), 390–399.

Sonnhammer, E.L.L. and Östlund, G. (2015). InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*, **43**(D1), D234–D239.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**(9), 1312–1313.

Stern, D.L. (2013). The genetic causes of convergent evolution. *Nature Reviews Genetics*, **14**(11), 751–764.

Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(3), 479–498.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G. et al (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(16), 6062–7.

Sweet, G. (1909). The eyes of Chrysochloris hottentota and C. asiatica. *Quarterly Journal of Microscopical Science*, **2**(210), 327–338.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P. et al (2017). The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, **45**(D1), D362–D368.

Thomas, G.W. and Hahn, M.W. (2015). Determining the null model for detecting adaptive convergence from genomic data: A case study using echolocating mammals. *Molecular Biology and Evolution*, **32**(5), 1232–1236.

Tillier, E.R. and Charlebois, R.L. (2009). The human protein coevolution network. *Genome Research*, **19**(10), 1861–1871.

Villar, D., Pignatelli, M., Rayner, T.F., Deaville, R., Murchison, E.P., Odom, D.T., Turner, J.M., Lukk, M., Jasinska, A.J., Park, T.J. et al (2015). Enhancer Evolution across 20 Mammalian Species. *Cell*, **160**(3), 554–566.

Walther, T.C., Alves, A., Pickersgill, H., Loïodice, I., Hetzer, M., Galy, V., Hülsmann, B.B., Köcher, T., Wilm, M., Allen, T. et al (2003). The conserved Nup107-160 complex is critical for nuclear pore complex assembly. *Cell*, **113**(2), 195–206.

Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., Xiong, X., Kagan, O., Kwan, J., Bezginov, A. et al (2015). Panorama of ancient metazoan macromolecular complexes. *Nature*, **525**(7569), 339–44.

Wang, Y., Cortez, D., Yazdi, P., Neff, N., Elledge, S.J. and Qin, J. (2000). BASC, a super complex of BRCA1-associated proteins involved in the recognition and repair of aberrant DNA structures. *Genes and Development*, **14**(8), 927–939.

Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, **18**(5), 691–699.

Wilkinson, M. (2012). Caecilians. *Current Biology*, **22**(17), R668–R669.

Wolfe, N.W. and Clark, N.L. (2015). ERC analysis: Web-based inference of gene function via evolutionary rate covariation. *Bioinformatics*, **31**(23), 3835–3837.

Xu, P.X., Zhang, X., Heaney, S., Yoon, A., Michelson, A.M. and Maas, R.L. (1999). Regulation of Pax6 expression is conserved between mice and flies. *Development (Cambridge, England)*, **126**(2), 383–95.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution*, **11**(9), 367–372.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**(8), 1586–1591.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y. and Wang, S. (2010). GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, **26**(7), 976–978.

Zhang, J. and Kumar, S. (1997). Detection of convergent and parallel evolution at the amino acid sequence level. *Molecular Biology and Evolution*, **14**(5), 527–536.

Zou, Z. and Zhang, J. (2015). No genome-wide protein sequence convergence for echolocation. *Molecular Biology and Evolution*, **32**(5), 1237–1241.