Computational Methods for Discovering Genetic Functions of Conserved Non-coding Elements with Comparative Genomics

by

Elysia Saputra

B. Eng., National University of Singapore, 2015

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

2023

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Elysia Saputra

It was defended on

July 17, 2023

and approved by

Nathan Clark, PhD, Associate Professor, Department of Human Genetics, University of Utah

Andreas Pfenning, PhD, Associate Professor, Computational Biology Department, Carnegie Mellon University

Adam Siepel, PhD, Professor, Cold Spring Harbor Laboratory School of Biological Sciences

Dennis Kostka, PhD, Associate Professor, Department of Developmental Biology, University of Pittsburgh

Dissertation Director: Maria Chikina, PhD, Assistant Professor, Department of Computational and Systems Biology, University of Pittsburgh

Computational Methods for Discovering Genetic Functions of Conserved Non-coding Elements with Comparative Genomics

Elysia Saputra, PhD

University of Pittsburgh, 2023

Unveiling the genetic encodings of complex phenotypes is a fundamental goal of biology. With the increasing availability of sequenced genomes, it has become possible to elucidate molecular adaptations that engender species diversity with evolutionary-based methods. Although morphological differences arise from changes in transcriptional regulation, regulatory non-coding elements are still insufficiently characterized, and there is a lack of phylogenetic tools that account for their evolutionary properties. This dissertation addresses this gap by developing new tools to perform unbiased genome-wide predictions of regulatory element adaptations that underlie convergent phenotypes. This dissertation introduces three new tools, discussed in the following chapters.

In Chapter 1, we introduce empirical strategies for calibrating phylogenetic signals against statistical biases that arise from phylogenetic, technical, and biological sources. We develop phylogenetically-constrained trait permutation strategies for binary and continuous traits, and benchmark them systematically on various methods and convergent phenotypes. This study demonstrates the effectiveness of phylogeny-aware permutation strategies for improving the statistical behavior and prediction specificity from phylogenetic analysis.

In Chapter 2, we build a maximum likelihood-based phylogenetic method tailored to characterizing the adaptation of conserved regulatory elements associated with phenotypic convergence. We benchmark the method using the classical case of convergent evolution of mammalian lineages to subterranean habitats and demonstrate the ability of the method to modularly identify phenotype-relevant local segments of regulatory elements. In Chapter 3, we apply the method to study the regulatory changes underlying the convergent adaptation of mammals to life at altitude. We release the tool as a software package that can be used by the research community.

In Chapter 4, we develop an alignment-free phylogenetic method for characterizing regulatory motif adaptations that underlie phenotypic convergence from orthologous, but not necessarily alignable sequences. Using a reference-free alignment dataset, we benchmark the method against competing alignment-based and alignment-free methods using the convergence case of vision loss in mammals and demonstrate the superior performance of the method. We finally apply the method to investigate the regulatory motif adaptation underlying the convergent evolution of longevity and increased body size in mammals. We make the tool publicly available to use for scalable computations of motif-level convergence signals.

Table of Contents

List of Tables

List of Figures

Preface

First, I would like to express my gratitude to my advisors Dr. Maria Chikina and Dr. Nathan Clark, whom I have had the fortune to work closely with in the past five years. Their consistent support and guidance in my scientific training have truly been invaluable. Thank you for pushing me to always think more critically and creatively. The past five years have truly been the most enriching learning experience that I have had.

I would also like to thank my thesis committee, Dr. Dennis Kostka, Dr. Andreas Pfenning, and Dr. Adam Siepel for their continuous support, feedback, and collaboration.

To all past and present members of the Chikina and Clark labs, especially to Amanda Kowalczyk, Javad Rahimikollu, Wayne Mao, Tugrul Balci, Raghav Partha, and Allie Graham, thank you for your friendships and continuous support in my learning over the years.

To Dr. Igal Szleifer, who believed in my potential before I did. Thank you for instilling in me the audacity to leave the traditional path that I was "supposed" to take and instead to seriously go into science. You have been the single factor that forever changed my life trajectory. As I move along in my career, I will try to always heed your advice to have the courage to believe in me and never to lose my naivety.

To the people who have become my family away from home: Dini, Uncle Khing, Aunt Melinda, Funita, Patrick, Ritchie, Wyatt, Wesley, Dante, Liviera, Albert, Richard, Cindy, Emily, my goddaughter Gianna, David, Li Ling, Hsien, and Ying, thank you for opening your home to me and embracing me as part of your family.

To my CPCB friends, Neha Cheemalavagu, Yutong Qiu, Daniel Yuan, Minxue Jia, Stefan Andjelkovic, and Tyler Lovelace, thank you for your invaluable friendships and camaraderie that kept me going throughout my PhD years.

Most importantly, I would not be where I am without the constant support of my family – my parents, Haddy and Lily, my sisters, Edina and Emilia, my brother-in-law, Benyamin, my nephew, Aiden, and Dariyah. Thank you for giving me the freedom to pursue what I love to do, despite it causing us to have to live far away. It breaks my heart that the distance and the unexpected pandemic had caused me to miss many crucial events in our family, but you love and understand me nonetheless. It means everything.

Finally, to my soul sister Shelvia, who despite being 9,533 miles away has been a constant source of companionship, joy, courage, and wisdom throughout all of the ups and downs of my PhD journey. Thank you for always celebrating my small wins and keeping me grounded when things spiral. None of this work would have been completed without you. This dissertation is dedicated to you.

1.0 Introduction

A fundamental pursuit of modern biology is to uncover the genetic encodings of phenotypic variations. With the advancement of high throughput sequencing technologies, the last decades have seen the flourishing of studies on genotype-to-phenotype mappings, which have demonstrated success in various areas [124, 175, 215, 217, 269]. The diversity of traits in the natural world also encompasses various depths of evolutionary timescales. For trait variations that evolve over shallow timescales, such as variations in a population, traditional population genetics approaches such as genome-wide association studies (GWAS) [246] can usually be used for identifying genetic changes that are associated with the trait. However, the utility of traditional genetics approaches decreases as we move to deeper evolutionary timescales. Fortunately, recent years have seen a dramatic increase in the number of species genomes that have been sequenced, such as the 240-way mammalian alignment produced by the Zoonomia Project [5], the 1,107-way phylogeny of Ascomycota fungi [226], and others. With this wealth of clade-level genomic data and the advancement of modern evolutionary theories, it becomes possible to design statistical comparative algorithms to infer genotype-to-phenotype mappings that are uniquely encoded through deep evolutionary time.

To uncover the genetic basis of an extreme trait, one could sequence the genome of the species with the extreme trait and identify the loci that have diverged in association with the trait. However, there can be millions of nucleotide differences between a species and its closest relatives, and it would still be difficult to disentangle trait-associated loci from lineage-specific nucleotide changes that are not relevant to the trait. One strategy to overcome this challenge is to leverage on *convergent* traits, which are traits that have independently evolved across multiple lineages in response to similar selection pressures. In the realization of convergent traits, certain genomic changes repeatedly occurred across independent clades over millions of years. These repeated occurrences thus provide “natural biological replicates” of the evolution of trait, giving us the statistical power to detect the convergent molecular signals associated with it.

This concept have given rise to numerous evolutionary-based comparative strategies for characterizing the genetic underpinnings of convergent phenotypes. Some of these approaches detect divergence at specific amino acid or nucleotide positions [48, 53, 75], while other strategies use evolutionary rate deviations of genomic regions from neutrality as a proxy for selection [110, 128, 172, 188, 190]. Despite differences in approaches, a commonality among these strategies is that they are built on the assumption that sequence conservation is reflective of functional importance. For example, genes that contribute to the fitness of a species in a new environment may evolve under increased selection constraint and exhibit a strong sequence conservation, because mutations in the sequence can possibly be deleterious. On the other hand, genes that provide a reduced contribution to the fitness of the species in the new environment may evolve under relaxed constraint and accumulate more mutations. This is a reasonable assumption for amino acid sequences or protein-coding regions for which structural conservation is critical. The application of evolutionary-based methods to analyze protein-coding sequences have indeed been largely successful [103, 128, 166, 187, 256].

However, other than the protein-coding regions themselves, other critical determinants of gene expression differences that give rise to phenotypic diversity are the regulatory non-coding elements that control transcription, such as enhancers and promoters. In fact, about 90% of single nucleotide polymorphisms (SNPs) identified by GWAS to be phenotype-associated are located in non-coding regions [61, 84, 107, 167]. Despite their importance, regulatory non-coding elements remain insufficiently characterized. The direct application of conservation-based comparative genomics algorithms to regulatory elements can be difficult because their working assumption of correlating conservation with function is often incompatible with the underlying grammar of regulatory elements. Typically, each enhancer or promoter unit is modularly composed of numerous transcription factor (TF) binding motifs. Each of these TFs may contribute differently to the activity of the element, which can also facilitate pleitropic functions [134]. The modular structure and function of TF motifs allow them to turn over rapidly [55, 171], possibly as a result of functional redundancies [267] or compensatory mechanisms [171]. Effectively, regulatory elements tend to have low sequence conservation [56, 220, 249] and homologous functional activity can still arise despite vast differences in TF binding patterns [186, 258]. There are indeed some types of regulatory elements that are perfectly conserved across species, such as ultraconserved enhancers [232].

Even so, mutagenesis experiments showed that these elements were able to retain their function at high levels of sequence mutations [232]. All these observations suggest that different segments of a regulatory element may have different phenotypic relevance, and therefore experience different evolutionary pressures. Overall, the structural, functional, and evolutionary properties of regulatory elements motivate an alternative perspective to the sequence conservation-based design of conventional comparative genomic methods.

The increase in sequenced genomes greatly improves our statistical power to perform phylogenetic inference, but it also introduces new sets of challenges. A larger number of species brings about a trade-off between high phylogenetic resolution and large variations in genome data quality [105, 230]. There are stochastic variations in sequence yields across sample libraries and true chromosomal changes such as insertions, deletions, invertions, or duplications that can introduce missing data in the resulting phylogenomic alignments. Missingness can introduce systematic errors that can lead to false conclusions [105, 230], and there is a lack of available methods for addressing this technical bias in a phylogenetic context.

Importantly, there is also a phylogenetic reasoning that motivates a rigorous handling of biases in phylogenomic analysis. The evolutionary history of species evolution across a phylogeny inherently contains a bias that is encoded by common ancestry. Species that share a longer evolutionary history will tend to have more similar characteristics than species that are more distantly related. As such, signals from phylogenomic analysis naturally contain a complex pattern of phylogenetic non-independence. While there are parametric statistical strategies that have been developed to address this phylogenetic bias [39, 73, 85, 88, 94, 106, 150], these strategies usually make strong assumptions about statistical distributions and the generative evolutionary models of the datasets. In cases where the assumptions do not match the true evolutionary process that produced the observed data, the application of such correction strategies can further propagate systematic errors.

Finally, additional biases can arise from variations in base compositions – particularly GC bias – be it among different genomes or across the chromosomes of a single genome [207]. One of the reasons for GC bias is GC-biased gene conversion, an evolutionary process in which the biochemical properties of nucleotides favor the conversion of A/T to G/C bases in a non-adaptive way. Effectively, this mechanism often increases the substitution rate at the affected genomic regions such as recombination hotspots, which can often be confounded for signals of positive selection [127, 207]. The finding that GC-biased substitution patterns also affected some annotated human accelerated regions (HARs) suggests that this problem also likely affects conserved non-coding elements [191]. All in all, the aforementioned sources of biases would likely have an inflated effect on regulatory elements, especially due to their short lengths (7-30bp) and their high tolerance for sequence variations.

This thesis seeks to address these issues by developing new statistically robust approaches to elucidate mappings between regulatory elements and convergent phenotypes. We start by introducing phylogenetic permulation, a set of novel statistical strategies for performing empirical corrections of signals from phylogenomic analysis against biases from phylogenetic, technical, and biological origins. These calibration strategies subsequently become the foundations to the development of two phylogenetic methods that modularly characterize TF-scale changes associated with the evolution of convergent traits. The first method, *phyloConverge*, is an "alignment-based" method that uses generative nucleotide substitution modeling to compute local convergent shifts in evolutionary rates of TF-scale nucleotide segments in conserved regulatory elements. The second method, *AFconverge*, is an "alignment-free" method that detects trait-associated motif gains and losses in the flexible sequence space of weakly conserved orthologs. In this dissertation, we will discuss the development and application of these methods on various convergence cases, and illustrate new angles for interrogating the adaptation patterns of regulatory elements to selection.

2.0 Phylogenetic Permulations: a statistically rigorous approach to measure confidence in associations in a phylogenetic context

2.1 Attribution statement

All of the work in this chapter was performed by myself, with the following exceptions:

- Development of permulations for continuous phenotypes, pathway permulations, and the related analyses were done by Amanda Kowalczyk, Ph.D.
- Benchmarking analysis with Forward Genomics was performed by Luisa Cusick.

2.2 Introduction

Despite the availability of complete genomes for many species, identifying the genetic elements responsible for a phenotype of interest is difficult because there are millions of genetic differences between almost every pair of species. One strategy to link genotypes and phenotypes is to take advantage of convergent evolutionary events in which multiple unrelated species have evolved similar characteristics. Such events represent natural biological replicates of evolution during which species may have experienced similar genetic changes driving similar phenotypic changes. When lineages independently evolve or lose a shared phenotype, convergent molecular signals can be used to identify specific genetic elements associated with the phenotypic shift.

Diverse analytic approaches have been developed to use convergent phenotypes to identify specific genetic elements underlying a trait. The methods include analyzing convergent amino acid substitutions [75] and convergent shifts in evolutionary rates [103, 110, 128, 195, 256], as well as investigating convergent gene loss [103, 166]. Methods that analyze convergent shifts in evolutionary rates (rather than convergence to any specific sequence) have been particularly successful. We have previously developed one such method called *RERconverge* [128, 188] to link genetic elements to convergently evolving phenotypes based on evolution across a sequence of interest. Our method has been successfully used to identify the genetic basis of adaptation to a marine habitat [40], regression of ocular structures in a subterranean habitat [187], and evolution of extreme lifespan and body size phenotypes [129] in mammals. Other groups have developed similar methods for identifying convergent shifts in evolutionary pressure. The *Forward Genomics* algorithm, which correlates percent sequence change along a phylogeny with phenotypic changes [103, 195], has been used to identify genetic elements underlying low levels of biliary phospholipid levels in horses and guinea pigs, the loss of ability to synthesize vitamin C in some primates, bats, and guinea pigs, as well as the loss of ocular structures in two independent subterranean mammals. Both *RERconverge* and *Forward Genomics* involve a phylogenetic inference step and a subsequent test for phenotype association. More sophisticated but computationally intensive methods that consider the phenotype at the phylogenetic inference step have also been developed, notably *PhyloAcc* [110], although these methods are difficult to scale to genome-wide analyses. A related but distinct approach is to assess the association between gene loss (the limiting case of relaxed evolutionary pressure) and convergent phenotypes. A recent study used phylogenetic generalized least squares (PGLS) [88] to compute associations between gene losses and diverse traits and found a large number of significant associations [195].

Importantly, these methods are often applied in a genome-wide discovery context. As such, the general approach can be summarized as using a statistical test to calculate the association between convergent phenotypes and some measure of molecular evolution (evolutionary rate or gene loss) across a large number of genomic regions, followed by multiple hypothesis testing corrections. If an enrichment of small p-values is observed, then it is presumed that some genes (or other genetic elements) are truly associated with the phenotype. This conclusion rests on the assumption that under the null hypothesis of no association, each data point is sampled independently from a common null distribution, in which case uniform p-values would be observed. However, when applied to genome-scale datasets, phylogenetic methods often show atypical statistical behavior in which the expected uniform distribution of p-values is not observed when using null phenotypes (Figure 2.1). For example, the standard *RERconverge* analysis is anti-conservative when applied to the marine phenotype but conservative when applied to the long-lived large-bodied phenotype. *Forward*

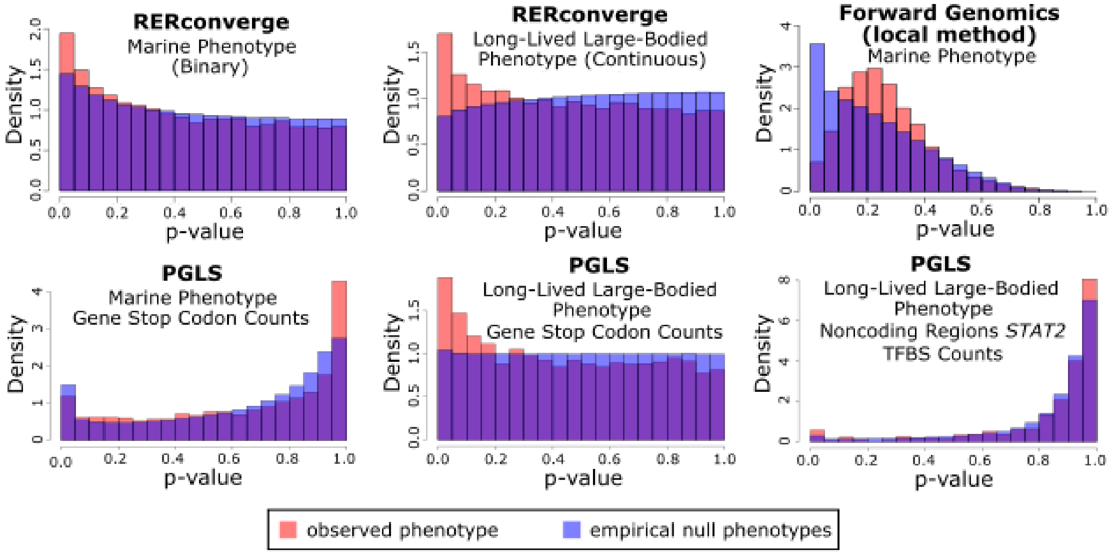


Figure 2.1: Parametric p-values from phylogenetic analyses deviate from the expected uniform distribution when assessed on null phenotypes. Histograms comparing p-values obtained using an observed phenotype (red) compared to p-values obtained from 500 (or more) null phenotypes from permulations. We evaluate a binary phenotype (marine) and a continuous phenotype (long-lived large-bodied) through *RERconverge*, a binary phenotype (marine) through *Forward Genomics*, and a binary phenotype (marine) and a continuous phenotype (long-lived and large-bodied) through PGLS with gene stop codon counts and noncoding element *STAT2* TFBS counts. In all cases, the empirical null from permulations (shown in blue) is non-uniform. Since null p-value distributions are often non-uniform (shown in blue), observed parametric p-values from standard statistical tests (shown in red) cannot be interpreted using traditional strategies.

Genomics likewise produces large deviations from the expected null. This issue exists for even the widely used PGLS method, which produces a near-uniform null when applied to gene loss in long-lived large-bodied mammals, but an extremely skewed distribution when applied to loss of transcription factor binding sites in the same phenotype.

The fact that a non-uniform null is observed for even the simple PGLS method demonstrates that deviations from the expected null cannot be explained by the phylogenetic structure of the data alone, but can also result from other sources of dependence that arise in the context of large multiple alignment datasets. Differences in genome quality [105], nucleotide frequencies [207], a mis-specified phylogeny, or other unknown systematic effects all create systematic biases which accumulate when the method is applied to thousands of genomic regions. As such, even if the tests can be proven to be theoretically valid under some assumptions (such as the well-understood PGLS model), they are not guaranteed to produce the expected uniform distribution when applied repeatedly to data from the same multiple sequence alignment. This deviation from the null expectation can result in overestimated statistical confidence and produce spurious genotype-phenotype associations.

The problem is further compounded when results from genetic elements are aggregated at the pathway level. Beyond the existing biases that arise from the nature of multiple sequence alignments, geneset analyses suffer additional non-independence induced by the evolutionary process itself. It is well established that genes that are functionally related experience correlated evolutionary pressure and thus evolve in a dependent fashion [42, 43, 118]. One extreme example of such coevolution is "reductive evolution", where losing a member of interacting proteins decreases the selection pressure for preserving its interacting partners [179]. As a result of coevolution, many functionally related genes "travel in packs" in association with a phenotype, meaning that if one gene in a group appears to be associated with a phenotype, the other genes in the group will as well because they do not evolve independently. The result is that a function could appear as associated with the phenotype due to random chance instead of actual involvement, causing an erroneous inference of enrichment.

The implication of coevolution is apparent when we apply standard pathway enrichment analysis to gain insight into which groups of functionally related genes are overrepresented among convergently evolving genes, as implemented in standard tools such as *Gorilla*, *GO::TermFinder*, and *RERconverge* enrichment functions [25, 58, 59, 128]. Figure 2.2 demonstrates how correlated evolutionary rates can cause problems in pathway enrichment analyses. When genes are ranked based on gene-phenotype associations, coevolving genes tend to have clustered ranks. Such clusters make it easier to observe enrichment of extreme ranks, or coevolving genes that all have either high or low ranks, due to chance alone, and

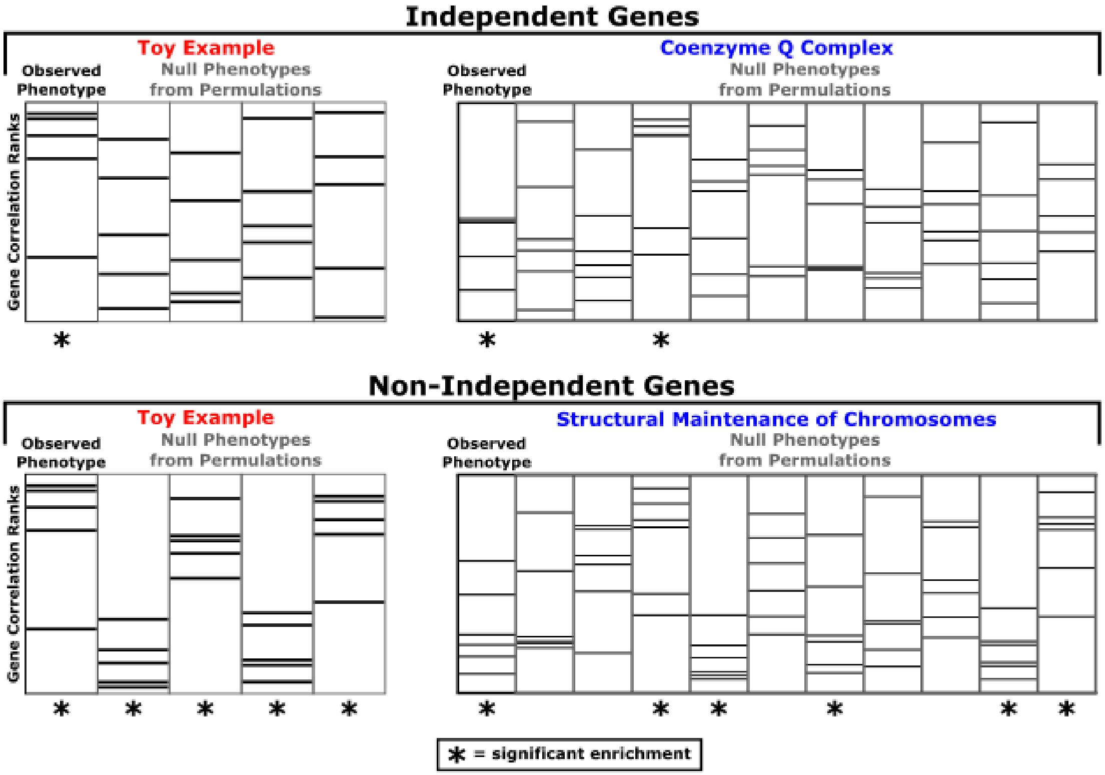

Independent Genes
Toy Example
Observed Phenotype
Null Phenotypes from Permulations
Coenzyme Q Complex
Observed Phenotype
Null Phenotypes from Permulations
Gene Correlation Ranks
Non-Independent Genes
Toy Example
Observed Phenotype
Null Phenotypes from Permulations
Structural Maintenance of Chromosomes
Observed Phenotype
Null Phenotypes from Permulations
Gene Correlation Ranks
* = significant enrichment

Figure 2.2: Pathway enrichment statistics from *RERconverge* long-lived large-bodied analyses demonstrate artificially inflated significance because genes in many pathways are non-independent. Accordingly, null phenotypes from permulations often show false signals of enrichment.

therefore the typical null expectation does not hold. Even when using a null phenotype, genes appear to cluster at the extremes of the ranked list. The clustering, and resulting enrichment, is caused by the genes “traveling in packs”, in which case simple enrichment tests assign undue confidence to an essentially spurious enrichment.

Rigorous statistical handling needs to be employed to address these sources of bias. Systematic solutions have been devised to correct issues with non-independence, both in the contexts of quantitative genetics [2] and phylogenetics [236]. However, these systematic approaches often make assumptions on the evolutionary process or other distributional assumptions, which may not accurately represent the data. We argue that an empirical approach that is grounded in the observed data can provide better calibration against sources of bias. In the context of gene expression, this problem is typically handled by performing label permutations [152, 205, 240] and in certain cases parametric adjustments [261]. However, simple label permutations are not applicable to associations involving a phylogeny as they would not preserve the underlying phylogenetic relationships, thereby producing false positives.

Here, we develop a novel strategy that combines permutations and phylogenetic simulations to generate null phenotypes, termed "permulations". The strategy addresses statistical non-independence empirically by generating phenotype permutations from phylogenetic simulations. In this way, the strategy preserves the underlying phylogenetic dependence by sampling permutations from the correct covariance structure. It also more accurately mimics the null expectation for a given phenotype by exactly matching the distribution of observed phenotype values for continuous phenotypes and exactly matching the number and structure of foreground branches (branches on which the phenotype changes) for binary phenotypes. We use these "permulated" phenotypes to calculate empirical p-values for gene-phenotype associations and pathway enrichment related to a phenotype. In doing so, we have created a statistical pipeline that accurately reports confidence in relationships between genetic elements and phenotypes at the level of both individual elements and pathways.

2.3 Materials and Methods

2.3.1 Permulations: A Hybrid Approach of Using Permutations and Phylogenetic Simulations to Generate Null Statistics

The goal of permulations is to empirically calibrate p-values from phylogenetic methods by producing permutations of the phenotype tree that account for the structure in the data. The permulation method requires a master species tree and a species phenotype (either continuous or binary). The method then returns a set of phenotypes that are random but preserve the phylogenetic dependence of the input phenotype. We typically generate 1,000 such permulated phenotypes, which are then used in the framework of a certain phylogenetic method (e.g., *RERconverge*) to compute gene-trait associations, resulting in 1,000 empirical null statistics for each gene. Similarly, we can also run enrichment analyses using the permulated phenotypes to produce 1,000 empirical null statistics for each pathway. Finally, for each gene or pathway, we calculate the empirical p-value as the proportion of empirical null statistics that are as extreme or more extreme than the observed parametric statistic for that gene or pathway. Since empirical null statistics capture the true null distributions for genes and pathways, the empirical p-values represent the confidence we have to reject the null hypotheses of no association, correlation, or enrichment given the underlying structure of our data. Note that permulations do not eliminate the need for multiple hypothesis correction; even with a corrected null model, the likelihood that false discoveries are made from performing multiple statistical inferences simultaneously still exists. Our permulation methods for binary and continuous phenotypes have been included in the publicly available *RERconverge* package for R [128] (published on github at `https://github.com/nclark-lab/RERconverge`), with a supplementary walkthrough also available as a vignette included in the *RERconverge* package.

2.3.1.1 Phylogenetic Permulation for Continuous Phenotypes

For continuous traits, generating permulated phenotypes is a two-step process. First, null phenotype values are simulated. Second, real phenotype values are assigned based on the simulated values. In step one, given the master tree with branch lengths representing average evolutionary rates and phenotype values for each species, we simulate a random phenotype using the Brownian motion model of evolution. The Brownian motion model takes a “random walk” down the master tree phylogeny to assign phenotype values. Since more closely related species are a shorter “walk” from each other, they are more likely to have more similar phenotype values than more distantly related species. In step two, real phenotype values are assigned to species based on ranks of the simulated values. The species with the highest simulated value is assigned the highest observed value, the species with the second-highest simulated value is assigned the second highest observed value, and so on.

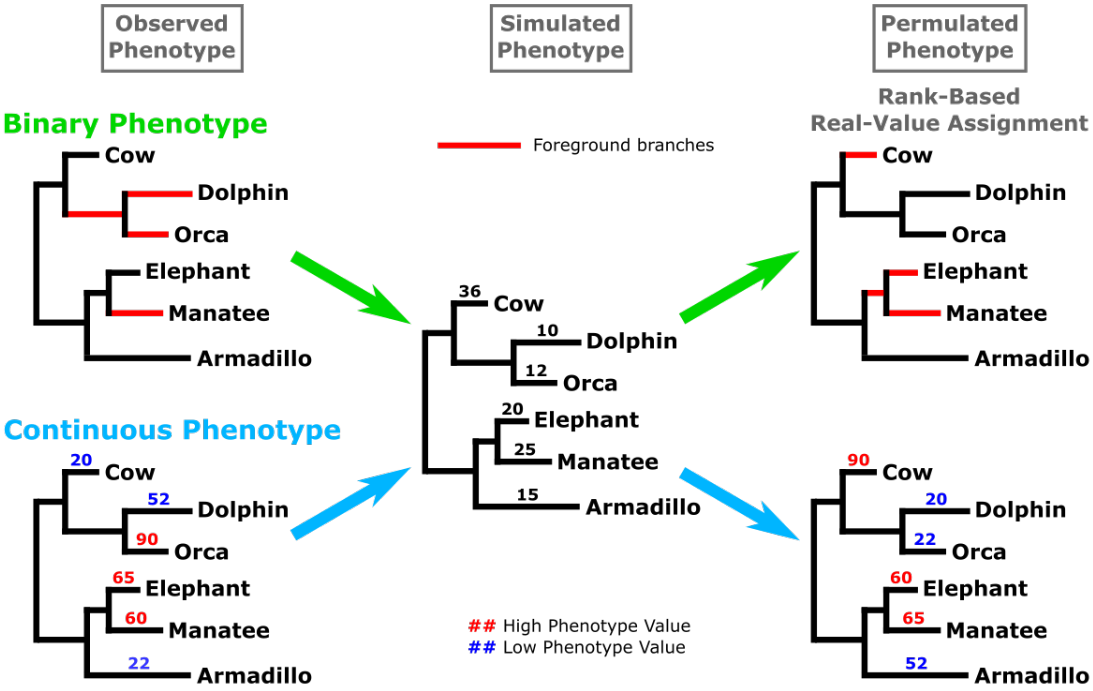


Figure 2.3: Permulated phenotypes were generated by simulating phenotypes and then assigning observed phenotype values based on the rank of simulated values. Simulations were performed using Brownian motion phylogenetic simulations and a phylogeny containing all mammals with branch lengths representing the average evolutionary rate along that branch genome-wide. For binary phenotypes, foreground branches for permulated phenotypes are assigned based on the highest-ranked simulated values while preserving the phylogenetic relationships between foregrounds. For continuous phenotypes, observed numeric values were assigned directly to species based on ranks of simulated values.

By doing so, observed phenotypes are shuffled among species with respect to the underlying phylogenetic relationships among the species. Since simulated values are more similar among more closely related species compare to distantly related species, the newly reassigned real values follow the same pattern (Figure 2.3).

2.3.1.2 Phylogenetic Permulation for Binary Phenotypes

For binary traits, the critical feature is the number of foreground species and their exact phylogenetic relationship, and hence the inferred number of phenotype-positive internal nodes or equivalently phenotypic transitions. The two-step process proposed above does not guarantee to perfectly preserve this structure. Instead, we employ a rejection sampling strategy where the simulation is used to propose phenotypes which are accepted only if they match the stricter requirements. Specifically, species are ranked based on simulated values, and a set of top-ranked species chosen to match the number of foreground species in the observed phenotype are proposed as a null phenotype. The proposed phenotype is only accepted if it preserves the phylogenetic relationships among chosen foregrounds, as observed in the actual foregrounds (Figure 2.3, Binary Phenotype). Using the simulation as the proposed distribution ensures that phylogenetically dependent phenotypes are generated and thus speeds up the construction of null phenotypes over what can be achieved from random selection.

We present two binary permulation strategies: the complete case (CC) method and the species subset match (SSM) method. The SSM method accounts for the fact that not all genes have orthologs in all species while the CC method ignores species presence/absence for simplicity. The strategies encompass the trade-off between computational feasibility and statistical exactitude—in some cases, it may not be possible to perform the SSM method, in which case the CC method is a viable alternative. The CC method is the first and simpler strategy. The CC method performs permulations using the master tree in which all species are present and therefore generates permulated trees that contain the complete set of species. Since not all species will have sequences available for all genes and the CC method produces one set of permulated phenotypes for all the genes, the exact number of foreground and background species per genetic element may not be preserved because of species presence/absence in those alignments (Figure 2.4). Thus, the CC method is an imperfect but fast method to generate null phenotypes, but we recommend use of the SSM method whenever feasible.

In contrast, the SSM method accounts for the presence/absence of species in different

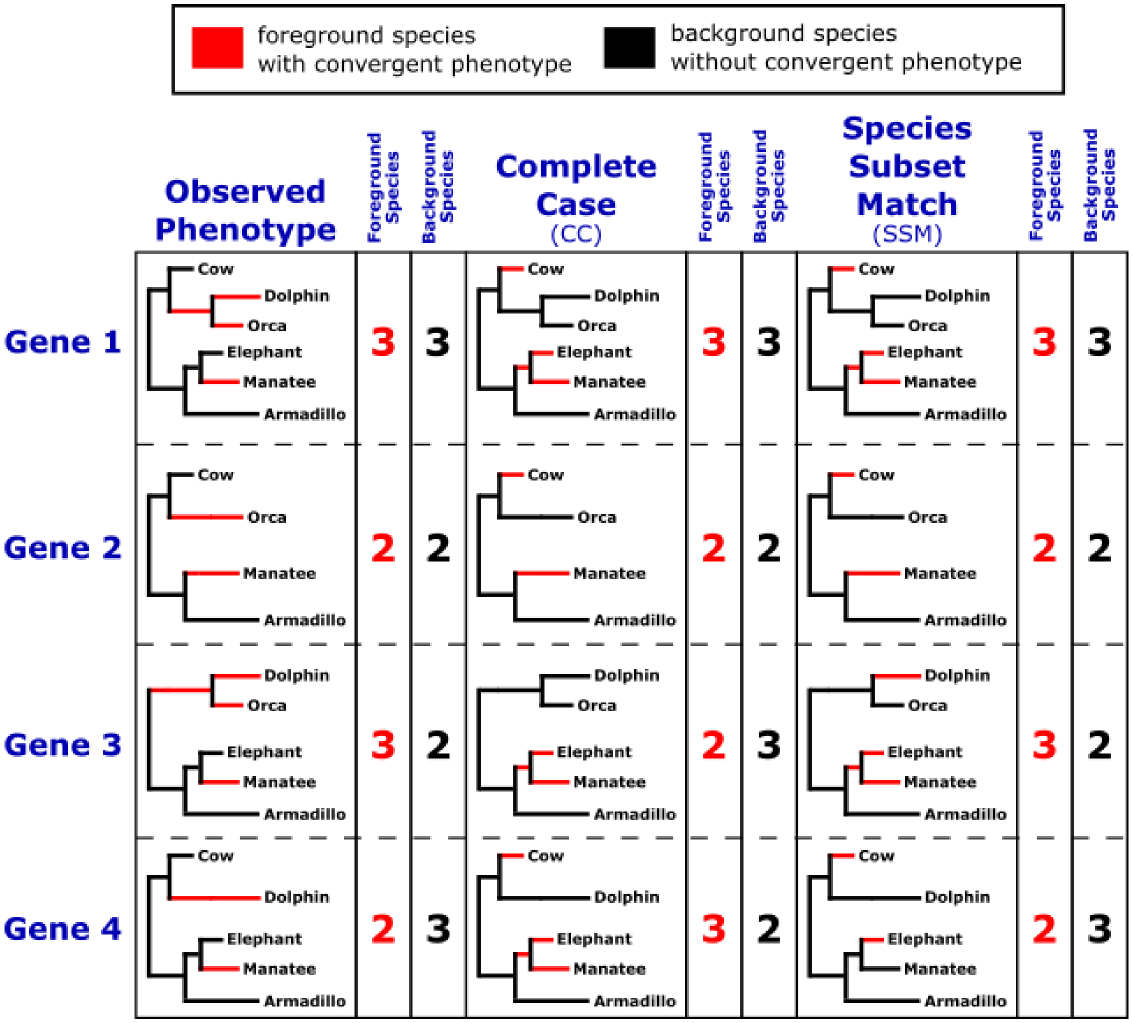


Figure 2.4: “Complete Case” and ”Species Subset Match” binary permulations. Examples of toy binary phenotypes permulated using the complete case (CC) method or the species subset match (SSM) method. For the CC method, top-ranked simulated values are assigned as foreground branches regardless of gene-specific species absence. For the SSM method, top-ranked simulated values are assigned as foreground branches after considering gene-specific species absence so the number of foreground and background species for each gene is consistent across every permulated phenotype. Note that in the case of genes with all species present (e.g., Gene 1), CC and SSM methods are identical.

gene trees. For each permulation, the SSM method generates separate null phenotypes for each tree in the set of genetic elements. Since genetic element-specific trees contain exactly the species that have that genetic element, the null phenotypes exactly match the observed phenotypes for that genetic element in terms of number of foreground and background species (Figure 2.4). Additionally, unlike the CC method, null phenotypes for a single permulation iteration are distinct, and potentially unique, from each other because they are generated on a genetic element-by-genetic element basis. Although the SSM method is statistically more ideal than the CC method, it is much more computationally intensive and may not be feasible for very large datasets. For example, the CC method took 7 seconds to produce 50 permulated traits for 200 genes, whereas the SSM method took 15.5 minutes.

2.3.2 Implementation of Permulation Methods

As shown in Figure 2.3, each permulated phenotype is generated by first performing a phylogenetic simulation using an established phylogenetic topology. To generate the master tree, whose branch lengths represent the average evolutionary rates of all genetic elements in the dataset for each species, the function *readTrees* in *RERconverge* can be used. Next, the master tree and the trait values (binary or continuous) are used to compute the expected variance of the phenotype per unit time, and subsequently perform a Brownian motion simulation to simulate branch lengths; the R package *GEIGER* [96] is used to perform both operations. Simulated values are then used in different ways for binary and continuous phenotypes to generate permulated phenotypes.

In *RERconverge*, CC and SSM permulations are performed using the *getPermsBinary* function, by setting the argument "*permmode*" to "*cc*" or "*ssm*", respectively. The function requires the user to supply information on the original foreground species and their relationships by specifying 1) the names of the extant (tip) foreground species and 2) an R list object containing pair(s) of sister species whose common ancestor(s) is to be included in the foreground set as well. Using these inputs, the function infers the original phenotype tree and assigns the phenotype values to the correct branches (1 for foreground, 0 for background), which is subsequently used as constraints for the permulation. Phylogenetic simulations are then run using the master tree to assign simulated branch lengths to the tree branches.

For the CC permulation, the n tip branches with the highest trait values from the simulation, where n is the number of observed tip foregrounds, are selected as the new foregrounds. The function then calls the *foreground2Tree* function in *RERconverge* with "*clade*" set to "*all*" to construct a binary tree with a foreground set that includes all branches (tip and internal) in the foreground clades. A valid permulation has the same number of internal and tip foreground branches as the original phenotype. Thus, permulated phenotypes with an incorrect foreground configuration are rejected and phenotype generation is repeated until the correct number of permulations is achieved. Note that the CC method uses the same permulated phenotype for every genomic element, so statistics for some genes will not be calculated for some permulations because of species presence/absence across genes. In other words, some genes will have fewer total permulations because of the way permulated phenotypes are constructed. The exact number of foreground and background species may also differ across each permulated phenotype for the same gene.

The SSM permulation matches the tree topology of the permulated phenotypes to the tree of individual genes. To do this, the SSM permulation follows the same steps as described above, with an additional step of trimming off branches that are missing in the gene tree. In this case, the m longest tip branches (where m is the number of observed tip foregrounds in the gene tree) are chosen as new tip foregrounds to run *foreground2Tree*. Thus, in the SSM method, genes with different tree topologies will have different sets of permulations. However, for each unique topology, the number and phylogenetic relationships of the foregrounds are preserved. Figure 2.4 shows examples of CC- and SSM-permulated trees for 4 genes with distinct topologies.

For the continuous phenotype, the function *simpermvec* generates a permulated phenotype given the original phenotype vector and the underlying phylogeny with appropriate branch lengths. The master tree from the *RERconverge readTrees* function is appropriate to use for simulations. In most cases, the user will not have to use the *simpermvec* function directly—instead, the *getPermsContinuous* function that calculates null empirical p-values for gene correlations and pathway enrichments will call *simpermvec* internally.

After calculating empirical null statistics and p-values, empirical p-values per gene are calculated by finding the proportion of null statistics from permulated phenotypes that are as extreme or more extreme than the statistic calculated using the real phenotype. This proportion represents the proportion of times that random chance produces a concordance between gene and phenotype evolution that is as strong as the observed statistic, given the underlying structure of the data. In *RERconverge*, the *permpvalcor* function calculates the empirical p-values for a given set of permulation association statistics. Note that since empirical p-values are a proportion of total permulations, the precision of empirical p-values is based on the total number of permulations performed. For example, with 1,000 permulations, the lowest reportable p-value is 0.001 and empirical p-values calculated as 0 must be reported as <0.001 because we only have precision to report p-values to the thousandths place.

Finally, to determine the number of permulations that can provide sufficient correction for systematic bias, the function *plotPositivesFromPermulations* can be used to plot how the number of significantly accelerated or conserved genetic elements changes with increasing number of permulations. From the generated plot, users can determine the minimum number of permulations by evaluating when the number of positives start to stabilize.

2.3.3 Empirical p-values for Pathway Enrichment

Empirical null statistics and p-values for pathways are calculated using the empirical null statistics and p-values for individual genes. For each set of empirical null statistics generated from a particular permulated phenotype, genes are assigned the log of the empirical null p-value times the sign of the empirical null statistic for that permulation. Empirical null pathway statistics are calculated for each permulation using those values with the *RERconverge* function *fastWilcoxGMTall* that performs a Wilcoxon Rank-Sum test comparing values from genes in a pathway to values in background genes. The function *getEnrichPerms* calculates null enrichment statistics given a set of null correlation statistics, or, alternatively, *getPermsBinary* and *getPermsContinuous* calculate both null correlation and null pathway enrichment statistics simultaneously by default for the binary and continuous phenotypes, respectively. Empirical p-values for pathway enrichment are then calculated as the proportion of empirical null statistics that are as extreme or more extreme than the observed enrichment statistic using the *permpvalenrich* function. Pathways that show significant parametric p-values and non-significant empirical p-values are likely cases of genes "moving in packs" and are not truly significantly enriched.

2.3.4 Phylogenetic Methods for Benchmarking

2.3.4.1 ***RERconverge***

RERconverge finds associations between genetic elements and phenotypes by detecting convergent evolutionary rate shifts in species with convergent phenotypes. The method operates on any type of genetic element and has been used successfully for both protein-coding and noncoding regions. Prior to running *RERconverge*, phylogenetic trees for each genetic element are generated using the Phylogenetic Analysis by Maximum Likelihood (PAML) program [268] or related method, with branch lengths that represent the number of substitutions that occurred between a species and its ancestor. Raw evolutionary rates are converted to relative evolutionary rates (RERs) using *RERconverge* functions *readTrees* and *getAllResiduals*, which normalize branches for average evolutionary rate along that branch genome-wide and correct for the mean-variance relationship among branch lengths [188]. RERs and phenotype information are then supplied to *correlateWithBinaryPhenotype* or *correlateWithContinuousPhenotype* functions to calculate element-phenotype associations. Kendall's τ associations are calculated for binary phenotypes, and Pearson correlation values are calculated for continuous phenotypes, both by default.

After calculating association statistics, signed log p-values for associations are used to calculate pathway enrichment using the rank-based Wilcoxon Rank-Sum test. The *fastWilcoxGMTAll* function in *RERconverge* calculates pathway enrichment statistics over a list of pathway annotations using all genes in a particular annotation set as the background.

2.3.4.2 Phylogenetic Generalized Least Squares (PGLS)

PGLS analyses were conducted through R as implemented in the "*nlme*" package using the *gls* function. Within-group correlation structure was defined using the *corBrownian* function from the "*ape*" package and a master tree with branch lengths representing genome-wide evolutionary rates per species.

2.3.5 Datasets for Method Evaluation

We evaluated the performance of our permulation methods by using *RERconverge* to find genetic elements that demonstrated convergent acceleration of evolutionary rates in association with convergent phenotypic adaptations that are well-characterized, namely the evolution of the marine mammal phenotype [40, 166], the subterranean mammal phenotype [187], and the long-lived large-bodied mammal phenotype [129]. For the remaining part of this article, we will refer to these phenotypes as the marine phenotype, the subterranean phenotype, and the long-lived large-bodied phenotype, respectively. We used the set of protein-coding gene trees across 63 mammalian species previously computed by Partha et al. [188]. These trees have the "Meredith+" tree topology [129] (Figure 2.5), a modification of the tree topologies published by Meredith et al. [165] and Bininda-Emonds et al. [22], resolved for their differences across various studies as originally reported by Meyer et al. [166].

For the binary marine phenotype, we set three independent lineages as foreground species that possessed the marine trait (blue branches in Figure 2.5, Binary Phenotype): pinnipeds (Weddell seal, walrus), cetaceans (bottlenose dolphin, killer whale, the cetacean ancestor), and sirenians (West Indian manatee) [40]. For the subterranean phenotype, we set as foregrounds three independent subterranean species for which high quality genomes were available in our dataset: naked mole-rat, star-nosed mole, and cape golden mole (red branches in Figure 2.5, Binary Phenotype).

Finally, for the continuous long-lived large-bodied phenotype, we used the "3L" trait as defined in previous work [129]. The numerical phenotype was constructed by calculating the first principal component (PC1) between body size and maximum lifespan across 61 mammal species (Figure 2.5, Continuous Phenotype). PC1 therefore represents the agreement between body size and lifespan—species like whales with long lifespans and large sizes have large phenotype values and species like rodents with short lifespans and small sizes have small

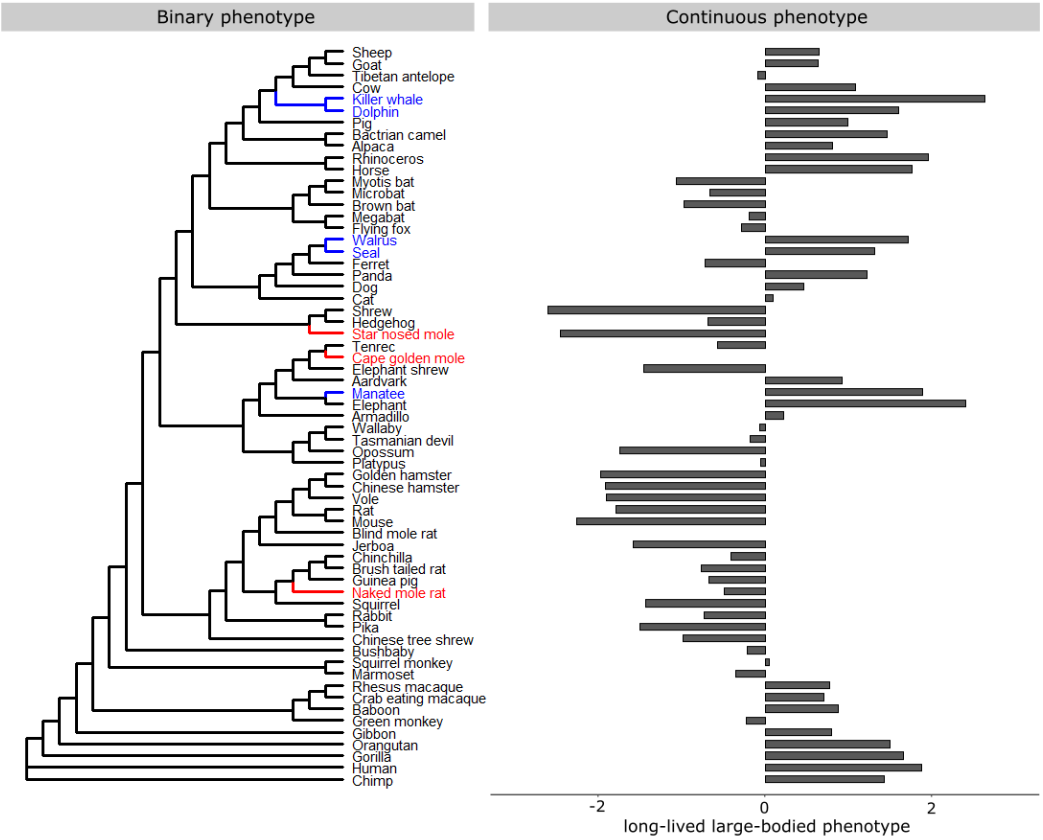


Figure 2.5: Meredith+ tree topology and the binary and continuous phenotypes evaluated. Binary phenotypes include the marine mammal phenotype and the subterranean mammal phenotype (foreground branches are indicated in blue and red, respectively). The continuous phenotype evaluated is the long-lived large-bodied phenotype as constructed based on the first principal component between species body size and maximum longevity.

phenotype values. For example, killer whale, elephant, and rhino have the highest values (2.63, 2.40, and 1.95) because they are both large and long-lived, whereas shrew, star-nosed mole, and mouse have the smallest values (-2.62, -2.46, and -2.27) because they are small and short-lived. Human, while longest-lived among the mammals included, has the fifth largest value (1.87) because humans are relatively small compared to the other mammals. Likewise, large grazing animals like cow also have smaller PC1 values (1.08, the 15th largest value) because although cows are large, they are not very long-lived given their body size.

Noncoding regions were identified based on evolutionary convergence from *phastCons* scores across the 63 mammal species as described here: `https://github.com/nclark-lab/RERconverge/blob/master/NoncodingRegionWorkflow`. Stop codon calls per gene were obtained from Meyer et al. [166] and were based on genome-wide calls across species.

TFBS calls were obtained using the HOCOMOCO *STAT2* binding site motif based on position weight matrix scores. Calls for 29,880 noncoding regions corresponding to human chromosome 1 were used for analyses. Of those regions, 560 had a sufficient number of calls and variation in calls across species to calculate PGLS statistics.

2.4 Results

2.4.1 Permulation of Binary Phenotypes Improved Power and Type I Error Control

To evaluate the performance of the permulation methods compared to the parametric method for binary phenotypes, we first used *RERconverge* to find genetic elements with convergently accelerated evolutionary rates in species with the marine phenotype. We considered three p-value calculation methods: parametric, complete case (CC) permulations, and species subset match (SSM) permulations. The resulting p-values were corrected for multiple hypothesis testing using Storey's correction [237]. We see in Figure 2.1 that the parametric p-values for the association of genes with the observed marine phenotype (red histogram) were enriched for small p-values. According to the standard parametric approach, which assumes a simple null hypothesis with uniformly distributed p-values, the enrichment of low p-values indicated the possible presence of genes with evolutionary rate shifts that were significantly correlated with marine adaptation. However, when we constructed the empirical null p-value distribution using 1,000 permulations of the marine phenotype, the null distribution of parametric p-values was not uniform. In fact, the enrichment of low p-values was also present in the null distribution (blue histogram), albeit a lesser enrichment than the observed, meaning that observing low p-values by chance was more likely than expected. Thus, if we used standard multiple testing procedures directly on the parametric p-values, we would identify more positive genes than the true number of positives, in other words causing an undercorrection of p-values.

To demonstrate that our permulation strategy effectively corrected for the background p-value distribution, we plotted similar histograms of the empirical p-values for the marine phenotype versus 1,000 permulated phenotypes, generated from both CC and SSM permulations. With permulations, we can see that while some enrichment of small empirical p-values was observed for the marine phenotype, the empirical p-values for the null phenotypes were almost perfectly uniform, meaning that our permulation methods were able to construct the correct null distribution (Figure 2.6). When we overlaid the p-value histograms of the parametric and empirical p-values for the marine phenotype, we can see that compared to the parametric method, the histograms for the CC and SSM permulations had steeper slopes at low p-values, indicating that the permulation methods had better Type I error control (Figure 2.7A). Furthermore, the histograms for the permulation methods plateaued at higher π_0 than the parametric method, consistent with the postulation that the parametric method would identify more (possibly false) positives. These findings were also observed when we defined genes with significant evolutionary acceleration in marine mammals (i.e., "marine-accelerated" genes) by setting a rejection threshold of Storey's false discovery rate (FDR) ≤ 0.4 (the high threshold was set considering the high minimum FDR from the parametric method), as shown in Figure 2.7B. For the permulation methods, as the number of permulations increased, the number of identified marine-accelerated genes increased and eventually stabilized after $\sim$400 permulations. The asymptotic numbers of marine-accelerated genes identified by permulations ($\sim$350 genes for CC permulation and $\sim$450 genes for SSM per-

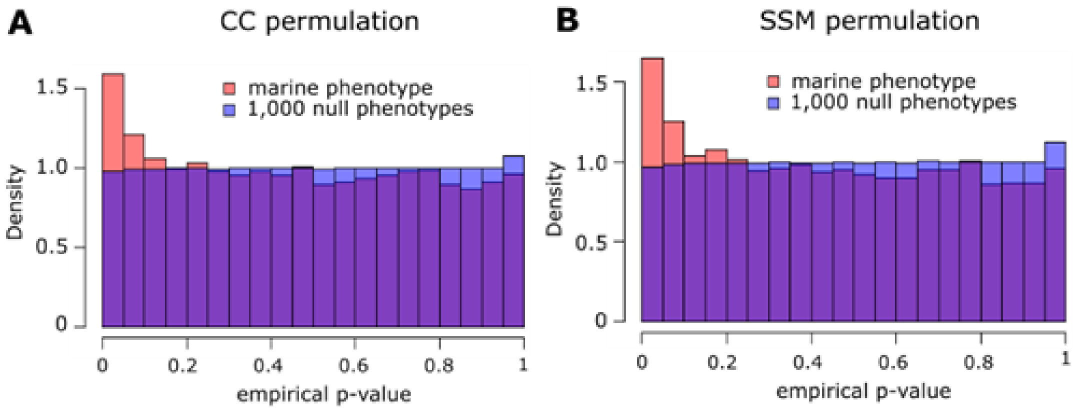


Figure 2.6: Histograms of empirical p-values computed for the marine phenotype (red) and 1,000 null phenotypes (blue) produced using (A) the Complete Case (CC) permulation and (B) the Species Subset Match (SSM) permulation methods.

mulation) were much smaller than the ~700 genes identified through parametric statistics, demonstrating improved Type I error control.

Surprisingly, while the permulation methods identified fewer significantly accelerated regions, we could have greater confidence in their significance. Figure 2.7C shows the minimum FDRs achieved by the permulation methods with increasing number of permulations. The figure shows that the permulation methods provided better control of FDRs compared to the parametric method with only a few permulations (above ~125 permulations). With increasing permulations, the minimum FDR continued to drop to reach levels below 0.1 at 1000 permulations, while the minimum FDR from parametric statistics was higher at above 0.3. Use of the permulation null substantially improved the statistical power of the method and provided much higher confidence in detecting true correlations between evolutionary rate shifts and the convergent phenotype of interest.

Lastly, we found that permulation methods could identify marine-accelerated genes that were missing in many species, i.e., genes with phylogenetic trees containing few species. In contrast, the parametric method failed to identify any such gene (Figure 2.7D).

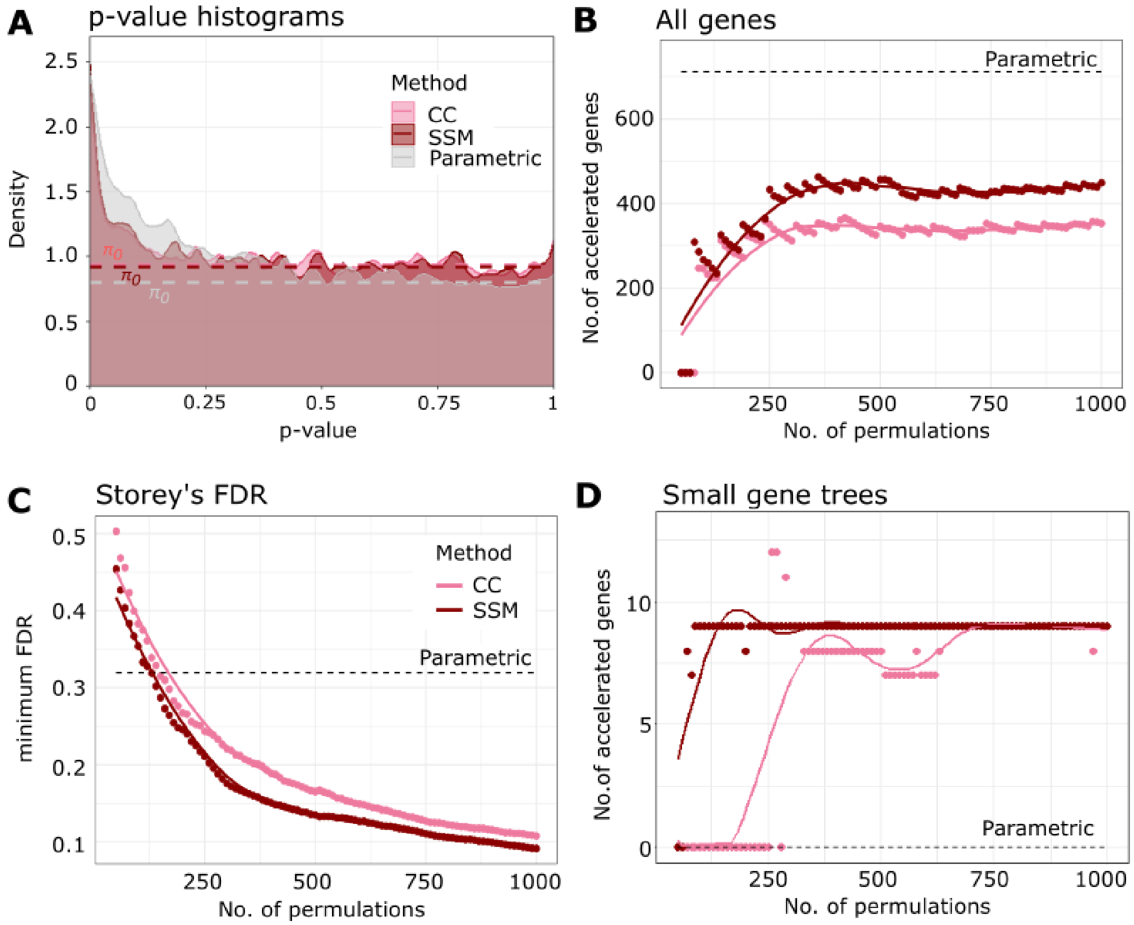


Figure 2.7: Permulation of binary phenotypes corrects for inflation of statistical significance in finding evolutionarily accelerated genes in marine mammals. (A) Histogram of parametric and permulation p-values for the marine phenotype from the parametric, the complete case (CC) permulation, and the species subset match (SSM) permulation methods. (B) Permulation methods identify fewer accelerated genes in marine mammals compared to the parametric method, correcting for the inflation of significance. The rejection region of the multiple hypothesis testing is set to be Storey's FDR ≤ 0.4, considering the weak power of the parametric method. (C) Binary permulation methods have greater statistical power compared to the parametric method, as shown by the minimum false discovery rate (FDR) calculated using Storey's method. (D) Permulation methods can identify accelerated genes that are missing in many species (gene tree size ≤ 30), whereas the parametric method fails to do so.

2.4.2 Binary Permulation Methods Improved Gene-level Detection of Functional Enrichment

We have demonstrated that the permulation methods showed favorable statistical properties based on the distribution of p-values. We expected that this approach also improved the biological signal of rate convergence analysis. To address this question, we asked if the marine-accelerated genes identified by binary permulations were enriched for functions that were consistent with the marine phenotype. Our group previously identified marine-specific pseudogenes that should be undergoing accelerated evolution in marine mammals due to relaxation of evolutionary constraint [166]. Putative pseudogenes associated with marine mammals were identified using *Bayes Traits* software [184] to find signals of coevolution between marine status and pseudogenization. In addition, our group also previously found that marine-accelerated genes that evolved under relaxed constraint were enriched for genes responsible for the loss of olfactory and gustatory functions [40]. Thus, to represent the "ground truth", we selected a collection of gene sets relevant to olfactory and gustatory functions from the Mouse Genome Informatics (MGI) database and top-ranking marine-specific pseudogenes with *Bayes Traits* FDR values less than 0.25.

We then performed the one-tailed Fisher's exact test to measure the enrichment of the functions in the marine-accelerated genes from the parametric and permulation methods. The Fisher's exact test odds ratios indeed showed that the CC and SSM permulation methods generally magnified or maintained the effect sizes of enrichment across the gene sets compared to the parametric method (Figure 2.8A). At worst, the permulation methods matched the performance of the parametric method (e.g., "taste/olfaction phenotype" gene set). The improved performance of the permulation methods was also demonstrated in the example precision-recall curves for the marine-associated pseudogenes in Figure 2.8B.

To see if this observation generalized to other phenotypes, we repeated the whole analysis to find genes that were accelerated in species with the subterranean phenotype. As subterranean-accelerated genes have been found to be enriched in ocular functions [187, 188, 195], we picked gene sets relevant to vision-related functions as the "ground truth". In general, the signals we obtained from *RERconverge* for the subterranean phenotype were

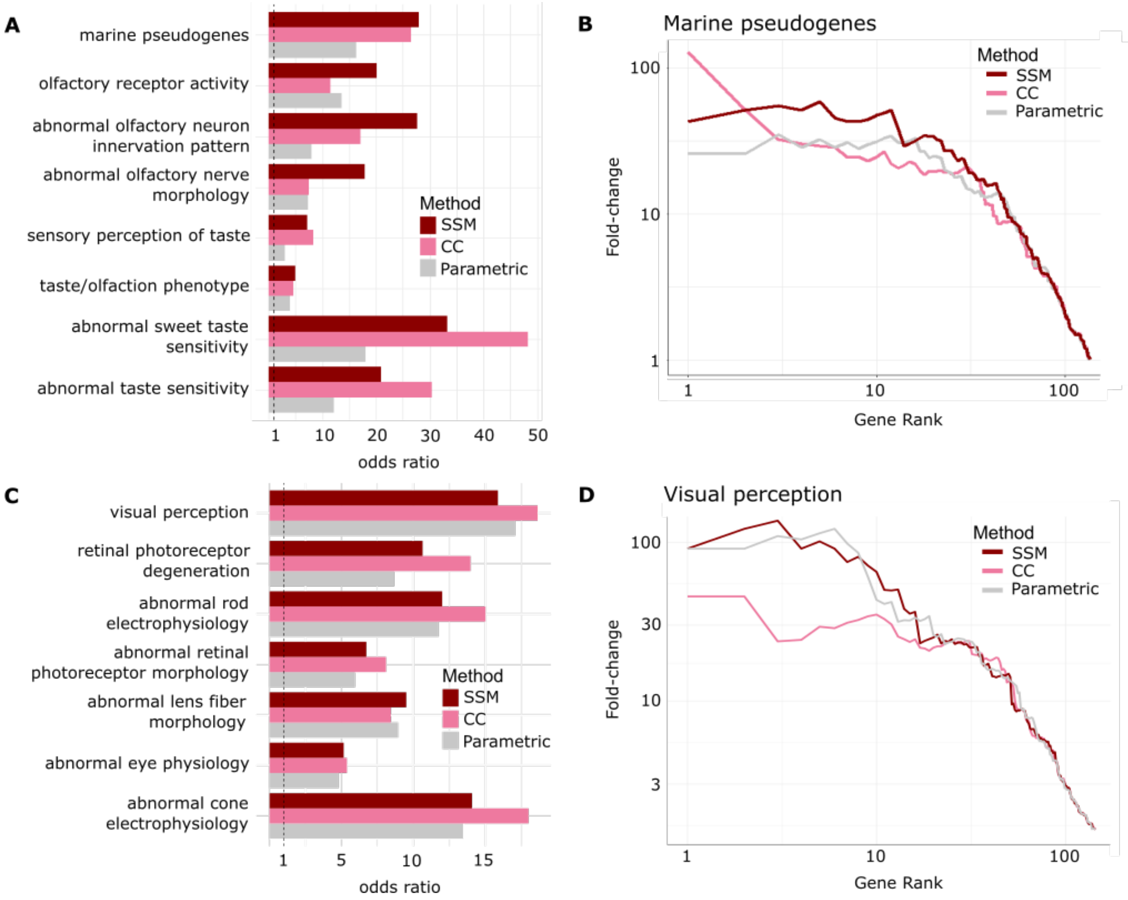


Figure 2.8: Binary permulation methods have matching or improved power compared to the parametric method in detecting enrichments of functions consistent with known phenotypes. (A) Fisher's exact test odds ratios showing that marine-accelerated genes identified by the permulation methods have greater enrichment of gustatory genes, olfactory genes, and marine pseudogenes, compared to the parametric method. (B) Precision-recall curves for the enrichment of the marine pseudogenes in the identified marine-accelerated genes. Greater area under the curve (curves that have higher values on the left side of the plot) have greater enrichment. (C) Fisher's exact test odds ratios showing that subterranean-accelerated genes identified by the permulation methods have greater or comparable enrichment of ocular genes, compared to the parametric method. (D) Precision-recall curves for the enrichment of the visual perception genes in the identified subterranean-accelerated genes.

much weaker than in the marine phenotype case, but the enrichment was still captured in the rankings of the genes. Similar to the marine phenotype, permulation methods generally improved or matched the performance of the parametric method (Figures 2.8C-D).

2.4.3 Binary Permulation Method Corrects for False Positives in Related Approaches

In addition to performing permulations using *RERconverge*, we tested our methods using *Forward Genomics* and PGLS. Other methods, such as *PhyloAcc*, would require tens of millions of computational hours to generate 500 permulations (from the analysis with *RERconverge*, the number of identified accelerated genes plateaued after 400-500 permulations were used (Figure 2.7B)), and thus permulations were not scalable to those analyses.

Forward Genomics [103, 195], like *RERconverge*, tests for an accelerated evolutionary rate in a set of foreground species by correlating a normalized substitution rate with phenotypes using Pearson correlation. It works only for binary phenotypes and has demonstrated success in coding and non-coding elements. *Forward Genomics*' "global method" uses substitution rate with respect to each tree's root to correlate with trait loss and identify convergent relaxed selection; therefore, it does not correct for evolutionary relatedness. The "local branch method", an improvement on the original approach, uses substitution rate with respect to the most recent ancestor to identify relaxed selection, which substantially improves its power [195]. We used the most recent version of both the global and the local methods to test for associations between gene evolutionary rates and the binary marine phenotype.

Both global and local *Forward Genomics* methods had unusual p-value distributions. The local method identified high proportion of positives with significant p-values (Figure 2.1), while p-values from the global method were highly concentrated around 0.5 (global p-values not shown). Adjusting for multiple testing further exaggerated this issue. For the global method, due to the number of genes with very low p-values, the lowest possible Benjamini-Hochberg (BH) corrected parametric p-value was 0.531, and for the local method, the lowest possible corrected p-value was 0.465. For the local method, out of 18,797 genes, more than half of the genes (12,438) had the lowest possible corrected parametric p-value. As such, it was impossible to designate a significance cut-off, because it would either include no genes or include most of the genes. Applying the permulation strategy to *Forward Genomics* output, we found that of the same set, 889 had corrected empirical p-values that were less than or equal to 0.465 (the minimum observed corrected parametric p-value), allowing for a more reasonable selection of a rejection threshold. Thus, permulation can improve statistical performance even for a statistic with known flaws.

We further investigated our results from *Forward Genomics* at the pathway level in addition to analyzing results at the individual gene level. We used the marine pseudogenes as a "ground truth" set of genes that should be undergoing accelerated evolution in marine species, to test our ability to detect pathway enrichment of these genes. As shown in Figure 2.9A, the global and local parametric test statistics showed slight enrichment for elements that were pseudogenized in marine mammals, and the difference was improved when empirical p-values were computed. Figure 2.9B shows the same data as precision-recall plots, clearly demonstrating that the permulation correction improved the predictive power of both methods.

Next, we tested the effect of permulations on PGLS results. PGLS tests for association between two traits across species while adjusting for the phylogenetic relationships among those species. In doing so, it numerically corrects for non-independence due to phylogenetic relatedness. Note that unlike *RERconverge* and *Forward Genomics*, PGLS does not require evolutionary rate information and is therefore a more generalized phylogenetic analysis. We tested PGLS using both the binary marine and the continuous long-lived large-bodied phenotype for coevolution with stop codon counts across genes. We additionally tested the continuous phenotype for coevolution with *STAT2* transcription factor binding site counts across noncoding regions.

Like other methods, PGLS demonstrated unexpected null behavior that varied across genomic datasets and phenotypes (Figure 2.1). Although the null distribution of p-values for associations between the long-lived large-bodied phenotype and the stop codon counts showed only a slight inflation of low p-values (5.2% of null p-values below 0.05) and otherwise nearly uniform distribution, tests using the marine phenotype and the transcription factor binding site counts showed much different behavior. Permulations for associations

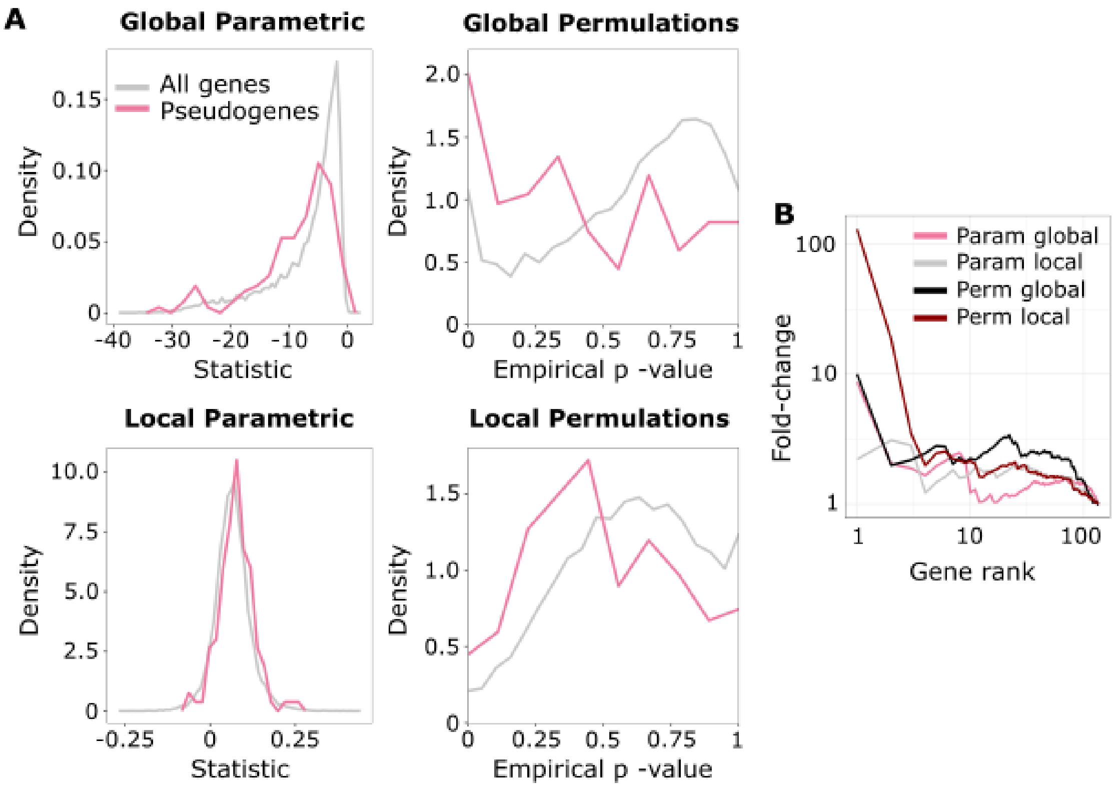


Figure 2.9: Binary permulation methods improve *Forward Genomics*' positive-predictive value and power. (A) Distributions of *Forward Genomics* statistics and corresponding permulation p-values for local and global methods. Both global and local statistics show slight shifts (to the left for global statistics and to the right for local statistics) indicating enrichment of marine mammal pseudogenes under accelerated evolution (global AUC=0.6235; local AUC=0.6196). Permulation p-values show a more dramatic shift toward significant values for marine pseudogenes under accelerated evolution for the global method (AUC=0.6653) and about the same shift for the local method (AUC=0.6086) compared to parametric statistics. (B) Precision-recall curves for the enrichment of pseudogenes in marine-accelerated genes using parametric statistics and permulation p-values for both local and global methods. Permulated values represent a unique ranking in which ties in permulation p-values for genes are broken based on parametric statistics. Permulation methods perform at least as well as both global and local methods, indicated by curves that are higher at the left side of the plot.

between the marine phenotype and stop codon counts revealed that, although there might appear to be a meaningful enrichment of low observed p-values, such enrichment was observed even when analyzing permulated phenotypes. Conversely, although the enrichment of low observed p-values appeared relatively less for associations between the long-lived large-bodied phenotype and transcription factor binding site counts in non-coding regions, such enrichment was indeed meaningful because it was greater than observed when analyzing permulated phenotypes. Together, these observations indicate that PGLS may exhibit aberrant statistical behaviors that the exact nature of the behaviors may vary greatly across datasets, and that permulations are a valid strategy to identify and correct those behaviors.

2.4.4 Permulations Improve Power to Detect Genes Correlated with a Continuous Phenotype

When we used *RERconverge* to evaluate the long-lived large-bodied mammal phenotype, a continuous phenotype, we observed that the Type I error rate was in fact too low. We demonstrated this by performing one thousand permulations to generate 1,000 null statistics and p-values for each gene, calculating empirical p-values as the proportion of null statistics that were as extreme or more extreme than the observed statistic per gene. As shown in Figure 2.1, the parametric null p-value distribution for genes associated with the long-lived large-bodied phenotype was non-uniform, and in fact sloped down at low p-values. This indicates that observing small p-values due to chance alone happened less often in our dataset than we would typically expect compared to the standard uniform expectation. In practice, the result of the non-uniform null was an overcorrection of parametric p-values using a standard multiple hypothesis testing correction. In other words, for this dataset, corrected parametric p-values were larger than they should be when using multiple hypothesis testing correction (such as a Benjamini-Hochberg correction) that assumed a uniform null. The null distribution of empirical p-values, however, did follow a standard uniform null by construction, so Benjamini-Hochberg corrected empirical p-values represented our true, higher confidence in a correlation between gene evolutionary rate and phenotypic evolution. We observed this increased confidence in our data—after multiple hypothesis testing correction, only 24 parametric p-values remained significant at an α threshold of 0.15, while 305 empirical p-values remained significant. Regardless of the increase in power, empirical p-values provide a more accurate representation of confidence in rejecting the null hypothesis, and thus are a more valid metric than parametric p-values.

2.4.5 Permulations Correct Pathway Enrichments for Genes with Correlated Evolutionary Rates

After generating null p-values and statistics from permulations for either binary or continuous traits, those values can be used to calculate null pathway enrichment statistics. Empirical p-values for pathways are then calculated as the proportion of null pathway enrichment statistics as extreme or more extreme than the observed statistic. This procedure corrects for gene sets with correlated evolutionary rates, that is genes whose rates will "travel in packs" regardless of any relation to the phenotype (e.g., Figure 2.2). Such groups of genes will tend to show enrichment more often than would be observed if the genes' rates were independent after conditioning on phenotype, resulting in false signals of pathway enrichment.

Permulations account for the non-independence problem by explicitly incorporating it into the null distribution used to calculate empirical p-values. In the demonstrated case of the Coenzyme Q Complex, only one permulation out of the ten depicted shows enrichment due to random chance (indicated by an asterisk * below the vertical bar in Figure 2.2), which would correspond to an empirical p-value of 0.1 in this toy example. This interpretation is identical to the standard p-value interpretation—-the proportion of times we expect to see a statistic as extreme or more extreme than observed assuming that the null expectation is true. In the case of permulations, we simply explicitly calculate the null expectation rather than using a predefined distribution (t-distribution, F-distribution, etc.). In the case of enrichment for a pathway with independent genes, the significance of the empirical p-value will agree with the significance of the parametric p-value because the null expectation from permulations agrees with the typical null expectation.

In the case of a pathway with genes with non-independent evolutionary rates, the empirical p-value will be larger than the parametric p-value because the empirical p-value will

Table 2.1: Top-enriched pathways with quickly evolving genes in association with the long-lived large-bodied phenotype according to parametric p-values. Note that due to the number of pathways, the lowest possible Benjamini-Hochberg corrected permulation p-value is 0.0913. Bolded values show significance at $\alpha = 0.25$. Note that many accelerated pathways that appear to be enriched based on parametric p-values are not enriched based on permulation p-values.

Pathway	**Statistic**	**p-adjusted**	**Perm p-adjusted**
Olfactory Signaling	0.217	9.25e-43	**0.199**
GPCR Signaling	0.0606	8.34e-7	0.596
Biological Oxidations	0.150	1.10e-6	0.276
Valine and Isoleucine Degradation	0.219	3.32e-5	0.354
Fatty Acid Metabolism	0.215	8.26e-5	0.352

penalize for non-independence. An example with “Structural Maintenance of Chromosomes” genes shows that, although there is an apparent enrichment based on the observed phenotype, half (5 out of 10) of permulated phenotypes show at least as strong enrichment for an empirical p-value of 0.5. Therefore, although the pathway does appear to be enriched from parametric statistics, its enrichment is actually not exceptional given the null expectation for that set of genes.

Empirical p-values are calculated for every pathway individually. Tables 2.1 and 2.2 shows top enriched pathways under accelerated evolution and decelerated evolution in association with the long-lived large-bodied phenotype. While most significantly enriched pathways under decelerated evolution based on parametric p-values also demonstrate significant empirical p-values, many pathways under significant acceleration show non-significant empirical p-values. Thus, this phenotype shows little evidence for accelerated pathway evolution associated with phenotypic evolution.

Table 2.2: Top-enriched pathways with slowly evolving genes in association with the long-lived large-bodied phenotype according to parametric p-values. Note that due to the number of pathways, the lowest possible Benjamini-Hochberg corrected permulation p-value is 0.0913. Bolded values show significance at $\alpha = 0.25$. Note that many accelerated pathways that appear to be enriched based on parametric p-values are not enriched based on permulation p-values.

Pathway	**Statistic**	**p-adjusted**	**Perm p-adjusted**
Cytokine-Cytokine Receptor Interaction	-0.181	3.40e-20	**0.0913**
Mitotic Cell Cycle	-0.132	6.03e-12	**0.213**
Immune System	-0.0600	1.54e-6	**0.0913**
DNA Replication	-0.122	2.81e-6	0.352
Fanconi Anemia	-0.212	4.45e-5	**0.221**

2.4.6 Comparison of Phylogenetic Simulations, Permutations, and Permulations

Alternatives to permulations include either permutations or simulations alone. Permutations involve randomly assigning phenotype values to species regardless of the underlying phylogenetic relationships among those species. Meanwhile, simulations refer to the first step of permulations—phenotype values are generated based on predicted phenotype evolution along the phylogenetic tree. However, unlike permulations, simulations do not include reassigning the observed values based on simulated values, and thus do not preserve the distribution of the original phenotype values.

At the pathway level, permulations result in p-values that are about equally as conservative as phylogenetic simulations alone and more conservative than permutations alone (Figure 2.10). Both permulations and simulations are preferred to permutations because null phenotypes generated from permulations or simulations reflect the underlying phylogenetic relationships among species, while null phenotypes from permutations do not. Therefore, the empirical null generated from permulations or simulations more closely represents the true null expectation for phenotype evolution. Although permulations and simulations show similar performance, we prefer permulations because permulated phenotypes exactly match the distribution of observed phenotypes, and thus create null phenotypes uniquely tailored to a particular continuous phenotype of interest. Such matching eliminates statistical anomalies that can arise due to discrepancies in range and distribution of permulated phenotypes compared to observed phenotypes.

2.5 Discussion

We present permulations, a set of novel empirical methods to address problems of non-independence and bias in phylogenetic analysis. The methods use phylogenetic relationships among species alongside known values of an observed phenotype to inform Brownian motion simulations from which permuted phenotypes are then generated. By doing so, the methods empirically construct the possibly composite null distribution and account for this complexity in multiple hypothesis testing. For permulation of binary phenotypes, the phylogenetic characteristics preserved are the number of foreground branches and the underlying relationships among foreground branches. For continuous phenotypes, the exact distribution of phenotype values is preserved in addition to the underlying phylogenetic relationships among species.

From testing the strategy on binary and continuous phenotypes, we find that our permulation strategy is an effective approach for overcoming challenges in multiple testing with composite nulls in comparative phylogenetic studies. We discuss with examples how our binary and continuous permulation methods fix issues of both undercorrection and overcorrection of p-values for specified phenotypes, and subsequently improve the quality and confidence of prediction. Note that although our examples demonstrate the usefulness of permulations, they are not necessarily representative of how empirical null distributions will deviate from the typical null for all phenotypes over all phylogenies for all sets of genetic

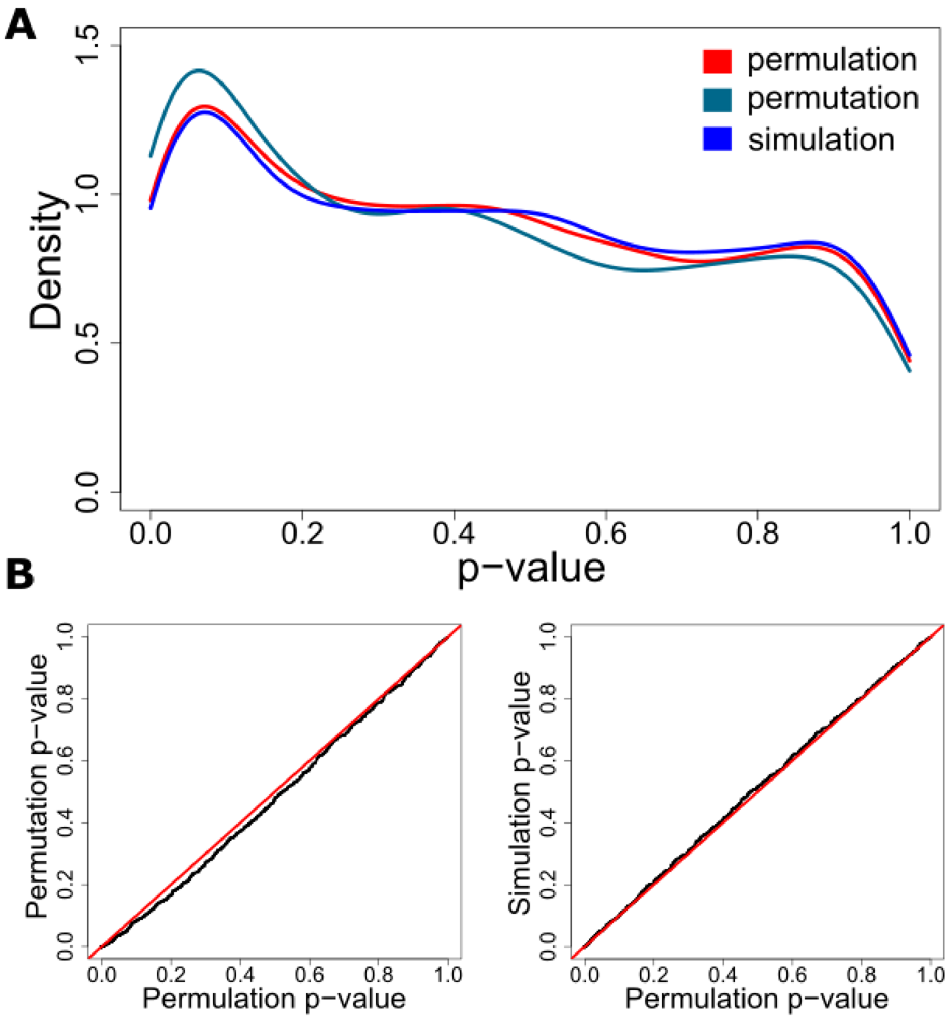


Figure 2.10: Permulations p-values are more conservative than permutation p-values and about equally as conservative as simulation p-values. All plots demonstrate enrichment for canonical pathways associated with the long-lived large-bodied phenotype. (A) Density plots representing the empirical p-value distributions for the three methods to generate null p-values. Permulation and simulation curves are very similar, while the permutation curve demonstrates a stronger enrichment of low p-values and therefore less conservative p-values. (B) Q-Q plots comparing empirical p-values from permulations to empirical p-values from simulations and permutations also demonstrate that permulation p-values are more conservative than permutation p-values and about equally as conservative as simulation p-values.

elements. In fact, we expect permulations to behave differently as those variables change, and thus the best way to determine how permulations will affect a particular data set is to run the permulation analyses.

Devising a systematic solution for such problems is difficult because the causes of complex null distributions in phylogenetic studies can be confounding. The necessity for incorporating phylogenetic information to correct for phylogenetic effects is well understood [73, 210, 236], and some systematic solutions have been designed to tackle the problem, including Phylogenetic Independent Contrast (PIC) [73], Phylogenetic Generalized Least Squares (PGLS) [88], phylogenetic autoregression [39, 85], and phylogenetic mixed models [94, 106, 150]. However, systematic solutions usually make phylogenetic or distributional assumptions that can lead to inaccuracies if the assumptions do not accurately represent the data. For example, PIC makes an assumption that the observed phenotype evolved by Brownian motion, and it can lead to overcorrection when the selection giving rise to the observed data did not actually cause strong phylogenetic effects [156]. In addition, phylogenetic mixed models usually assume that evolution along the phylogeny follows a Brownian motion process and that the resulting phenotype values are normally distributed. Without fully understanding the underlying evolutionary mechanism, incorrect assumptions can lead to overcorrection or undercorrection of statistical confidence. Empirically correcting p-values using permulation methods allows us to circumvent the need to artificially deconstruct this unknown correlation structure in the data. Importantly, while our permulation methods are based on Brownian motion simulations, the simulated trait values themselves are not incorporated in the null phenotypes, and instead are only used as a way to incorporate phylogenetic dependencies in informing how trait values should be permuted across the phylogeny. In this sense, the choice of simulation model is not important.

For binary phenotypes, our permulation methods choose permuted foreground sets by matching the number of foregrounds and their underlying relationships to those observed in the actual phenotype. This approach of defining null phenotypes can be justified by phylogenetic non-independence, a notion that arises from the implications of shared ancestry [73]. At the time of divergence, closely related species diverging from a common ancestor are likely to experience similar selective pressures as the ancestor as well as similar genetic predispositions to respond to the selection pressures. With progressing evolutionary time, the daughter species will evolve independently in response to their respective environments. Such similarities in environmental pressures and genetic predispositions diminish with increasing evolutionary distance between species, meaning that the variance in phenotype values will increase with increasing divergence in evolutionary time. Considering this phylogenetic non-independence and that adaptations to selection pressures are often assumed to be reflected in evolutionary rates, it is reasonable to preserve the pattern of divergence between foreground species to construct hypothetical null phenotypes, in finding correlations between evolutionary rates and phenotypes. It is impossible to pick a new set of foreground branches with perfectly matching divergence times, but matching divergence patterns can serve as a justifiable workaround because the general implications of shared ancestry on phylogenetic non-independence among the new set of foregrounds would apply in a similar way.

We developed two versions of permulation methods for binary phenotypes. The complete case (CC) algorithm produces one permuted phenotype from the master tree to apply for all genes simultaneously, while the species subset match (SSM) algorithm produces distinct permuted trees for each gene, accounting for the differences in species membership in different gene trees. This makes the CC method statistically imperfect. For example, a gene that is missing in some species will have a phylogenetic tree that is missing some branches. Because the CC method produces permuted trees from the master tree that contains all species, it may not conserve the number and relationships of foregrounds across the permulations of the example gene (e.g., genes 3 and 4 in Figure 2.4). In contrast, the SSM method accounts for differences in numbers and patterns of foregrounds among different genes and addresses each gene independently. This means that the SSM method is the ideal implementation of our concept of binary permulations. However, the CC method is both computationally much faster and accounts for the fact that existing comparative genomics methods take in phenotype inputs in different forms. For example, *Forward Genomics* requires one phenotype tree to apply for all genes, while *HyPhy RELAX* requires multiple phenotype trees with matching topology to each gene. Regardless of the statistical flaw, our results demonstrate that applying the CC method on *Forward Genomics* is beneficial for improving prediction (Figure 2.9). The CC method is significantly faster than the SSM method because it only produces one permuted tree for each permulation, instead of a heterogeneous set of permuted trees applying to different genes. Therefore, in the case of limited computational resources or very large datasets in which using the SSM method is infeasible, the CC method can serve as a good alternative.

Our results also demonstrate that binary permulations improve the sensitivity of RERconverge to identify significantly accelerated genes that are missing in many species (Figure 2.7D), i.e., genes with small trees. Because of lower species numbers, genes with small trees suffer from lower statistical power compared to genes with large trees (for example, the number of ways to permute a small tree is much fewer compared to a large tree). As such, pooling all the p-values together to perform multiple testing correction unfairly penalizes genes with small trees. Calculating empirical p-values from multiple empirical permulations is a way to correct for this imbalance in power by indirectly incorporating important covariates, which accounts for the number of foregrounds, backgrounds, and the ratio and phylogenetic relationship between them. Indeed, the pooled null empirical p-values have a uniform distribution (Figure 2.6), establishing the validity of applying standard multiple testing methods to identify significant divergence in evolutionary rates. Future work can evaluate if such benefits are similarly observed when applied to other comparative genomics methods.

Permulations grant increased power to detect genes associated with a continuous phenotype as suggested by the shape of the empirical null distribution (Figure 2.1). When p-values from permulations are compared with permutations or simulations of trait values, we find that permulation p-values are more conservative than p-values from permutations alone, and equally as conservative as p-values from simulations alone. This suggests that permulations offer a valid alternative to phylogenetic simulations. Importantly, permulations preserve the exact distribution and range of phenotype values, a critical characteristic related to the power of the correlation calculated between gene evolution and phenotype evolution. Thus, permulations more accurately match the power between observed and permulated statistics compared to observed and simulated statistics.

Although many of our tests of the permulation strategy were performed using RERconverge, permulations are applicable to any similar methods. When using permulations to calculate empirical p-values using *Forward Genomics*, an alternative evolutionary rates-based method, we show that we can quantify a realistic confidence level at which we believe a gene is under accelerated evolution in a subset of species. Even when using the *Forward Genomics* global method, a deprecated method that does not account for phylogenetic relationships among species, permulations improved the ability to detect accelerated evolution in marine pseudogenes (Figure 2.9). The improvement is likely due to permulations indirectly capturing phylogenetic information through their construction. For the *Forward Genomics* local method, permulations captured realistic confidence levels without losing the ability to detect accelerated evolution in marine pseudogenes. Theoretical p-values directly from the *Forward Genomics* method (Figure 2.1) show over half of the genome under significantly accelerated evolution related to the marine phenotype (12,438 out of 18,797 genes with the lowest possible Benjamini Hochberg corrected p-value), which is biologically highly unlikely [66, 67, 68, 130]. Permulations reduce the number of genes under significantly accelerated evolutionary rates to a more modest number (889 genes if using the same confidence level cut-off) to more accurately reflect both the biology of the system and our confidence in identifying genes with significant evolutionary rate shifts.

Our permulations also reveal aberrant statistical behavior in PGLS. Designed to correct for phylogenetic relatedness when testing for coevolution of traits, PGLS indeed demonstrates a near-uniform empirical p-value distribution for one set of tests for coevolution of the long-lived large-bodied phenotype and gene stop codon counts. However, the method's behavior is dramatically different when testing for coevolution of gene stop codon counts with the binary marine phenotype. It likewise shows undesirable behavior when testing for coevolution of *STAT2* transcription factor binding site counts across non-coding regions. In addition to revealing a non-uniform null, the exact identity of non-coding regions with significant observed and permulation p-values is different, completely altering analysis results. These findings suggest that phylogenetic methods may behave in unexpected ways, and permulations are a valid strategy to investigate those behaviors and perform appropriate statistical corrections.

Finally, permulations demonstrate a crucial correction to pathway enrichment statistics that corrects for coevolution among genes in a pathway of interest. Since pathways often contain functionally related genes that evolve at similar rates, performing pathway enrichment treating each gene as an independent observation is statistically incorrect and will result in erroneous conclusions. Performing permulations at the pathway level identifies pathways that are falsely shown to be enriched and correctly quantifies the confidence at which we may state that a pathway is enriched. We argue that a strategy like permulations is essential in virtually all cases of pathway enrichment calculations to account for gene non-independence driven by correlated evolutionary trends.

Overall, permulations are an important statistical consideration that should be undertaken to accurately report results from evolutionary rates-based analyses as presented here. Regardless of whether permulation allows for greater or fewer null hypothesis rejections at a given threshold, they are an accurate depiction of statistical power given a data structure. In the absence of a known parametric null that accurately represents a data set, a permulation-style approach is an important tool to calculate statistical confidence.

3.0 *phyloConverge*: Prediction of local convergent shifts in evolutionary rates underlying convergent phenotypes

3.1 Attribution statement

A pre-print of this chapter was posted on *bioRxiv* on May 4, 2022. All of the work in this chapter was performed by myself, with the following exception:

- Identification of conserved transcription factor motif coordinates was performed by Weiguang Mao, Ph.D.

3.2 Introduction

Decoding the genetic basis of complex phenotypes is a central goal of biology, and one strategy for learning genotype-to-phenotype associations is by studying the genetic basis of morphological adaptation. When species transition to a new environment, accompanying shifts in selection pressures can cause numerous molecular changes that give rise to phenotypic alterations at the organismal level. Morphological and physiological adaptations are enabled by changes in both protein-coding elements and regulatory elements that play key roles in determining gene expression patterns in different contexts [33, 260].

With the wealth of sequenced species genomes that has been produced by high-throughput sequencing, it is possible to identify the functional associations of genetic elements by comparing the sequences of species with an extreme phenotype with orthologous sequences in other species. Convergent evolution is a useful phenomenon that allows us to distinguish phenotype-associated evolutionary processes from lineage- or species-specific changes that cannot be attributed to specific selection pressure. When independent lineages convergently adapt to a common selection pressure, genetic elements that control the selected phenotypes are likely to undergo similar selective shifts. Some genetic elements that experience stronger selective constraints would shift to a slower evolutionary rate, while other genetic elements, such as those supporting functionality no longer needed in the new environment, may experience relaxed constraints and accumulate more divergence. This relationship between selection and sequence conservation has given rise to parameter models for detecting lineage-specific rate shifts [119, 228], as a successive step toward convergent rate shifts across disjoint clades. The relationship between convergent phenotypes and convergent rate shifts can be exploited to associate genetic elements with high-level phenotypic adaptations. The utility of this comparative framework has been successfully demonstrated in numerous studies [110, 129, 166, 187, 195, 208] and engendered several computational algorithms [103, 110, 128, 155, 195]. Of the existing methods, *Forward Genomics* [103, 195] and *RERconverge* [128, 188] stand out as having been applied at genome-wide scale to a variety of different phenotypes.

The methods have demonstrated success in identifying genome-wide phenotypic associations for both protein-coding and non-coding elements, but their application to non-coding regions is limited because such methods require a defined unit of non-coding sequence to operate on. The typical strategy for defining non-coding units is to use *PhastCons* [227], which segments the alignment into conserved regions. This approach produces a set of conserved non-coding elements (CNEs) that represent putative regulatory elements (REs) and have a size range of 50-500bp, much larger than a single transcription factor binding site (TFBS). This disconnect between the CNE unit and the TFBS, which is the atomic unit of sequence activity, poses specific challenges for evolutionary analysis.

REs typically contain multiple TFBS for different TFs (though often with some repetition) [145]. Detailed experiments on dissection of well-characterized REs have revealed that the relationship between individual TFBS and the functional output of the RE is complex. Ablating TFBSs may eliminate activity, change it, or have no effect [116, 176, 189, 223, 232]. Moreover, RE activity is itself multifactorial as many REs are pleiotropic and can drive expression in seemingly unrelated contexts. These pleiotropic effects can occur via identical TFs binding to identical sites, different TFs binding to identical sites, and different site usage. All three scenarios have been observed [235]. From the perspective of genome-wide evolutionary analysis of CNEs (which are computationally identified putative REs), individual TFBSs may have different and possibly context-specific contributions to regulatory activity and thus have different evolutionary pressure and histories. It is thus quite likely that phenotype-driven changes in evolutionary rate may be more localized than the typical CNE length. As such, the information content across a given CNE may not be uniform, making it necessary to interrogate genetic elements at a higher resolution. There is a need for a computational strategy that allows us to scan a multiple sequence alignment (MSA) and identify functional units of REs without prior definition of non-coding element units.

The common approach that is used by most alignment-based comparative genomics methods for identifying genomic elements underlying phenotypic convergence is to correlate phenotype values with metrics that quantify changes in substitution rates from neutrality. The *Forward Genomics* branch method captures substitution rate shifts by computing sequence divergence between each pair of parent-daughter nodes in the phylogeny and uses Pearson correlation to measure association with phenotype changes at each branch [195]. *RERconverge* [128, 188] and the *phyloP* framework in *PHAST* [190] both use maximum likelihood estimation of evolutionary rates across the phylogeny and detect convergent rate shifts by comparing element-specific trees against a null model, with different approaches. In *phyloP*, a single neutral evolution model is used as a reference point for a likelihood ratio test (LRT) performed to compare two maximum likelihood-estimated models, one that allows convergent rate shifts in a subset of branches and one that assumes no convergent rate shifts exist. Meanwhile, *RERconverge* estimates evolutionary tree models for each individual element and quantifies rate shifts as the residuals from regressing out the tree against the null tree averaged from all the elements. Unlike both of these methods that use a single reference null model, *PhyloAcc* allows the estimation of variations in shift patterns by using a hierarchical Bayesian modeling approach [110].

However, each of these methods comes with limitations. Apart from *phyloP*, most methods are not implemented in ways that make them computationally efficient to compute local segments of an input element. For example, to score every nucleotide in a CNE, the other methods would require constructing a multiple sequence alignment for each nucleotide, followed by all the downstream computations for estimating the convergence signal of the nucleotide. Therefore, these methods have limited capacity to make scalable and unsupervised predictions of functional segments of CNEs. While the *phyloP* framework serves as a good foundation for building a scalable tool for performing large-scale scanning of regulatory elements, there is still a necessity to improve the robustness of the statistical predictions of the method. A case in point is the fact that *phyloP*'s LRT is not expected to produce well-behaved p-values for the simple reason that the single parameter scaling model is an oversimplification. In reality, there is considerable variation in evolutionary rates across the genome. Thus, when scaling with respect to a single neutral tree, the most accurate model would give each branch its own scaling parameter to account for local variation. Consequently, increasing the number of parameters from one to two will often produce a significantly better fit even in the absence of a specific foreground signal. Previous findings also highlighted that *phyloP* is unable to differentiate strong signals produced by a single branch versus strong signals from the convergence of multiple weaker branches [110], and that it is not sufficiently powerful for analyzing short segments [190].

In addition, it has been previously noted that phylogenetic inference methods can produce highly skewed statistics when testing the same hypothesis across a large collection of genetic elements. This phenomenon is not specific to particular inference methods, and indeed even occurs in the context of phylogenetic generalized least squares (PGLS), but is a general problem that arises from the hypothesis having shared bias structure [94, 129, 210, 212, 236]. Such biases can arise from failing to completely account for phylogenetic dependence [73] or systematic variation across genomes either of biological [64, 207, 245] or technical [105] origin. When testing a single hypothesis genome-wide, these subtle effects induce test dependency that results in highly skewed p-value distributions. *Forward Genomics* and *PhyloAcc* have incorporated strategies to account for phylogenetic biases; *Forward Genomics* removes phylogenetic non-independence by computing branch-specific sequence identities, whereas *PhyloAcc* uses a Markov chain to estimate branch-specific changes in conservation state such that the conservation state of a branch is only dependent on its parent branch. However, a parametric approach can still fail at producing a healthy distribution of statistics if the assumptions of the approach do not fit the true generative evolutionary process that produces the observed data [212].

To address these challenges, we present *phyloConverge*, a fast comparative genomics method that performs fine-grained local convergence analysis to identify genomic regions associated with phenotypic convergence. Our method combines explicit parameterization of evolutionary rate shifts and a phylogeny-aware trait permutation strategy to produce unbiased convergent rate shift scores calibrated to the local context of the chromosomal region. We show from benchmarking experiments using the convergence case of mammalian adaptation to subterranean habitats that *phyloConverge* produces convergence predictions with superior statistical robustness. We demonstrate that by computing local convergence signals at TFBS motif-level, we learn functional signals in greater detail, capturing variations in convergent rate shifts across fragments of a CNE in support of their possible involvement in regulating different functions.

3.3 Materials and Methods

3.3.1 Design of *phyloConverge*

We propose a method called *phyloConverge* that combines the generative nucleotide substitution modeling capability of *phyloP* with an empirical strategy for correcting statistical biases that have not been effectively captured by the two-parameter model, given a MSA of a region (or a nucleotide position) of interest, a phylogenetic model of neutral nucleotide substitution, and a defined set of convergent species (i.e., "foregrounds"). In *phyloP*, a convergent rate shift is inferred by performing maximum likelihood estimation of two branch scaling factors (Figure 3.1A). The first scaling factor ρ, which measures the phylogeny-wide rate shift relative to the provided neutral tree, is analogous to the parameter used to compute the widely used *phyloP* conservation track. An additional λ parameter measures evolutionary rate shifts that occur exclusively among the foregrounds relative to the entire phylogeny. Given these definitions, evidence for rate convergence is quantified by a LRT comparing the null hypothesis of constant scaling across both foreground and background (i.e., all branches are uniformly scaled by ρ_o) against the alternative hypothesis that the foreground branches are scaled by λ, in addition to the background scaling ρ_1. After estimating these parameters and performing hypothesis testing, the conservation/acceleration score is finally defined

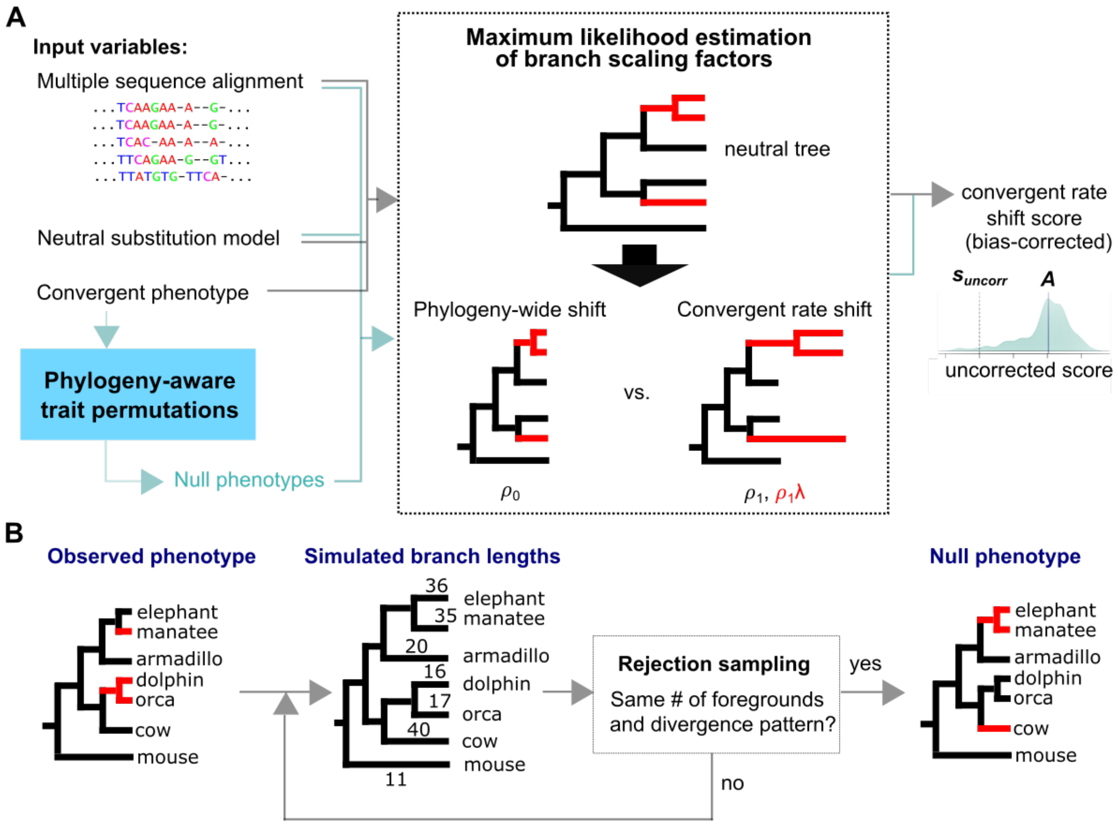


Figure 3.1: Workflow of *phyloConverge*. (A) Given input variables that include the set of species with convergent phenotype ("foregrounds"), the neutral model of evolution, and the multiple sequence alignment, *phyloConverge* combines generative nucleotide substitution modeling and phylogeny-aware trait permutation to compute convergent rate shift scores that are empirically corrected for statistical biases.(B) *phyloConverge* uses phylogenetic permulation to produce null phenotypes that preserve the covariates in the observed phenotype, namely the number of foregrounds and the foreground phylogenetic dependence.

by computing the negative log-likelihood of the LRT p-value, noting the magnitude of λ (conservation if $\lambda < 1$, acceleration if $\lambda > 1$).

To correct for biases, we previously developed a phylogenetic trait permutation method called permulation, a *portmanteau* of permutation and simulation [212]. Permulation is a rejection sampling approach that uses Brownian motion simulations to produce multiple "fake" (null) traits by selecting new sets of foreground species that are matched to the true observed trait in terms of number of species and phylogenetic dependence (Figure 3.1B). With these null traits, we can perform the equivalent of permutation tests to correct the test statistics. We incorporate this trait permulation strategy into *phyloConverge* to produce n null traits and use *phyloP* to compute the convergence scores for both the observed convergent trait and the set of n null traits. Finally, we measure the corrected significance of rate shift by computing an empirical p-value p_{corr}, defined as the proportion of the null *phyloP* scores that are as extreme or more extreme than the *phyloP* score of the true phenotype. The corrected convergent rate shift score s_{corr} is then defined as the negative logarithm of p_{corr}, signed by the direction of rate shift (deceleration or acceleration). Specifically, $s_{corr} > 0$ denotes stronger foreground conservation, whereas $s_{corr} < 0$ denotes foreground acceleration.

While such a permutation test is important for accurately calibrating our confidence in the identified genotype-phenotype associations, the main drawback is that it necessitates a large number of computations to achieve a high p-value resolution, which increases running time significantly. To overcome this drawback, we adopted an adaptive permutation strategy previously applied in expression quantitative trait loci (eQTL) analysis [253], which balances p-value resolution against running time by pruning the number of permutations if a certain significance threshold has been crossed and computing an adaptive p-value. This adaptive approach indeed reduces the computational overhead greatly, without incurring a loss in accuracy within the controlled significance level (Figure 3.2). *phyloConverge* also measures the robustness of the convergence signals using a leave-one-out approach, in which the parametric scoring with *phyloP* is repeated by removing one foreground species for each repetition. Convergence signals are determined to be robust if the removal of one foreground species does not immediately erase the signals or flip their direction.

3.3.2 Implementation details of *phyloConverge*

The input of *phyloConverge* includes a multiple sequence alignment (MSA), a phylogenetic model of neutral nucleotide substitution (which can be estimated from sites that are expected to undergo neutral evolution, e.g., fourfold-degenerate sites), and the list of species

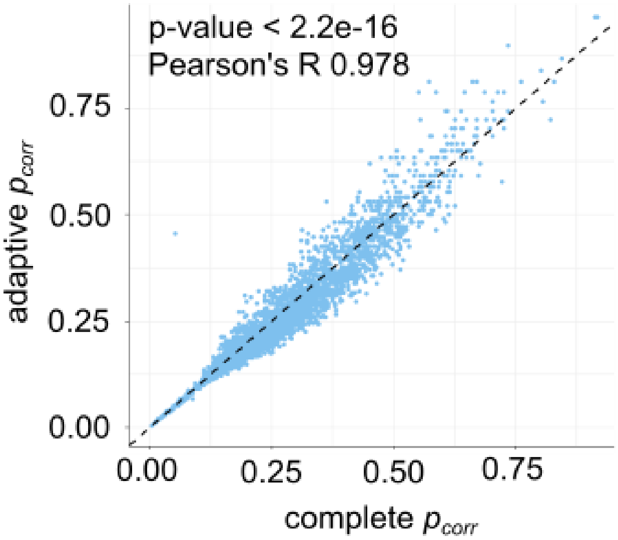


Figure 3.2: Correlation between empirical p-values computed with adaptive permulations versus the complete permulations, with maximum permulations and controlling for significance level α of 0.05.

with the convergent phenotype. To quantify the uncorrected association score between the evolutionary rate of a genetic element and the convergent phenotype, *phyloConverge* uses the *phyloP* function in the *RPHAST* package [190, 111], specifying the "*LRT*" option as the hypothesis testing method and the "*CONACC*" setting for the scoring method (positive scores denote conservation, negative scores denote acceleration).

To empirically calibrate for statistical biases, *phyloConverge* performs permulations [212] to produce numerous null or "fake" phenotypes that are phylogenetically constrained to the generative model of the true phenotype (Figure 3.1B). We previously developed two permulation strategies for binary phenotypes: the 'complete case' (CC) method, which produces null phenotype trees from the complete topology, and the 'species subset match' (SSM) method, which accounts for missing sequences in a particular MSA. While the SSM method is more stable and accurate, the CC method is significantly faster, with comparable if slightly less accuracy. For tractability, *phyloConverge* currently makes use of the CC method. After numerous valid null phenotypes are obtained, *phyloP* is used to compute scores for each of the null phenotypes, such that a null distribution of *phyloP* scores for the given MSA is obtained.

In such a permutation test, the significance of deviations from the expected value is typically measured by computing empirical p-values that are defined as the proportion of the null statistics that are as extreme or more extreme than the observed test statistic. In a two-tailed test, these extreme values make up the area under the curve beyond the observed statistic and the negative of the observed statistic. As *phyloP* defines acceleration as a negative score and conservation as a positive score following the "*CONACC*" scoring mode, the two tails of the null score distribution signify opposing directions of rate shift, where the lower tail denotes acceleration and the upper tail denotes deceleration. Because of this directionality and because the null distribution is not necessarily trivial or symmetric (e.g., histogram in Figure 3.1A), we calculated the two-tailed empirical p-value p_{corr} using the two-sided conditional p-value approach described by Kulinskaya [133], which transforms one-sided p-values into equivalent, weighted two-sided p-values for symmetric or asymmetric distributions. Suppose the distribution of null *phyloP* scores follows a strictly increasing continuous cumulative distribution function F. Then, p_{corr} is computed as follows:

$$\begin{aligned} p_{corr} = & \frac{F(s_{uncorr})}{F(A)}\mathbf{1}(s_{uncorr} \leq 0)\mathbf{1}(s_{uncorr} \leq A) + \\ & \frac{1-F(s_{uncorr})}{1-F(A)}\mathbf{1}(s_{uncorr} \geq 0)\mathbf{1}(s_{uncorr} \geq A). \end{aligned} \tag{1}$$

where s_{uncorr} is the uncorrected score (computed by *phyloP*) for the observed phenotype and A is the value that the null distribution is centered on. For our purposes, A was chosen as the median of the null scores, such that the weights at both the left and right sides of A were equal. Note that according to Equation 1, p_{corr} is only computed if there is agreement between the placements s_{uncorr} relative to A and zero, respectively. Subsequently, the bias-corrected conservation/acceleration score s_{corr} is computed as the negative logarithm of p_{corr}, signed by the relative position of s_{uncorr} with respect to A, as follows:

$$s_{corr} = -\log_{10} p_{corr}\text{sign}(s_{uncorr}). \tag{2}$$

Finally, to improve the computational tractability of permulations, we incorporated a simple strategy to adaptively terminate permulations when a target significance threshold had been reached. For example, suppose we would like to control for significance threshold $\alpha = 0.05$ with a maximum of 1000 permutations. For a genetic element to be significantly associated with the convergent trait, there can only be a maximum of 50 null scores that are as extreme or more extreme than the observed uncorrected score. Formally, suppose we want to control the test for a significance level of α, and we set a maximum of N permulations. Denoting $S^{'}$ as the set of computed null statistics, for a hypothesis to be statistically significant at α significance level, the maximum number of null statistics that are as extreme or more extreme than the true statistic, s_{uncorr}, is therefore αN, defined as the "pruning" threshold. At every permulation iteration i, we track whether the pruning threshold has been reached, given the value of the median of the null distribution at iteration i, A_i. If the threshold has been reached, the adaptive p_{corr} is computed by modifying Equation 1 as follows:

$$\begin{aligned} p_{corr} =& \frac{\min\left(\alpha N + 1,\, \sum_{s' \in S'} \mathbf{1}(s' \leq s_{uncorr})P + \mathbf{1}(s' \geq s_{uncorr})Q + 1\right)}{\min\left(N + 1,\, \sum_{s' \in S'} \mathbf{1}(s' \leq A_i)P + \mathbf{1}(s' \geq A_i)Q + 1\right)}, \\ \text{where} \quad & P = \mathbf{1}(s_{uncorr} \leq A_i)\mathbf{1}(s_{uncorr} \leq 0) \\ & Q = \mathbf{1}(s_{uncorr} \geq A_i)\mathbf{1}(s_{uncorr} \geq 0) \end{aligned} \tag{3}$$

The addition of "+1" to each term is done to correct the tail ends of the distribution. The approach indeed offers remarkable improvements in speed – parallelizing over 60 cores on one compute node with 95GB memory, the scoring of ∼36,000 CNEs with 500 permulations can be completed in ∼1.5 hours. Using ∼5,000 randomly selected subset of the CNEs dataset and the subterranean foregrounds, the empirical p-values calculated from all 500 permulations (p_{total}) and the adaptive empirical p-values (p_{corr}) computed to control significance levels α of 0.05 correlate very well (Pearson's R = 0.978, p-value $< 2.22e-16$), with negligible loss in resolution within the significance level that is controlled (Figure 3.2). However, we note the necessity for weighing the trade-off between the maximum number of permulations and the α level to control for. For example, with a maximum of 500 permulations, setting $\alpha = 0.01$ means that only a maximum of 5 extreme null scores are allowed such that the computations may be prematurely terminated. In such cases, the performance of adaptive permulation may suffer, because stochasticity can cause premature termination of the permulations. The R implementation of *phyloConverge* is available on GitHub (`https://github.com/ECSaputra/phyloConverge`).

3.3.3 Dataset construction

To benchmark our method, we used a dataset previously produced by Roscito et al. [208], which contains a multiple genome alignment for 24 species. The 24-way phylogeny contains 4 subterranean mammal lineages: naked mole rat, cape golden mole, star-nosed mole, and the blind mole rat (Figure 3.3). Roscito et al. identified 491,576 conserved non-coding elements (CNEs) by using the *PhastCons* [227] tool to identify conserved elements that aligned well among at least 15 species. We extracted the MSA of each CNE from the multiple genome alignment using the *sub.msa* function in the *RPHAST* package. Additionally, to construct MSAs of genes from the alignment, we obtained the CDS coordinates of the genes in the *mm10* NCBI RefSeq annotations. The *sub.msa* function was similarly used to extract the alignments corresponding to the CDS coordinates, and the CDS alignments of each gene were concatenated using the *concat.msa* function in *RPHAST*.

To prepare CNE-specific trees for benchmarking with the *RERconverge* software, we used *phangorn* [218] to estimate the maximum likelihood tree for each CNE. The *readTrees* function in *RERconverge* was then used to read the CNE-specific trees into a *multiPhylo* object and compute a master tree with branch lengths that were averaged from the corresponding branches across all CNE trees.

3.3.4 Transcription factor binding site (TFBS) motif calling

Genome-wide scanning for possible TFBS motifs was performed using *PWMScan* [3]. The parameters for *PWMScan* include the genome assembly of interest, the position weight matrix (PWM) of the motif of interest, and a threshold cutoff for calling the TFBS motif. We obtained the PWMs of 771 TFBS motifs from the HOCOMOCO database (version 11) [132]. We then used *motifDiverge* [126] to compute the background frequency of each nucleotide

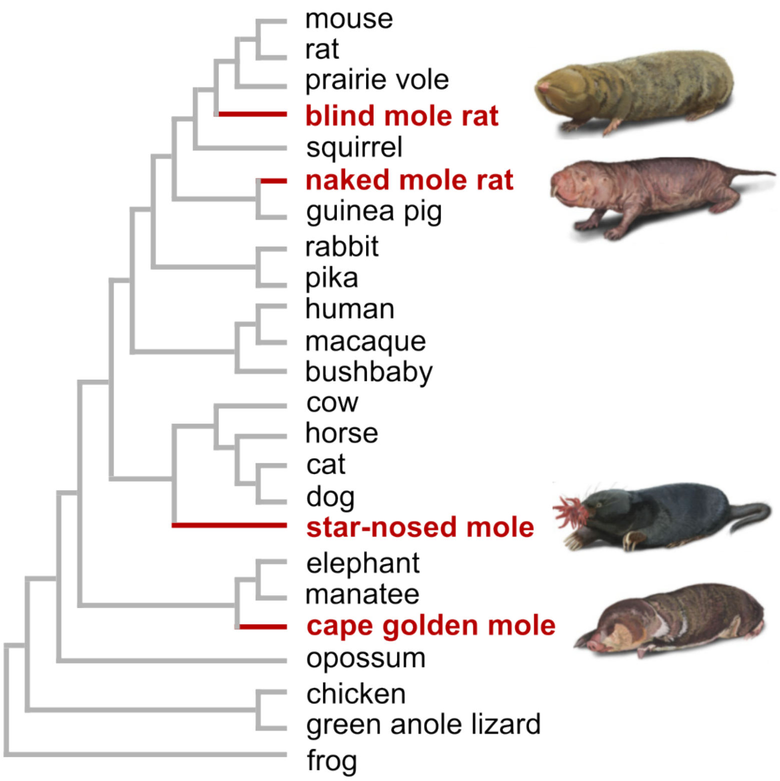


Figure 3.3: Phylogeny of benchmarking dataset.

based on the probability matrix of a given TFBS and infer the PWM matrix of each TFBS and the TFBS calling cutoff to use (set to control Type I error rate below 10^{-5}). Using these input parameters, we previously made the genome-wide TFBS calls for other uses with the human *hg19* coordinate from the UCSC Genome Browser [121]. These genome-wide calls were then lifted over to the mouse *mm10* coordinate using the *liftOver* tool from UCSC [131]. Finally, the set of conserved TFBS motifs were identified by intersecting the TFBS coordinates with the CNE coordinates using *BEDTools* [199].

3.3.5 Identification of tissue-specific "marker" open chromatin regions (OCRs)

To evaluate the functional enrichments of top-ranking subterranean-accelerated CNEs, we computed the correlations between the CNEs with tissue-specific, "marker" open chromatin regions (OCRs) in mouse tissues. For mouse embryonic tissues, we compiled publicly available ATAC-seq datasets (see Table 3.1 for identifiers). The marker OCRs for the whole eye, retina, and lens were taken directly from Supplementary Data 16 of Roscito et al [208]. For the remaining tissues, the datasets with multiple replicates were first pre-processed by identifying consensus regions (regions that were present across at least 2 replicates) using the *GenomicRanges* package in R [137]. Subsequently, the marker OCRs of each given tissue were obtained by subtracting regions that were open in any other tissue from the tissue of interest using *BEDTools*.

We also used the chromatin accessibility atlas across adult mouse tissues [144]. Given that the dataset was presented in the format of consensus peaks, we first identified the OCRs in each tissue by setting the 80th percentile of the read count distribution as a threshold. Then, we identified the marker OCRs of each given tissue by subtracting regions that were open in at least 80% of the other tissues. Finally, the regions were lifted over from the *mm9* to the *mm10* coordinates.

3.3.6 Benchmarking *phyloConverge* against existing methods

We first used *phyloConverge*, *phyloP*, and *RERconverge*+permulation to compute the convergence scores of the set of CNEs produced by Roscito et al. For *phyloConverge* and

Table 3.1: List of publicly available datasets used for validation.

Dataset	**Source**
Mouse embryonic ATAC-seq, whole eye E11.5	Roscito et al. [208]
Mouse embryonic ATAC-seq, retina E14.5	Roscito et al. [208]
Mouse embryonic ATAC-seq, lens E14.5	Roscito et al. [208]
Mouse embryonic ATAC-seq, midbrain E11.5	Roscito et al. [208]
Mouse embryonic ATAC-seq, limb E11.5	Roscito et al. [208]
Mouse embryonic ATAC-seq, kidney E14.5	Roscito et al. [208]
Mouse embryonic ATAC-seq, liver E14.5	Roscito et al. [208]
Mouse embryonic ATAC-seq, heart E14.5	Roscito et al. [208]
Adult mouse ATAC-seq	Liu et al. [144], Count matrix obtained from: `https://doi.org/10.6084/m9.figshare.c.4436264.v1`
Mouse retinal single nuclei ATAC-seq	Norrie et al. [177], scATAC-seq GEO Accession number: GSM4995565, scRNA-seq GEO Accession number: GSE164044
Retinal tissue marker genes	Macosko et al. [151] (Table S4)

RERconverge+permulation, 500 null phenotypes were used. For each method, the top ∼9,400 CNEs were identified by selecting the appropriate threshold that would result in a set with a comparable size to the set produced by Roscito et al., specifically 9,428 CNEs with p-value $\leq$ 0.032 for *phyloConverge*, 9,455 CNEs with p-value $\leq$ 0.05 for *RERconverge*, and 9,325 CNEs with s_{uncorr} $\leq$ -5.1 for *phyloP*. For the coding region analysis with *phyloConverge*, the same threshold of p-value $\leq$ 0.032 was used to select the top-ranking coding regions.

We applied the random subsampling strategy previously used by Roscito et al. to compute correlations between subterranean-accelerated CNEs with marker OCRs. Before computing correlations, we merged nearby subterranean-accelerated CNEs that were within 50bp apart to correct for inflation of significance resulting from multiple CNEs that were very close together. Afterwards, for each tissue, we used *BEDTools* to find the number of intersections between the marker OCRs of the tissue and the subterranean-accelerated CNEs. We then subsampled 1,000 matched-sized sets of randomly selected CNEs from the total set of CNEs and similarly found the number of intersections with the marker OCRs to obtain the null distribution. The strength of correlation between the subterranean-accelerated CNEs and the marker OCRs were quantified as the Z-score computed with respect to the null distribution. This analysis was performed for the top-ranking subterranean-accelerated CNEs from the four methods tested.

To quantify the agreement between two sets of top-ranking subterranean-accelerated CNEs identified by two different comparative methods, we first noted the number of overlapping CNEs between the two sets using *BEDTools*. Then, we subsampled two sets of randomly selected CNEs from the total set, containing matching numbers of CNEs as the two actual sets, and similarly noted the number of overlapping CNEs. We performed the random subsampling 1,000 times to obtain a null distribution of the number of overlapping CNEs between two randomly selected sets of CNEs with the given sizes. The actual number of overlaps was then converted to a Z-score with respect to the null distribution.

3.3.7 Functional enrichment analysis

To associate subterranean-accelerated CNEs with genes, we used the Genomic Regions Enrichment of Annotations Tool (*GREAT*) [162], specifically with the R package *rGREAT* [90]. *GREAT* associates genes with proximal or distal CNEs using a default association rule called "basal-with-extension". *GREAT* first determines a "basal regulatory region" around each gene, defined as the window within 1kb downstream and 5kb upstream of the transcription start site (TSS) regardless of overlaps with neighboring genes. Then, the regulatory domain of the gene is extended until it overlaps the basal regulatory region of neighboring genes, up to 1Mb both upstream and downstream. Afterwards, using the set of subterranean-accelerated CNEs as the 'foreground regions' and the total CNEs as the 'background regions', *GREAT* performs hypergeometric tests to compute the enrichments for foreground regions in each gene's regulatory domain, relative to the superset of background regions. We used *GREAT* to evaluate the enrichments for 21,395 Ensembl genes and set the significance cutoff as Benjamini-Hochberg adjusted p-value ≤ 0.05.

After identifying genes that were significantly enriched for subterranean-accelerated CNEs, we performed enrichment analysis on 1,330 genesets in the canonical pathways using the gene-CNE associations. First, for each given geneset, we determined the number of foreground CNEs and background CNEs that were associated with members of the geneset. Then, as in *GREAT*, we used Fisher's exact test to compute the probability of observing the number of geneset-associated foregrounds and geneset-associated backgrounds given the total number of foreground and background CNEs. We set p-value ≤ 0.05 as a significance cutoff and used Storey's q-value method [237], with the *qvalue* package in R, to compute the corresponding false discovery rate.

We also performed enrichment analysis on the canonical pathways for the coding regions. Fisher's exact test was used to compute the probability of observing the number of accelerated coding regions and background coding regions that overlapped members of a geneset given the total number of accelerated coding regions. Empirical correction was performed by permuting the foreground coding regions 1,000 times and performing Fisher's exact test. Similar to the CNEs, the significance cutoff was set as permutation p-value ≤ 0.05.

3.3.8 Enrichment analysis on retinal cell-type-specific marker genes and marker OCRs

We performed enrichment analysis on the top-ranking genes that were subterranean-accelerated in the coding regions using the retinal tissue-specific marker genes produced in Macosko et al. [151] as validation datasets. To perform enrichment analysis on the top-ranking subterranean-accelerated CNEs, we used a dataset of single cell ATAC-seq regions across different retinal tissues [177]. Clustering of single cells was performed using *Seurat* [95] and *Signac* [239] for the single cell RNA-seq and single cell ATAC-seq data, respectively, and tissue type assignments were made by integrating the multimodal datasets and transferring the single cell RNA-seq cluster labels to the corresponding single cell ATAC-seq clusters [238]. Cell type-specific marker OCRs were finally defined by finding the differentially accessible ATAC-seq peaks for the five resulting clusters (rods, cones, bipolar cells, amacrine cells, Müller glia). For both the coding and non-coding analysis, enrichment analysis was performed using the hypergeometric test. The cutoff for significant enrichment was set as Benjamini-Hochberg adjusted hypergeometric p-value ≤ 0.05.

3.3.9 TFBS motif-level convergence analysis

We used *phyloConverge* to compute convergence scores for individual conserved TFBS motifs that overlapped CNEs. Setting permulation p-value threshold ≤ 0.05 and with leave-one-out filtering, we identified 42,477 significantly accelerated motifs and 81,101 significantly decelerated motifs. To identify CNEs that underwent significant changes in motif content due to selection pressures ("motif-enriched"), we performed enrichment analysis to identify 3 categories of CNEs: (1) CNEs that were significantly enriched for accelerated motifs ("motif-accelerated CNEs"), (2) CNEs that were significantly enriched for decelerated motifs ("motif-decelerated CNEs"), and (3) CNEs that were significantly accelerated for both accelerated and decelerated motifs ("mixed-motif CNEs"). Enrichment analysis was performed using Fisher's exact test. Specifically, if we define the foreground motifs to be (1) the significantly accelerated motifs for the motif-accelerated category, (2) the significantly decelerated motifs for the motif-decelerated category, and (3) the significantly accelerated and decelerated motifs for the mixed-motif category, we tested the probability of observing m foreground motifs out of the n motifs in a CNE, given the total number of foreground motifs and the total number of conserved motifs. The significance threshold was set as Fisher's p-value ≤ 0.05. After the sets of motif-enriched CNEs were identified, functional enrichment analysis for each CNE set was performed using the *GREAT* tool. For this analysis, we used the Gene Ontology Biological Process, Cell Component, and Molecular Functions annotations, and significantly enriched annotations were identified by setting FDR ≤ 0.1 for both the binomial test and the hypergeometric test performed by *GREAT*.

To identify TFs whose binding motifs exhibit global convergence signals, we used Fisher's exact test to compute the enrichment for a given TF in the set of 42,477 significantly accelerated motifs and the set of 81,101 significantly decelerated motifs, respectively. For example, to compute the enrichment for TFBS A in the set of all significantly accelerated motifs, we computed the probability of observing n_A significantly accelerated motif A out of the 42,477 significantly accelerated motifs, given that there were a total of N_A TFBS calls for motif A and N_T total TFBS calls genome-wide. Motifs with global convergence signals were identified by setting Fisher's p-value ≤ 0.05. Pathway enrichment analysis of the motifs with global acceleration/deceleration signal was performed using *GREAT*, specifically using the Reactome pathway annotations. Significantly enriched annotations were identified by setting FDR ≤ 0.05 for both the binomial test and the hypergeometric test.

Finally, motif-specific functional enrichment analysis was also performed using *GREAT* with the Reactome pathway annotations, with FDR ≤ 0.05 for both the binomial and hypergeometric tests. Correlations between significantly enriched annotations were identified by empirically computing the probability of observing n number of overlapping genes between a pair of annotations, relative to the null distribution of overlaps between a randomly selected pair of gene sets with matching sizes to the annotations of interest. The significance threshold for the correlations was set as empirical p-value ≤ 0.05.

3.3.10 Unsupervised prediction of TFBS-scale segments with convergent rate shift

We use the *scanWithPhyloConverge* function in the *phyloConverge* software to scan each CNE with a sliding window of ±5bp and compute the convergence signal of each window. In other words, each nucleotide is scored using an 11bp window surrounding it, to approximate the size of a TFBS motif. To define significantly accelerated or decelerated segments, we first identified the nucleotides that have permulation p-values ≤ 0.05 and are robust based on leave-one-out filtering. Then, each unit of a contiguous segment is identified by combining consecutive nucleotide positions that passes these filters, and then extending the segment by ±5bp.

Of the identified segments, the sets of known and new accelerated/decelerated segments were identified by using the intersect function of *BEDTools*. *De novo* motif discovery on the new accelerated/decelerated segments was performed using the *STREME* tool in the *MEME* suite [13]. The training stage was conducted using the default approach of using a p-value threshold of 0.05, while the "*patience*" parameter was set at 5000, meaning that the discovery process was terminated if 5000 consecutive motifs had p-values > 0.05. Significantly enriched motifs were identified by setting an E-value threshold of 0.5. Finally, the *TomTom* motif comparison tool [92] was used to evaluate whether the enriched motifs significantly matched known consensus motifs.

3.4 Results

3.4.1 Benchmarking *phyloConverge* on the convergent adaptation of subterranean mammals

Because it is not well understood how convergent adaptations interact with other factors to engender the sequence patterns observed in REs, it is not possible to construct simulated datasets that can reliably represent real sequences. Thus, we benchmark *phyloConverge* using a well-characterized convergent trait, the subterranean mammal habitat. We use a dataset that was previously analyzed by Roscito et al. [208], which contains a MSA of 24 species including 4 subterranean mammal lineages: the naked mole rat, the cape golden mole, the star-nosed mole, and the blind mole rat (Figure 3.3). Using the *PhastCons* tool [227], Roscito et al. had previously computed 491,576 conserved non-coding elements (CNEs) that align well among at least 15 species in the phylogeny and used the Forward Genomics "branch" method [195] to identify 9,364 "subterranean-accelerated" CNEs. Using this dataset, we benchmark the performance of *phyloConverge* against three competing methods: *phyloP*'s "branch" method, *RERconverge* with permulation, and *Forward Genomics* branch method. Because some of the other methods cannot score a given sequence in segments, in this benchmarking experiment, *phyloConverge* scores are computed by fitting the entire CNE sequence to ensure fair comparison.

We score the 491,576 CNEs for acceleration using *phyloConverge*, *phyloP*, and *RERconverge*. To benchmark the CNE rankings produced by the different methods, we identify the top ∼9,400 subterranean-accelerated CNEs computed by each method to approximately match the size identified by Roscito et al. Comparing the four size-matched sets of top-ranking CNEs identified by the four methods, only 609 CNEs (∼6.5% of each set) are commonly identified by all four methods (Figure 3.4A). Notably, *phyloConverge* and *RERconverge* identify 5,240 (∼55.5%) common CNEs, which is considerably higher than any other pair (17.6% -28%). *phyloConverge* and *RERconverge* are quite different in their model specifications, while the statistical testing procedures with *RERconverge* and *Forward Genomics* are more conceptually similar. The main point of similarity between *RERconverge* and *phyloConverge* in this analysis is that both rely on maximum likelihood estimates of evolutionary rates and both use the permulation bias correction, suggesting that these features drive the observed overlap. We highlight that given the large sample space of 491,576 CNEs, the overlaps between the sets are statistically significant compared to random chance (Figure 3.4B). Additionally, we note that among these top-ranking CNEs, the permulation p-values of *phyloConverge* hits are smaller than the *RERconverge* hits (Figure 3.4C). Given that the same set of permulated phenotypes are used in both the *phyloConverge* and *RERconverge* analysis, this means that the chance of wrongly rejecting the null hypothesis of no phenotype association is smaller for top *phyloConverge* hits than *RERconverge*, suggesting that *phyloConverge*

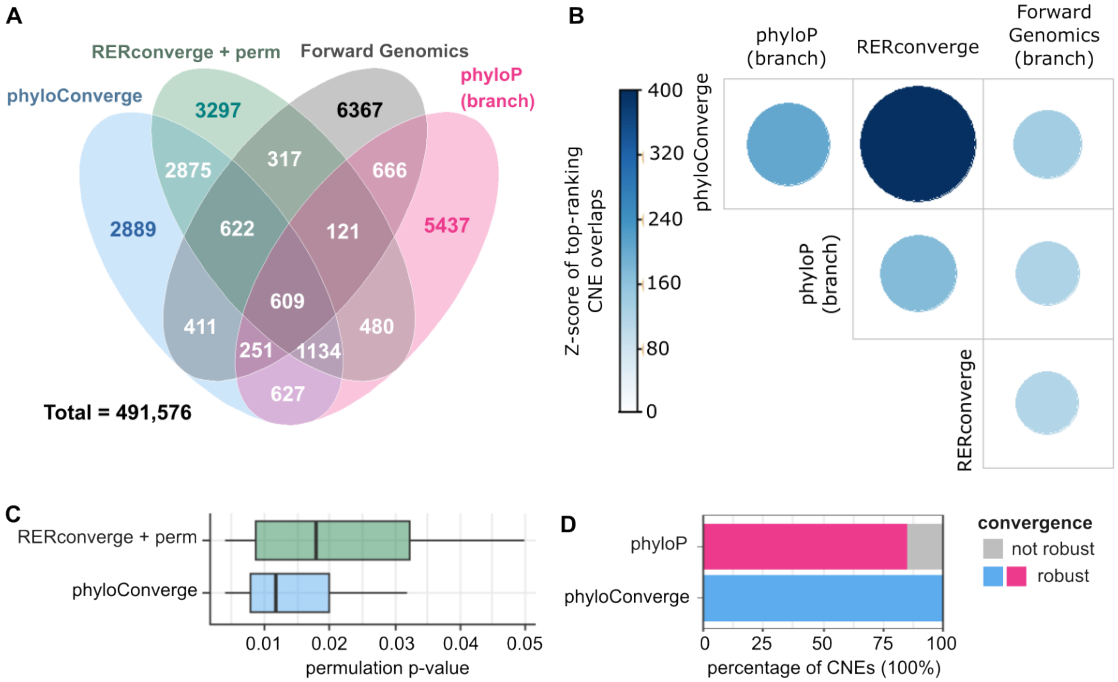


Figure 3.4: Benchmarking *phyloConverge*'s statistical performance. (A) Venn diagram showing overlaps among the top ∼9,400 subterranean-accelerated conserved noncoding elements (CNEs) identified by *phyloConverge*, *phyloP*'s "branch" method, *RERconverge*+permulation, and *Forward Genomics*'s "branch" method. (B) Overlaps between top-ranking subterranean-accelerated CNEs identified by the four methods show very strong statistical significance. (C) Boxplots showing the distribution of permulation p-values of the top ∼9,400 CNEs from *phyloConverge* and *RERconverge*. (D) Convergence signals predicted by *phyloConverge* are more robust than *phyloP* "branch" method.

provides better Type I error control compared to *RERconverge*. We also compare the robustness of convergence signals of the top accelerated CNEs identified by *phyloConverge* and *phyloP*. We find robust convergence signals for 99.4% of the accelerated CNEs identified by *phyloConverge*, and for 84.5% of that identified by *phyloP* (Figure 3.4D).

Next, we benchmark the methods for their specificity against confounders. One of the challenges of using comparative genomics to identify regions associated with a specific phenotype is that the phenotype-unaware conservation signal is already highly associated with functional data. Thus, we evaluate the correlation between the global scaling factor ρ_o (equivalent to computing the phylogeny-wide conservation score) and the absolute magnitude of convergent rate shift scores from the different methods (Figure 3.5A). *phyloP*, which uses no bias correction, indeed shows a negative correlation between ρ_o and the magnitude of convergence scores (Figure 3.5A for top-ranking CNE set, Figure 3.5B for the entire CNE set). This suggests that without statistical calibration, stronger convergence signals are given to elements that are more strongly conserved. Some bias is observed to a lesser extent for *Forward Genomics*, which controls for phylogenetic non-independence by computing branch-specific sequence identity values, thus making them independent [195]. In comparison, the empirical correction approach utilized by *phyloConverge* and *RERconverge* seems to completely remove this bias.

Repeating this experiment with CNE length as another covariate, we find a similar trend with longer regions being more likely to produce strong convergence scores for *phyloP* and *Forward Genomics* (Figure 3.5C). *RERconverge* signals do not show positive correlation with region length (although some non-uniformity is observed at extreme values), while *phyloConverge* shows a completely uniform distribution of scores across elements of different lengths. All in all, these observations suggest that *phyloConverge* has leading performance in statistical behaviors compared to the competing methods.

To quantify how these statistical properties affect biological inference, we evaluate the associations of the top subterranean-accelerated CNEs with tissue-specific open chromatin regions (OCRs), hereby termed “marker OCRs”, across several mouse embryonic and adult tissues. Among the common traits shared by subterranean mammals is that reduced reliance on vision results in degenerated visual structures. These species have small eyes and are either effectively blind or have minimal vision capacity [34, 101, 211, 242]. We expect that eye-related regulatory regions would experience relaxed selection and exhibit greater divergence. By computing the number of intersections between subterranean-accelerated CNEs and marker OCRs, we observe that all methods produced sets of accelerated CNEs with strong enrichments (permutation p-value ≤ 0.05 or $Z > 1.96$) for the marker OCRs of the embryonic whole-eye, retina, and lens, relative to 1,000 randomly selected size-matched null

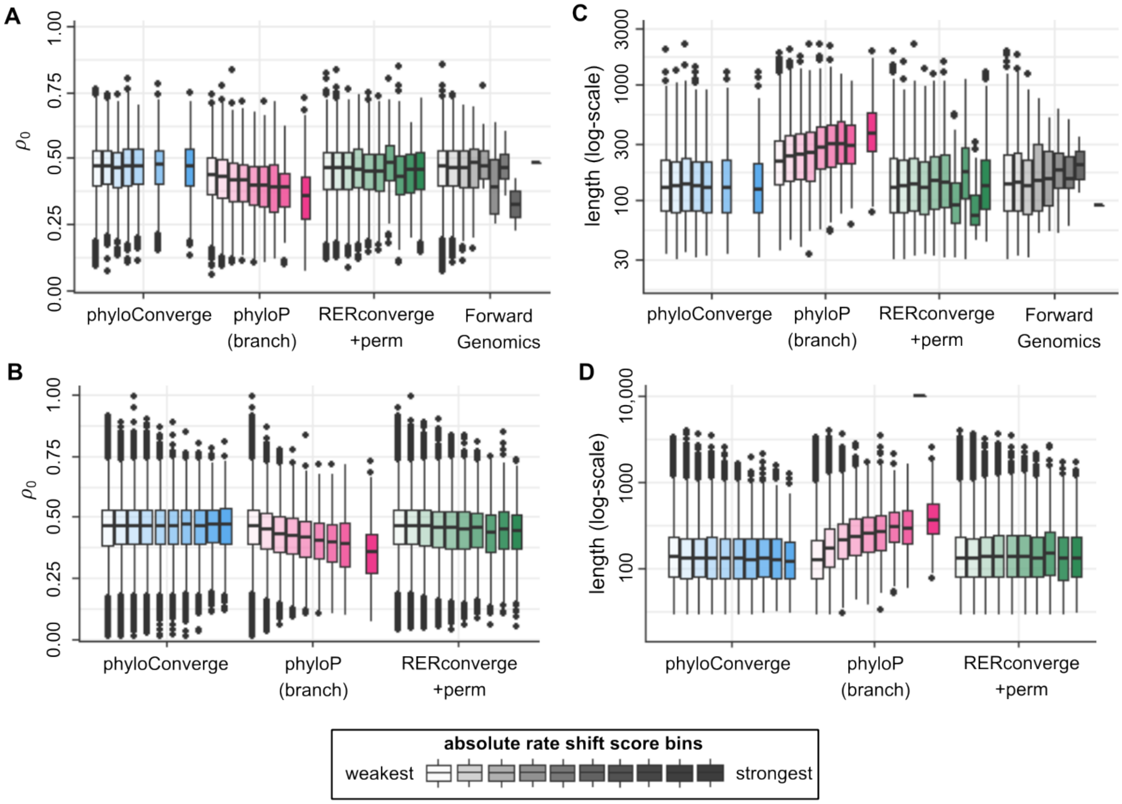


Figure 3.5: *phyloConverge* corrects association statistics from confounders. (A) Correlations between conservation (smaller ρ_o) and the absolute values of rate shift scores (grouped by equidistant score binning) among the top ∼9,400 CNEs, and (B) among all CNEs. (C) Correlations between CNE lengths (in bp) and the absolute values of rate shift scores among the top ∼9,400 CNEs, and (D) among all CNEs. Missing boxplots arise from different method's discretizing extreme values.

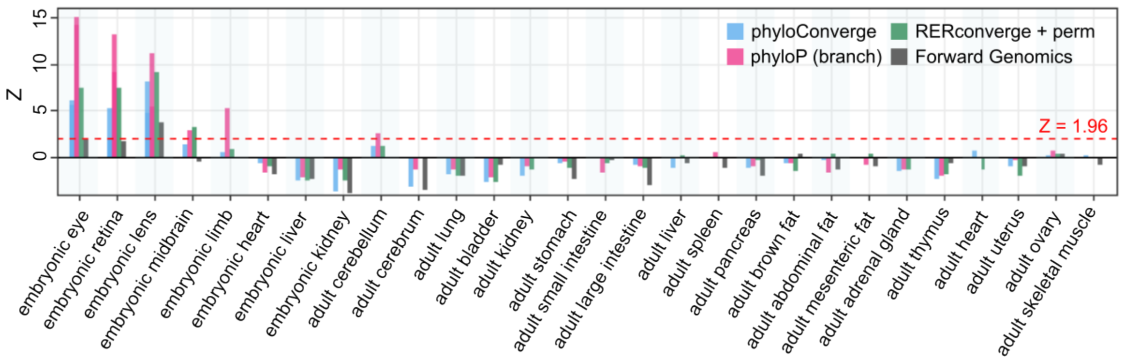


Figure 3.6: Functional enrichments for mouse tissue-specific open chromatin regions in top-ranking subterranean-accelerated CNEs, plotted as Z-scores relative to 1,000 null CNE sets.

sets of CNEs (Figure 3.6), which agrees with our expectation. For these eye tissues, *phyloConverge* produces strong signals that almost match *RERconverge*, and are much stronger than *Forward Genomics*. We also observe that not correcting for biases (i.e., *phyloP*) greatly magnifies these signals and also produces very strong associations with tissues for which excessive relaxation of genetic elements is not expected, such as embryonic limb, midbrain, and adult cerebellum. Meanwhile, the remaining three methods show no enrichment for the marker OCRs of the control non-ocular tissues for which an enrichment is not expected.

Interestingly, *phyloConverge*, *phyloP*, and *RERconverge* produce moderate to strong correlation with embryonic midbrain marker OCRs, which is not detected by *Forward Genomics*. The presence of correlation is consistent with the observation that specific midbrain structures that receive direct optical input (superior colliculus and lateral geniculate nucleus) are highly atrophied in subterranean mammals, although the sizes of most structures in the midbrain are comparable to mice [44, 45]. A similar effect is observed in the cave-dwelling Pachón ecotype of the Mexican tetra fish, which possesses degenerated visual structures [170]. In addition, among the adult tissues, the marker OCRs of the adult cerebellum moderately correlates with *phyloConverge* and *RERconverge*, and strongly correlates with *phyloP*. The cerebellum is a major structure in the hindbrain that regulates motor coordination [203], cognitive and emotional processing [219], as well as ocular motor control [122]. In naked mole rats, the cerebellar region involved in visual signal processing is indeed degenerated, coinciding with the expansion of the region for the somatosensory system [157] that facilitates the processing of tactile cues for navigation [214]. These observations lend support to the hypothesis that eye degeneration is concomitant with complementary changes in brain structures that is detectable as reduced constraint on some brain-specific regulatory regions. We note that accelerated regions are enriched for parts of the brain that are relatively constrained in connectivity and function, such as the midbrain and cerebellum, and strongly depleted for cerebrum, which is plastic. Overall, these results demonstrate that *phyloConverge* can predict phenotypic associations with high specificity.

3.4.2 Subterranean-accelerated elements are enriched for distinct functions from accelerated coding regions

To evaluate the functional associations of the top subterranean-accelerated CNEs identified by *phyloConverge* in detail, we use the Genomic Regions Enrichment of Annotations Tool (*GREAT*) to associate the ∼9,400 top-ranking CNEs with genes based on distance, and compute CNE enrichments for each gene. We find that 76 out of 21,395 genes in the Ensembl database are significantly enriched for the subterranean-accelerated CNEs (hypergeometric test FDR ≤ 0.05) (Table 3.2). We then use the information on these enriched genes and gene-CNE associations to evaluate the enrichment for canonical pathways in the accelerated CNEs. Out of 1,330 canonical pathways, we identify 10 significantly enriched pathways (enrichment p-value ≤ 0.05, corresponding to Storey's FDR ≤ 0.13) (Figure 3.7A). Notably, some of the top-ranking pathways form a cluster of interrelated processes that regulate ocular and/or neuronal functions. The top-enriched pathway, the calcium/calmodulin-dependent (Ca-CaM) protein kinase activation pathway, plays a role in photoreceptor-regulated light adaptation and maintains the circadian rhythmicity of the mammalian retina [125]. One of the genes in the Ca-CaM pathway around which significant distribution of subterranean-accelerated CNEs are found, *CAMK2D*, has indeed been found to regulate choroidal and retinal neovascularization in mice [9]. The second-ranked hit, the paired-like homeodomain transcription factor 2 (Pitx2) pathway, is a downstream effector of the Wnt/nuclear β-

Table 3.2: Ensembl genes that are significantly enriched for subterranean accelerated CNEs.

SLC7A6OS	*BRCC3*	*PRL2C5*	*SLC7A6*	*GPR137B*
ZFP687	*KLHL4*	*DACH2*	*PPA2*	*G0S2*
NEUROG2	*SAP30L*	*SPHKAP*	*CD226*	*IKBKE*
SRGAP2	*SYNE1*	*DSCAM*	*PACRG*	*PID1*
H2AFB3	*AGAP1*	*QK*	*CAMK2D*	*TMEM261*
PTPRD	*CD34*	*PLXNA2*	*PITX2*	*CAMK1G*
DGKK	*EMCN*	*GRIA1*	*NMUR2*	*COL25A1*
LEF1	*PABPC6*	*GLRA2*	*PROX1*	*SLC24A2*
RAB28	*F9*	*5730508B09RIK*	*RPS6KC1*	*IL1RAPL1*
VRK1	*3110018I06RIK*	*MLLT3*	*SGCD*	*5730480H06RIK*
CNKSR2	*ADAMTSL1*	*BNC2*	*DIAPH2*	*CCDC171*
PI4KB	*AGO2*	*PCP4*	*GBX2*	*ANK2*
PDHA2	*ETNPPL*	*APOB*	*GPATCH2*	*VBP1*
SGMS2	*TET2*	*SERTAD4*	*GM10097*	*CXXC4*
GEMIN8	*MYC*	*SLIT2*	*IRX1*	*MCTP2*
8030423J24RIK				

catenin pathway [26] that is critical for eye morphogenesis, and mutations in *PITX2* can cause eye defects and neurodegeneration [37]. The peroxisome proliferator-activated receptor γ coactivator-1 α (PGC1α) pathway regulates energy metabolism in photoreceptors and similarly manages light susceptibility [62], retinal angiogenesis [209], and circadian clock [143]. These pathways are correlated with neuronal pathways including AMPA receptor trafficking and ERBB signaling, the latter of which plays a role in neural development, myelination, and synaptic plasticity [164]. Other enriched hits include pathways related to the extracellular matrix and the immune system.

We then use *phyloConverge* to score the acceleration of 19,816 protein-coding regions genome-wide and examine the contrast between pathway enrichment in genes accelerated in

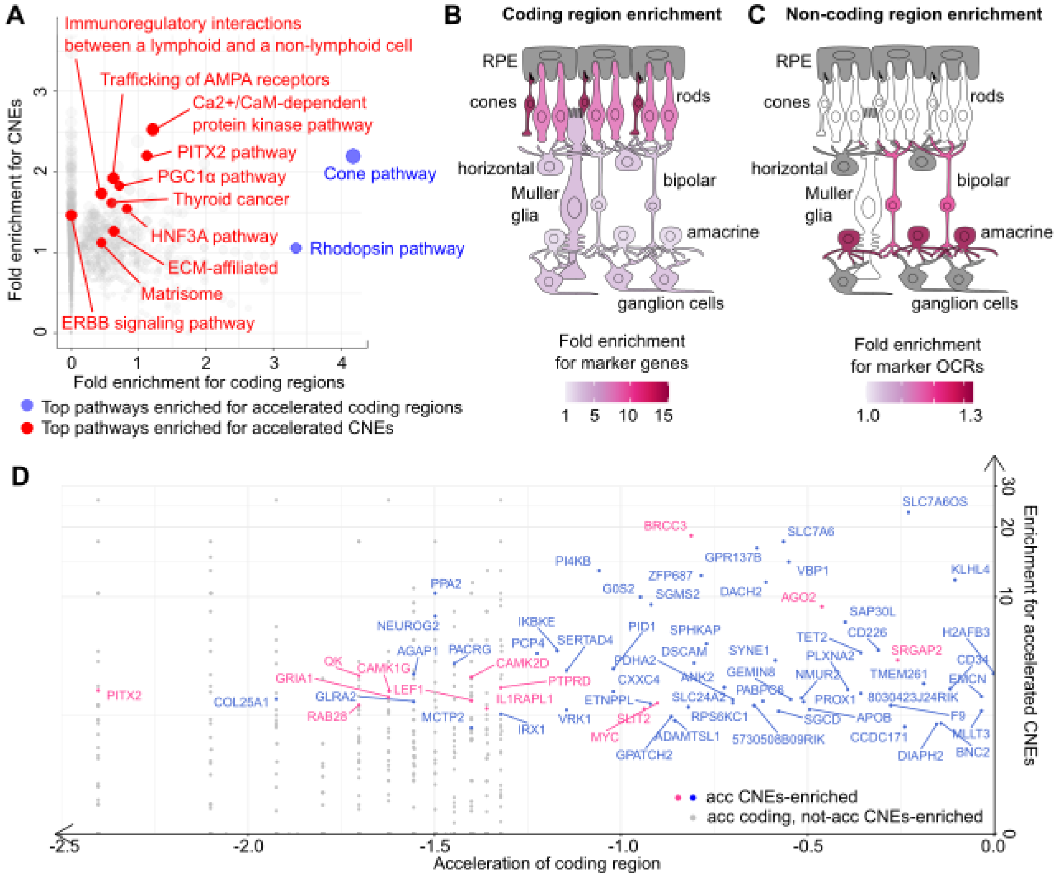


Figure 3.7: Protein-coding region versus CNE acceleration occurs across distinct biological functions. (A) The top-ranking canonical pathways enriched for subterranean-accelerated protein-coding regions (blue) or CNEs (red). (B) Genes that are strongly accelerated in the coding regions show the strongest enrichment for cone and rod photoreceptor marker genes. (C) Strong enrichments for subterranean-accelerated CNEs are found in retinal cells in the inner nuclear layer, namely the amacrine and bipolar cells. Tissues in (B) and (C) for which the Benjamini-Hochberg adjusted enrichment p-value > 0.05 are colored in white, while tissues for which genetic or genomic annotations are not available are colored in grey. (D) Genes that are enriched with accelerated CNEs (blue and pink dots) can have varying evolutionary rate acceleration in the protein-coding regions, while some genes can have strongly accelerated coding regions without enrichment for accelerated CNEs (grey dots). Negative score (s_{corr}) denotes stronger acceleration.

coding regions and genes accelerated in CNEs (Figure 3.7A). We find that the top-ranking pathways with the strongest enrichment for accelerated coding regions (denoted in blue) are specific to phototransduction. Importantly, the neuronal and developmental pathways enriched for accelerated CNEs are not enriched for accelerated coding regions. We provide detailed network views of the top pathways enriched for both acceleration types in Figures 3.8 and 3.9.

A similar trend emerges when we examine enrichment for retinal cell-type-specific marker genes and marker OCRs from single cell sequencing experiments. Using a curated dataset of tissue-specific marker genes across different retinal cell types [151], we observe that genes with accelerated coding regions are significantly enriched for all retinal cell types, but the cone and rod photoreceptors show drastically stronger fold-enrichment than other tissues (Figure 3.7B). In contrast, using genomic annotations of marker OCRs in five retinal cell types [177], we find that the photoreceptor layer and the Müller glia exhibit no enrichment, while cells in the inner nuclear layer of the retina (the amacrine and bipolar cells) show strong, statistically significant enrichment (Figure 3.7C). While photoreceptors, amacrine, and bipolar cells are all neuronal cell-types, the photoreceptors are highly specialized for phototransduction, while amacrine and bipolar cells are specialized inter-neurons that perform signal transduction and signal processing function. Bipolar cells are solely responsible for relaying information from the photoreceptors to the inner layers of the retina and performing specific transformation of neuronal signals [65], while amacrine cells relay signals from the bipolar cells to the ganglion cells and control the temporal regulation of visual signals [244]. The observation of weak enrichment for accelerated coding regions and strong enrichment for accelerated CNEs in the amacrine and bipolar cells lends further evidence to the argument that subterranean adaptation is accompanied by transformation of neuronal functions, which are driven by changes in transcriptional control rather than the genes themselves.

On the individual gene level, there is a subset of genes that are significantly enriched with accelerated CNEs but are not strongly accelerated in the coding regions (Figure 3.7D). The lack of concordance between protein-coding acceleration and enrichment for accelerated CNEs for this set of genes may reflect the role of CNEs in regulating the expression of pleiotropic genes that experience relaxed selection on certain functions but are other-

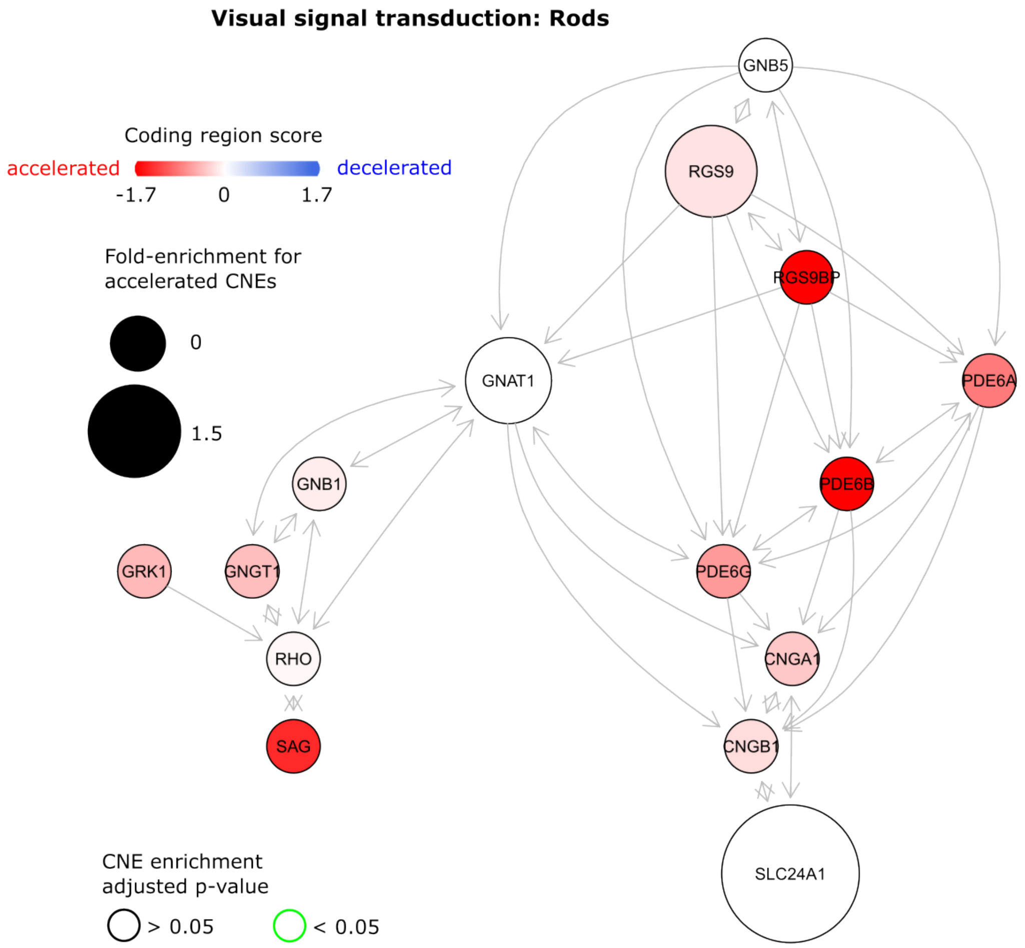


Figure 3.8: Evolutionary rate shifts in protein coding regions and CNEs in an example top-ranking pathway enriched for protein-coding regions acceleration. While none of the individual genes are significantly enriched for accelerated CNEs, they are accelerated in the coding regions (except for a small handful that are neutral). This observation suggests that for pathways that are highly specific to vision, selection pressures tend to act mainly on the coding regions.

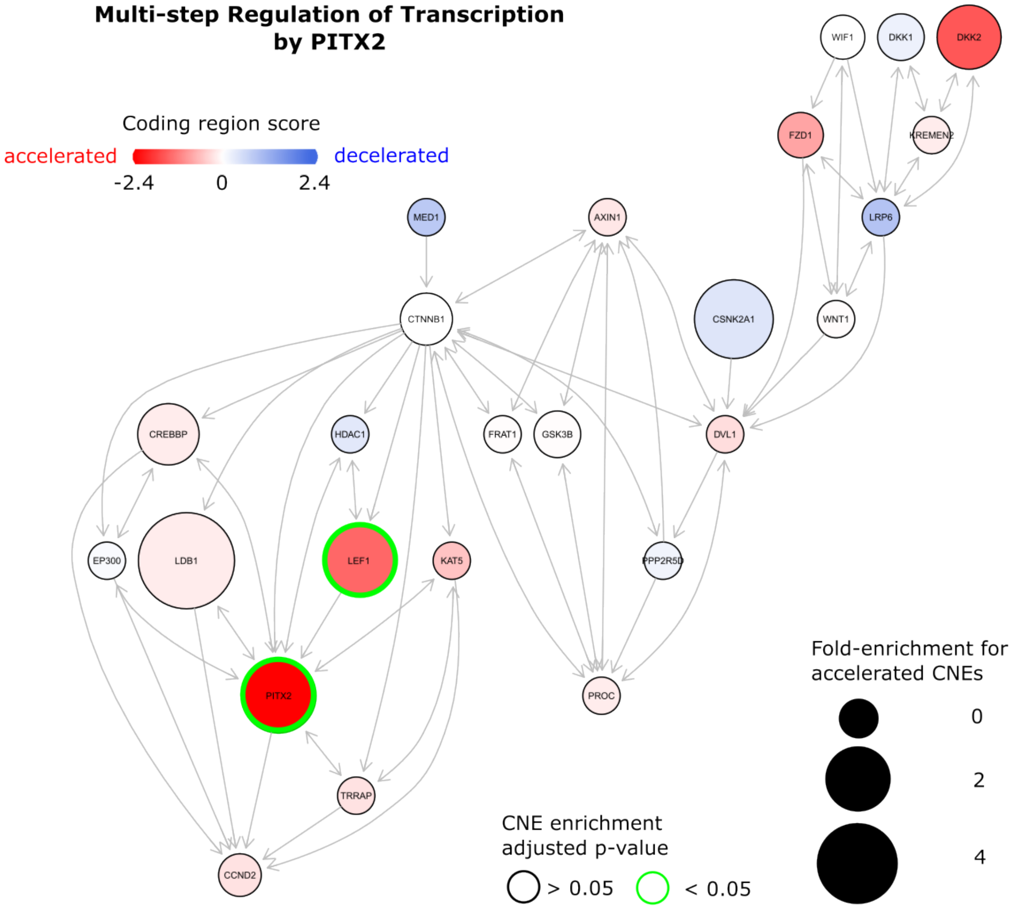


Figure 3.9: Evolutionary rate shifts in protein coding regions and CNEs in an example top-ranking pathway enriched for CNE acceleration. The pathway contains a mix of genes with accelerated or decelerated coding regions that are variably enriched for accelerated CNEs. This observation suggests that for pathways with pleiotropic functions (e.g., general neuronal functions), selection can occur in both the coding and non-coding regions.

wise still critical for survival. Thus, their protein-coding portions did not accelerate and remain under constraint, but their regulatory elements specific to the vision functions could be under relaxed constraint. The top-enriched genes in this set include genes that encode for amino acid transporters (*SLC7A6*), probable nuclear localization of RNA polymerase II (*SLC7A6OS*), regulators of DNA damage response (*BRCC3* and *G0S2*), and a chaperone protein for the Von Hippel-Lindau tumor suppressor gene product (*VBP1*). These functions are general cellular processes that are involved across many tissues, and thus strong conservation of their protein sequences (but not necessarily regulatory elements) would be expected. Examples of eye-related pleiotropy in the set include *DIAPH2*, which has been associated with both age-related macular degeneration and ovarian development [250], and *PROX1*, which is involved in the development of not only the lens and the central nervous system, but also the liver, pancreas, and heart [30, 136, 181, 234, 257]. For these pleiotropic genes, changes that drive the convergent phenotype may have occurred in the regulatory elements that control their expression.

In summary, using pathway and marker enrichment analyses, we find that coding region and CNE acceleration is concentrated in distinct biological functions. While coding region acceleration is observed in genes whose functions are specific to visual signal transduction, non-coding acceleration is enriched for a broader set of developmental and neuronal processes. These observations support the hypothesis that relaxation of selection in the coding regions is concentrated in highly specialized genes, while pleiotropic genes that contribute to the development and function of the visual system but have additional non-vision related roles experience mostly non-coding relaxation.

3.4.3 Transcription factor motif-scale convergence signals reveal the modular evolution of regulatory elements

The general framework of *phyloConverge* has the capacity to fit the convergent rate shifts model at arbitrarily small, even base-pair, resolution, which allows for a deeper inquiry into the information content of different parts of a given CNE. To understand how convergent shifts are reflected in the transcription factor binding site (TFBS) profiles, we

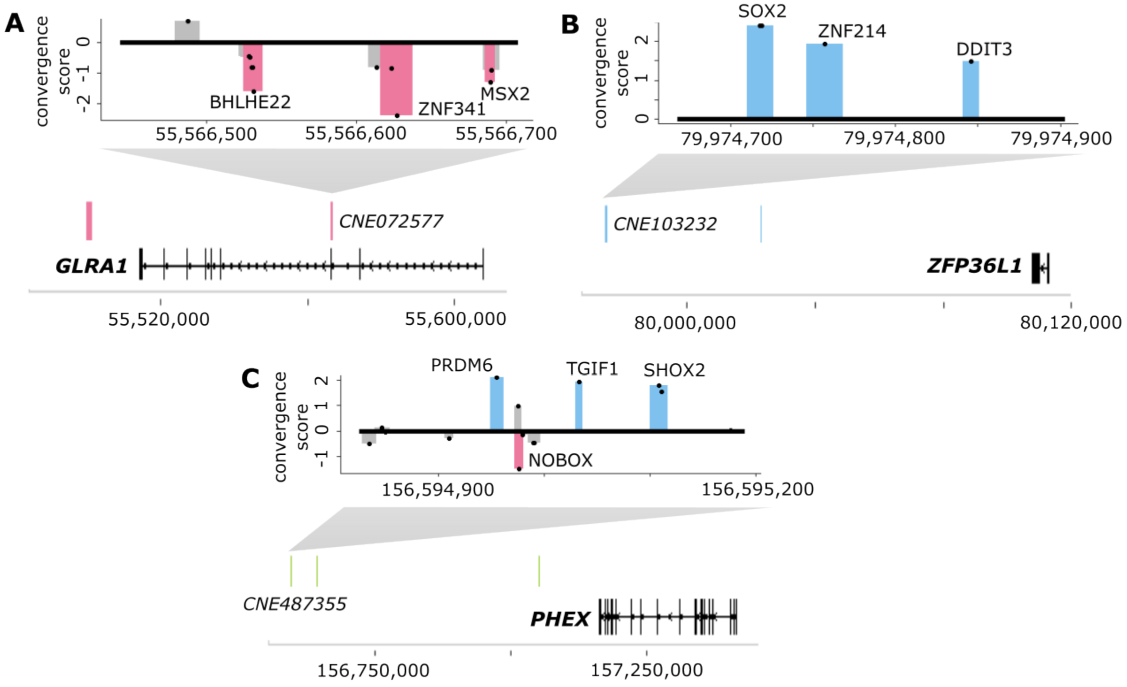


Figure 3.10: Local convergence signals of transcription factor binding site (TFBS) motifs in a CNE highlight modularity of CNE function. (A) CNE072577 is enriched for significantly accelerated motifs. (B) CNE103232 is enriched for significantly decelerated motifs. (C) CNE487355 is enriched for significantly accelerated and decelerated motifs.

identify 1,761,185 TFBS matches within the 491,576 CNEs and compute their motif-level convergence scores (Figures 3.10A-C shows the TFBS convergence profiles of three example CNEs). Using an empirical p-value threshold of 0.05 and filtering for signal robustness, we identified 42,477 significantly accelerated motifs and 81,101 significantly decelerated motifs (2.4% and 4.6% of the total number of conserved motifs, respectively).

To identify CNEs that undergo significant changes in motif content in response to selection, we compute the enrichment for non-redundant, significantly accelerated or decelerated motifs in each CNE. We identify 3 categories of CNEs: 1,579 CNEs that are enriched for significantly accelerated motifs (e.g., Figures 3.10A), 2,288 CNEs that are enriched for significantly decelerated motifs (e.g., Figures 3.10B), and 381 CNEs that are enriched for

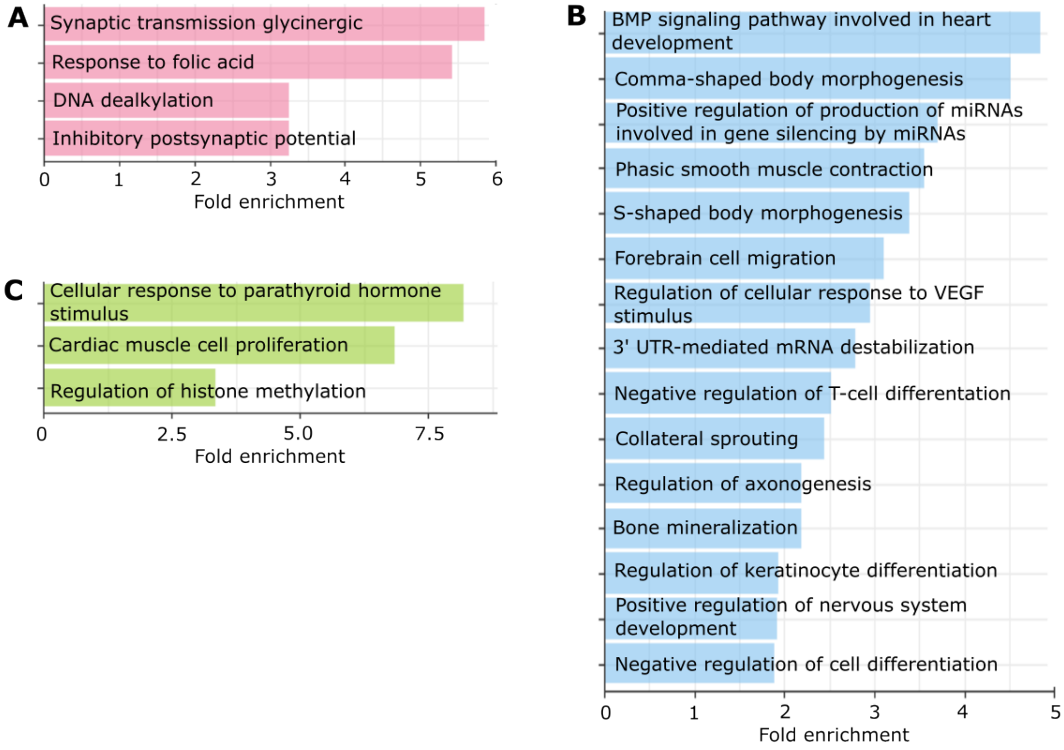


Figure 3.11: Gene Ontology enrichments for "motif-accelerated" CNEs, "motif-decelerated" CNEs, and "mixed-motif" CNEs. (A) Gene Ontology terms that are associated with CNEs enriched for significantly accelerated motifs, (B) significantly decelerated motifs, and (C) both significantly accelerated and decelerated motifs.

both significantly accelerated and decelerated ("mixed") motifs (e.g., Figures 3.10C). Figures 3.11A, B, and C show the Gene Ontology (GO) terms that are significantly enriched for the motif-accelerated CNEs, motif-decelerated CNEs, and mixed-motif CNEs, respectively (Benjamini-Hochberg FDR ≤ 0.1 for both binomial test and hypergeometric test from the *GREAT* tool). While the enriched GO terms for the motif-accelerated and mixed-motif CNEs are specific to a limited number of systems, the enriched annotations for the decelerated-motif CNEs comprise a wider range of functions.

We first take *CNE072577* as an example motif-accelerated CNE that is located in the intron of *GLRA1* (Figures 3.10A). *GLRA1* encodes for glycine receptor $\alpha 1$, which facilitates the transmission of postsynaptic currents specifically in OFF-cone bipolar cells and A-type retinal ganglion cells of the mammalian retina [251]. Meanwhile, from motif-level convergence scores in *CNE072577*, we can see that significant acceleration of the CNE occurs locally in 3 segments that correspond to known binding motifs for *BHLHE22*, *ZNF341*, and *MSX2*. Interestingly, *BHLHE22* has also been suggested to be an expression marker of the OFF-cone bipolar cells [225]. *MSX22*, on the other hand, plays a role in affecting the cell fate commitment and differentiation of retinal ganglion cells [115]. Not only is the acceleration of motifs that regulate retinal development consistent with the degradation of eye structures in subterranean mammals, but the agreement between the annotated functions and transcriptional activities of glycine receptor $\alpha1$, *BHLHE22*, and *MSX22* suggests that local *phyloConverge* scores can provide some insight about the phenotypic associations and possible tissue specificities of specific segments of a CNE.

Next, one of the genes annotated in the "Regulation of keratinocyte differentiation" GO term is *ZFP36L1*, which is located close to *CNE103232* (Figures 3.10B). *ZFP36L1*, or Butyrate Response Factor 1 (BRF1), is known to be involved in the developmental processes of several tissues, including keratinocytes [6, 93] and the paracrine system [153]. The local segments that undergo significant convergent rate deceleration in *CNE103232* correspond to binding sites for *SOX2*, *ZNF214*, and *DDIT3*. *DDIT3* (Chop) is a downstream effector of the unfolded protein response (UPR), a mechanism that is activated to rescue cells from endoplasmic reticulum (ER) stress arising from hostile environmental stressors, such as hypoxia [52]. Indeed, epidermal keratinocyte differentiation is also mediated by the UPR pathway, and changes in Chop levels during keratinocyte differentiation have been observed [24, 159, 241]. Convergent changes in keratinocyte differentiation mechanisms are possibly related to the fact that subterranean mammals have adapted to the burrowing demands of life underground by developing thick footpads. Meanwhile, the association between *SOX2* and *ZFP36L1* are likely due to their role in mediating paracrine signaling. *SOX2*-positive mouse pituitary stem cells that express BRF1 have been found to exhibit paracrine functions, which mediates the commitment and differentiation of progenitors to pituitary cell types [154]. Curiously, paracrine signaling, *SOX2*, and BRF1 have all been suggested to play a role in regulating circadian rhythmicity [38, 120, 160]. More specifically, *SOX2* and paracrine signaling have been found to control circadian pacemaking in the suprachiasmatic nucleus (SCN) neurons, the central clock of biological circadian rhythms [38, 160]. These findings suggest the possible involvement of *ZFP36L1* and the *SOX2* binding site of *CNE103232* in facilitating changes in the circadian machinery of subterranean mammals. Although subterranean mammals live in the dark and have regressed ocular structures, their circadian machineries have been found to be conserved, although their regulation patterns have changed compared to mouse [10, 82].

Finally, Figures 3.10C shows the association between *CNE487355* with *PHEX*, a gene in the "Cellular response to parathyroid hormone stimulus" GO annotation. Changes in parathyroid signaling in subterranean mammals are likely driven by the fact that their habitats lack solar exposure and thus are naturally deficient of vitamin D. Several subterranean mammals have indeed been found to be naturally vitamin D-deficient [168, 222]. When vitamin D levels are low, the body compensates by secreting higher levels of parathyroid hormones [97], which causes changes to systems affected by calcium metabolism, including bone metabolism. Changes in the physiology and mineralization of bone and teeth in subterranean mammals have been reported [168]. *PHEX* is predominantly expressed in odontoblasts, osteoblasts, and osteocytes, and is regulated by PTHrP(1-34), a parathyroid hormone-related protein that is active in osteoblasts during development [247]. Interestingly, one of the segments undergoing significant deceleration in *CNE487355* correspond to the binding site for *SHOX2*, which controls stylopod development in mice specifically by regulating chondrocyte maturation and endochondral ossification [270].

Thus, local convergence analysis with *phyloConverge* shows that regulatory elements likely evolve in modular units, resulting in non-uniform convergence signals across the element. Segmenting an element into modules allows us to have a more nuanced perspective on how pleiotropic elements respond to selection pressures. The identified phenotype-associated segments serve as testable hypotheses that can be further tested in experimental settings. Additionally, the identified phenotype associations also signify that *phyloConverge* can detect TFBS motif-scale convergence signals with high fidelity.

3.4.4 Global patterns of motif adaptation highlight regulatory rewiring associated with subterranean adaptation

Next, we investigate the functional signals that explain the patterns of motif-level adaptation genome-wide. We first ask if there are specific motifs that experience global convergence shifts in correlation with subterranean adaptation. We find a significant enrichment for 43 motifs in the set of 42,477 significantly accelerated motifs, and for 117 motifs in the set of 81,101 significantly decelerated motifs, with 11 motifs experiencing both global acceleration and deceleration. Among these motifs are motifs that are known for their involvement in eye and nervous system development, including *SOX2*, *ALX1*, *BHLHE22*, *ARX*, and *TBR1*.

Next, we use the *GREAT* tool to evaluate the functional enrichment of the 42,477 significantly accelerated motifs and the 81,101 significantly decelerated motifs for annotations in the Reactome pathway database. We find that the accelerated and decelerated motifs are enriched for functions that are strongly interrelated, shown in clusters in Figure 3.12A. One cluster reflects the regulatory changes that drive the adaptation of neuronal functions, with some functions enriched for only the accelerated motifs, only the decelerated motifs, or both, in the case of the "Neurexins and neuroligins" and the "Neurotransmitter receptors and postsynaptic signal transmission" annotations. This finding suggests that the adaptation of critical and pleiotropic functions in subterranean mammals is largely driven by regulatory rewiring that involves gains and losses of different motifs across CNEs, in contrast with ocular-specific functions that are mainly driven by coding region. Another enriched cluster reflects changes in SUMOylation and TFAP2 TF family. The enrichment of these functions may represent the regulatory adaptation to the hypoxic environment of subterranean habitats. SUMOylation, in which the Small Ubiquitin-related MOdifier (SUMO) attaches covalently to proteins, plays a role in regulating the hypoxic response, and the deSUMOylation of TFAP2A has been found to increase the transcriptional activity of Hypoxia-Inducible Factor 1 (HIF-1) under hypoxia [36]. Additionally, SUMOylation and TFAP2 are involved in regulating physiological circadian rhythm [109, 138], which is known to undergo changes in subterranean mammals [16].

Finally, given the functional promiscuity of TFs, we ask if motif-level convergence signals

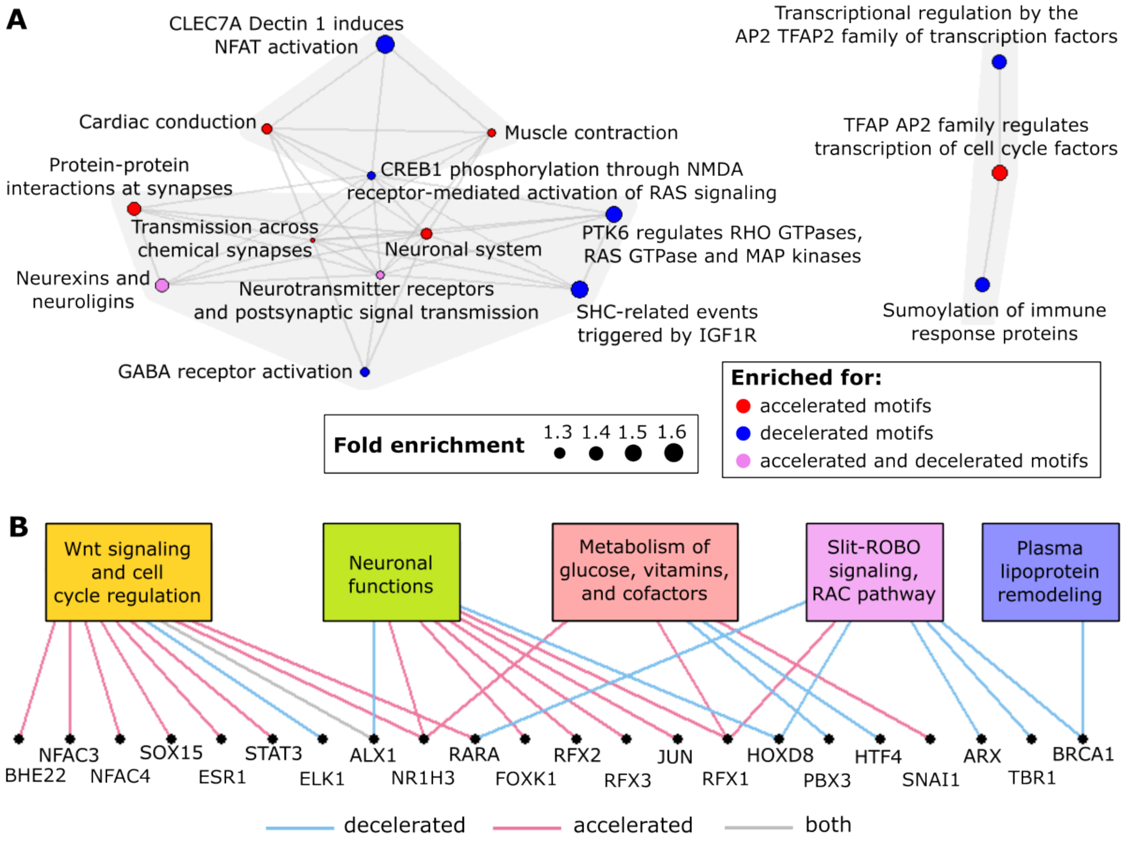

A
CLEC7A Dectin 1 induces NFAT activation
Cardiac conduction
Muscle contraction
Protein-protein interactions at synapses
CREB1 phosphorylation through NMDA receptor-mediated activation of RAS signaling
Transmission across chemical synapses
Neuronal system
PTK6 regulates RHO GTPases, RAS GTPase and MAP kinases
Neurexins and neuroligins
Neurotransmitter receptors and postsynaptic signal transmission
SHC-related events triggered by IGF1R
GABA receptor activation
Transcriptional regulation by the AP2 TFAP2 family of transcription factors
TFAP AP2 family regulates transcription of cell cycle factors
Sumoylation of immune response proteins
Fold enrichment
1.3 1.4 1.5 1.6
Enriched for:
accelerated motifs
decelerated motifs
accelerated and decelerated motifs
B
Wnt signaling and cell cycle regulation
Neuronal functions
Metabolism of glucose, vitamins, and cofactors
Slit-ROBO signaling, RAC pathway
Plasma lipoprotein remodeling
BHE22
NFAC3
NFAC4
SOX15
ESR1
STAT3
ELK1
ALX1
NR1H3
RARA
FOXK1
RFX2
RFX3
JUN
RFX1
HOXD8
PBX3
HTF4
SNAI1
ARX
TBR1
BRCA1
decelerated
accelerated
both

Figure 3.12: Convergent motif-level changes highlight regulatory rewiring associated with subterranean adaptation. (A) Reactome pathway annotations enriched for accelerated and decelerated motifs genome-wide. (B) Motif-specific functional enrichment analysis reveals functions associated with acceleration or deceleration of specific motifs that have global convergence signals.

could highlight the specific functional changes that occur with the acceleration or deceleration of specific motifs. We use *GREAT* to perform motif-specific functional enrichment analysis for the acceleration and deceleration of individual motifs that were identified as having global convergence signals, given the genome-wide distribution of each motif. With FDR ≤ 0.05, we identify the enrichment for 43 unique and interrelated pathways across 22 motifs. To illustrate the interpretation of these results using the *RFX3* motif; while the genome-wide distribution of the motif is associated with various phenotypes including neuronal functions, cardiac functions, and ion homeostasis, our results show that the convergent acceleration of *RFX3* motifs in subterranean mammals is strongly specific to neuronal functions only. This suggests that when selection pressures act only on specific functions of pleiotropic transcription factors, changes in transcription factor activities can be mediated by convergent rate shifts occurring in specific locations that are associated with the functions. Clustering the enriched annotations based on pathway correlations, we find that motif-specific regulatory changes in subterranean adaptation mainly occurs in 5 functional categories, namely "Wnt signalling and cell cycle regulation", "neuronal functions", "metabolism of glucose, vitamins and cofactors", "Slit-ROBO signaling and RAC pathway", "plasma lipoprotein remodeling" (Figure 3.12B). Consistent with the previous results, most of the motif changes associated with neuronal functions and Wnt signaling are in the accelerated direction.

3.4.5 Unsupervised scanning of conserved elements predicts segments with potential association with subterranean phenotype

Having established that *phyloConverge* signals can capture TFBS-scale convergence, we ask if *phyloConverge* can be used to make predictions of phenotype-relevant TFBS-scale segments without requiring users to supply prior definition of known TFBS coordinates to score. We use *phyloConverge* to scan each nucleotide in the CNEs and compute the score from a window of ±5bp around the nucleotide, considering that ~10bp is the approximate scale of TFBS motifs. We find that the scanning output can highlight strongly accelerated and decelerated segments that correspond to known TFBS motifs. The top of Figure 3.13A shows the example scanning output for *CNE327067*, which is located close to *SLC24A2*, a cation/calcium ion exchanger that maintains the homeostasis of sodium, potassium, and calcium ion levels in the brain, retinal ganglion cells, and the retinal cone photoreceptors [224]. We can see that the segment with the strongest acceleration signal correspond to a known TFBS for *POU4F2*, a canonical retinal marker whose expression together with *ISL1* has been found to be sufficient for specifying the retinal ganglion cell fate [262]. We also observe that there are other strongly accelerated or decelerated segments that do not correspond to known TFBS motifs.

After performing the scanning on all the CNEs, we identified nucleotide segments that exhibit significant acceleration or deceleration signals. Out of the identified segments, only about 5.4% of them overlap with known TFBS coordinates that we previously called (Figure 3.13B). We then perform *de novo* motif discovery analysis for the newly identified segments, using the *STREME* tool [13]. Setting an E-value threshold of 0.5, we found enrichment for 5 motifs in the new accelerated segments (Figure 3.13C), and 10 motifs in the new decelerated segments (Figure 3.13D). We used the *TomTom* motif comparison tool [92] to test whether the enriched motifs are similar to known consensus motifs. Using an E-value threshold of 1 for the *TomTom* test, we characterize the known motifs that are significantly similar to the enriched motifs (Tables 3.3 and 3.4). Interestingly, there are 4 motifs that are enriched in both the accelerated and decelerated new segments. Some of these hits are associated with hypoxia response; *FOXD2* (motif 5 in the decelerated set and motif 2 in the accelerated set) have been previously found to be enriched in the binding sites of HIF-2α (Hypoxia Inducible Factor 2 alpha) in HepG2 cells [231], *SP1/2* (motif 4 in the decelerated set and motif 5 in the accelerated set) are found in the binding sites of HIF-1α in RCC4 and HKC-8 cells [231], and *STAT4* (motif 3 in the decelerated set and motif 1 in the accelerated set) are upregulated in the primary human macrophages under hypoxia [71]. Another motif that is also enriched in both the accelerated and decelerated sets is *SOX10*, which is highly expressed in the brain. During development, *SOX10* is expressed exclusively in oligodendrocyte precursor cells and is critical for controlling the maturation of oligodendrocytes [192]. These findings further characterize the regulatory changes in neuronal development and hypoxia response that occur with subterranean adaptation.

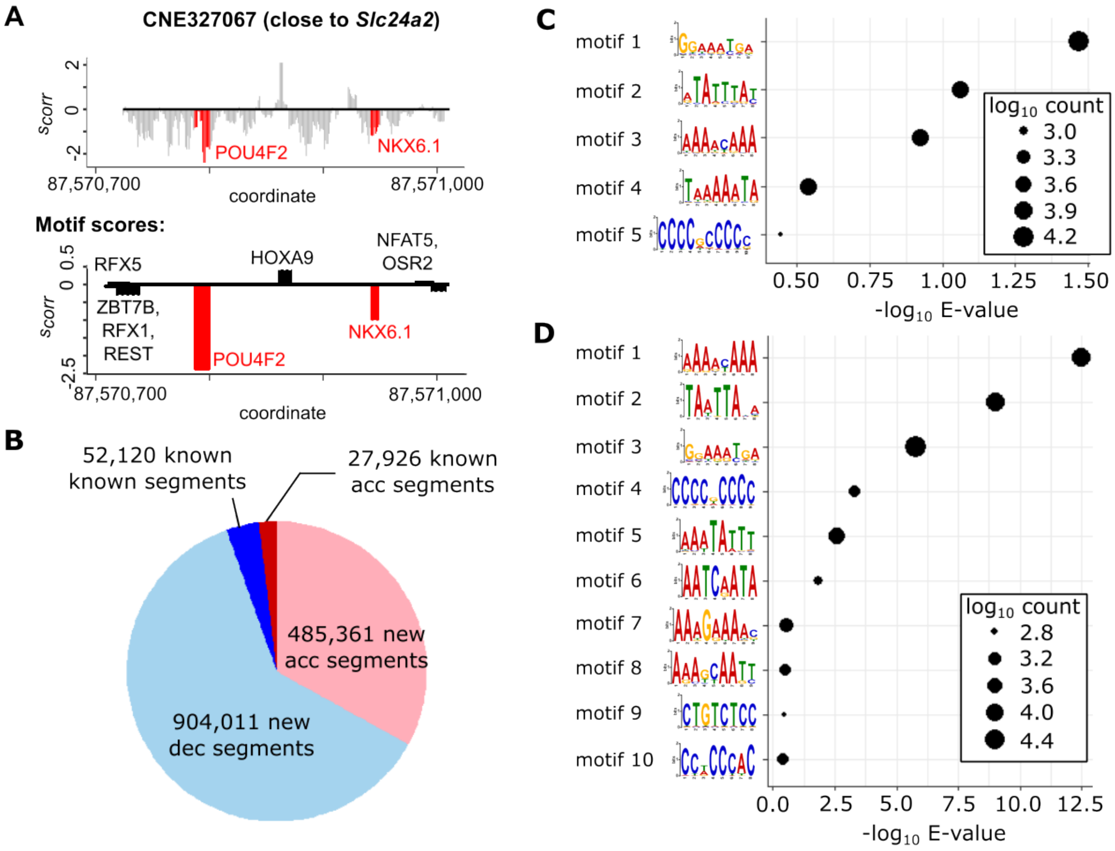


Figure 3.13: Unsupervised scanning of CNEs proposes potential motifs undergoing significant convergent changes in subterranean adaptation. (A) The scanning output of an example CNE, *CNE327067*, where a sliding window of 11bp is used to compute the convergence signals at every nucleotide position. The bottom plot shows the corresponding scores for known TFBS motif coordinates. (B) The number of identified segments with significant convergent rate shifts from scanning. (C) *De novo* motif discovery analysis on accelerated segments that are newly identified without prior knowledge of motif coordinates, and (D) the same analysis for newly identified decelerated segments.

Table 3.3: *De novo* motif discovery analysis results for segments predicted to be convergently accelerated

Motif name	**STREME E-value**	**TomTom match**	**E-value**	**Database**
Motif 1	0.034	STAT4	0.65	JASPAR2022 core vertebrates non-redundant v2
Motif 2	0.087	FOXD2	0.707	Jolma2013
Motif 3	0.12	SOX10	0.207	JASPAR2022 core vertebrates non-redundant v2
Motif 4	0.29	MEF2A, MEF2B, MEF2C, MEF2D	0.0386 to 0.488	JASPAR2022 core vertebrates non-redundant v2
Motif 5	0.36	KLF, SP, PATZ	$2.65e-5$ to 0.447	JASPAR2022 core vertebrates non-redundant v2

Table 3.4: *De novo* motif discovery analysis results for segments predicted to be convergently decelerated

Motif name	**STREME E-value**	**TomTom match**	**E-value**	**Database**
Motif 1	$3.2e-13$	SOX10, FOXL1, SOX15	0.185 to 0.803	JASPAR2022 core vertebrates non-redundant v2, Uniprobe
Motif 2	$9.2e-10$	SOX14, SOX21, SRY	0.181 to 0.333	Uniprobe
Motif 3	$1.6e-6$	STAT4	0.746	JASPAR2022 core vertebrates non-redundant v2
Motif 4	$4.9e-4$	KLF, SP, PATZ	$4.1e-5$ to 0.358	JASPAR2022 core vertebrates non-redundant v2
Motif 5	0.0025	FOXD2	0.979	Jolma2013
Motif 6	0.015	ONECUT, CUX, SOX3	0.0148 to 0.565	JASPAR2022 core vertebrates non-redundant v2, Jolma 2013
Motif 7	0.28	-	-	-
Motif 8	0.32	MSX, BARX1	0.348 to 0.508	Jolma2013 and Uniprobe
Motif 9	0.34	-	-	-
Motif 10	0.4	GLI, SP, KLF	0.11 to 0.931	JASPAR2022 core vertebrates non-redundant v2, Jolma2013

3.5 Discussion

We introduce *phyloConverge*, a new comparative genomics method that combines explicit estimation of nucleotide substitution rates and adaptive calibration of test statistics to identify the phenotypic associations of genetic elements. For a phenotype of interest, *phyloConverge* quantifies the amount of local rate convergence signal via a maximum likelihood estimation of a two-parameter neutral tree scaling model. The MLE statistics are calibrated with an empirical p-value, which dramatically reduces multiple sources of bias.

Benchmarking our method using an empirical dataset that was previously analyzed for rate convergence in subterranean mammals [208], we find that *phyloConverge* identifies CNEs that exhibit strong associations with ocular functions–which satisfies our expectations for the phenotype–and discover that the regression of ocular functions may be accompanied by changes in neuronal functions and development. We also demonstrate that *phyloConverge* can analyze a given CNE in segments and provide insights about its pleiotropic activity in the specific phenotypic context. Importantly, *phyloConverge* produces unbiased signals because it corrects for biological and technical confounders.

phyloConverge offers a scalability to perform rapid, calibrated scoring at flexible resolution. We have demonstrated this flexibility by applying *phyloConverge* in three complementary ways: scoring entire CNEs for aggregate CNE-level acceleration, scoring TFBS for TFBS-level convergent rate shifts, and dissecting aggregate acceleration signals with high resolution scoring. This highly flexible framework allows for rapid convergent acceleration scanning with less computational overhead than competing methods. For example, to score 1 million elements, *RERconverge* would require the pre-estimation of 1 million element-specific trees, and *Forward Genomics* would require computations of the neutral tree model as well as local (branch-specific) or global (relative to the root) percent identity values per branch per element. For the same analysis, *phyloConverge* would only require the estimation of one neutral tree model, while the scoring of the 1 million elements would be performed through a small number of parameter estimations and hypothesis testing. While adding the permulation step incurs additional computational cost, we have demonstrated previously the advantage of calibrating the resulting statistics via permutations is not method-dependent.

phyloConverge can be tractably extended to perform convergence scans genome-wide, generating convergent rate tracks similar to the *phyloP* conservation score. This provides the option of scanning entire genome alignments to detect coordinates with significant convergent shifts in evolutionary rates without needing prior knowledge about the coordinates and definitions of the functional regulatory units of the elements. We propose a possible strategy for performing unsupervised predictions of regions with convergence signals and the possible associated motifs (Figure 3.13), but the optimal approach for such unsupervised genome-wide scanning remains to be determined because we still lack a thorough understanding of non-coding regions to inform our interpretation of significant rate shifts in non-coding elements. For example, there are types of non-coding elements whose conservation patterns are less well-understood, including long non-coding RNAs, which tend to have a highly conserved promoter region but a less conserved transcribed region [117], and microRNAs, which can also have varying conservation patterns [193]. Investigation into understanding conservation in non-coding elements and how it can inform the design of unsupervised genome-wide scanning for convergent rate shifts can be pursued in future work.

It is important to note that predictions generated by sequence alignment-based methods such as *phyloConverge* should be interpreted with some caveats. It is increasingly understood that some enhancers can have homologous functional activity across distantly related species despite lacking enhancer-wide sequence conservation. For example, characterization of the putative Islet-Spacer enhancers in sponge, fish, mouse, and human revealed that functionally homologous enhancers can have high variability in compositions, orientations, numbers, and alignments of a common set of TFBS [258]. The quality of predictions also hinges upon the global alignability of sequences, which can deteriorate with increasing evolutionary distance. For such enhancers, sequence alignment-based methods would likely fail. In this instance, "alignment-free" methods that compare sequences in some functional readout space may be appropriate. Nonetheless, sequence alignment-based methods would be sufficiently powerful to analyze promoter regions and strongly conserved enhancers that are often critical for developmental processes. We also note that the statistical power of permutation-based methods such as permulation would only increase with the number of permutations used. Depending on the compute power available to the user and the size of the dataset, user may be limited to use a relatively small number of permulated phenotypes (e.g., our study uses 500 permulations). Consequently, the resulting p-values may not have a sufficiently high resolution for traditional multiple testing approaches. However, permutation tests have been widely used as a strategy for correcting for multiple testing, not by taking a family-wise correction approach, but by calibrating each individual statistic through constructing the null distribution while preserving irregularities that may exist in the data. As a result, permutation tests would produce well-calibrated rankings of top hits, while avoiding the issue of over-correction or under-correction of statistics that is inevitable with family-wise approaches.

4.0 Convergent evolution of protein-coding and regulatory non-coding regions underlying mammalian adaptation to high altitudes

4.1 Attribution statement

All of the work in this chapter was performed by myself, with the following exception:

- Curation of gene hits from high altitude-related population genetics studies and analysis on amino acid convergence was performed by Allie Graham, Ph.D.

4.2 Introduction

Oxygen is a critical fuel of eukaryotic life. Since the atmospheric oxygen concentration increased ∼450 million years ago to approximately the current level [140], land-living eukaryotes adapted to the newly oxic environment and developed energy metabolism machineries that can exploit oxygen as an efficient energy source. With oxygen as the final electron acceptor in the mitochondrial respiration of mammals, the yield of ATP per glucose molecule is 7.5 times greater than that from the anaerobic respiration via fermentation [173, 204]. The dramatic increase in energy production from aerobic respiration provided cells with the ability to perform 1,000-fold biochemical reactions compared to anaerobic respiration [202], which eventually enabled cells to achieve higher levels of complexity like compartmentalization and multicellularity. It is therefore not surprising that low oxygen environments, known as hypoxia, can cause aberrations to biological functions. In fact, hypoxia is pathological in many human diseases, including ischaemia reperfusion injury, cancer, pre-eclampsia, endometriosis, heart diseases, stroke, and more.

However, some species have evolved to survive in hypoxic conditions. Species that have adapted to high altitude environments, for instance, are able to thrive in hypobaric hypoxia – a condition where the partial pressure of oxygen in the atmosphere is greatly decreased – in addition to other stressors including high ultraviolet exposure, extreme cold, and dryness.

Uncovering the genetic basis of adaptation to chronic hypoxia at high altitudes could possibly offer insights on the mechanisms of oxygen transport and metabolism that engender hypoxia tolerance, which could be informative for designing treatments for hypoxic diseases.

Numerous population genetics studies have been conducted to look into this question, many of which highlighted the hypoxia-inducible factor (HIF) pathway as the master controller of hypoxia response at high altitude [18, 21, 169, 229]. At high oxygen levels, HIF-1α is hydroxylated in an oxygen-dependent manner, resulting in its degradation. Under hypoxia, this degradation is halted, which results in the stabilization of HIF-1α levels. The dimerization of HIF-1α and HIF-1β then activates downstream pathways that control response to hypoxia, including reducing mitochondrial biogenesis, regulating red blood cell production, and others [20]. Given the central role of the HIF pathway in regulating hypoxia response, it is reasonable to see consistent signals of selection for this pathway across different high altitude populations. However, beyond the HIF pathway, previous reports have documented that altitude-associated loci from different populations can have few or no overlaps [19] and heritability patterns of altitude-associated loci can differ across populations [17], suggesting that the convergent physiological adaptation to high altitudes in these populations occurred through independent mechanisms. While there can be local selection pressures that uniquely act on each population, more needs to be done to characterize loci that show robust signals of selection for high altitude adaptation.

Fortunately, as we zoom out from a population level to a macroevolutionary level, acclimatization of lowland species to high altitudes have occurred repeatedly across independent clades [79, 123, 174, 194, 197]. This gives us the statistical power to perform a comparative analysis to characterize the genetic and epigenetic sequence adaptations that have repeatedly been selected over millions of years to give rise to high altitude phenotypes. Thus, this work aims to decode the genetic basis of high altitude adaptation using convergence analysis, with evolutionary rate as a proxy for selection. We first conduct a meta-analysis of population genetics studies on mammalian high altitude adaptation to establish our understanding on the expected functions and pathways that undergo changes during adaptation to high altitudes. Then, we perform phylogenetic analyses to identify proteins and regulatory elements that experience decelerated or accelerated evolutionary rates in association with convergent adaptation to high altitude. By grounding our phylogenetic predictions on the meta-analysis results, we distinguish pathways and mechanisms that have likely evolved under purifying or positive selection during adaptation to high altitude.

4.3 Materials and Methods

4.3.1 Construction of amino acid and conserved non-coding region alignments for a 120-way mammalian phylogeny

We used a multiple genome alignment dataset for 120 mammals recently produced by Hecker and Hiller [98] (Figure 4.1). We introduced a minor correction to the phylogenetic tree to consolidate the relative placements of the clades according to the common structure of the mammalian phylogeny. This topology was used for all tree estimations performed for all genes and conserved non-coding elements (CNEs). To compute the neutral substitution model for the updated tree topology, we first identified the fourfold-degenerate (4D) sites across the entire genome. Then, the *phyloFit* function from *PHAST* [111] was used to estimate the nucleotide substitution model from the concatenation of all the identified 4D sites.

To construct the amino acid alignment dataset from the 120-way mammalian genome alignment, we first obtained the coding region coordinates for a total of 19,610 genes in the NCBI RefSeq gene annotations for the *hg38* assembly, and extracted the multiple sequence alignments (MSAs) for the coding regions of the genes using *RPHAST* [111]. We then used a custom codon model to convert the nucleotide sequence of each gene orthologs to the corresponding amino acid sequence, and the amino acid orthologs were then aligned using MUSCLE. The resulting MSAs were used to compute gene-specific evolutionary trees using *phangorn* [218], and the *readTrees* function in *RERconverge* was used to compute a master gene tree with branch lengths that were averaged from the corresponding branches across all gene trees.

To define conserved non-coding elements (CNEs), we took the conserved elements previ-

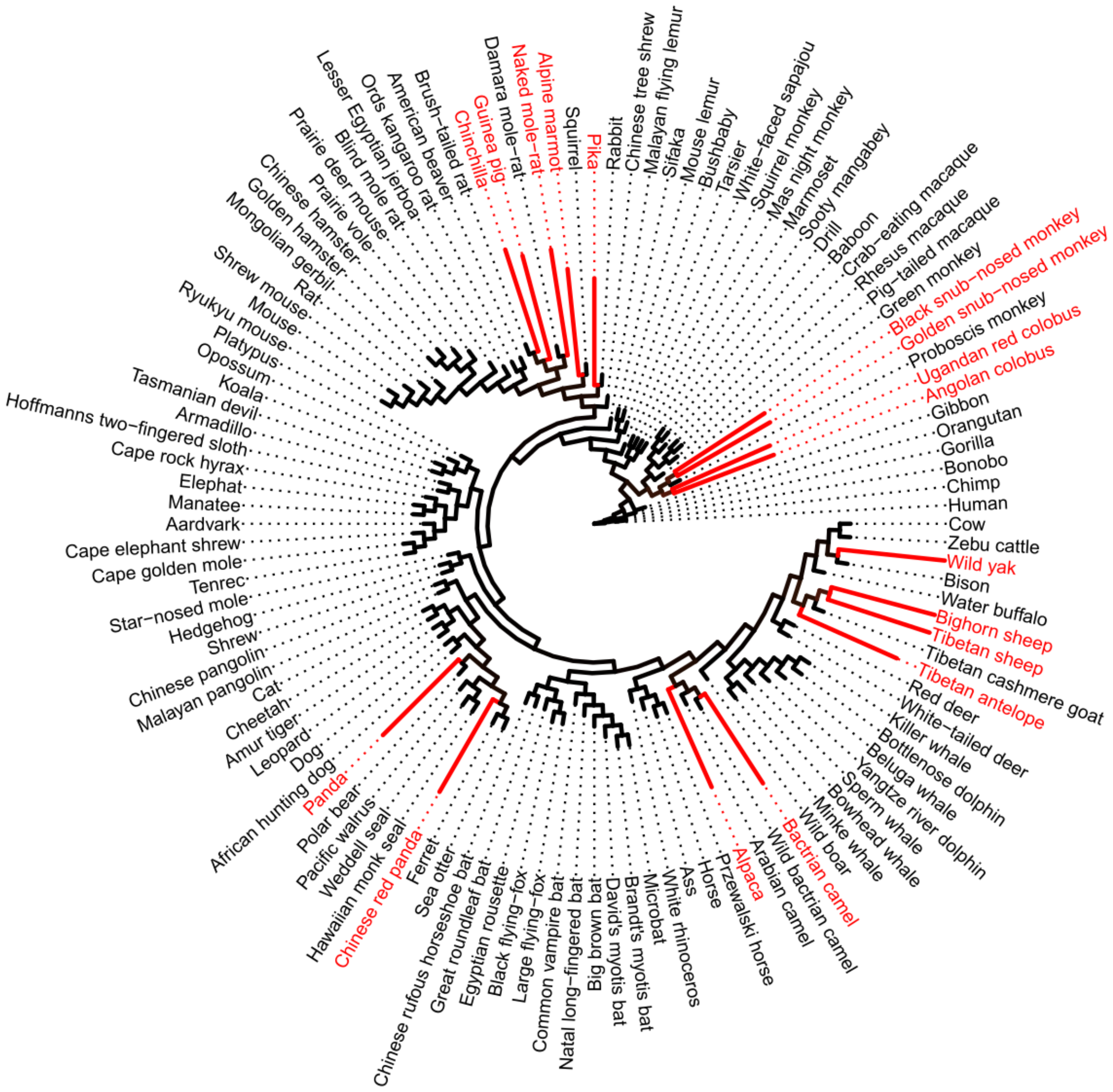


Figure 4.1: 120-way updated mammalian phylogeny used for convergent high altitude analysis. High altitude species (red tip branches) are defined as species that live exclusively at altitudes ≥1,000 meters.

ously identified from the 120-way mammalian alignment using *GERP++* [49] and removed the subset that overlapped exons and were less than 30bp in length, resulting in a total of 1,050,080 CNEs. The MSAs of the CNEs were then used to compute CNE-specific nucleotide trees using *phangorn*, with the "Generalized Time Reversible" model as the nucleotide substitution model for the estimation. Finally, the *readTrees* function in *RERconverge* was used to read the CNE-specific trees into a *multiPhylo* object and compute a master CNE tree with branch lengths that were averaged from the corresponding branches across all CNE trees.

4.3.2 Annotation of high altitude species in the 120-way mammalian alignment

To define the set of "foreground" species that have adapted to high altitude environments, we first curated information on the range of altitudes occupied by all the species in the phylogeny from various sources. We then assigned species that were known to exclusively occupy altitudes no lower than 1,000 meters to be high altitude species (red branches in Figure 4.1). This criterion designated 17 species as high altitude species, including pika (*Ochotona princeps*), naked mole rat (*Heterocephalus glaber*), guinea pig (*Cavia porcellus*), chinchilla (*Chinchilla lanigera*), Alpine marmot (*Marmota marmota*), Angolan colobus (*Colobus angolensis palliatus*), Ugandan red colobus (*Piliocolobus tephrosceles*), black snub-nosed monkey (*Rhinopithecus bieti*), golden snub-nosed monkey (*Rhinopithecus roxellana*), Tibetan antelope (*Pantholops hodgsonii*), wild yak (*Bos grunniens mutus*), sheep (*Ovis aries*), bighorn sheep (*Ovis canadensis*), Bactrian camel (*Camelus bactrianus*), alpaca (*Vicugna pacos*), lesser panda (*Ailurus fulgens styani*), and panda (*Ailuropoda melanoleuca*).

4.3.3 Meta-analysis of genes associated with high altitude adaptation from population genetics studies

To perform a meta-analysis of altitude-associated gene hits from population genetics studies, we compiled the results from a total of 23 studies, comprising findings from populations of humans [63, 86, 112, 216, 264, 266], primates [41], dogs [87, 142], pika [80], ungulates [57, 80, 100, 233], pigs [1, 141], and marmots [12]. We identified significant genes from each study and counted the number of instances that each gene was identified as significant. Genes that were significant in at least three instances were collated and evaluated for functional enrichment.

4.3.4 Detection of altitude-associated convergent rate shifts in amino acids and conserved non-coding elements (CNEs)

Convergence analyses on amino acids and CNEs were performed using the *RERconverge* package [128, 188]. *RERconverge* detects the convergent rate shift of a genomic elements by computing the correlation between a convergent phenotype of interest and the relative evolutionary rates (RERs) of the orthologs of the element across species in the phylogeny. RERs are defined as the relative substitution rate along each branch of an element-specific tree normalized against the "neutral" branch length averaged genome-wide. In the *RERconverge* framework, RERs are quantified by computing the residuals from regressing the element-specific tree against the neutral tree, and then correcting them for heteroscedasticity.

The specific steps performed in *RERconverge* were as follows. First, we used the *getAllResiduals* function to compute the RERs of the gene orthologs from the gene-specific trees and the pre-computed master gene tree. Then, the *correlateWithBinaryPhenotype* function was used to compute the associations between the RERs with the high altitude phenotype. To correct for statistical and phylogenetic biases, we used permulation [212] to produce null phenotypes that were sampled from a generative model inferred from the observed high altitude phenotype. The null phenotypes were then used to compute null correlation statistics, which were subsequently used to compute empirical/permulaton p-values of phenotype associations. The permulation p-value of a gene was defined as the proportion of the null statistics that were equal to or more extreme than the observed statistics. The same steps were also used to compute the convergence scores of the CNEs.

To identify the set of significantly accelerated or decelerated proteins or CNEs, we set a permulation p-value threshold of ≤ 0.05. We also conducted a "leave-one-out" robustness filter in which the computation was repeated by excluding one foreground species each time. A convergence signal was only considered robust if removing one species did not eliminate or reverse the convergence signal.

4.3.5 Transciption factor motif-scale convergence analysis

We also performed convergence analysis by scoring conserved transcription factor (TF) motifs that overlapped the CNEs. We previously made motif calls for 771 TF motifs from the HOCOMOCO database (version 11) [132] for the human *hg19* coordinate. First, we obtained the position weight matrices (PWMs) for all of the motifs. Then, we computed the background nucleotide frequencies for each PWM using *motifDiverge* [126]. The background frequencies were then used to infer the PWM score cutoff to control Type I error rate of the motif calling at below 10^{-5}. Given the inferred score thresholds and the PWMs, we used *PWMScan* to perform motif calling on the *hg19* assembly. The motif calls were then lifted over to the *hg38* coordinates with the *liftOver* tool from UCSC [131], and the coordinates of the conserved motifs were identified by intersecting the motif coordinates with the CNE coordinates using *BEDTools* [199]. Finally, the convergence signals of the conserved motifs were computed using *phyloConverge* [213], which also performed leave-one-out robustness tests. Significantly decelerated or accelerated motifs were identified by setting permulation p-value ≤ 0.05 and using the robustness filter.

4.3.6 Functional enrichment analysis

For the amino acid hits and the meta-analysis of population genetics gene hits, functional enrichment analysis was performed using Fisher's exact test with permutation tests for bias correction. We constructed the null distribution of Fisher's exact test odds ratio from sampling 500 sets of randomly selected hits with a matching set size as the observed hits. Then, we computed the empirical p-values of enrichment by calculating the fraction of the null odds ratio that were as extreme or more extreme than the observed odds ratio. The fold enrichment value was quantified using the observed odds ratio. To control for the strictest false discovery rate (FDR) threshold in this resolution, we set a permutation p-value threshold of ≤ 0.002, which is the resolution of the permutation test. This corresponded to a Benjamini-Hochberg FDR of 0.16, 0.12, and 0.38 for the population genetic gene hits, the decelerated amino acids, and the accelerated amino acids, respectively.

Functional enrichment analysis for the CNEs and the TF motifs were performed using the Genomic Regions Enrichment of Annotation Tool (*GREAT*) [162], specifically using the *rGREAT* wrapper package in R [90]. We used *GREAT*'s "twoClosest" option for associating the regions (CNEs or motifs) with the two genes that they were the closest to. Defining the significantly accelerated or decelerated regions as "foreground" regions and the entire pooled regions as "background" regions, *GREAT* used the gene-region association information to perform hypergeometric tests and binomial tests to compute the functional enrichment of gene set annotations. We set a Benjamini-Hochberg false discovery rate (FDR) threshold of 0.05 for both the binomial and hypergeometric tests to identify significantly enriched annotations. All enrichment analysis were performed on the Gene Ontology, the Reactome pathway, and the Cell Markers Augmented (2021) annotations.

4.3.7 Organizing enriched annotations

The correlation between each pair of annotations were computed by tracking the number of overlaps between foreground gene hits that were members of the annotations. The correlation p-value was then computed empirically using thousands of randomly sampled genes matching the numbers of gene hits in each annotation. Correlated enriched annotations were represented in an undirected graph, where significantly correlated annotations were connected by edges.

For larger graphs that were overly dense, clusters of correlated annotations were identified by accounting for the number of shortest paths between each pair of annotations. We use the Relative Forest Accessibility (RFA) index to quantify the "global proximity" of each pair of annotations in the correlation graph that was pre-computed as previously described above. Let L denote the graph Laplacian of the correlation graph. The RFA matrix P of the graph was computed as follows:

$$P = (I + L)^{-1}, \tag{4}$$

where $P \in [0, 1]$. Each entry of matrix P, p_{rc}, is a metric of correlation between annotations r and c as it could be interpreted as the probability that a spanning forest of the correlation graph would include a tree rooted at r that would have a path to c. A probability threshold for the entries in P could then be set to tune the density of the correlation matrix P. The resulting sparse matrix P was used as an adjacency matrix to construct a graph, and clusters of related annotations were identified using community clustering algorithms (e.g., Louvain community clustering).

4.4 Results

4.4.1 Meta-analysis of population genetics studies on high altitude adaptation

Many studies that reported genotype changes associated with high altitude adaptation looked at short evolutionary timescales at the level of populations. As such, we still lack a thorough undertanding of genotype-phenotype mappings that robustly explain species adaptation to high altitude across independent clades. We start this study by performing a meta-analysis of gene hits from collated population genetics studies on high altitude adaptation. We collect gene hits from 23 population genetics studies on high altitude and identify 307 genes that are highlighted in at least three studies. We then performed functional enrichment analysis on these genes to learn the estimated functional categories and pathways that likely experience a gain-of-function in high altitude adaptation.

Figure 4.2 shows the Gene Ontology annotations that are significantly enriched for the 307 genes. Using a graphical clustering approach, we identify that the enriched annotations are generally categorizable to several classes, including phosphodiesterase activity, neuronal functions, thyroid functions, kinase and transferase activity, and two giant clusters containing terms related to immune response, cell migration, development and morphogenesis. Meanwhile, Figure 4.3 shows the significantly enriched Reactome pathway annotations. The enriched pathways contain a lot of kinases, including hypoxia-modulated MET and MAPK, as well as Erythropoietin, a glycoprotein hormone that promotes the production of red blood cells [221]. In the next section, we will use the findings from this meta-analysis as an anchor to evaluate the outcomes of the phylogenetic convergence analysis.

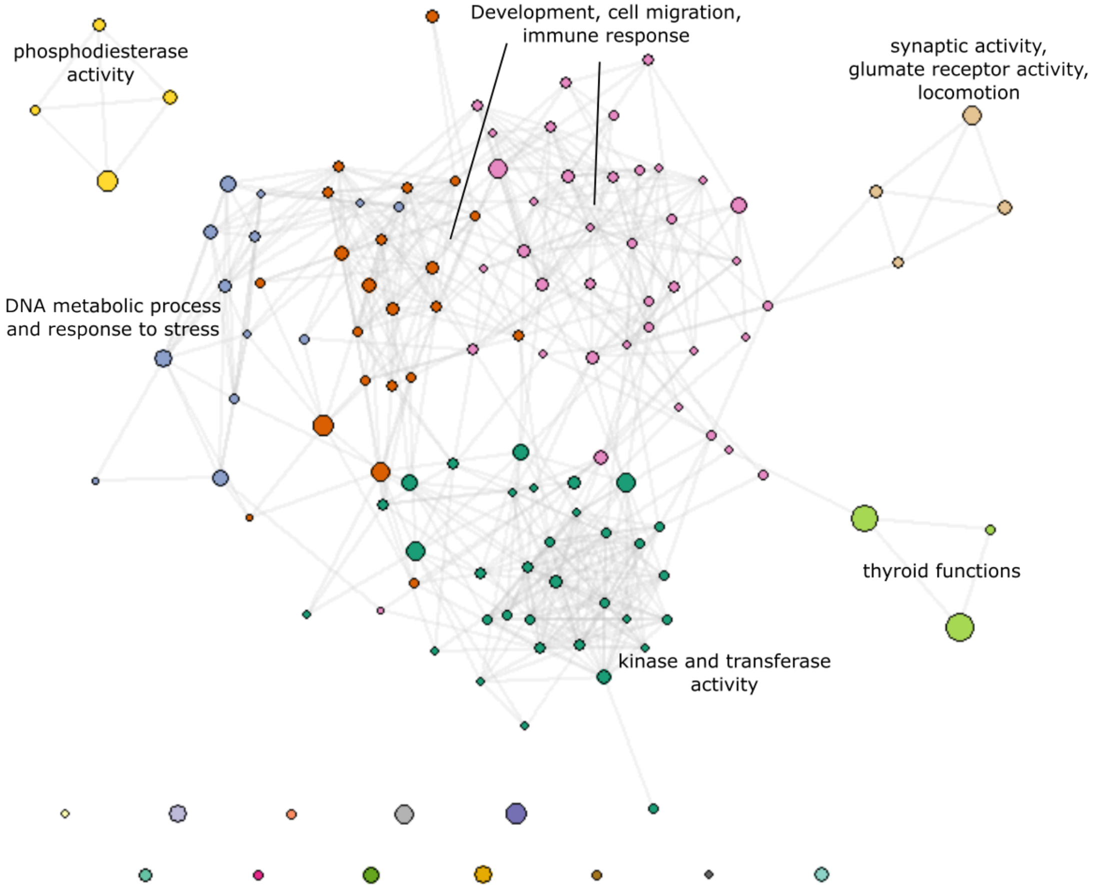

phosphodiesterase activity
Development, cell migration, immune response
synaptic activity, glumate receptor activity, locomotion
DNA metabolic process and response to stress
thyroid functions
kinase and transferase activity

Figure 4.2: Gene Ontology enrichment for population genetics gene hits associated with high altitude.

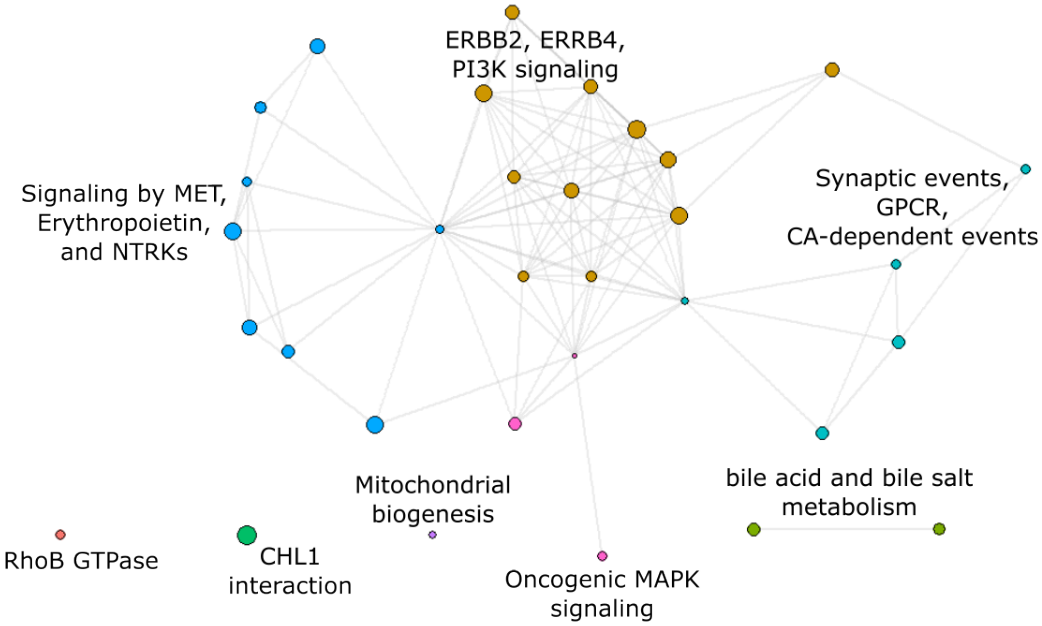


Figure 4.3: Reactome pathway enrichment for population genetics gene hits associated with high altitude.

4.4.2 Correlations among findings from convergence analysis and meta-analysis of population genetics studies on high altitude adaptation

After establishing the functional enrichment from the meta-analysis of high altitude-related population genetics studies, we evaluate the agreement between the population genetics findings and the conclusions from our convergence analyses of proteins and regulatory elements associated with high altitude adaptation. We compute the convergence signals of proteins and non-exonic conserved non-coding elements (CNEs) that are at least 30bp in length. Setting a significance threshold of permulation p-value ≤ 0.05, we identify that high altitude adaptation is associated with 443 significantly decelerated and 586 significantly accelerated proteins out of a total of 19,137 proteins, and 25,041 significantly decelerated and 59,975 significantly accelerated CNEs out of a total of 1,050,080 CNEs. We also evaluate the convergence of regulatory elements at the level of transcription factor (TF) motifs by identifying conserved TF motif coordinates in CNEs and computing their convergence signals. Setting permulation p-value ≤ 0.05, we identify 131,501 significantly decelerated motifs and 95,867 significantly accelerated motifs, out of a total of 2,933,078 conserved motifs.

We perform enrichment analyses on the sets of decelerated and accelerated proteins, CNEs, and motifs using the Gene Ontology, Reactome pathway, and Cell Marker annotations, and evaluate how well the enriched terms from these analysis agree with the findings from the population genetics meta-analysis (Figure 4.4). We find that the enriched terms from population genetics have significant positive correlations with the enriched terms from convergently decelerated and accelerated proteins, CNEs, and motifs (all with Benjamini-Hochberg adjusted Fisher's exact p-values ≤ 0.01). A rate deceleration may suggest that the element evolves under an increased constraint from purifying selection, whereas a rate acceleration may signify positive selection that gives rise to innovative adaptation, or a relaxation of constraint. The agreement between the population genetics findings and the convergent protein signals provides strong evidence that the enriched functions are critical for high altitude adaptation. Interestingly, we observe a much stronger correlations between the population genetics findings and convergence signals at the TF motif-level in both the accelerated and decelerated directions. This observation suggests that the functional adap-

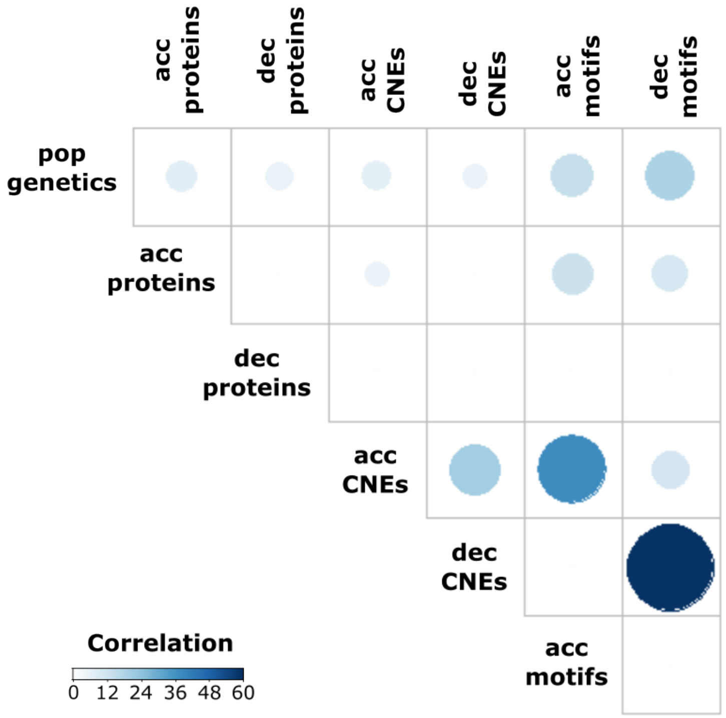


Figure 4.4: Correlations among enriched annotations population genetics hits and convergence analyses on high altitude adaptation. Significantly enriched annotations from each analysis was first identified, and correlations between enriched annotations of each pair of analysis were computed using Fisher's exact test. Plotted dots represent significant correlations with Benjamini-Hochberg adjusted p-values ≤ 0.01. Correlations are measured using Fisher's exact odds ratio.

tations to high altitude also involve regulatory remodeling that is strongly driven by changes in TF binding on regulatory elements.

Figure 4.5 shows the terms that are enriched in the population genetics hits and are also enriched among the convergently *decelerated* proteins, CNEs, or motifs. We observe variations in functions that undergo decelerated evolution in proteins only, in regulatory elements only, or in both proteins and regulatory elements. The deceleration of evolutionary rates in proteins and genomic elements involved in these functions suggest that they evolved under increased selection constraint that disfavors alterations in their sequence. In other words,

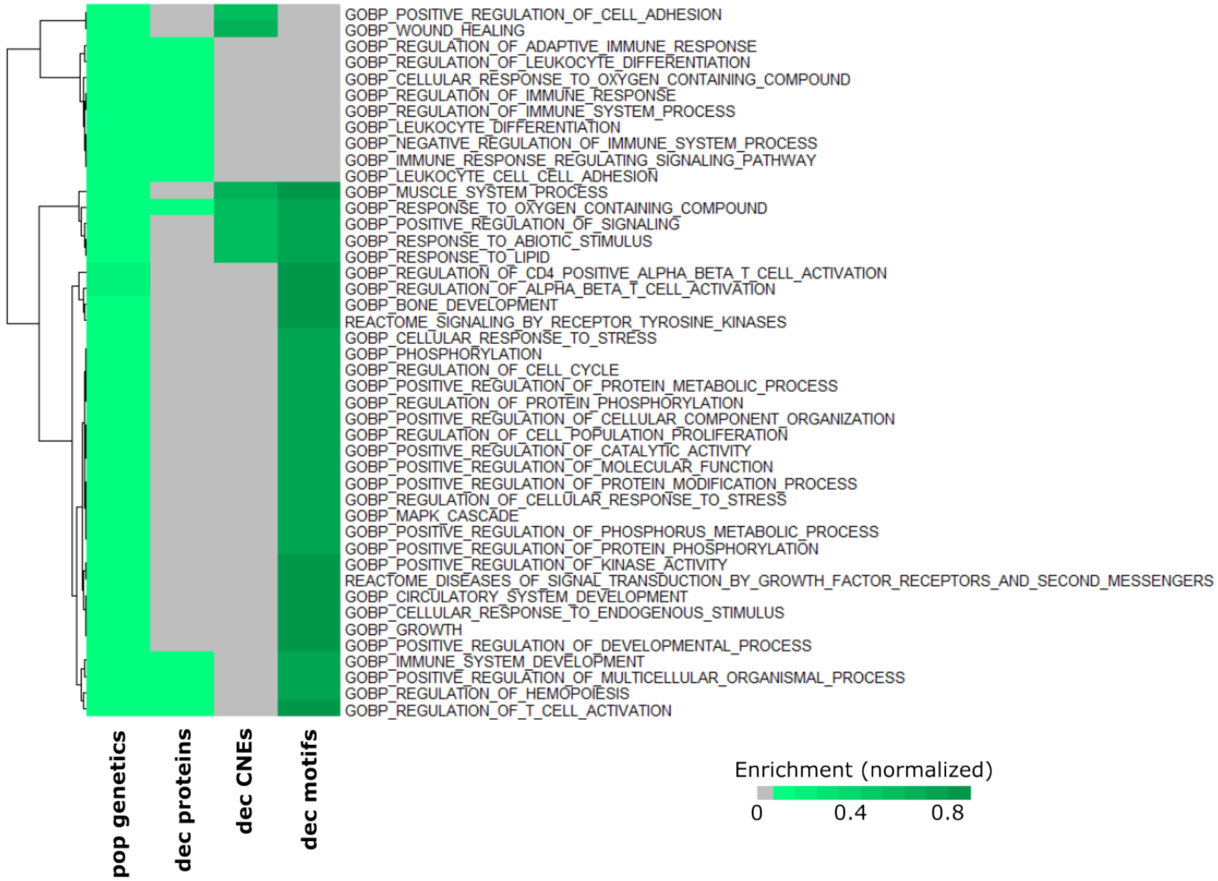


Figure 4.5: Correlations between significantly enriched annotations from population genetics, decelerated proteins, and decelerated regulatory elements. Non-significant terms are shown in grey. Because the enrichment analysis methods are different for different types of analysis, fold enrichment values are normalized using the maximum enrichment value in each analysis.

they likely evolved under purifying selection. Functions that are enriched for decelerated proteins are predominantly related to the immune system, with additional terms including regulation of hemopoiesis, response to oxygen-containing compounds, circulatory system development, and regulation of multicellular organismal process. These are key mechanisms that are critical to fitness and survival at hypoxic conditions. Indeed, changes in the immune system and red blood cell count and physiology have been observed in populations adapting to high altitude [11].

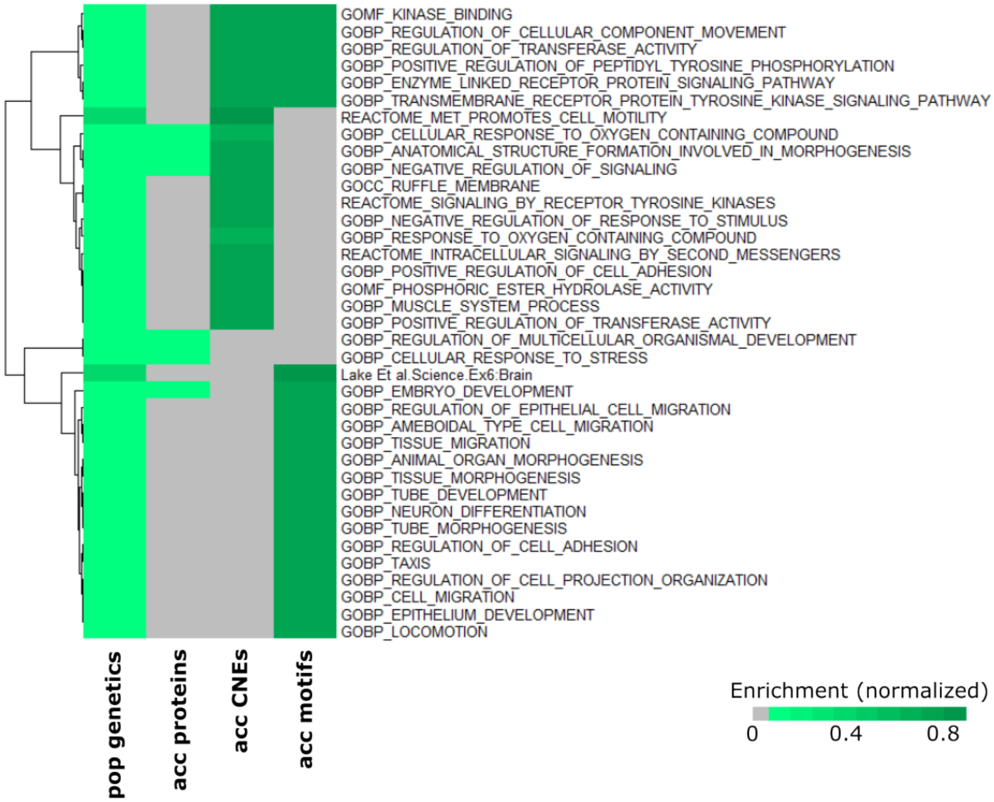


Figure 4.6: Correlations between significantly enriched annotations from population genetics, accelerated proteins, and accelerated regulatory elements. Non-significant terms are shown in grey. Because the enrichment analysis methods are different for different types of analysis, fold enrichment values are normalized using the maximum enrichment value in each analysis.

Meanwhile, there are functions that do not undergo purifying selection in the protein-coding regions, but show deceleration in the regulatory elements. These functions include the MAPK pathway and other kinase activities, growth and development, cell adhesion, and cellular response to stress. This category encompasses general functions that are more pleiotropic in nature. Because of this pleiotropy, the increased selection constraint likely acts not on the protein-coding regions themselves, but on binding sites of TFs that regulate the specific function under selection.

We then evaluate the correlations between the enrichments from population genetics and the *accelerated* proteins and regulatory elements (Figure 4.6). Because the population genetics hits represent single nucleotide polymorphisms that gave rise to a gain-of-function in high altitude adaptation, the convergent acceleration of these set of functions may reflect a positive selection in which a faster evolution occurred in favor of promoting the beneficial allele. Many of the functions in this category are similar to those that evolve under purifying selection of regulatory elements only, including kinase activities, cell adhesion, cellular response to stress, and development. However, this set also includes additional terms related to cell motion and migration, cell projection, neuron differentiation, embryo development, and tissue morphogenesis. A likely commonality among these functions is the critical involvement of the actin cytoskeleton in facilitating their mechanism.

The small set of terms that are enriched for both the population genetics hits and the accelerated proteins are mostly related to embryonic development and morphogenesis. This observation is interesting because chronic hypoxia at high altitude is known to cause preeclapsia and intrauterine growth restriction [169]. It is possible that high altitude species have evolved advantageous variants that are positively selected to boost their reproductive system under hypoxia.

Interestingly, from Figures 4.5 and 4.6, there is only one function that shows significant correlations with the population genetics outcome in both the acceleration and deceleration direction, which is the "Response to oxygen-containing compound" annotation. It is clearly expected that the machinery for oxygen metabolism would be a core mechanism that undergo substantial convergent changes in facilitating adaptation to hypoxia.

4.4.3 Divergence of conserved non-coding elements (CNEs) underlie altitude-associated changes in the renin-angiotensin-aldosterone system

From the convergence analysis at the CNE-level, we find strong enrichment for low p-values in the phenotype association p-value distribution for accelerated CNEs associated with high altitude, whereas that of decelerated CNEs shows no enrichment (Figure 4.7A). This observation suggests that at the entire CNE unit-level, regulatory adaptations to high altitude are mainly driven by CNE divergence. Using the Reactome pathway annotations to evaluate the pathway enrichments of the accelerated CNEs, we find significant enrichment for 6 pathways, 3 of which are related to collagen metabolism (Figure 4.7B). Meanwhile, results from the Gene Ontology (GO) annotations show the most pervasive enrichment for terms related to neuronal functions, vascularization, kidney development, the extracellular matrix (ECM), and receptor tyrosine kinase activity (Figure 4.7C). The observed functional enrichment hints at a regulatory rewiring of the renin-angiotensin-aldosterone axis as an adaptive mechanism to hypobaric hypoxia at high altitude.

Given the prominence of collagen-related annotations in the pathway enrichment analysis results, we investigate the functions that collagen metabolism may facilitate in underlying adaptation to high altitude. We identify CNE-enriched GO annotations that are significantly correlated with the collagen-related Reactome annotations, namely "collagen formation", "collagen chain trimerization", and "collagen biosynthesis and modifying enzymes". Figure 4.8 illustrates the correlations among the Reactome and GO terms, where neighboring nodes that are connected by an edge are significantly correlated (empirical p-value ≤ 0.002). We find that immediate "first neighbors" of the collagen-related pathways are GO terms that are related to the ECM. To identify the systems that are affected by ECM changes in high altitude species, we evaluate the first neighbors of the ECM-related GO terms and find that they are predominantly terms related to vascularization, especially in the kidney.

The importance of collagen-related pathways in high altitude adaptation have previously been highlighted by Qi et al. [196], in which they performed comparative transcriptomics analysis on multiple tissues of yaks living at different altitudes, as well as a lowland control. They found that collagen genes *COL1A2*, *COL3A1*, *COL5A2*, *COL14A1*, and *COL15A1* were differentially expressed in at least 5 (out of 7) different tissues in response to hypoxia at high altitude. They also highlighted the enrichment for ECM- and collagen-related pathway and GO annotations among genes that are positively correlated with high altitude, particularly in lung and heart tissues. Additionally, the role of ECM in vascularization is also well-documented [200]. Besides providing a structural framework for blood vessel walls, ECM can also control the migration, growth, and healing of vascular cells. Meanwhile, the kidney is a highly vascularized structure that also has a large demand for oxygen, making it very fragile to hypoxia [77]. A moderate-term mouse experiment at high altitude

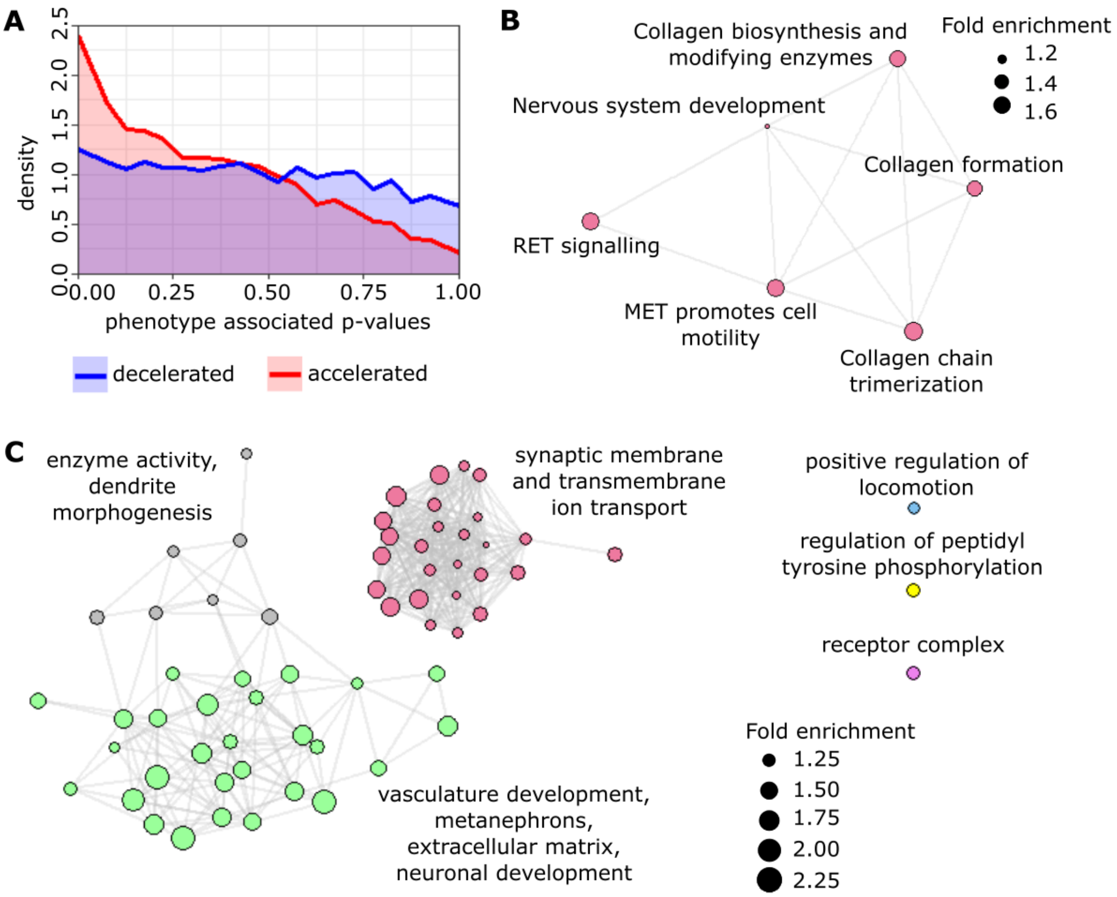


Figure 4.7: Altitude-associated convergent divergence of conserved non-coding elements (CNEs) are enriched for functions related to the renin-angiotensin-aldosterone axis. (A) Empirical p-value distribution for association between CNEs and high altitude phenotype. (B) Reactome pathway annotations that are enriched for CNEs that are significantly diverged in high altitude species, and (C) the same plot for Gene Ontology annotations (Benjamini-Hochberg adjusted p-value ≤ 0.05).

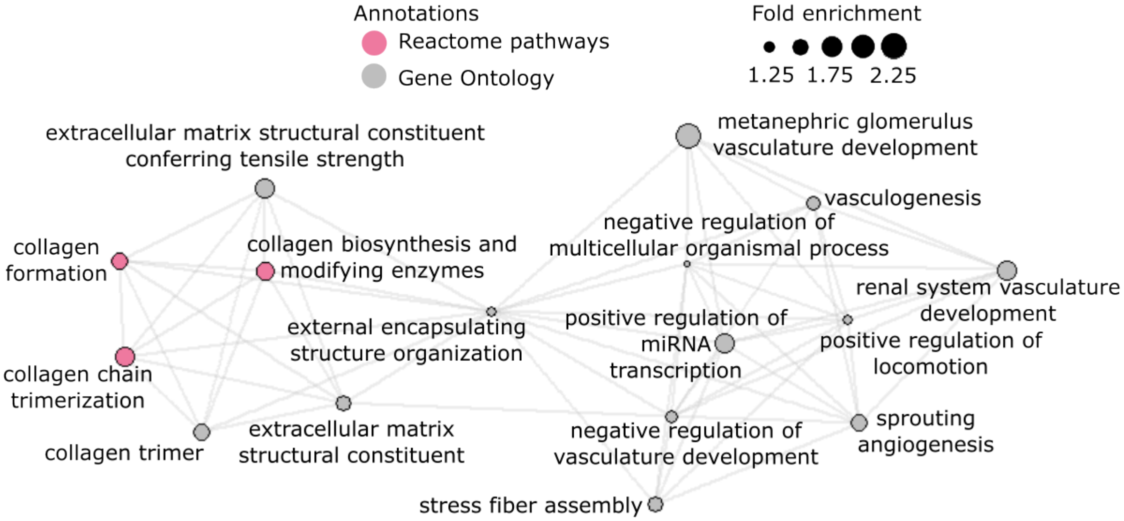

Annotations
Reactome pathways
Gene Ontology
Fold enrichment
1.25
1.75
2.25
extracellular matrix structural constituent conferring tensile strength
collagen formation
collagen biosynthesis and modifying enzymes
collagen chain trimerization
collagen trimer
external encapsulating structure organization
extracellular matrix structural constituent
metanephric glomerulus vasculature development
vasculogenesis
negative regulation of multicellular organismal process
renal system vasculature development
positive regulation of miRNA transcription
positive regulation of locomotion
negative regulation of vasculature development
sprouting angiogenesis
stress fiber assembly

Figure 4.8: Altitude-associated adaptation of collagen metabolism is related to vascularization. Graph shows correlations among the collagen-specific Reactome pathway annotations and Gene Ontology annotations that are all enriched for convergently diverged CNEs in high altitude mammals.

demonstrated that the renin-angiotensin-aldosterone system facilitated a protective mechanism against hypoxic conditions, in which efferent arterioral vasoconstriction was activated to increase glomerular filtration rate [99].

4.4.4 Transcription factor motif-scale analysis highlight the involvement of G protein-coupled receptor signaling in high altitude adaptation

Finally, we perform convergence analysis on individual TF motifs that intersect CNEs (Figure 4.9A). Unlike at the CNE-scale, we find strong enrichments of low p-values for both decelerate and accelerated motifs (Figure 4.9B). We then use the Gene Ontology annotations and the Reactome pathway annotations to find the functions that are enriched for the accelerated and decelerated motifs (Figure 4.9C). We find that the enriched functions in both directions are highly interrelated. For the accelerated motifs, most of the enriched functions are consistent with the CNE-level results, with the addition of the hypoxia-modulated platelet-derived growth factor (PDGF) signaling pathway and the MAPK pathway. The PDGF pathway plays a role in mediating the remodeling of pulmonary vasculatures under hypoxia [76].

Meanwhile, the decelerated motifs are enriched for pathways related to G protein-coupled receptors (GPCR). Different types of GPCRs have indeed been found to be involved in facilitating hypoxia response. For example ligands like β-adrenoreceptor agonists, ET-1, and lysophosphatidic acid can activate GPCRs that will then increase and stabilize HIF-1 activity [31, 108, 139]. Additionally, the hypoxia-induced mitogenic factor (HIMF) facilitates pulmonary circulation vasoconstruction by a mechanism that involves Gαq [69].

4.5 Discussion

In this work, we perform convergence analysis on the protein-coding and regulatory non-coding adaptations that are associated with the convergent evolution of mammalian lineages to high altitude. We compare the conclusions of the convergence analysis with that of a

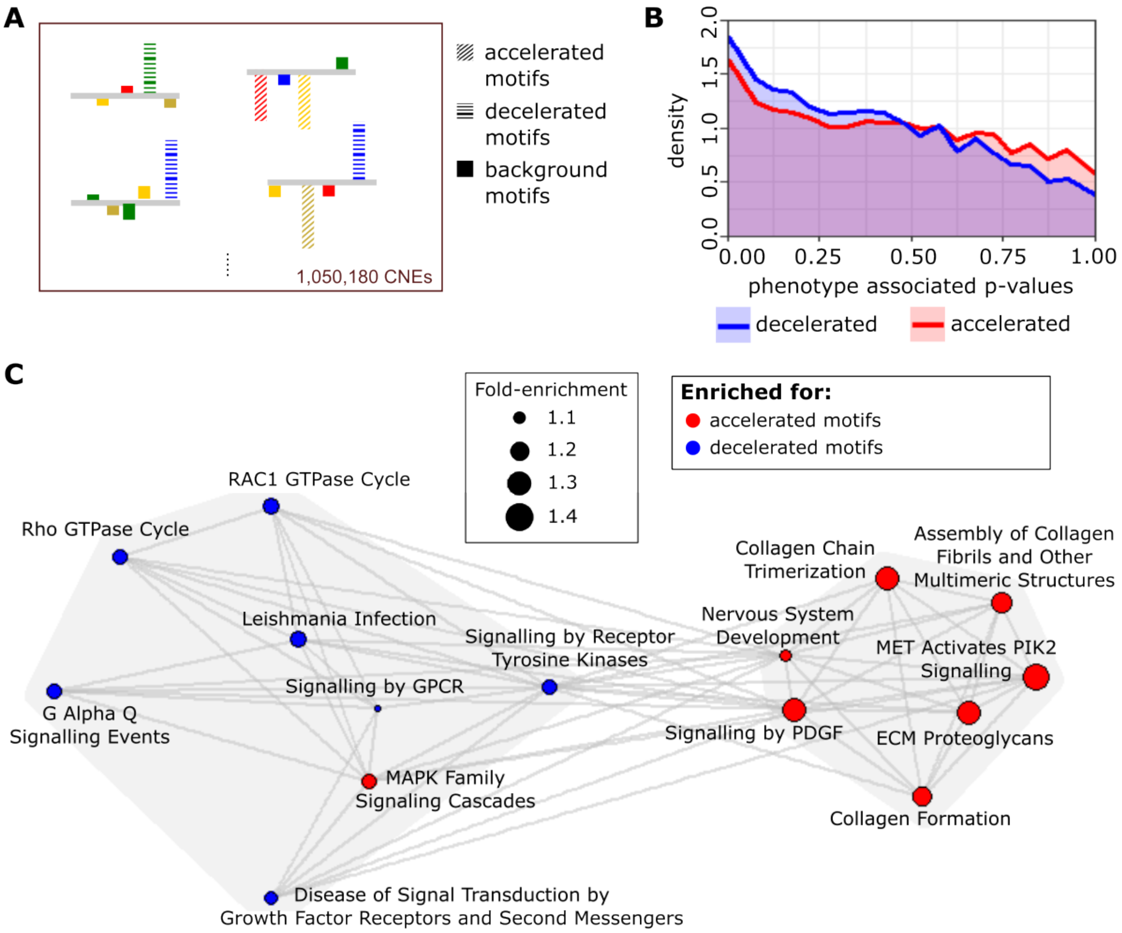


Figure 4.9: Transcription factor (TF) motif changes highlight the involvement of G protein-coupled receptors in high altitude adaptation. (A) Individual TF coordinates in CNEs are scored for convergence. (B) Motif-level phenotype association p-value distributions. (C) Reactome pathway annotations that are enriched for accelerated and decelerated motifs.

meta-analysis on high altitude population genetics studies and find that they are largely in agreement. By anchoring our analysis on the meta-analysis, we identify several functional categories that likely evolve in different modes in response to high altitude stressors.

In the first category, core functions that are absolutely critical for survival at altitude evolve under strong purifying selection that acts on the protein-coding regions. These functions include regulation of oxygen metabolism, hemopoiesis, the immune system, and circulatory system development. In the second category of functions, rate acceleration or deceleration only occur in the non-coding elements, while the protein-coding regions evolve neutrally. This category includes mechanisms that are ubiquitously active in different systems, including kinase activity, growth and development, stress response, and cell adhesion. The third category is characterized only by rate acceleration of non-coding regions. This mechanism possibly signifies a positive selection that acts on sequence motifs to innovate new transcription factor (TF) binding patterns that can engender new important functions. The final category includes functions that are accelerated in the coding region, possibly as a sign of positive selection. This category mainly includes functions related to embryonic development.

In interpreting the results of the meta-analysis, we note that there is an inherent bias that stems from variations in the methodology of study and reporting that was used by different studies. For instance, some studies only reported the top n gene hits, whereas others gave a full accounting for their analysis. Some studies also focused their analysis on certain pathways of interest, such as the HIF pathway. In addition, as the meta-analysis is conducted across species, there may be differences of genes that are present across species, and different gene background sets would have been used as well in each study. As such, conclusions from the meta-analysis should not be taken at face value, and instead should be treated as a rough estimate.

5.0 AFconverge: alignment-free phylogenetic method for predicting convergent evolution of regulatory elements

5.1 Attribution statement

All of the work in this chapter was performed by myself, with the following exception:

- The function for performing motif convolution was developed by Ali Tugrul Balci.

5.2 Introduction

One of the major pursuits of modern biology is to understand how complex phenotypes arise from genetic differences. It is thought that phenotypic diversity stems from differences in gene expression patterns, which are increasingly attributed more to changes in non-coding regulatory elements (REs) than protein sequences. One strategy to characterize REs that underlie a phenotype is to identify REs that evolve in association with the evolution of the phenotype. In particular, comparative genomics algorithms that predict the associations between DNA (or amino acid) sequence evolution and the convergent evolution of phenotypes have become widely used and largely demonstrated success (e.g., [110, 128, 135, 155]).

However, there is a lack of algorithms that are suitable for addressing the evolutionary mechanisms of REs. Many of the existing phylogenetic methods take a “top-down” approach of computing element-level signals that are often computed from multiple sequence alignments of RE orthologs. This dependence on sequence alignment is incongruous with the structural and functional properties of REs. REs are composed of multiple transcription factor (TF) binding sites that work modularly and in combination with one another, creating an exponential number of possible motif combinations with varying levels of cooperativity and redundancy. Additionally, many TFs are pleiotropic, and specific motifs that underlie a phenotype under selection may experience stronger selection pressures than other motifs in a given element. Recent studies have also described that REs such as enhancers are not under strict sequence conservation [249] and could even retain functional homology across deep evolutionary time regardless of extensive variations in motif frequency, composition, and ordering [258]. As such, to understand how selection acts on regulatory machineries, it is necessary to go beyond studying patterns at the aggregate level of an entire RE and examine the correlation patterns of motif selection, a challenge that is still largely unsolved by current approaches.

One existing algorithm that takes an alignment-free approach in evaluating the phenotypic association of REs is Regulatory Element forward genomics (*REforge*) [135]. *REforge* uses a user-defined set of TFs expected to be associated with a phenotype of interest to assign each sequence with a “collective binding” score of the TFs on the sequence, computes the difference of scores between each parent-child node pairs (“branch scores”), and finds the association between these branch scores and trait loss/preservation. Although *REforge* scores individual motifs, by design, *REforge* is unable to provide a TF-level convergence signal. It is also limited to detecting RE divergence, and thus would not be able to predict analogous RE or TF turnovers that occur independently (e.g., [54]). Additionally, *REforge* requires prior knowledge of phenotype-associated TFs, which is difficult to determine for complex phenotypes that are not well-characterized.

Another approach for alignment-free comparative genomics has also been introduced by Kaplow et al. [113] in their TACIT (Tissue-Aware Conservation Inference Toolkit) model. In TACIT, functional genomics datasets are used to train a convolutional neural network (CNN) to predict tissue-specific open chromatin regions (OCRs) from sequences. The model can then be used to predict OCR signals across many species for which functional data is not available, and correlate the signals with phenotype. While CNNs are highly effective for learning complex sequence features, the challenges with TACIT lie with the fact that validating the integrity of cross-species model prediction is difficult in the absence of ground truth data, and that prior knowledge on expected tissues that are affected by the phenotype is needed.

Here, we introduce *AFconverge* (*alignment-free* converge), a “bottom-up”, TF-centric phylogenetic method that predicts the patterns of regulatory motif adaptations underlying phenotypic evolution. Unlike *REforge* and TACIT that use hypothesis-driven approaches, *AFconverge* takes a hypothesis-free approach, with inference of functional signals performed in downstream analyses. We first benchmark our method using the classical case of the convergent loss of vision in mammalian lineages. Then, we apply *AFconverge* to study the promoter adaptations that occur with a less well-characterized phenotype, the evolution of extended lifespan and large body size in mammalian lineages, and demonstrated the flexibility of the application of *AFconverge* for deciphering the complexity of regulatory adaptations at multiple scales. To our knowledge, *AFconverge* is the first algorithm that computes TF motif-level convergence signals in an alignment-free manner.

5.3 Materials and Methods

5.3.1 Introduction to the *AFconverge* framework

Figure 5.1 illustrates the overall schematics of the *AFconverge* workflow. Given a set of *comparable* (but not necessarily alignable) orthologous sequences of a DNA region, *AFconverge* first performs 1D-convolution on one-hot-encoded sequences to scan for the strongest evidence for TF binding sites, using the position weight matrices (PWMs) of a set of motif features as convolutional filters. A "motif score" of a feature in a sequence is defined as the max-pooled value of the convolution output. *AFconverge* also performs motif calling using *PWMScan* [3] to filter out signals from features for which no strong evidence of TF binding is observed in at least n of the orthologs. If the analysis is performed on multiple regions, the output of this motif convolution step would be a sparse, three-way tensor of motif scores, with the number of regions, features, and genomes as the three dimensions.

Then, the "phenotype association score" of each feature in each region is calculated. Specifically, if $\mathbf{X}$ is the motif score tensor, phenotype association of feature m in region r is quantified by firstly taking the tensor fiber $\boldsymbol{x}_{rm}$ along the genome axis, and then correlating its elements with the corresponding convergent phenotype values using Spearman's rank correlation test. Importantly, *AFconverge* employs a phylogeny-aware bias correction strategy called *permulation*, which uses Brownian motion phylogenetic simulations to produce null phenotypes that preserve the phylogenetic dependence and value distribution of the true phenotype [212]. These null phenotypes are then used to compute empirical p-values, defined as the proportion of the null statistics that are as extreme or more extreme than the true phenotype statistic. Finally, the phenotype association score of the feature is defined as the negative logarithm of the empirical p-value, multipled by the sign of the raw correlation statistic (negative for feature loss and positive for feature gain). Because the distribution of null statistics are often asymmetric and non-trivial, we use the conditional p-value calculation strategy proposed by [133] to compute two-sided p-values that are equivalent to the corresponding one-sided test, while maintaining a healthy p-value distribution.

Thus, *AFconverge* takes a motif-centric approach to score the phenotype associations of motif features in a modular way, and can be employed with either binary or quantitative phenotypes. Figure 5.2 depicts how motif-level convergence scores computed by *AFconverge* can be interpreted. In this illustration, two TF motifs (blue and red) in orthologs of a certain enhancer are evaluated for their respective associations with mammalian lifespan as the phenotype of interest. From the illustration, increasing binding affinity of the blue motif is positively correlated with increasing lifespan, which we define as a phenotype-associated "motif gain". In contrast, decreasing binding affinity of the red motif is correlated with increasing lifespan, which we define as a phenotype-associated "motif loss".

In computing motif-level convergence signals, we have to ensure that the statistical significance of a "motif gain" actually detects the convergent appearance of the motif according to a certain significance threshold, instead of merely representing a numeric correlation that is not reflected in the actual appearance of the motif. Likewise, a detected "motif loss" should reflect the convergent disappearance of motifs that were actually present in ancestral species according to a certain significance threshold. *AFconverge* controls this issue by performing motif calling to filter out motif convergence scores that do not pass these criteria. Specifically, suppose we want to evaluate whether there is a convergent gain/loss of motif m in the orthologs of element A. *AFconverge* first uses *PWMscan* to call motif m in each ortholog of A, according to pre-defined p-value threshold (default 10^{-5}). Then, the total number of motif m called across the orthologs of A is counted, denoted by n_m^T. If the phenotype is continuous, the only filter user needs to specify is a minimum threshold for n_m^T to make sure

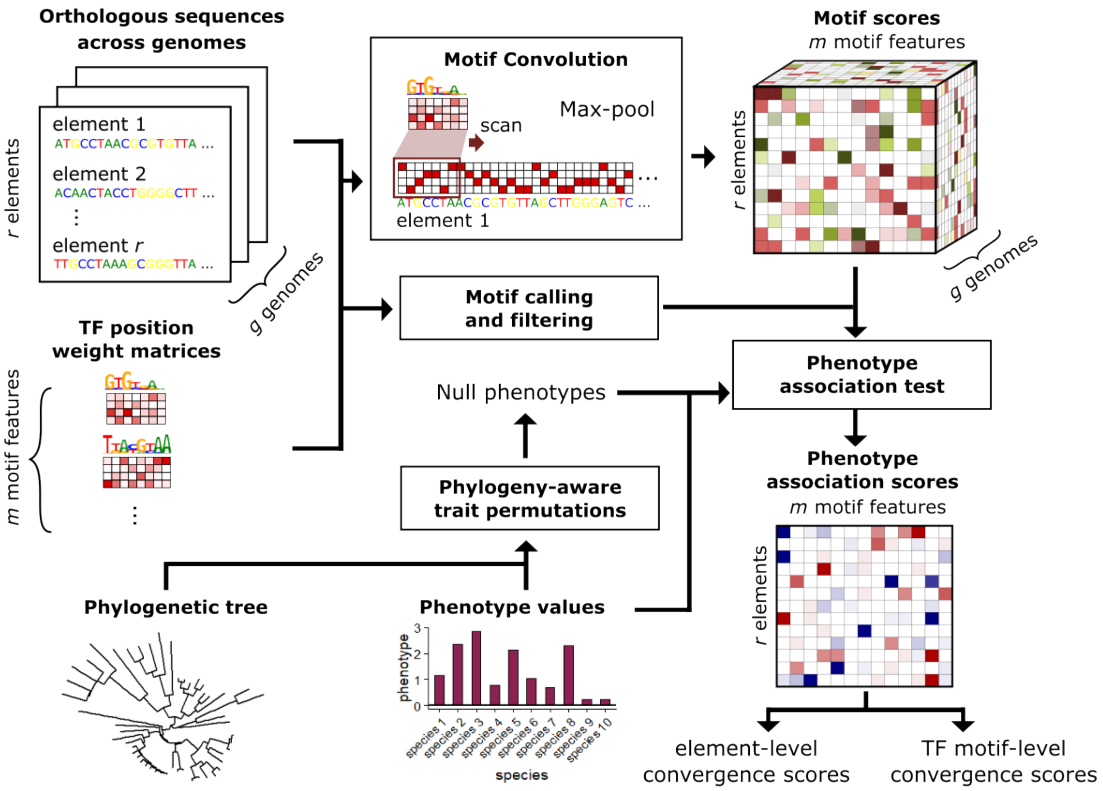


Figure 5.1: Workflow of *AFconverge*. *AFconverge* performs motif calling and 1D-convolution on comparable orthologous sequences across species genomes using transcription factor (TF) position weight matrices as filters, resulting in a three-way tensor of motif scores quantifying TF binding. Correlation between TF binding strength per element and the phenotype values is then computed, using Spearman's rank correlation test that is corrected for biases with a phylogeny-aware trait permutation method.

that the convergence signal of the motif is detected from a sufficient number of species.

If the phenotype is binary, the filtering is conditioned on whether the convergence signal is positive (motif gain) or negative (motif loss). Let N^c denote the total number of orthologs of element A of the convergent species only, and n_m^c denote the number of motif m called among the orthologs of A of the convergent species only. If the motif convergence score for motif m in element A has a positive sign (motif gain), user will need to set minimum threshold for n_m^c in addition to n_m^T (e.g., setting $n_m^c \geq 2$ means that a convergent motif gain requires that a statistically significant appearance of the motif is observed in the orthologs of at least 2 convergent species). If the motif convergence score is negative (motif loss), user will need to set a minimum threshold of N^c in addition to n_m^T (e.g., setting $N^c \geq 2$ means that at least two of the orthologs of the convergent species must exist, to avoid detecting a "motif loss" due to the absence of the entire orthologs in the convergent species). Motif convergence scores that do not pass these conditions for binary or continuous phenotypes are filtered out. For the analysis in this work, we set $n_m^c \geq 2$, $N^c \geq 2$, and $n_m^T \geq 10$.

5.3.2 Constructing a dataset of reference-free promoter orthologs from the Zoonomia mammalian phylogeny

We constructed a dataset of orthologous DNA sequences for 19,565 promoters from the 241-way reference-free mammalian alignment produced by the Zoonomia Consortium [81]. Specifically, we defined "reference-free promoter orthologs" as windows that were similarly proximal to the transcription start sites (TSS) of genes across genomes, obtained by extending a window of ±250bp from the TSS. The first step in constructing the promoter dataset was to identify "anchor" positions of each gene promoter that can be "lifted over" across species genomes with high fidelity. For the human *hg38* assembly, the anchors were defined as the TSS of the genes. Then, we defined a ±50bp window around each TSS in the human assembly, and then lifted the window over to other assemblies using *halLiftover* [102]. We also lifted over the TSS position coordinate specifically. Using the lifted-over coordinates of the TSS and the 50bp window, we reconstructed the orthologs of the 101bp *hg38*-window in all other species using *HALPER* [274] and defined the successfully transferred orthologs as

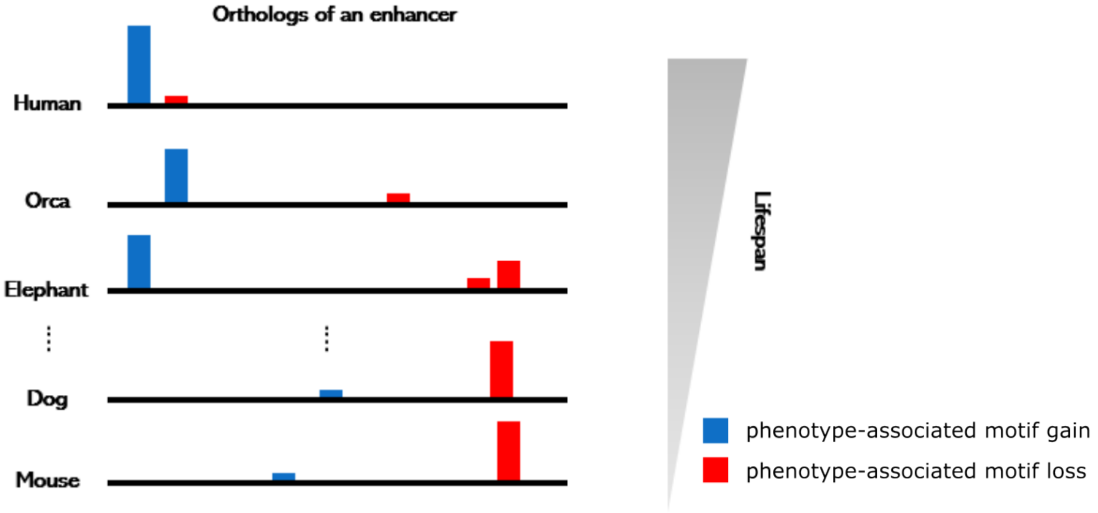

Orthologs of an enhancer
Human
Orca
Elephant
Dog
Mouse
Lifespan
phenotype-associated motif gain
phenotype-associated motif loss

Figure 5.2: Interpretation of motif-level phenotype association scores computed by *AFconverge*. Each bar represents the binding affinity score of a motif. Gain of the blue TF motif is positively associated with the phenotype (e.g., lifespan), whereas loss of the red motif is negatively associated.

regions that were contiguous and within 50bp-1000bp long, with a 25bp "protection" buffer around the summit in either direction. The orthologous anchors were determined to be the summit coordinates of the successfully transferred orthologous windows. For promoters whose 101bp window were successfully lifted over but whose TSS was not, the orthologous anchor coordinates were determined to be the lifted-over coordinate whose original coordinate in the *hg38* assembly was the closest to the TSS in *hg38*. All other promoters that did not pass these two criteria were discarded. Finally, promoter regions were obtained by extending 250bp upstream and downstream from the anchors. We specifically used the NCBI RefSeq transcript dataset for *hg38*, from which we identified 56,698 sets of promoter orthologs.

5.3.3 Motif dataset for convolutional filters

As convolutional filters, we used a repository of 693 non-redundant TF motif archetypes in the human genome that were clustered from >4,000 motifs [248]. Specifically, we used version 2.0-beta, which can be found on the following website: `https://resources.altius.org/~jvierstra/projects/motif-clustering-v2.0beta/`. Additionally, we also computed the GC ratio of the sequences, as well as CG and GC patterns (the average-pooled value from convolving the sequence with "CG" and "GC" patterns, respectively).

5.3.4 Identification of outlier species

To identify outliers species in the dataset, we first took ~5000 randomly selected sets of promoter orthologs, and performed 1D-convolution on these orthologs using the PWMs of the TF motifs. This step produced a genome-by-motif matrix of motif scores (or TF binding scores) for each set of promoter orthologs. Then, we removed phylogenetic bias in each matrix by de-correlating the motif scores with a standard statistical whitening transformation, using the covariance matrix computed from the neutral phylogenetic tree model. Outliers in each promoter's matrix were identified by computing the Mahalanobis distance of each species from the center of the multivariate distribution in the principal component space of the de-correlated matrix, setting a threshold of Benjamini-Hochberg χ^2 FDR ≤ 0.01. Pooling all the ∼5000 promoters together, the final set of outlier species were defined as the species that were assigned as outliers in at least 10% of the promoters. This process assigned 23 species as outliers that were then removed from the analysis.

5.3.5 Phenotypes for evaluation

As there are many non-trivial factors that introduce biases to real sequences, it is not plausible to simulate sequences that could reliably represent the ground truth. Thus, to benchmark our method, we used the classical example of convergent vision loss in mammals (Figure 5.3A). The set of extant "foreground" species (i.e., species with the convergent phenotype) in the Zoonomia dataset include several species of moles and mole rats (*Heterocephalus glaber*, *Fukomys damarensis*, *Ellobius talpinus*, *Ellobius lutescens*, *Nannospalax galili*, *Sorex araneus*, *Condylura cristata*, *Scalopus aquaticus*, *Chrysochloris asiatica*), echolocating bats (*Noctilio leporinus*, *Myotis davidii*), and *Rhinolophus sinicus.* To enable permulation, we simplified the large echolocating bat clade by pruning the remaining species other than *Noctilio leporinus* and *Myotis davidii.* In addition to removing 23 outlier species, this results in 200 total species. The ancestors of *Heterocephalus glaber* and *Fukomys damarensis*, *Ellobius talpinus* and *Ellobius lutescens*, and *Noctilio leporinus* and *Myotis davidii* were included as foregrounds for permulation.

Subsequently, we applied *AFconverge* to investigate the regulatory adaptations underlying extended lifespan in mammals. Our group had previously defined the 'long-lived, large-bodied' (3L) phenotype as the evolution of long lifespan that correlates positively with body size, quantified by taking the first principal component between the log-transformed maximum lifespan and the log-transformed adult weight of the species [129] (Figure 5.3B). We took a subset of 167 species for which the phenotype annotations are available in the Animal Aging and Longevity Database (AnAge) [243] and removed the outliers, resulting in 144 total species.

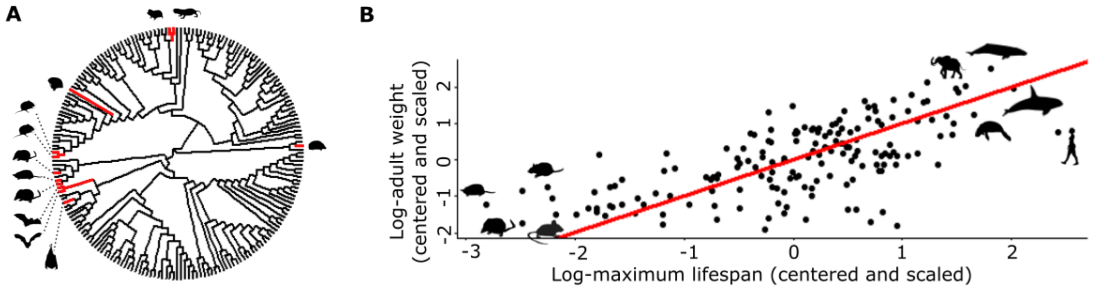


Figure 5.3: Phenotypes for evaluation. (A) We benchmark *AFconverge* with the convergent case of vision loss among independent mammalian lineages. In a 200-way mammalian phylogeny, 12 extant species (red branches) have independently lost their visual structures, including several species of bats, moles, mole rats, and the shrew. (B) We apply *AFconverge* to analyze the convergent evolution of mammalian longevity and increased body size, quantified as the first principal component between the log-transformed maximum lifespans and the log-transformed body sizes of the mammals.

5.3.6 Alternative methods for benchmarking analysis

We compared the performance of *AFconverge* against two alternative methods: *RERconverge*, an alignment-based algorithm [128, 188], and *REforge*, an alignment-free algorithm [135]. *RERconverge* measures the associations between convergent phenotypes and convergent shifts in relative evolutionary rates (RERs). RERs are quantified by computing the residuals from regressing the length of each branch in an element-specific tree against the average length of corresponding branch genome-wide, followed by correction for heteroscedasticity. The RERs are then correlated with binary phenotypes using Kendall's τ test. Meanwhile, *REforge* uses a set of pre-defined motifs for TFs that are expected to be associated with the phenotype of interest. The motifs are used to quantify binding scores on each sequence, from which a "collective binding" score is computed for the sequence. Finally, changes in collective binding scores per branch are computed and correlated with trait loss or preservation per branch.

5.3.7 Data preparation for alternative methods used in benchmarking analysis

To prepare the input dataset for *RERconverge*, we first used the MUSCLE (MUltiple Sequence Comparison by Log-Expecation) tool [60] to create a multiple sequence alignment (MSA) for each set of promoter orthologs. Afterwards, the promoter MSAs were used to infer promoter-specific evolutionary trees using *phangorn* [218], using the "General Time Reversible" nucleotide substitution model. Finally, the *readTrees* function in *RERconverge* was used to store the promoter-specific trees into a *multiPhylo* R object and compute the "average" tree, with branch lengths that were averaged from all the promoter-specific trees.

To prepare the input datasets for *REforge*, we used *PRANK* [146] to perform phylogenetic reconstruction of ancestral sequences for each set of promoter orthologs and estimate promoter-specific evolutionary trees. As motif priors for the scoring with *REforge*, we used the list of eye-related transcription factor motifs used by Langer et al. [135] in the original *REforge* publication. Out of the 28 motifs that Langer et al. used, we identified 25 motifs which we could find the motifs for in the JASPAR, Uniprobe, and cisBP databases (Table 5.1).

5.3.8 Detecting element-level phenotype-associated divergence

To measure convergent divergence of an entire element unit from *AFconverge*'s motif-level scores, we used Wilcoxon rank-sum test to detect whether there was a significant deviation in the number of convergent motif losses in the element, relative to the null distribution of motif-level convergence scores. Specifically, we defined the vector of motif-level convergence scores of the element of interest as the test group, and the pooled motif-level convergence scores of the remaining elements as the control group for Wilcoxon rank-sum test. Negative values of the test statistic signified convergent divergence of the element.

5.3.9 Functional enrichment analysis

Functional enrichment analysis was conducted using Fisher's exact test, corrected for biases from variations in gene set sizes using permutation tests. Specifically, the null distri-

Table 5.1: Ocular transcription factor motifs used as priors for *REforge* in benchmarking experiments

Motif name	**Label**	**Source**	**Motif name**	**Label**	Source
Dmbx1	MA0883.1	JASPAR	Pax2	MA0067.2	JASPAR
Emx1	MA0612.2	JASPAR	Pax5	MA0014.3	JASPAR
Esx1	MA0644.2	JASPAR	Pax6	MA0069.1	JASPAR
Hes1	MA1099.2	JASPAR	Six1	MA1118.1	JASPAR
Nrl	MA0842.2	JASPAR	Mitf	MA0620.3	JASPAR
Prox1	MA0794.1	JASPAR	Mitf	MA1899.1	JASPAR
Sox21	MA0866.1	JASPAR	Sox1	MA0870.1	JASPAR
Vsx1	MA0725.1	JASPAR	Bhlhb2	Bhlhb2_1274_015681.bml	Uniprobe
Nanog	NANOG+M6357_1.02+D	cisBP	Lhx2	MA0700.2	JASPAR
Crx	MA0467.2	JASPAR	Pitx2	MA1547.2	JASPAR
Gbx2	MA0890.1	JASPAR	Hsf1	MA0486.2	JASPAR
Hoxa6	MA1497.1	JASPAR	Hsf1	MA0319.1	JASPAR
Hoxc5	Hoxc5_2639.2	Uniprobe			

bution of Fisher's exact test odds ratio was obtained by performing the test on 500 match-sized sets of randomly sampled elements. The empirical p-values were then computed by calculating the fraction of the null odds ratio values that were greater than or equal to the odds ratio of the true phenotype. Functional annotations that were used for analysis in this study include the CellMarker Augmented (2021) annotations [273], the Reactome pathway annotations, and the Gene Ontology annotations as stated.

5.3.10 Computing enrichment of gained and lost motif features

We first set a Type I error threshold of 0.01 to identify gained and lost motif features that were strongly correlated with the convergent phenotype. Then, to compute the enrichment of feature A in the set of gained (or lost) features, we used Fisher's exact test to calculate the probability of identifying n number of feature A in the set of gained (or lost) features, given the total number of feature A. To correct for statistical biases due to variations in the total number of calls across features, we performed permutation tests to construct the null distribution of odds ratio by randomly selecting null sets of motif calls, where each set contained the same number of calls as the total number of calls for feature A. The empirical p-value of the enrichment for feature A was then computed as the fraction of null odds ratios that were greater than or equal to the true odds ratio of feature A. Correction for multiple testing was finally performed using Benjamini-Hochberg FDR to adjust the empirical p-values, and fold-enrichment was quantified using the true odds ratio.

5.3.11 Learning latent correlations in motif selection

We used Empirical Bayes Matrix Factorization (EBMF) [254] to decompose the phenotype association matrix into latent variables (LVs). EBMF estimates the optimal number of LVs and amount of sparsity in the learned representation by learning prior distributions from the data directly. To encourage interpretability, we constrained the prior distributions of the LV loadings to follow a mixed non-negative uniform distribution. EBMF was performed with the *flashr* package in R, using a greedy forward and backward fitting algorithm. The top-ranking promoters represented by each LV were determined by reconstructing the distribution of loading values from the fit model, and identifying the promoters with loading values above the 95th percentile of the distribution. The feature gains and losses represented by each LV were obtained by reconstructing the distribution of factor values from the fit model, and then identifying the features whose factor values lay outside of the 95% confidence interval of the distribution.

5.4 Results

5.4.1 *AFconverge* outperforms competing methods in predicting convergent divergence of ocular-related promoters in blind mammals

We first benchmarked *AFconverge* against two competing methods – *REforge*, which was an alignment-free comparative method, and *RERconverge*, which was alignment-based – using the convergence case of loss of vision in mammalian lineages as the benchmarking phenotype (Figure 5.3A). Given the degradation of ocular structures in these species, we expect that promoters related to ocular functions would be convergently diverged because they would be evolving under decreased selection constraints. Because the two alternative methods were only able to compute convergence scores at an entire element-level, we used *AFconverge*'s motif-level convergence scores to compute element-level convergence scores, specifically by using Wilcoxon rank-sum test to identify promoters that have convergently lost a significantly large number of motif features due to relaxation of constraint. Figure 5.4A shows the phenotype association p-value distributions produced by the three methods. From the p-value distributions, it is evident that *AFconverge* was able to produce a strong enrichment of low p-values, meaning that it was able to identify promoters that were convergently diverged in association with the loss of visual structures in the foreground species. Meanwhile, both *REforge* and *RERconverge* produced substantially weaker enrichment of low p-values compared to *AFconverge*.

We then evaluated whether the top-ranking promoters identified by the three methods as convergently diverged in blind mammals were indeed associated with eye-specific functional

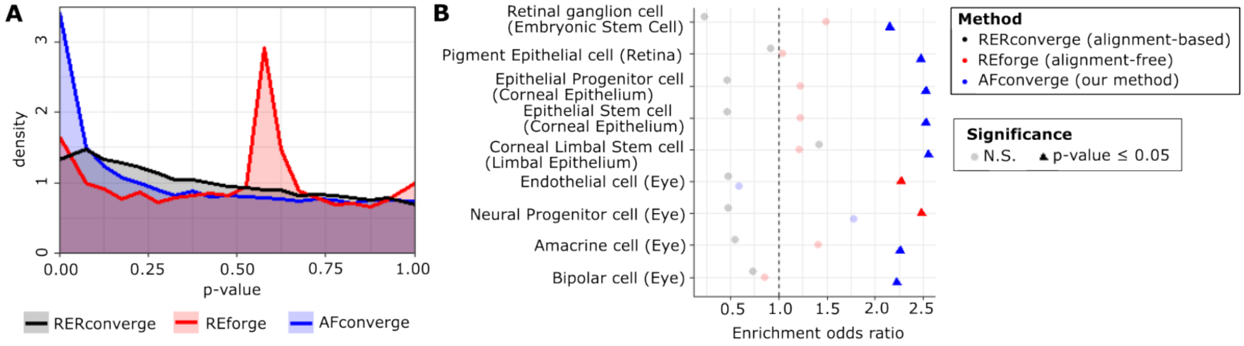


Figure 5.4: Benchmarking analysis on the convergent loss of vision in mammals. (A) Histograms showing the distributions of phenotype association p-values computed by *AFconverge*, *RERconverge* (alignment-based), and *REforge* (alignment-free). (B) Enrichment analysis showing the correlation between top-ranking convergently diverged promoters predicted by the three methods with ocular-specific terms in the CellMarker Augmented (2021) annotations.

annotations. To make a fair comparison across the methods, we first selected approximately equal numbers of top-ranking promoters with the strongest signals for convergent divergence, specifically 1,636 promoters for *AFconverge* (permulation p-value ≤ 0.01), 1,618 promoters for *REforge* (Pearson's p-value ≤ 0.068), and 1,625 promoters for *RERconverge* (Kendall's p-value ≤ 0.046). Then, we evaluated whether the three sets of promoters were enriched for cell type-specific markers of ocular tissues curated by the CellMarker Augmented (2021) annotation [273]. Setting a threshold for enrichment p-value ≤ 0.05 for all three analysis, we found that the promoters identified by *AFconverge* as having the greatest convergent alterations in their global TF profile were indeed significantly enriched for markers of multiple eye tissues (Figure 5.4B). Moreover, the functional enrichment produced by *AFconverge* was substantially larger than that of alternative approaches. *REforge*, which was also an alignment-free method, was able to pick up enrichment for two annotations. Interestingly, although *AFconverge*'s convergence analysis was conducted in a hypothesis-free manner, it was able to outperform *REforge*'s hypothesis-driven analysis, in which a pre-defined set of known eye-related TFs were specified to compute eye-specific promoter divergence. Meanwhile, the alignment-based method, *RERconverge*, failed to identify enrichment for any eye-related terms in this analysis.

5.4.2 *AFconverge* highlights that convergently diverged promoters in blind mammals are most enriched for neuronal and ocular functions

After benchmarking on eye-specific annotations, we evaluated the enrichment for all CellMarker Augmented annotations among the promoters predicted as significantly diverged by *AFconverge*. Out of 1,079 annotations, the top-ranking convergently diverged promoters were significantly enriched for 78 annotations (empirical p-value ≤ 0.05) (Figure 5.5). Notably, a large proportion of these hits were associated with neuronal or ocular tissues ($\sim$36% of the top-ranking hits shown in Figure 5.5). In fact, most of the neuronal hits heavily occupy the top of the list when ranked according to fold enrichment. This is consistent with the fact that vision loss in blind species is often accompanied by the remodeling or degradation of neuronal cell types. In subterranean mammals, the superior colliculus and lateral geniculate nucleus, which are the components of the midbrain that receives optical signals from the eye, are degenerated compared to mice [44, 45]. Additionally, the cerebellum of naked mole rats have been reported to undergo a remodeling in which the region for the somatosensory system for processing tactile cues for navigation is expanded, while the region for visual system is degraded [157].

We note that the 78 significantly enriched hits reflect developmental processes across different systems. This is consistent with previous reports that morphological convergence can arise from molecular convergence of development regulatory mechanisms that can involve pleiotropic TFs [259]. When we accounted for significant overlaps among annotations, we indeed found that the enriched annotations were significantly interrelated (Figure 5.6). In the graph representation in Figure 5.6, the densest region represent stem and precursor cells in different developmental tissues.

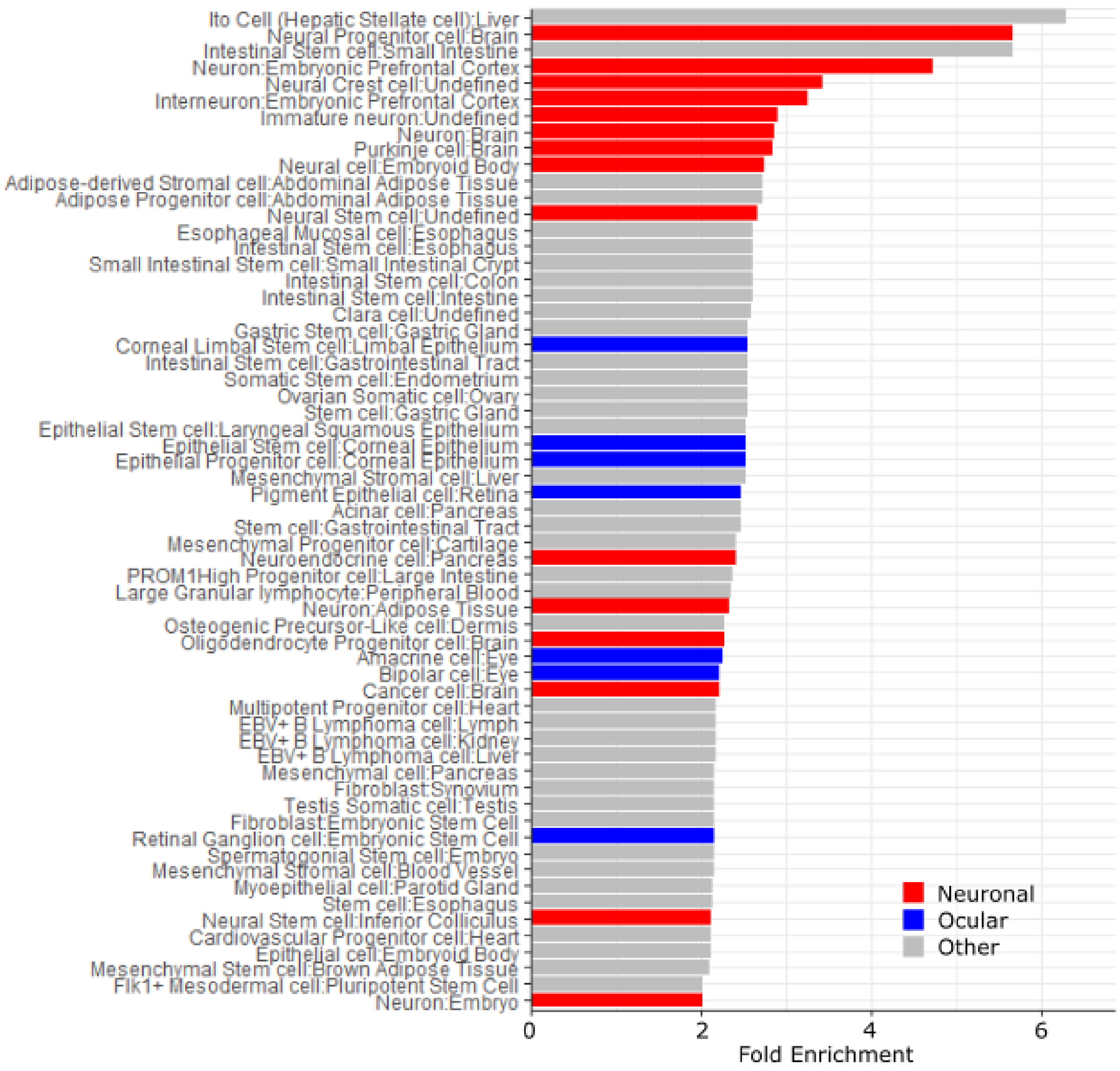


Figure 5.5: Top-ranking CellMarker Augmented (2021) annotations that are significantly enriched (p-value ≤ 0.05) for promoters that are predicted by *AFconverge* to be significantly diverged in blind mammals. The plotted annotations are the ones with fold enrichment ≥ 2.

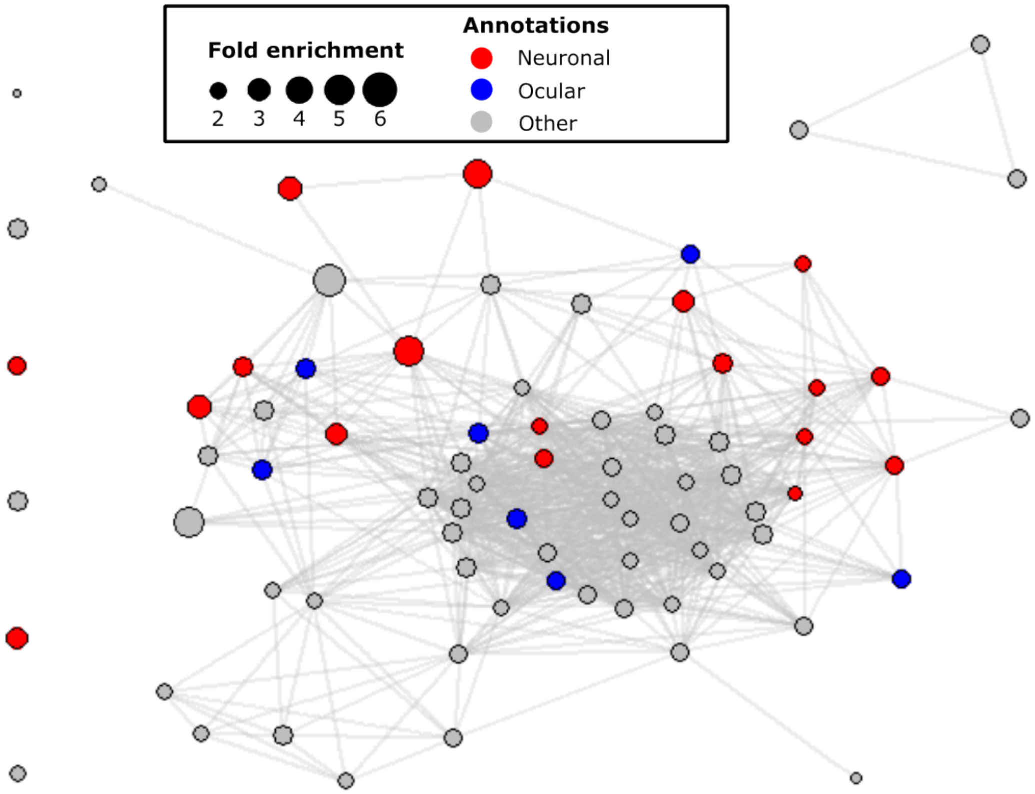


Figure 5.6: Enriched cell marker annotations represent significantly interrelated developmental tissues. Significantly correlated annotations are connected with edges (empirical p-value ≤ 0.001).

5.4.3 *AFconverge* predicts global convergent losses of transcription factor motifs relevant to ocular phenotype

Switching from a promoter-level to a *motif*-level perspective, we ask whether *AFconverge*'s motif-level convergence signals could highlight specific TFs whose binding sites underwent significant convergent changes across all promoters due to selection. Setting a permulation p-value threshold of ≤ 0.01, we first identified 119,330 "promoter-motifs" that were convergently lost across promoters, out of 1,712,466 total scored promoter-motifs. We found that this set of lost promoter-motifs were significantly enriched for the 34 motif features with Benjamini-Hochberg FDR ≤ 0.01, out of the total of 696 features evaluated (Figure 5.7). Notably, among this set of features that experienced widespread convergent losses were a number of known regulators of eye development, neuronal development, and circadian rhythmicity, including binding sites for PAX6, a master regulator of eye development [89]; MECP2, a neuronal TF that caused a deterioration in visual acuity when knocked out in mice [272], and circadian rhythm regulators CLOCK, HEY, HES1, and MAX. Convergent losses of binding motifs for circadian regulators suggest a remodeling of circadian control machinery, which has indeed been reported in subterranean mammals relative to mouse [10, 82]. We also note the loss of three unique motifs for ZBTB14, suggesting its importance in the phenotype. ZBTB14 has been found to be a biomarker for vitreous seeding retinoblastoma [78] and plays a role in retinal differentiation [27]. Meanwhile, there were 33,031 convergently gained promoter-motifs, but they were not significantly enriched for any individual motif feature.

5.4.4 *AFconverge* identifies widespread gains and losses of motif features associated with the evolution of longevity

After benchmarking the method, we then used *AFconverge* to investigate the regulatory adaptations underlying the convergent evolution of a complex phenotype that is less well-understood, which is the evolution of extended lifespan in mammals. Our group previously distinguished two extended lifespan traits – the 'long-lived, large-bodied' (3L) phenotype describes the evolution of long lifespan that correlates positively with body size, while the

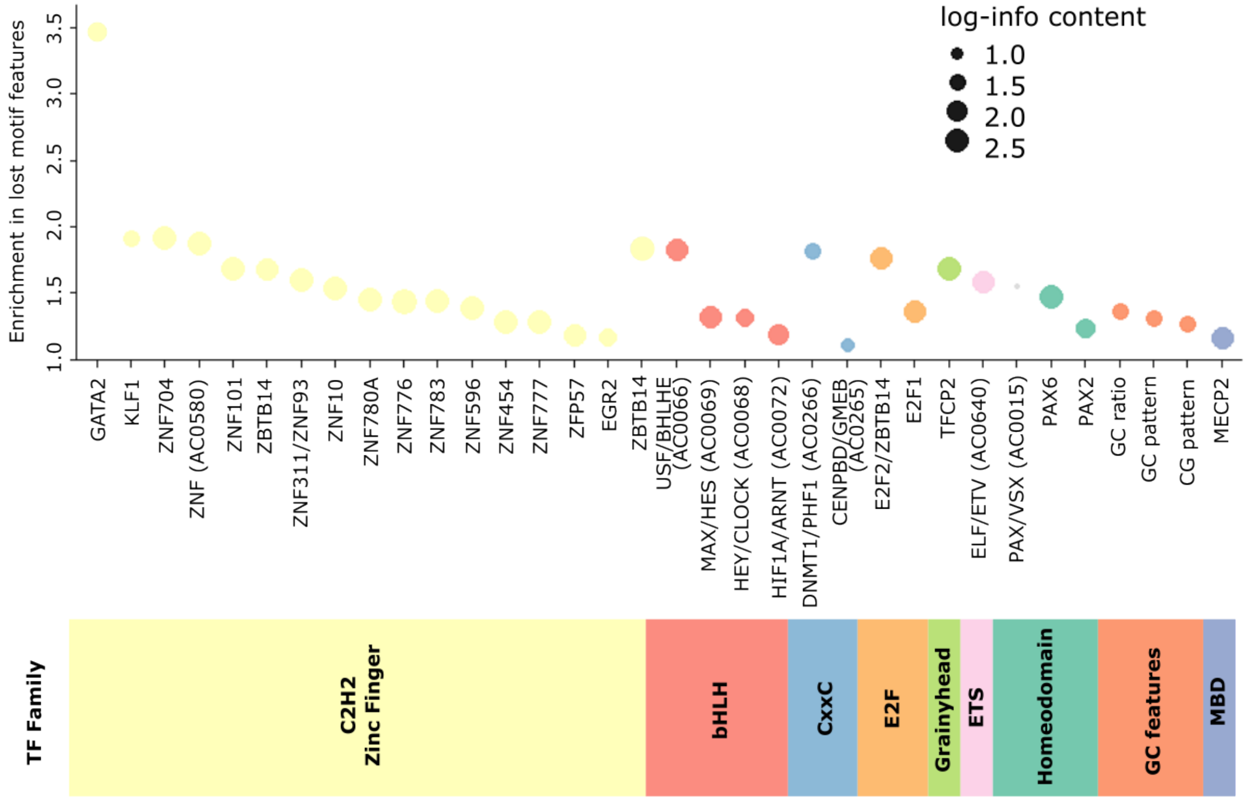


Figure 5.7: Motif features that are globally lost across promoters in association with convergent vision loss.

'exceptionally long-lived given body size' (ELL) phenotype describes extended longevity that is corrected for body size [129]. This work focuses on the 3L phenotype (Figure 5.3B).

After computing the association scores for all promoter-motif features, we observed a strong enrichment for motif gains, and a weaker signal for motif losses (Figure 5.8A). Setting a permulation p-value threshold of ≤ 0.01, we identified 31,683 gained features and 13,488 lost features that were strongly correlated with the evolution of the 3L phenotype, out of a total of 2,762,582 promoter-motifs (1.15% and 0.49% of the total, respectively). We then evaluated whether there were significant widespread gains of motifs across promoters in association with the evolution of the 3L phenotype. We indeed discovered significant enrichment of 103 motif features in the set of 3L-associated motif gains, with Benjamini-Hochberg FDR ≤ 0.01 (Figure 5.8B). Many of the features that were enriched in the set of motif gains were binding sites for TFs that have been documented for their involvement in regulating longevity, or in mechanisms known to be associated with longevity. Importantly, AFconverge identified widespread gains of motifs for Forkhead O (FOXO) TFs (motif cluster AC0036), which were known to be master regulators of longevity and modulate lifespan via the insulin/insulin-like growth factor pathway [23]. We also identified widespread gains of motifs for tumor suppressors SMAD4 (AC0597) and PRDM4 (AC0287). This observation supports the hypothesis that a tighter cancer control machinery is the underlying explanation for Peto's paradox, in which increasing body size does not correlate with increased cancer risk [35]. Indeed, the expansion of SMAD4 activity has been observed in long-living turtles relative to other vertebrates [198].

Another interesting finding was the identification of significant widespread gains of motifs for pluripotency regulators, including SOX2 (AC0659), NANOG (AC0636), and ZSCAN4 (AC0530). This observation is consistent with recent findings from comparative transcriptomics that reported that genes whose expression positively correlated with maximum lifespan in mammals were controlled by pluripotency regulators, including SOX2 and NANOG [147]. Finally, we also identified motifs for TFs that are involved in regulating mechanisms related to longevity, including regulation of T-cell activity and development (NFAT TFs), innate immune system (IRF3), and cell cycle (MYBL1, MYBL2).

Switching gears to motif losses, there were 6 motif features that were significantly en-

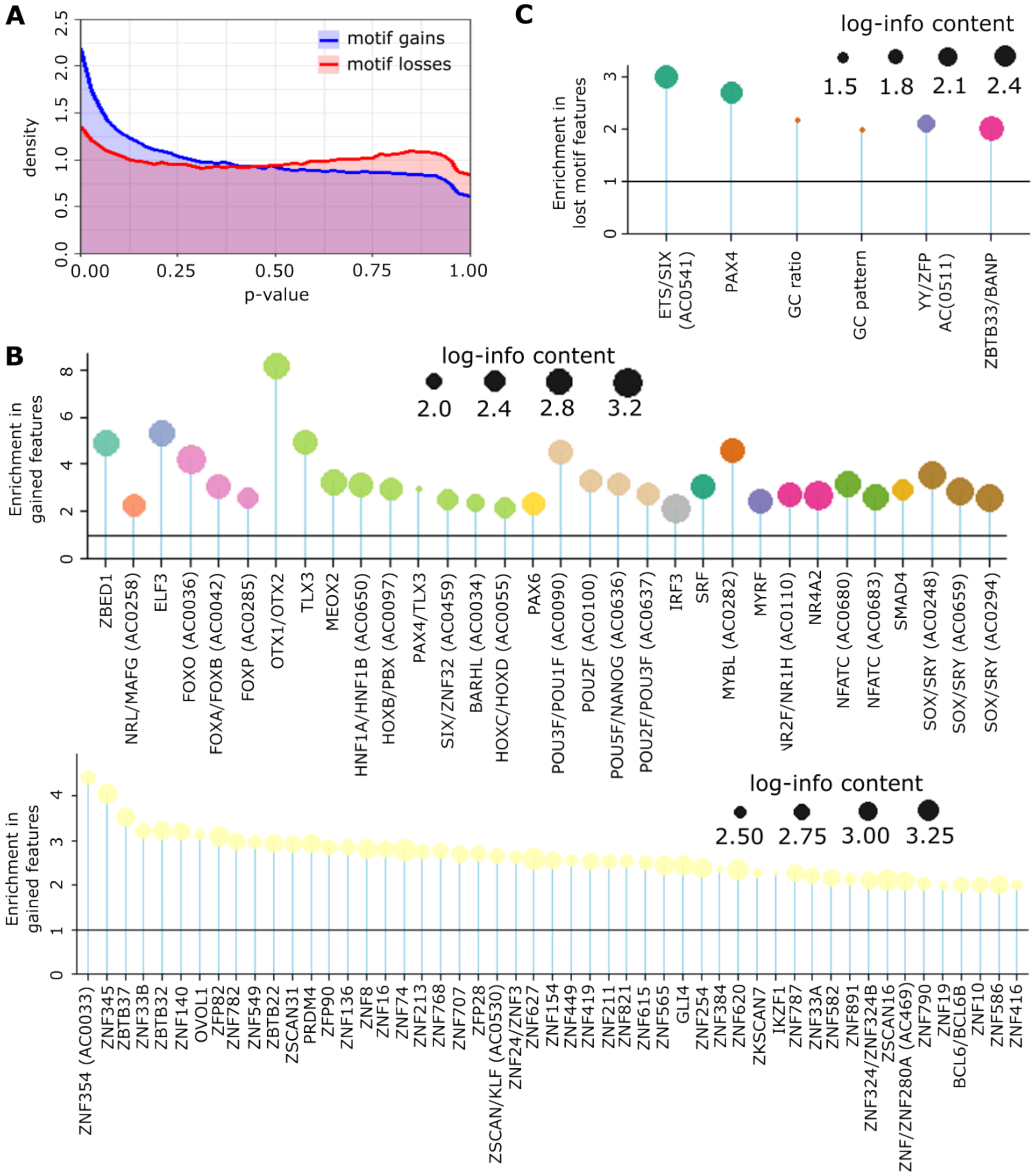


Figure 5.8: Motif features with global convergence signals in association with longevity. (A) Motif-level phenotype association p-value distributions. (B) Features with global longevity-associated gains across promoters (plotted hits have fold enrichment ≥ 2). (C) Features with global longevity-associated losses across promoters.

riched in the set of 3L-associated motif losses globally (Benjamini-Hochberg FDR ≤ 0.1) (Figure 5.8C). Among these hits were motifs for TFs whose aberration has been previously characterized to increase mammalian lifespan, including SIX5 and ETS1. The introduction of SIX4 and SIX5 knockout alleles have been found to prolong lifespan and enhance the regeneration of skeletal muscles in mouse models of Duchenne Muscular Dystrophy [265]. Meanwhile, ETS1 has been suggested to control the down-regulation of the ribosome pathway in long-living humans, and ETS1 knockdown causes the reduction of cellular senescence in embryonic lung fibroblast and human dermal fibroblast cells [263].

5.4.5 Correlations in motif adaptation highlight immunity, germline development, and cancer control as core mechanisms underlying longevity

Finally, we hypothesized that the evolution of the 3L phenotype was driven by a handful of core mechanisms that were facilitated by specific sets of TFs. We therefore asked whether the convergent shift in these core mechanisms could be observed in, and therefore inferred from, the correlation patterns of motif convergence signals genome-wide. To learn the correlation patterns, we used Empirical Based Matrix Factorization (EBMF) to decompose the phenotype association matrix into loading and factor matrices, resulting in 200 latent variables (LVs) with good reconstruction quality (Pearson's R = 0.85, p-value $< 2.2e-16$) (Figure 5.9A). Out of the 200 LVs, we focused on 2 pairs of strongly anticorrelated LVs that captured a major proportion of variance explained (Figure 5.9B). LV1 and LV2, which were significantly anticorrelated (Pearson's R -0.79, pval $< 2.2e-16$) (Figure 5.9C), captured 11.8 and 12.3 times, respectively, the proportion of variance explained by each LV on average. Meanwhile, LV3 and LV4, which were significantly anticorrelated (Pearson's R -0.70, pval $< 2.2e-16$) (Figure 5.9D), captured 4.5 and 4.8 times, respectively, the proportion of variance explained by each LV on average (note that the proportions of variance explained computed by EBMF do not add up to 100% as the LVs are not orthogonal).

Although LV1 and LV2 represented distinct sets of promoters, their strongly anti-correlated latent factor values mean that gains and losses of many motifs in promoters represented by the two LVs could be interpreted to occur in opposite ways. Because we set a non-negative

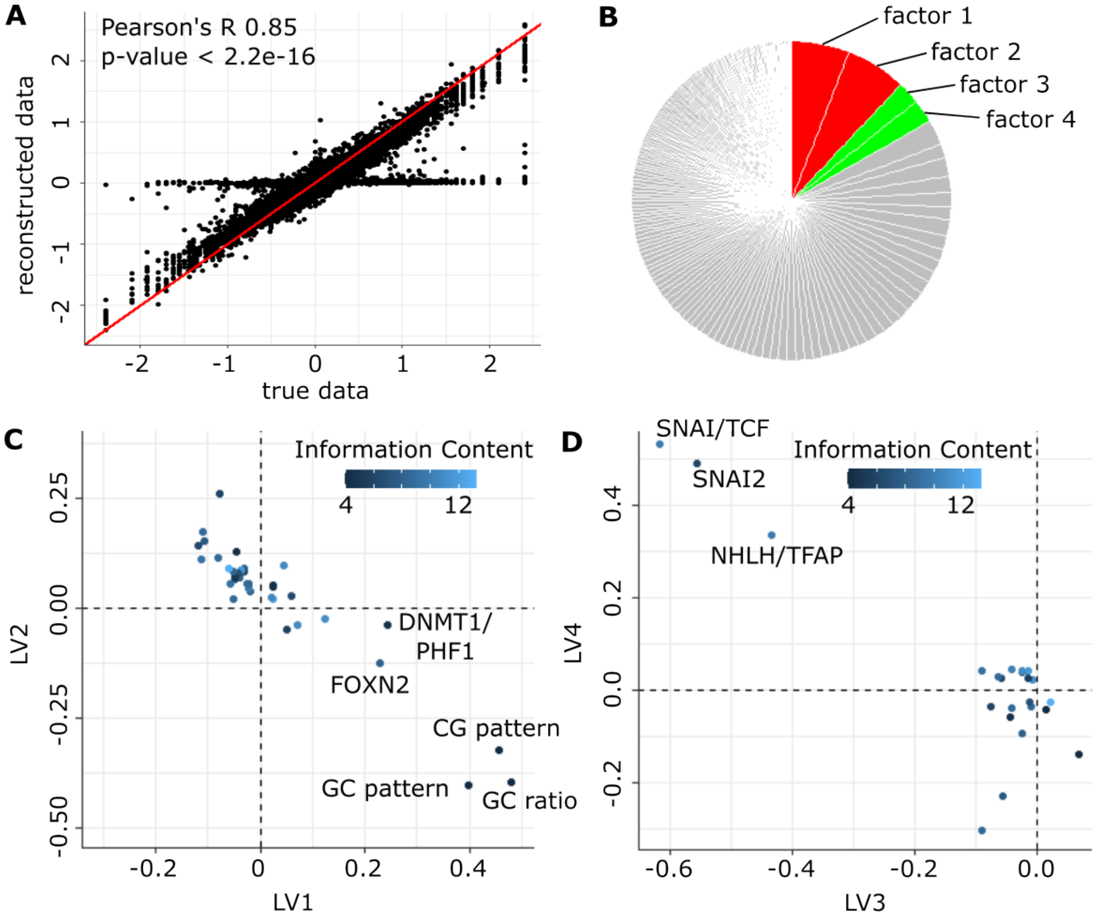


Figure 5.9: Latent factorization highlights major drivers of longevity. (A) Reconstructed data significantly correlates with true data. (B) Latent variables (LV) 1-4 capture a substantial proportion of variance explained. (C) Motif features commonly represented by LV1 and LV2. (D) Motif features commonly represented by LV3 and LV4.

constraint on the loading matrix, positive (negative) factor values could be interpreted as feature gains (losses). Looking at the set of represented features that were shared by LV1 and LV2 (Figure 5.9C), it is evident that the strongest drivers of these LVs were features related to GC content. Specifically, promoters represented by LV1 experienced convergent GC gains, whereas promoters represented by LV2 experienced GC losses. Many studies have indeed linked CpG density in promoters with lifespan and ageing, particularly because CpG sites are targets of methylation with which epigenetic regulation is facilitated [158, 161]. These studies have outlined that CpG densities in specific sets of promoters were predictive of lifespan in vertebrates.

We then performed functional enrichment analysis on the combined set of 3,874 promoters represented by LV1 and LV2, using Cell Marker annotations and Reactome pathway annotations. We found that the significantly enriched Cell Marker annotations (Type I error ≤ 0.01) include terms related to adaptive immunity, germline development, placental development, and intestinal tissues (Figure 5.10). The enrichment for adaptive immunity supports the popular hypothesis that adaptive immunity co-evolved with longevity to equip species with the ability to fight off pathogens over a long lifespan [180]. Immunity and longevity also share many common regulatory mechanisms that need to be tightly controlled to balance the benefits of adaptive immunity against its metabolic costs in trade-off with sustaining longevity [182]. Meanwhile, many studies have documented a tight coupling between germline development and longevity regulation in multiple species, the mechanisms of which are still poorly understood [7]. For example, in *C. elegans* and *Drosophila*, the loss of germline cells or germline stem cells promoted longevity [8, 74], but transplanting adult mice with young ovaries prolonged lifespan, postulated to be the result of unknown "life enhancing factors" produced by mammalian gonads [32].

Figure 5.10 also illustrates significant correlations between the enriched Cell Marker and Reactome pathway annotations. We observe that the Cell Marker annotation for "retinoic acid signaling-responsive fetal germ cell (Fetal Gonad)" is significantly associated with the Reactome pathway annotations "MAP2K & MAPK activation" and "interleukin-37 signaling". It is known that the MAPK pathway has a conserved role in regulating the development of male and female germ cells [47, 149], but the mechanistic role of interleukin-37 (IL-37) in

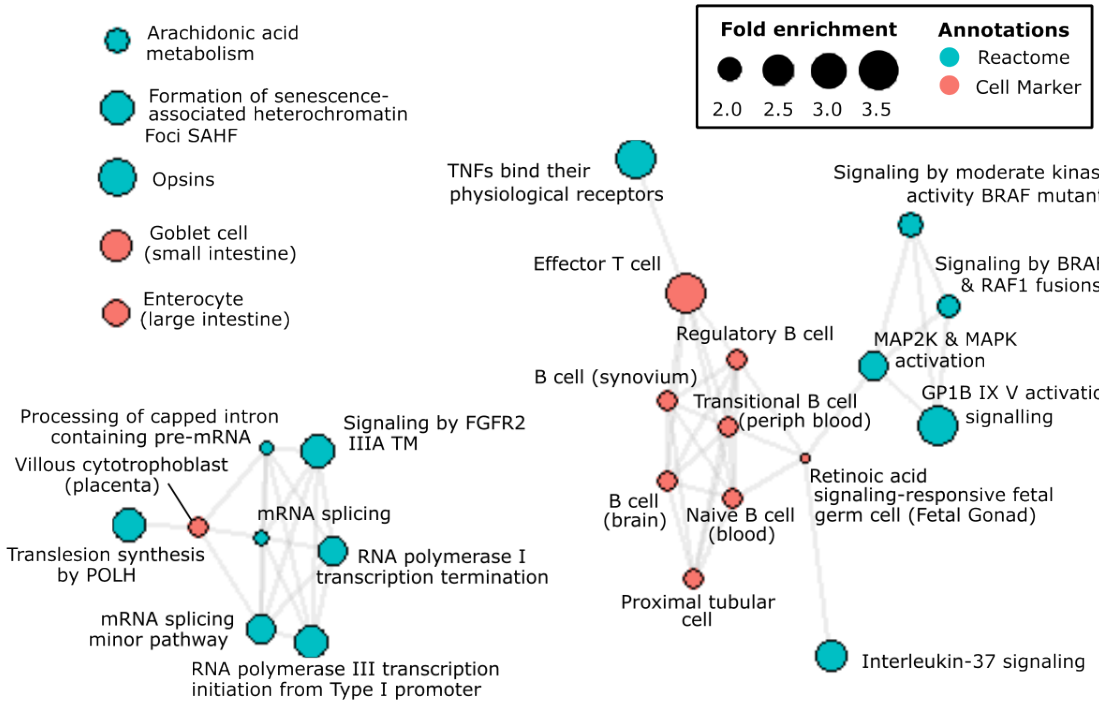

Arachidonic acid metabolism
Formation of senescence-associated heterochromatin Foci SAHF
Opsins
Goblet cell (small intestine)
Enterocyte (large intestine)
Fold enrichment
2.0
2.5
3.0
3.5
Annotations
Reactome
Cell Marker
TNFs bind their physiological receptors
Signaling by moderate kinase activity BRAF mutants
Effector T cell
Signaling by BRAF & RAF1 fusions
Regulatory B cell
MAP2K & MAPK activation
B cell (synovium)
Transitional B cell (periph blood)
GP1B IX V activation signalling
Processing of capped intron containing pre-mRNA
Signaling by FGFR2 IIIA TM
Villous cytotrophoblast (placenta)
mRNA splicing
Retinoic acid signaling-responsive fetal germ cell (Fetal Gonad)
B cell (brain)
Naive B cell (blood)
Translesion synthesis by POLH
RNA polymerase I transcription termination
Proximal tubular cell
mRNA splicing minor pathway
Interleukin-37 signaling
RNA polymerase III transcription initiation from Type I promoter

Figure 5.10: Cell Marker and Reactome pathway annotations enriched for promoters represented by Latent Variables 1 and 2 (empirical p-value ≤ 0.01). Significantly correlated terms are connected by edges.

the reproductive system is less well-documented. IL-37 is an anti-inflammatory member of the commonly pro-inflammatory IL-1 family, and different isoforms of IL-37 are expressed in different tissues including the uterus. IL-37 has been found to have anti-inflammatory and anti-cancer effects in endometrial and cervical cancers [114, 183, 252], and it also suppressed inflammation mediated by polycystic ovary syndrome (PCOS) [70]. Interestingly, the anti-inflammatory activity of human IL-37 expressed in mice has also been documented to give rise to anti-aging effects, protecting against metabolic diseases [15], colitis [163], hepatitis [29], and neuronal injury [4]. Additionally, in a human population genetics study, the ratio between IL-37 and pro-inflammatory markers was positively correlated with indicators of healthspan [28]. Referring back to the connection between the reproductive system and longevity, our findings suggest the possibility that the pro-longevity adaptation occurring in the germline is mediated by changes in interleukin-37 signaling pathway.

Switching to LV3 and LV4, Figure 5.9D shows that the strongest drivers of the latent variables were motif clusters that corresponded to several regulators of mesenchymal stem cells (MSCs), including SNAI1, SNAI2, ZEB1, ASCL1, and TCF4. Many of these TFs interact with longevity regulator FOXO3 in their mechanism of action in both cooperative and inhibitory ways [91, 104, 255]. Looking at the functional enrichment of the 3,636 promoters represented by LV3 and LV4 (Figure 5.11), the significantly enriched Cell Marker annotations were highly specific to immune cells, including T cells, dendritic cells, and Natural Killer (NK) cells. The relationship between MSCs and immune cells is often discussed in the context of the role of MSCs in the tumor microenvironment. MSCs are attractive as a cancer therapy strategy because they can have anti-cancer effects, partially through the modulation of the immune system [206]. For example, T cells have immunomodulatory effects on the anti-inflammatory activity of MSCs, specifically by releasing interferon γ cytokines [206]. MSCs have also been found to modulate the activity of innate immune cell types in cancer, including dendritic cells and NK cells [83, 148, 271]. Interestingly, some of the enriched Reactome pathway annotations that were associated with the immune-related Cell Marker terms were related to purinergic signaling (Figure 5.11). Consistent with immune modulation of cancer control, purines such as adenosine triphosphates (ATP) and adenosines (ADO) are present in large amounts in the tumor microenvironment [51]. The concentrations of ATP

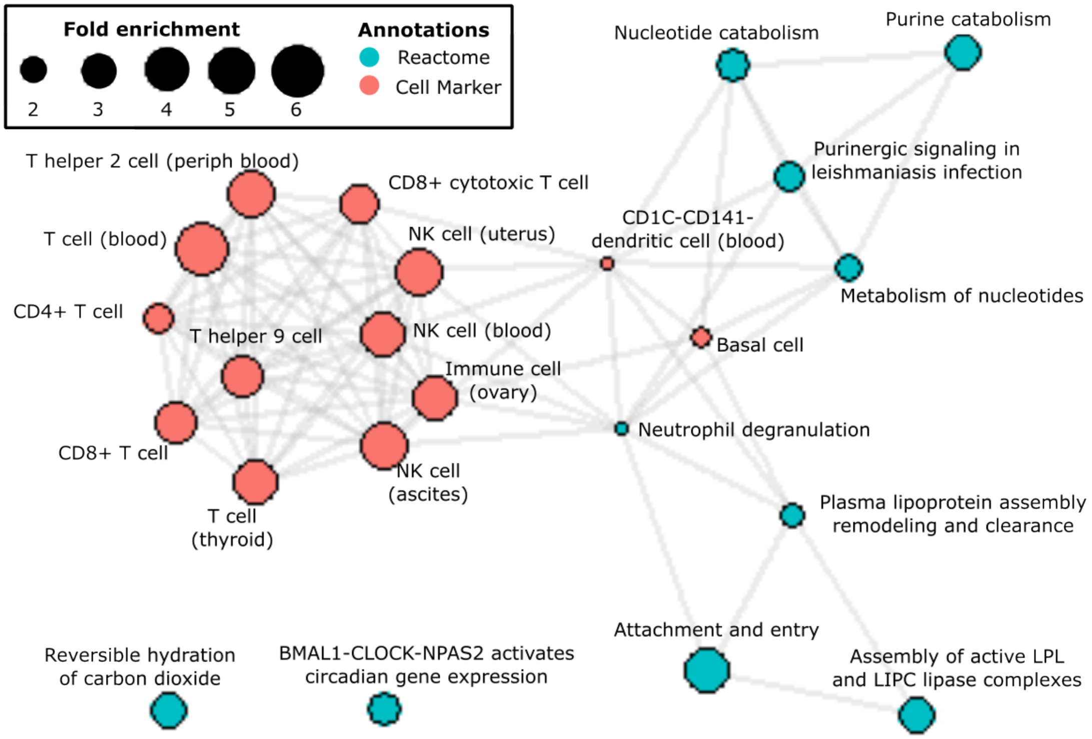

Fold enrichment
2
3
4
5
6
Annotations
Reactome
Cell Marker
Nucleotide catabolism
Purine catabolism
Purinergic signaling in leishmaniasis infection
CD1C-CD141- dendritic cell (blood)
Metabolism of nucleotides
T helper 2 cell (periph blood)
CD8+ cytotoxic T cell
T cell (blood)
NK cell (uterus)
CD4+ T cell
T helper 9 cell
NK cell (blood)
Basal cell
Immune cell (ovary)
Neutrophil degranulation
CD8+ T cell
NK cell (ascites)
T cell (thyroid)
Plasma lipoprotein assembly remodeling and clearance
Attachment and entry
Reversible hydration of carbon dioxide
BMAL1-CLOCK-NPAS2 activates circadian gene expression
Assembly of active LPL and LIPC lipase complexes

Figure 5.11: Cell Marker and Reactome pathway annotations enriched for promoters represented by Latent Variables 3 and 4 (empirical p-value ≤ 0.01). Significantly correlated terms are connected by edges.

and ADO in the tumor microenvironment control the extent of anti-tumor versus pro-cancer immune response, while in turn, the endonucleotidases released by immune cells (among others) modulate the levels of ATP and ADO [50]. These findings suggest the involvement of the immune-mediated cancer control machinery in giving rise to the longevity phenotype.

5.5 Discussion

In this work, we introduce *AFconverge*, a novel alignment-free comparative genomics method that predicts convergent regulatory adaptations associated with phenotypic convergence. *AFconverge* performs motif calling and convolution to measure the binding affinity of TFs on orthologous sequences of regulatory elements (REs), and computes the correlation between these motif scores and phenotype values. We benchmarked our method with a well-characterized convergent phenotype, vision loss in mammals, and demonstrated that *AFconverge* outperformed competing alignment-free and alignment-based approaches at correctly predicting divergence of promoters that were associated with ocular functions. Besides detecting element-level signals, *AFconverge* was also able to predict widespread convergent losses of TF motifs involved in eye development, neuronal development, and circadian rhythm, consistent with expectations that the ocular degradation in blind mammals are often accompanied by extensive rewiring in neuronal functions and circadian rhythm. We then applied our method to elucidate patterns of regulatory adaptations underlying the evolution of mammalian longevity, highlighting pluripotency regulation, cancer control, germline development, and immunity as key axes of extended lifespan.

AFconverge is a new addition to the set of comparative genomics methods that are designed to predict functional homology from a flexible sequence space, which also includes *REforge* and TACIT. However, to our knowledge, *AFconverge* is the first comparative method that quantifies convergence signals at the motif level in an alignment-free manner. As opposed to the element-focused perspective of other methods, our motif-focused perspective explicitly measures the unique evolutionary pressures experienced by different TF binding sites in different elements. Our approach does not make prior assumptions that specific TF motifs would exhibit global shifts in response to selection pressures, but instead allows for context-specific modularity. We demonstrated that this motif-centric strategy offers the flexibility of analyzing regulatory adaptations at multiple scales, ranging from predicting phenotype-relevant TFs with global convergence signals, detecting divergence of entire elements, to inferring correlations among co-evolving motifs.

Both the existing alignment-free comparative genomics methods, *REforge* and TACIT, are hypothesis-driven in which certain prior expectations have to be encoded to base the convergence analysis on. Specifically, *REforge* requires the specification of a set of TFs presumed to be associated with the phenotype of interest, whereas TACIT requires chromatin accessibility datasets from tissues of interest expected to be implicated in the phenotype.

These requirements limit our ability to evaluate complex phenotypes that are not very well-characterized, and are prone to errors when false priors are given. In contrast, *AFconverge* takes a hypothesis-free approach in which convergence signals are detected purely from sequence changes, while functional inference is only made in downstream analysis. This unbiased approach allows us to characterize functional adaptations that may not be as obvious, or may be more subtle and systemic in nature. For example, our analysis of the vision loss phenotype was able to capture the regulatory adaptation to hypoxia, likely to be driven by the predominantly subterranean convergent species defined as foregrounds. In addition, our latent factor analysis on the longevity phenotype was able to identify the core mechanisms that are likely to be the major drivers of long lifespan, in agreement with hypotheses from other works.

There are several avenues that future extensions of *AFconverge* can develop. Although *AFconverge* makes no prior assumptions on phenotype-relevant TFs, in its current implementation, *AFconverge* featurizes sequences using a set of pre-defined consensus motifs. In reality, there can be many functionally important sequence features that are still unknown. Future work can explore new strategies for featurizing sequence motifs *de novo*. There are several recent works that utilize deep learning models to learn new sequence features with emphasis on interpretability. The *tiSFM* ("totally interpretable sequence to function model") model, which combines motif convolution and attention layers to predict chromatin accessibility signals, uses pre-defined position weight matrices (PWMs) as motif priors, but allows tuning of the PWMs to relax these constraints and possibly learn new features [14]. Another example is the ExplaiNN model, which applies the "neural additive model" strategy combining convolutional neural networks and an interpretable linear model to learn *de novo* motif features affecting TF binding and chromatin accessibility [178]. Given functional readouts across tissues and possibly species, there is an opportunity to develop similar machine learning-based methods for conducting tissue-agnostic, motif-centric convergence analysis.

Finally, *AFconverge* uses max-pooling to represent the maximum evidence of TF binding on a sequence. However, in reality, we may not expect the relationship between the strength of TF binding and the regulated gene expression to always be monotonic. For example, the degree of TF cooperativity and competition in a given promoter, which affects the number of overlapping binding sites, can introduce noise in gene expression levels [185]. In fact, regulatory elements of developmental genes have been found to contain an optimal distribution of strong and weak binding sites, which can be a result of TF cooperativity [46, 72, 201]. This is another area in which a machine learning-based sequence featurization that is grounded on functional readouts can be immensely helpful to learn the regulatory syntax for specific contexts. All in all, *AFconverge* lays the groundwork for a motif-centric approach to study regulatory sequence adaptations underlying phenotypic convergence.

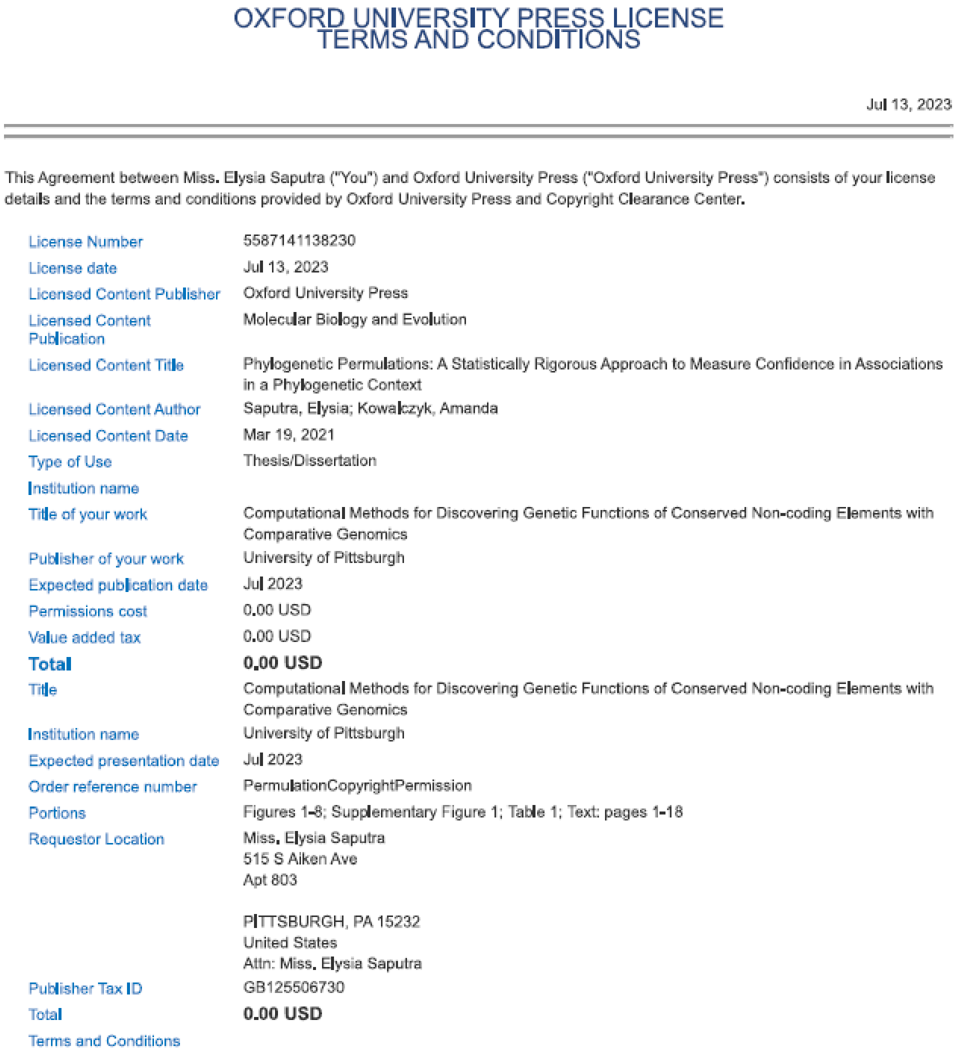
6.0 Conclusions

Evolutionary-based strategies provide an opportunity to understand genotype-phenotype mappings that are not easily accessible with traditional genetics. This thesis contributes new approaches for applying phylogenetic strategies to understand the functional mappings of regulatory non-coding elements, an area that is still insufficiently addressed by existing comparative genomics algorithms. In chapter 1, we propose phylogenetic permulations, a set of phylogenetically-constrained calibration methods that empirically correct signals from phylogenomic analysis against sources of biases. We illustrate the pervasive issue of statistical non-independence in phylogenomic analysis and the insufficiency of parametric calibration strategies in correcting it. We demonstrate (in this and other chapters) that our proposed empirical strategies are effective at improving the statistical robustness and specificity of predictions, in a manner that is not method-dependent. Our empirical approach also addresses an implicit issue in the current state of comparative datasets, which is that with the increasing number of sequenced genomes comes a substantial bias that arises from large variations in genome data quality. As data quality catches up, we believe that our empirical methods can be useful to the community as a strategy to establish confidence in the accuracy of genotype-phenotype mapping predictions from these datasets. When working with binary traits in which the foreground species set contains very large and complex clades, we find that binary permulation strategies can find it difficult to converge and produce null phenotypes with matching dependency structures. Future work can evaluate ways that the rejection sampling conditions can be relaxed without compromising the phylogenetic signal in the shared ancestry patterns.

In the remaining chapters, we describe the development and application of scalable methods for studying motif-level adaptations underlying convergent traits. In chapter 2, we present *phyloConverge*, a maximum likelihood-based algorithm that combines generative modeling of nucleotide substitutions and permulation-based calibration to detect local convergence shifts in evolutionary rates of conserved regulatory elements. We demonstrate the ability of our method to detect transcription factor motif-level convergence signals with high fidelity, which allows us to segmentize a given conserved enhancer into functional units, decode how pleiotropic enhancers respond to selection pressures, and perform genome-scale scanning for and *de novo* discovery of phenotype-relevant non-coding segments. *phyloConverge* offers a new perspective for understanding the evolutionary process of regulatory elements, zooming in from a high-level element-centric perspective to a low-level motif-centric perspective. Future expansions of *phyloConverge* can include adapting it to work with quantitative, ordinal, or categorical phenotypes. Additionally, chapter 3 demonstrates the preliminary application of *phyloConverge* in combination with other methods to investigate the protein and genomic adaptations associated with mammalian colonization of high altitude environments. Grounded on meta-analysis findings from population genetics studies on high altitude adaptation, we illustrate the improvement in predictive power offered by motif-centric analysis compared to element-centric analysis. By evaluating correlations between protein adaptation and motif adaptation, we describe the different evolutionary mechanisms in which regulatory changes can support pathway-level changes. Future developments of this work may explore a deeper investigation into how motif adaptation occurs in relation to purifying selection, positive selection, or relaxed selection at the pathway level.

Finally, in chapter 4, we introduce *AFconverge*, an alignment-free comparative algorithm that predicts the phenotype association of sequence features in regulatory elements. We perform analyses of promoter adaptations underlying a binary trait and a continuous trait – mammalian vision loss and longevity, respectively – and illustrate how motif-level convergence signals can be used to infer different forms of regulatory convergent shifts, including degeneration of entire promoters, global gains and/or losses of specific motifs, and co-evolutions of correlated motifs. Contrary to *phyloConverge* that is limited to conserved elements, *AFconverge* contributes an approach for studying regulatory elements for which functional homology can arise without sequence conservation. It can therefore be used to work with the increasing availability of novel, reference-free hierarchical alignments that can comprehensively account for structural rearrangements of genomic regions. However, the most substantial challenge with alignment-free comparative analysis lies in the implications of reference-free genome alignments themselves, and how they inform our interpretations of the convergence signals. For instance, reference-free alignments can encode duplications of certain genomic regions, which can result in overlaps among mapped orthologs of different elements. In our interaction with the promoter regions from Zoonomia mammalian alignment, out of the total number of promoters per species, we identify ∼8% to ∼18% redundancies. It is unclear whether these redundancies represent true replication events in which a certain gene or genomic element undergoes expansion in certain species (and therefore can be legitimate signals), or whether they should be discarded as confounders. Additionally, expanding the application of *AFconverge* to weakly conserved enhancers would require a better understanding on how orthologous enhancers can be reasonably identified in the absence of sequence conservation.

Appendix A



OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Jul 13, 2023

This Agreement between Miss. Elysia Saputra ("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.

License Number	5587141138230
License date	Jul 13, 2023
Licensed Content Publisher	Oxford University Press
Licensed Content Publication	Molecular Biology and Evolution
Licensed Content Title	Phylogenetic Permulations: A Statistically Rigorous Approach to Measure Confidence in Associations in a Phylogenetic Context
Licensed Content Author	Saputra, Elysia; Kowalczyk, Amanda
Licensed Content Date	Mar 19, 2021
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	Computational Methods for Discovering Genetic Functions of Conserved Non-coding Elements with Comparative Genomics
Publisher of your work	University of Pittsburgh
Expected publication date	Jul 2023
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	**0.00 USD**
Title	Computational Methods for Discovering Genetic Functions of Conserved Non-coding Elements with Comparative Genomics
Institution name	University of Pittsburgh
Expected presentation date	Jul 2023
Order reference number	PermulationCopyrightPermission
Portions	Figures 1-8; Supplementary Figure 1; Table 1; Text: pages 1-18
Requestor Location	Miss. Elysia Saputra 515 S Aiken Ave Apt 803 PITTSBURGH, PA 15232 United States Attn: Miss. Elysia Saputra
Publisher Tax ID	GB125506730
Total	**0.00 USD**
Terms and Conditions	

Figure A1: Molecular Biology and Evolution license/copyright permission to reuse content for Chapter 1 based on the paper Saputra et al. (2021).

Bibliography

[1] Huashui Ai, Bin Yang, Jing Li, Xianhua Xie, Hao Chen, and Jun Ren. Population history and genomic signatures for high-altitude adaptation in Tibetan pigs. *BMC Genomics*, 15(1):834., October 2014.

[2] David B. Allison, Gary L. Gadbury, Moonseong Heo, José R. Fernández, Cheol-Koo Lee, Tomas A. Prolla, and Richard Weindruch. A mixture model approach for the analysis of microarray gene expression data. *Comput. Statist. Data Anal.*, 39(1):1–20, March 2002.

[3] Giovanna Ambrosini, Romain Groux, and Philipp Bucher. PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics*, 34(14):2483–2484, July 2018.

[4] Jesus Amo-Aparicio, Alba Sanchez-Fernandez, Suzhao Li, Elan Z. Eisenmesser, Cecilia Garlanda, Charles A. Dinarello, and Ruben Lopez-Vales. Extracellular and nuclear roles of IL-37 after spinal cord injury. *Brain Behav. Immun.*, 91:194–201, January 2021.

[5] Gregory Andrews, Joel C. Armstrong, Matteo Bianchi, Bruce W. Birren, Kevin R. Bredemeyer, Ana M. Breit, Matthew J. Christmas, Hiram Clawson, Joana Damas, Federica Di Palma, Mark Diekhans, Michael X. Dong, Eduardo Eizirik, Kaili Fan, Cornelia Fanter, Nicole M. Foley, Karin Forsberg-Nilsson, Carlos J. Garcia, John Gatesy, Steven Gazal, Diane P. Genereux, Linda Goodman, Jenna Grimshaw, Michaela K. Halsey, Andrew J. Harris, Glenn Hickey, Michael Hiller, Allyson G. Hindle, Robert M. Hubley, Graham M. Hughes, Jeremy Johnson, David Juan, Irene M. Kaplow, Elinor K. Karlsson, Kathleen C. Keough, Bogdan Kirilenko, Klaus-Peter Koepfli, Jennifer M. Korstian, Amanda Kowalczyk, Sergey V. Kozyrev, Alyssa J. Lawler, Colleen Lawless, Thomas Lehmann, Danielle L. Levesque, Harris A. Lewin, Xue Li, Abigail Lind, Kerstin Lindblad-Toh, Ava Mackay-Smith, Voichita D. Marinescu, Tomas Marques-Bonet, Victor C. Mason, Jennifer R. S. Meadows, Wynn K. Meyer, Jill E. Moore, Lucas R. Moreira, Diana D. Moreno-Santillan, Kathleen M. Morrill, Gerard Muntané, William J. Murphy, Arcadi Navarro, Martin Nweeia, Sylvia Ortmann, Austin Osmanski, Benedict Paten, Nicole S. Paulat, Andreas R. Pfenning, BaDoi N. Phan, Katherine S. Pollard, Henry E. Pratt, David A. Ray, Steven K. Reilly, Jeb R. Rosen, Irina Ruf, Louise Ryan, Oliver A. Ryder, Pardis C. Sabeti, Daniel E. Schäffer, Aitor Serres, Beth Shapiro, Arian F. A. Smit, Mark Springer, Chaitanya Srinivasan, Cynthia Steiner, Jessica M. Storer, Kevin A. M. Sullivan, Patrick F. Sullivan, Elisabeth Sundström, Megan A. Supple, Ross Swofford, Joy-El Talbot, Emma Teeling, Jason Turner-Maier, Alejandro Valenzuela, Franziska Wagner, Ola Wallerman, Chao Wang,

Juehan Wang, Zhiping Weng, Aryn P. Wilder, Morgan E. Wirthlin, James R. Xue, and Xiaomeng Zhang. A genomic timescale for placental mammal evolution. *Science*, 380(6643):eabl8189, April 2023.

[6] Chiara Angiolilli, Emmerik F. A. Leijten, Cornelis P. J. Bekker, Ella Eeftink, Barbara Giovannone, Michel Olde Nordkamp, Marlot van der Wal, Judith L. Thijs, Sebastiaan J. Vastert, Femke van Wijk, Timothy R. D. J. Radstake, and Jorg van Loosdregt. ZFP36 Family Members Regulate the Proinflammatory Features of Psoriatic Dermal Fibroblasts. *J. Invest. Dermatol.*, 142(2):402–413, February 2022.

[7] Adam Antebi. Regulation of longevity by the reproductive system. *Exp. Gerontol.*, 48(7):596–602, July 2013.

[8] Nuno Arantes-Oliveira, Javier Apfeld, Andrew Dillin, and Cynthia Kenyon. Regulation of Life-Span by Germ-Line Stem Cells in Caenorhabditis elegans. *Science*, 295(5554):502–505, January 2002.

[9] Sadaf Ashraf, Samuel Bell, Caitriona O'Leary, Paul Canning, Ileana Micu, Jose A. Fernandez, Michael O'Hare, Peter Barabas, Hannah McCauley, Derek P. Brazil, Alan W. Stitt, J. Graham McGeown, and Tim M. Curtis. CAMKII as a therapeutic target for growth factor-induced retinal and choroidal neovascularization. *JCI Insight*, 4(6):e122442., March 2019.

[10] Aaron Avivi, Henrik Oster, Alma Joel, Avigdor Beiles, Urs Albrecht, and Eviatar Nevo. Circadian genes in a blind subterranean mammal II: conservation and uniqueness of the three Period homologs in the blind subterranean mole rat, Spalax ehrenbergi superspecies. *Proc. Natl. Acad. Sci. U.S.A.*, 99(18):11718–11723, September 2002.

[11] Jun Bai, Lijuan Li, Yanhong Li, and Liansheng Zhang. Genetic and immune changes in Tibetan high-altitude populations contribute to biological adaptation to hypoxia. *Environ. Health Preventive Med.*, 27, 2022.

[12] Liang Bai, Baoning Liu, Changmian Ji, Sihai Zhao, Siyu Liu, Rong Wang, Weirong Wang, Pu Yao, Xuming Li, Xiaojun Fu, Haiyan Yu, Min Liu, Fengming Han, Ning Guan, Hui Liu, Dongyuan Liu, Yuanqing Tao, Zhongdong Wang, Shunsheng Yan, Greg Florant, Michael T. Butcher, Jifeng Zhang, Hongkun Zheng, Jianglin Fan, and Enqi Liu. Hypoxic and Cold Adaptation Insights from the Himalayan Marmot Genome. *iScience*, 11:519–530, January 2019.

[13] Timothy L. Bailey. STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, 37(18):2834–2840, September 2021.

[14] Ali Tugrul Balci, Mark Maher Ebeid, Panayiotis V. Benos, Dennis Kostka, and Maria Chikina. An intrinsically interpretable neural network architecture for sequence to function learning. *bioRxiv*, 2023.01.25.525572., March 2023.

[15] Dov B. Ballak, Vienna E. Brunt, Zachary J. Sapinsley, Brian P. Ziemba, James J. Richey, Melanie C. Zigler, Lawrence C. Johnson, Rachel A. Gioscia-Ryan, Rachel Culp-Hill, Elan Z. Eisenmesser, Angelo D'Alessandro, Charles A. Dinarello, and Douglas R. Seals. Short-term interleukin-37 treatment improves vascular endothelial function, endurance exercise capacity, and whole-body glucose metabolism in old mice. *Aging Cell*, 19(1), January 2020.

[16] Andrew David Beale, David Whitmore, and Damian Moran. Life in a dark biosphere: a review of circadian physiology in "arrhythmic" environments. *J. Comp. Physiol. B*, 186(8):947–968, December 2016.

[17] Cynthia M. Beall. Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. *Integr. Comp. Biol.*, 46(1):18–24, February 2006.

[18] Cynthia M. Beall, Gianpiero L. Cavalleri, Libin Deng, Robert C. Elston, Yang Gao, Jo Knight, Chaohua Li, Jiang Chuan Li, Yu Liang, Mark McCormack, Hugh E. Montgomery, Hao Pan, Peter A. Robbins, Kevin V. Shianna, Siu Cheung Tam, Ngodrop Tsering, Krishna R. Veeramah, Wei Wang, Puchung Wangdui, Michael E. Weale, Yaomin Xu, Zhe Xu, Ling Yang, M. Justin Zaman, Changqing Zeng, Li Zhang, Xianglong Zhang, Pingcuo Zhaxi, and Yong Tang Zheng. Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. *Proc. Natl. Acad. Sci. U.S.A.*, 107(25):11459–11464, June 2010.

[19] Abigail W. Bigham. Genetics Of Human Origin and Evolution: High-Altitude Adaptations. *Curr. Opin. Genet. Dev.*, 41:8, December 2016.

[20] Abigail W. Bigham and Frank S. Lee. Human high-altitude adaptation: forward genetics meets the HIF pathway. *Genes Dev.*, 28(20):2189–2204, October 2014.

[21] Abigail W. Bigham, Xianyun Mao, Rui Mei, Tom Brutsaert, Megan J. Wilson, Colleen Glyde Julian, Esteban J. Parra, Joshua M. Akey, Lorna G. Moore, and Mark D. Shriver. Identifying positive selection candidate loci for high-altitude adaptation in Andean populations. *Hum. Genomics*, 4(2):79, 2009.

[22] Olaf R. P. Bininda-Emonds, Marcel Cardillo, Kate E. Jones, Ross D. E. MacPhee, Robin M. D. Beck, Richard Grenyer, Samantha A. Price, Rutger A. Vos, John L. Gittleman, and Andy Purvis. The delayed rise of present-day mammals. *Nature*, 446:507–512, March 2007.

[23] Ekin Bolukbasi, Nathaniel S. Woodling, Dobril K. Ivanov, Jennifer Adcott, Andrea Foley, Arjunan Rajasingam, Lauren M. Gittings, Benjamin Aleyakpo, Teresa Niccoli, Janet M. Thornton, and Linda Partridge. Cell type-specific modulation of healthspan by forkhead family transcription factors in the nervous system. *Proceedings of the National Academy of Sciences*, 118(8), February 2021.

[24] Anne-Sophie Borowiec, Philippe Delcourt, Etienne Dewailly, and Gabriel Bidaux. Optimal Differentiation of In Vitro Keratinocytes Requires Multifactorial External Control. *PLoS One*, 8(10):e77507, October 2013.

[25] Elizabeth I. Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J. Michael Cherry, and Gavin Sherlock. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, December 2004.

[26] Paola Briata, Cristina Ilengo, Giorgio Corte, Christoph Moroni, Michael G. Rosenfeld, Ching-Yi Chen, and Roberto Gherzi. The Wnt/beta-catenin–>Pitx2 pathway controls the turnover of Pitx2 and other unstable mRNAs. *Mol. Cell*, 12(5):1201–1211, November 2003.

[27] Matthew J. Brooks, Holly Y. Chen, Ryan A. Kelley, Anupam K. Mondal, Kunio Nagashima, Natalia De Val, Tiansen Li, Vijender Chaitankar, and Anand Swaroop. Improved Retinal Organoid Differentiation by Modulating Signaling Pathways Revealed by Comparative Transcriptome Analyses with Development In Vivo. *Stem Cell Rep.*, 13(5):891–905, November 2019.

[28] Vienna E. Brunt, Akpevweoghene P. Ikoba, Brian P. Ziemba, Dov B. Ballak, Alexander Hoischen, Charles A. Dinarello, Marissa A. Ehringer, and Douglas R. Seals. Circulating interleukin-37 declines with aging in healthy humans: relations to healthspan indicators and IL37 gene SNPs. *Geroscience*, 45(1):65–84, February 2023.

[29] Ana-Maria Bulau, Michaela Fink, Christof Maucksch, Roland Kappler, Doris Mayr, Kai Wagner, and Philip Bufler. In vivo expression of interleukin-37 reduces local and systemic inflammation in concanavalin A-induced hepatitis. *ScientificWorldJournal*, 11(2480-90.):;, 2011.

[30] Zoë Burke and Guillermo Oliver. Prox1 is an early specific marker for the developing liver and pancreas in the mammalian foregut endoderm. *Mech. Dev.*, 118(1):147–155, October 2002.

[31] Valentina Caprara, Silvia Scappa, Emirena Garrafa, Valeriana Di Castro, Laura Rosanò, Anna Bagnato, and Francesca Spinella. Endothelin-1 regulates hypoxia-inducible factor-1α and -2α stability through prolyl hydroxylase domain 2 inhibition in human lymphatic endothelial cells. *Life Sci.*, 118(2):185–190, November 2014.

[32] Shelley L. Cargill, James R. Carey, Hans-Georg Müller, and Gary Anderson. Age of ovary determines remaining life expectancy in old ovariectomized mice. *Aging Cell*, 2(3):185–190, June 2003.

[33] Sean B. Carroll. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134(1):25–36, July 2008.

[34] K. C. Catania. A nose that looks like a hand and acts like an eye: the unusual mechanosensory system of the star-nosed mole. *J. Comp. Physiol. A*, 185(4):367–372, October 1999.

[35] Aleah F. Caulin and Carlo C. Maley. Peto's Paradox: Evolution's Prescription for Cancer Prevention. *Trends Ecol. Evol.*, 26(4):175, April 2011.

[36] Georgia Chachami, Nicolas Stankovic-Valentin, Angeliki Karagiota, Angeliki Basagianni, Uwe Plessmann, Henning Urlaub, Frauke Melchior, and George Simos. Hypoxia-induced Changes in SUMO Conjugation Affect Transcriptional Regulation Under Low Oxygen∗[S]. *Mol. Cell. Proteomics*, 18(6):1197–1209, June 2019.

[37] Lisheng Chen and Philip J. Gage. Heterozygous Pitx2 Null Mice Accurately Recapitulate the Ocular Features of Axenfeld-Rieger Syndrome and Congenital Glaucoma. *Invest. Ophthalmol. Visual Sci.*, 57(11):5023, September 2016.

[38] Arthur H. Cheng, Pascale Bouchard-Cannon, Sara Hegazi, Christopher Lowden, Samuel W. Fung, Cheng-Kang Chiang, Rob W. Ness, and Hai-Ying Mary Cheng. SOX2-Dependent Transcription in Clock Neurons Promotes the Robustness of the Central Circadian Pacemaker. *Cell Rep.*, 26(12):3191–32028, March 2019.

[39] James M. Cheverud and Malcolm M. Dow. An autocorrelation analysis of genetic variation due to lineal fission in social groups of rhesus macaques. *Am. J. Phys. Anthropol.*, 67(2):113–121, June 1985.

[40] Maria Chikina, Joseph D. Robinson, and Nathan L. Clark. Hundreds of Genes Experienced Convergent Shifts in Selective Pressure in Marine Mammals. *Mol. Biol. Evol.*, 33(9):2182–2192, September 2016.

[41] Kenneth L. Chiou, Mareike C. Janiak, India A. Schneider-Crease, Sharmi Sen, Ferehiwot Ayele, Idrissa S. Chuma, Sascha Knauf, Alemayehu Lemma, Anthony V. Signore, Anthony M. D'Ippolito, Belayneh Abebe, Abebaw Azanaw Haile, Fanuel Kebede, Peter J. Fashing, Nga Nguyen, Colleen McCann, Marlys L. Houck, Jeffrey D. Wall, Andrew S. Burrell, Christina M. Bergey, Jeffrey Rogers, Jane E. Phillips-Conroy, Clifford J. Jolly, Amanda D. Melin, Jay F. Storz, Amy Lu, Jacinta C. Beehner, Thore J. Bergman, and Noah Snyder-Mackler. Genomic signatures of high-altitude adaptation and chromosomal polymorphism in geladas. *Nat. Ecol. Evol.*, 6(5):630–643, May 2022.

[42] Nathan L. Clark, Eric Alani, and Charles F. Aquadro. Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Res.*, 22(4):714–720, April 2012.

[43] Nathan L. Clark, Eric Alani, and Charles F. Aquadro. Evolutionary Rate Covariation in Meiotic Proteins Results from Fluctuating Evolutionary Pressure in Yeasts and Mammals. *Genetics*, 193(2):529–538, February 2013.

[44] H. M. Cooper, M. Herbin, and E. Nevo. Visual system of a naturally microphthalmic mammal: the blind mole rat, Spalax ehrenbergi. *J. Comp. Neurol.*, 328(3):313–350, February 1993.

[45] Samuel D. Crish, Christine M. Dengler-Crish, and Kenneth C. Catania. Central visual system of the naked mole-rat (Heterocephalus glaber). *Anat. Rec. A Discov. Mol. Cell. Evol. Biol.*, 288(2):205–212, February 2006.

[46] Justin Crocker, Namiko Abe, Lucrezia Rinaldi, Alistair P. McGregor, Nicolás Frankel, Shu Wang, Ahmad Alsawadi, Philippe Valenti, Serge Plaza, François Payre, Richard S. Mann, and David L. Stern. Low Affinity Binding Site Clusters Confer Hox Specificity and Regulatory Robustness. *Cell*, 160(0):191, January 2015.

[47] Debabrata Das and Swathi Arur. Regulation of oocyte maturation: Role of conserved ERK signaling. *Mol. Reprod. Dev.*, 89(9):353–374, September 2022.

[48] K. T. J. Davies, J. A. Cotton, J. D. Kirwan, E. C. Teeling, and S. J. Rossiter. Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence. *Heredity*, 108(5):480–489, May 2012.

[49] Eugene V. Davydov, David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput. Biol.*, 6(12):e1001025, December 2010.

[50] Paola de Andrade Mello, Robson Coutinho-Silva, and Luiz Eduardo Baggio Savio. Multifaceted Effects of Extracellular Adenosine Triphosphate and Adenosine in the Tumor-Host Interaction and Therapeutic Perspectives. *Front. Immunol.*, 8:1526., November 2017.

[51] Francesco Di Virgilio, Alba Clara Sarti, Simonetta Falzoni, Elena De Marchi, and Elena Adinolfi. Extracellular ATP and P2 purinergic signalling in the tumour microenvironment. *Nat. Rev. Cancer*, 18(10):601–618, October 2018.

[52] Paula Díaz-Bulnes, María Laura Saiz, Carlos López-Larrea, and Ramón M. Rodríguez. Crosstalk Between Hypoxia and ER Stress Response: A Key Regulator of Macrophage Polarization. *Front. Immunol.*, 10:2951., January 2020.

[53] Susanne Dobler, Safaa Dalla, Vera Wagschal, and Anurag A. Agrawal. Community-wide convergent evolution in insect adaptation to toxic cardenolides by substitutions in the Na,K-ATPase. *Proc. Natl. Acad. Sci. U.S.A.*, 109(32):13040–13045, August 2012.

[54] Sabina Domené, Viviana F. Bumaschny, Flávio S. J. de Souza, Lucía F. Franchini, Sofía Nasif, Malcolm J. Low, and Marcelo Rubinstein. Enhancer turnover and conserved regulatory function in vertebrate evolution. *Phil. Trans. R. Soc. B*, 368(1632):20130027, December 2013.

[55] Scott W. Doniger and Justin C. Fay. Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput. Biol.*, 3(5):e99., May 2007.

[56] Robin D. Dowell. Transcription factor binding variation in the evolution of gene regulation. *Trends Genet.*, 26(11):468–475, November 2010.

[57] Zewdu Edea, Hailu Dadi, Tadelle Dessie, and Kwan-Suk Kim. Genomic signatures of high-altitude adaptation in Ethiopian sheep populations. *Genes Genomics*, 41(8):973–981, August 2019.

[58] Eran Eden, Doron Lipson, Sivan Yogev, and Zohar Yakhini. Discovering Motifs in Ranked Lists of DNA Sequences. *PLoS Comput. Biol.*, 3(3):e39, March 2007.

[59] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinf.*, 10(1):1–7, December 2009.

[60] Robert C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinf.*, 5(1):1–19, December 2004.

[61] Stacey L. Edwards, Jonathan Beesley, Juliet D. French, and Alison M. Dunning. Beyond GWASs: Illuminating the Dark Road from Association to Function. *Am. J. Hum. Genet.*, 93(5):779, November 2013.

[62] Anna Egger, Marijana Samardzija, Vithiyanjali Sothilingam, Naoyuki Tanimoto, Christina Lange, Silvia Salatino, Lei Fang, Marina Garcia-Garrido, Susanne Beck, Michal J. Okoniewski, Albert Neutzner, Mathias W. Seeliger, Christian Grimm, and Christoph Handschin. PGC-1α Determines Light Damage Susceptibility of the Murine Retina. *PLoS One*, 7(2):e31272, February 2012.

[63] Christina A. Eichstaedt, Tiago Antão, Luca Pagani, Alexia Cardona, Toomas Kivisild, and Maru Mormina. The Andean Adaptive Toolkit to Counteract High Altitude Maladaptation: Genome-Wide and Phenotypic Analysis of the Collas. *PLoS One*, 9(3):e93314, March 2014.

[64] Pär G. Engström, Shannan J. Ho Sui, Oyvind Drivenes, Thomas S. Becker, and Boris Lenhard. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.*, 17(12):1898–1908, December 2007.

[65] Thomas Euler, Silke Haverkamp, Timm Schubert, and Tom Baden. Retinal bipolar cells: elementary building blocks of vision. *Nat. Rev. Neurosci.*, 15:507–519, August 2014.

[66] Adam Eyre-Walker and Peter D. Keightley. High genomic deleterious mutation rates in hominids. *Nature*, 397:344–347, January 1999.

[67] Adam Eyre-Walker, Peter D. Keightley, Nick G. C. Smith, and Daniel Gaffney. Quantifying the Slightly Deleterious Mutation Model of Molecular Evolution. *Mol. Biol. Evol.*, 19(12):2142–2149, December 2002.

[68] Adam Eyre-Walker, Megan Woolfit, and Ted Phelps. The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans. *Genetics*, 173(2):891, June 2006.

[69] Chunling Fan, Qingning Su, Yun Li, Lihua Liang, Daniel J. Angelini, William B. Guggino, and Roger A. Johns. Hypoxia-induced mitogenic factor/FIZZ1 induces intracellular calcium release through the PLC-IP3 pathway. *American Journal of Physiology - Lung Cellular and Molecular Physiology*, 297(2):L263, August 2009.

[70] Yan-Yan Fan, Hong-Yu Chen, Wei Chen, Yi-Nan Liu, Yan Fu, and Li-Na Wang. Expression of inflammatory cytokines in serum and peritoneal fluid from patients with different stages of endometriosis. *Gynecol. Endocrinol.*, 34(6):507–512, June 2018.

[71] Hsin-Yu Fang, Russell Hughes, Craig Murdoch, Seth B. Coffelt, Subhra K. Biswas, Adrian L. Harris, Randall S. Johnson, Hongxia Z. Imityaz, M. Celeste Simon, Erik Fredlund, Florian R. Greten, Jordi Rius, and Claire E. Lewis. Hypoxia-inducible factors 1 and 2 are important transcriptional effectors in primary macrophages experiencing hypoxia. *Blood*, 114(4):844–859, July 2009.

[72] Emma K. Farley, Katrina M. Olson, Wei Zhang, Daniel S. Rokhsar, and Michael S. Levine. From the Cover: Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proc. Natl. Acad. Sci. U.S.A.*, 113(23):6508, June 2016.

[73] Joseph Felsenstein. Phylogenies and the Comparative Method on JSTOR. *Am. Nat.*, 125(1):1–15, January 1985.

[74] Thomas Flatt, Kyung-Jin Min, Cecilia D'Alterio, Eugenia Villa-Cuesta, John Cumbers, Ruth Lehmann, D. Leanne Jones, and Marc Tatar. Drosophila germ-line modulation of insulin signaling and lifespan. *Proc. Natl. Acad. Sci. U.S.A.*, 105(17):6368–6373, April 2008.

[75] Andrew D. Foote, Yue Liu, Gregg W. C. Thomas, Tomáš Vinař, Jessica Alföldi, Jixin Deng, Shannon Dugan, Cornelis E. van Elk, Margaret E. Hunter, Vandita Joshi, Ziad Khan, Christie Kovar, Sandra L. Lee, Kerstin Lindblad-Toh, Annalaura Mancia, Rasmus Nielsen, Xiang Qin, Jiaxin Qu, Brian J. Raney, Nagarjun Vijay, Jochen B. W. Wolf, Matthew W. Hahn, Donna M. Muzny, Kim C. Worley, M. Thomas P. Gilbert, and Richard A. Gibbs. Convergent evolution of the genomes of marine mammals. *Nat. Genet.*, 47:272–275, March 2015.

[76] Henrik ten Freyhaus, Markus Dagnell, Maike Leuchs, Marius Vantler, Eva M. Berghausen, Evren Caglayan, Norbert Weissmann, Bhola K. Dahal, Ralph T. Schermuly, Arne Ostman, Kai Kappert, and Stephan Rosenkranz. Hypoxia enhances platelet-derived growth factor signaling in the pulmonary vasculature by down-

regulation of protein tyrosine phosphatases. *Am. J. Respir. Crit. Care Med.*, 183(8):1092–1102, April 2011.

[77] Qiangwei Fu, Sean P. Colgan, and Carl Simon Shelley. Hypoxia: The Force that Drives Chronic Kidney Disease. *Clin. Med. Res.*, 14(1):15, March 2016.

[78] Angela Galardi, Marta Colletti, Chiara Lavarello, Virginia Di Paolo, Paolo Mascio, Ida Russo, Raffaele Cozza, Antonino Romanzo, Paola Valente, Rita De Vito, Luisa Pascucci, Hector Peinado, Angel M. Carcaboso, Andrea Petretto, Franco Locatelli, and Angela Di Giannatale. Proteomic Profiling of Retinoblastoma-Derived Exosomes Reveals Potential Biomarkers of Vitreous Seeding. *Cancers*, 12(6), June 2020.

[79] Ri-Li Ge, Qingle Cai, Yong-Yi Shen, A. San, Lan Ma, Yong Zhang, Xin Yi, Yan Chen, Lingfeng Yang, Ying Huang, Rongjun He, Yuanyuan Hui, Meirong Hao, Yue Li, Bo Wang, Xiaohua Ou, Jiaohui Xu, Yongfen Zhang, Kui Wu, Chunyu Geng, Weiping Zhou, Taicheng Zhou, David M. Irwin, Yingzhong Yang, Liu Ying, Haihua Bao, Jaebum Kim, Denis M. Larkin, Jian Ma, Harris A. Lewin, Jinchuan Xing, Roy N. Platt Nd, David A. Ray, Loretta Auvil, Boris Capitanu, Xiufeng Zhang, Guojie Zhang, Robert W. Murphy, Jun Wang, Ya-Ping Zhang, and Jian Wang. Draft genome sequence of the Tibetan antelope. *Nat. Commun.*, 4(1858.):;, 2013.

[80] Ri-Li Ge, Qingle Cai, Yong-Yi Shen, A. San, Lan Ma, Yong Zhang, Xin Yi, Yan Chen, Lingfeng Yang, Ying Huang, Rongjun He, Yuanyuan Hui, Meirong Hao, Yue Li, Bo Wang, Xiaohua Ou, Jiaohui Xu, Yongfen Zhang, Kui Wu, Chunyu Geng, Weiping Zhou, Taicheng Zhou, David M. Irwin, Yingzhong Yang, Liu Ying, Haihua Bao, Jaebum Kim, Denis M. Larkin, Jian Ma, Harris A. Lewin, Jinchuan Xing, Roy N. Platt, David A. Ray, Loretta Auvil, Boris Capitanu, Xiufeng Zhang, Guojie Zhang, Robert W. Murphy, Jun Wang, Ya-Ping Zhang, and Jian Wang. Draft genome sequence of the Tibetan antelope. *Nat. Commun.*, 4(1858):1–7, May 2013.

[81] Diane P. Genereux, Aitor Serres, Joel Armstrong, Jeremy Johnson, Voichita D. Marinescu, Eva Murén, David Juan, Gill Bejerano, Nicholas R. Casewell, Leona G. Chemnick, Joana Damas, Federica Di Palma, Mark Diekhans, Ian T. Fiddes, Manuel Garber, Vadim N. Gladyshev, Linda Goodman, Wilfried Haerty, Marlys L. Houck, Robert Hubley, Teemu Kivioja, Klaus-Peter Koepfli, Lukas F. K. Kuderna, Eric S. Lander, Jennifer R. S. Meadows, William J. Murphy, Will Nash, Hyun Ji Noh, Martin Nweeia, Andreas R. Pfenning, Katherine S. Pollard, David A. Ray, Beth Shapiro, Arian F. A. Smit, Mark S. Springer, Cynthia C. Steiner, Ross Swofford, Jussi Taipale, Emma C. Teeling, Jason Turner-Maier, Jessica Alfoldi, Bruce Birren, Oliver A. Ryder, Harris A. Lewin, Benedict Paten, Tomas Marques-Bonet, Kerstin Lindblad-Toh, Elinor K. Karlsson, and Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature*, 587(7833):240–245, Nov 2020.

[82] Soumyaditya Ghosh, Kaitlyn N. Lewis, Richa Tulsian, Artem A. Astafev, Rochelle Buffenstein, and Roman V. Kondratov. It's about time; divergent circadian clocks in livers of mice and naked mole-rats. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 35(5):e21590, May 2021.

[83] Tithi Ghosh, Subhasis Barik, Avishek Bhuniya, Jesmita Dhar, Shayani Dasgupta, Sarbari Ghosh, Madhurima Sarkar, Ipsita Guha, Koustav Sarkar, Pinak Chakrabarti, Bhaskar Saha, Walter J. Storkus, Rathindranath Baral, and Anamika Bose. Tumor-associated mesenchymal stem cells inhibit naïve T cell expansion by blocking cysteine export from dendritic cells. *Int. J. Cancer*, 139(9):2068–2081, November 2016.

[84] Hector Giral, Ulf Landmesser, and Adelheid Kratzer. Into the Wild: GWAS Exploration of Non-coding RNAs. *Front. Cardiovasc. Med.*, 5:412556, December 2018.

[85] John L. Gittleman and Mark Kot. Adaptation: Statistics and a Null Model for Estimating Phylogenetic Effects. *Syst. Biol.*, 39(3):227–241, September 1990.

[86] Guido A. Gnecchi-Ruscone, Paolo Abondio, Sara De Fanti, Stefania Sarno, Mingma G. Sherpa, Phurba T. Sherpa, Giorgio Marinelli, Luca Natali, Marco Di Marcello, Davide Peluzzi, Donata Luiselli, Davide Pettener, and Marco Sazzini. Evidence of Polygenic Adaptation to High Altitude from Tibetan and Sherpa Genomes. *Genome Biol. Evol.*, 10(11):2919–2930, November 2018.

[87] Xiao Gou, Zhen Wang, Ning Li, Feng Qiu, Ze Xu, Dawei Yan, Shuli Yang, Jia Jia, Xiaoyan Kong, Zehui Wei, Shaoxiong Lu, Linsheng Lian, Changxin Wu, Xueyan Wang, Guozhi Li, Teng Ma, Qiang Jiang, Xue Zhao, Jiaqiang Yang, Baohong Liu, Dongkai Wei, Hong Li, Jianfa Yang, Yulin Yan, Guiying Zhao, Xinxing Dong, Mingli Li, Weidong Deng, Jing Leng, Chaochun Wei, Chuan Wang, Huaming Mao, Hao Zhang, Guohui Ding, and Yixue Li. Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. *Genome Res.*, 24(8):1308, August 2014.

[88] A. Grafen. The Phylogenetic Regression on JSTOR. *Philos. Trans. R. Soc. London, Ser. B*, 326(1233):119–157, December 1989.

[89] Timothy Grocott, Estefania Lozano-Velasco, Gi Fay Mok, and Andrea E. Münsterberg. The Pax6 master control gene initiates spontaneous retinal development via a self-organising Turing network. *Development*, 147(24):dev185827., December 2020.

[90] Zuguang Gu and Daniel Hübschmann. rGREAT: an R/bioconductor package for functional enrichment on genomic regions. *Bioinformatics*, 39(1):btac745, January 2023.

[91] Xiaowei Guo, Zhuojie Li, Xiaojie Zhu, Meixiao Zhan, Chenxi Wu, Xiang Ding, Kai Peng, Wenzhe Li, Xianjue Ma, Zhongwei Lv, Ligong Lu, and Lei Xue. A coherent FOXO3-SNAI2 feed-forward loop in autophagy. *Proc. Natl. Acad. Sci. U.S.A.*, 119(11):e2118285119, March 2022.

[92] Shobhit Gupta, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biol.*, 8(2):1–9, February 2007.

[93] Christine Hacker, Ralitsa Valchanova, Stephanie Adams, and Barbara Munz. ZFP36L1 is regulated by growth factors and cytokines in keratinocytes and influences their VEGF production. *Growth Factors*, 28(3):178–190, June 2010.

[94] J. D. Hadfield and S. Nakagawa. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *J. Evol. Biol.*, 23(3):494–508, March 2010.

[95] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck Rd, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–358729, June 2021.

[96] Luke J. Harmon, Jason T. Weir, Chad D. Brock, Richard E. Glor, and Wendell Challenger. GEIGER: investigating evolutionary radiations. *Bioinformatics*, 24(1):129–131, January 2008.

[97] R. P. Heaney. Toward a physiological referent for the vitamin D requirement. *J. Endocrinol. Invest.*, 37(11):1127–1130, November 2014.

[98] Nikolai Hecker and Michael Hiller. A genome alignment of 120 mammals highlights ultraconserved element variability and placenta-associated enhancers. *GigaScience*, 9(1):giz159, January 2020.

[99] Imed Helal, Godela M. Fick-Brosnahan, Berenice Reed-Gitomer, and Robert W. Schrier. Glomerular hyperfiltration: definitions, mechanisms and clinical implications. *Nat. Rev. Nephrol.*, 8:293–300, May 2012.

[100] Sher L. Hendrickson. A genome wide study of genetic adaptation to high altitude in feral Andean Horses of the páramo. *BMC Evol. Biol.*, 13(1):1–13, December 2013.

[101] John R. Hetling, Monica S. Baig-Silva, Christopher M. Comer, Machelle T. Pardue, Dalia Y. Samaan, Nasser M. Qtaishat, David R. Pepperberg, and Thomas J. Park. Features of visual function in the naked mole-rat Heterocephalus glaber. *J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol.*, 191(4):317–330, April 2005.

[102] Glenn Hickey, Benedict Paten, Dent Earl, Daniel Zerbino, and David Haussler. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*, 29(10):1341–1342, May 2013.

[103] Michael Hiller, Bruce T. Schaar, Vahan B. Indjeian, David M. Kingsley, Lee R. Hagey, and Gill Bejerano. A "Forward Genomics" Approach Links Genotype to Phenotype using Independent Phenotypic Losses among Related Species. *Cell Rep.*, 2(4):817–823, October 2012.

[104] Diana Hoogeboom, Marieke A. G. Essers, Paulien E. Polderman, Erik Voets, Lydia M. M. Smits, and Boudewijn M. Th. Burgering. Interaction of FOXO with β-Catenin Inhibits β-Catenin/T Cell Factor Activity $*$. *J. Biol. Chem.*, 283(14):9224–9230, April 2008.

[105] Peter A. Hosner, Brant C. Faircloth, Travis C. Glenn, Edward L. Braun, and Rebecca T. Kimball. Avoiding Missing Data Biases in Phylogenomic Inference: An Empirical Study in the Landfowl (Aves: Galliformes). *Mol. Biol. Evol.*, 33(4):1110–1125, April 2016.

[106] Elizabeth A. Housworth, Emília P. Martins, and Michael Lynch. The phylogenetic mixed model. *Am. Nat.*, 163(1):84–96, January 2004.

[107] Barbara Hrdlickova, Rodrigo Coutinho de Almeida, Zuzanna Borek, and Sebo Withoff. Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochim. Biophys. Acta*, 1842(10):1910–1922, October 2014.

[108] Heng-tong Hu, Qing-yong Ma, Dong Zhang, Su-gang Shen, Liang Han, Ya-dong Ma, Ruo-fei Li, and Ke-ping Xie. HIF-1alpha links beta-adrenoceptor agonists and pancreatic cancer cells under normoxic condition. *Acta Pharmacol. Sin.*, 31(1):102–110, January 2010.

[109] Yang Hu, Alejandra Korovaichuk, Mariana Astiz, Henning Schroeder, Rezaul Islam, Jon Barrenetxea, Andre Fischer, Henrik Oster, and Henrik Bringmann. Functional Divergence of Mammalian TFAP2a and TFAP2b Transcription Factors for Bidirectional Sleep Control. *Genetics*, 216(3):735, November 2020.

[110] Zhirui Hu, Timothy B. Sackton, Scott V. Edwards, and Jun S. Liu. Bayesian Detection of Convergent Rate Changes of Conserved Noncoding Elements on Phylogenetic Trees. *Mol. Biol. Evol.*, 36(5):1086–1100, May 2019.

[111] Melissa J. Hubisz, Katherine S. Pollard, and Adam Siepel. PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings Bioinf.*, 12(1):41–51, January 2011.

[112] Emilia Huerta-Sánchez, Michael DeGiorgio, Luca Pagani, Ayele Tarekegn, Rosemary Ekong, Tiago Antao, Alexia Cardona, Hugh E. Montgomery, Gianpiero L. Cavalleri, Peter A. Robbins, Michael E. Weale, Neil Bradman, Endashaw Bekele, Toomas Kivisild, Chris Tyler-Smith, and Rasmus Nielsen. Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations. *Mol. Biol. Evol.*, 30(8):1877–1888, August 2013.

[113] # Irene M. Kaplow, # Alyssa J. Lawler, # Daniel E. Schäffer, Chaitanya Srinivasan, Heather H. Sestili, Morgan E. Wirthlin, BaDoi N. Phan, Kavya Prasad, Ashley R. Brown, Xiaomeng Zhang, Kathleen Foley, Diane P. Genereux, Zoonomia Consortium∗∗, Elinor K. Karlsson, Kerstin Lindblad-Toh, Wynn K. Meyer, Andreas R. Pfenning, Gregory Andrews, Joel C. Armstrong, Matteo Bianchi, Bruce W. Birren, Kevin R. Bredemeyer, Ana M. Breit, Matthew J. Christmas, Hiram Clawson, Joana Damas, Federica Di Palma, Mark Diekhans, Michael X. Dong, Eduardo Eizirik, Kaili Fan, Cornelia Fanter, Nicole M. Foley, Karin Forsberg-Nilsson, Carlos J. Garcia, John Gatesy, Steven Gazal, Diane P. Genereux, Linda Goodman, Jenna Grimshaw, Michaela K. Halsey, Andrew J. Harris, Glenn Hickey, Michael Hiller, Allyson G. Hindle, Robert M. Hubley, Graham M. Hughes, Jeremy Johnson, David Juan, Irene M. Kaplow, Elinor K. Karlsson, Kathleen C. Keough, Bogdan Kirilenko, Klaus-Peter Koepfli, Jennifer M. Korstian, Amanda Kowalczyk, Sergey V. Kozyrev, Alyssa J. Lawler, Colleen Lawless, Thomas Lehmann, Danielle L. Levesque, Harris A. Lewin, Xue Li, Abigail Lind, Kerstin Lindblad-Toh, Ava Mackay-Smith, Voichita D. Marinescu, Tomas Marques-Bonet, Victor C. Mason, Jennifer R. S. Meadows, Wynn K. Meyer, Jill E. Moore, Lucas R. Moreira, Diana D. Moreno-Santillan, Kathleen M.

Morrill, Gerard Muntané, William J. Murphy, Arcadi Navarro, Martin Nweeia, Sylvia Ortmann, Austin Osmanski, Benedict Paten, Nicole S. Paulat, Andreas R. Pfenning, BaDoi N. Phan, Katherine S. Pollard, Henry E. Pratt, David A. Ray, Steven K. Reilly, Jeb R. Rosen, Irina Ruf, Louise Ryan, Oliver A. Ryder, Pardis C. Sabeti, Daniel E. Schäffer, Aitor Serres, Beth Shapiro, Arian F. A. Smit, Mark Springer, Chaitanya Srinivasan, Cynthia Steiner, Jessica M. Storer, Kevin A. M. Sullivan, Patrick F. Sullivan, Elisabeth Sundström, Megan A. Supple, Ross Swofford, Joy-El Talbot, Emma Teeling, Jason Turner-Maier, Alejandro Valenzuela, Franziska Wagner, Ola Wallerman, Chao Wang, Juehan Wang, Zhiping Weng, Aryn P. Wilder, Morgan E. Wirthlin, James R. Xue, and Xiaomeng Zhang. Relating enhancer genetic variation across mammals to complex phenotypes using machine learning. *Science*, 380(6643):eabm7993., April 2023.

[114] Jian Fa Jiang, Song Shu Xiao, and Min Xue. Decreased expression of interleukin-37 in the ectopic and eutopic endometria of patients with adenomyosis. *Gynecol. Endocrinol.*, 34(1):83–86, January 2018.

[115] Shao-Yun Jiang and Jian-Tao Wang. Msx2 alters the timing of retinal ganglion cells fate commitment and differentiation. *Biochem. Biophys. Res. Commun.*, 395(4):524–529, May 2010.

[116] Granton A. Jindal and Emma K. Farley. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Dev. Cell*, 56(5):575–587, March 2021.

[117] Per Johnsson, Leonard Lipovich, Dan Grandér, and Kevin V. Morris. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1840(3):1063–1071, March 2014.

[118] David Juan, Florencio Pazos, and Alfonso Valencia. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci. U.S.A.*, 105(3):934–939, January 2008.

[119] Pollard K. S., Salama S. R., N. Lambert, Lambot M. A., S. Coppens, Pedersen J. S., S. Katzman, B. King, C. Onodera, A. Siepel, Kern A. D., C. Dehay, H. Igel, M. Ares, Jr., P. Vanderhaeghen, and D. Haussler. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, 443(7108):167–172, August 2006.

[120] J. D. Keene. Biological clocks and the coordination theory of RNA operons and regulons. *Cold Spring Harbor Symp. Quant. Biol.*, 72(157-65.):;, 2007.

[121] W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. The human genome browser at UCSC. *Genome Res.*, 12(6):996–1006, June 2002.

[122] Amir Kheradmand and David S. Zee. Cerebellum and Ocular Motor Control. *Front. Neurol.*, 2:11499, September 2011.

[123] Eun Bae Kim, Xiaodong Fang, Alexey A. Fushan, Zhiyong Huang, Alexei V. Lobanov, Lijuan Han, Stefano M. Marino, Xiaoqing Sun, Anton A. Turanov, Pengcheng Yang, Sun Hee Yim, Xiang Zhao, Marina V. Kasaikina, Nina Stoletzki, Chunfang Peng, Paz Polak, Zhiqiang Xiong, Adam Kiezun, Yabing Zhu, Yuanxin Chen, Gregory V. Kryukov, Qiang Zhang, Leonid Peshkin, Lan Yang, Roderick T. Bronson, Rochelle Buffenstein, Bo Wang, Changlei Han, Qiye Li, Li Chen, Wei Zhao, Shamil R. Sunyaev, Thomas J. Park, Guojie Zhang, Jun Wang, and Vadim N. Gladyshev. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature*, 479:223–227, November 2011.

[124] Robert J. Klein, Caroline Zeiss, Emily Y. Chew, Jen-Yue Tsai, Richard S. Sackler, Chad Haynes, Alice K. Henning, John Paul SanGiovanni, Shrikant M. Mane, Susan T. Mayne, Michael B. Bracken, Frederick L. Ferris, Jurg Ott, Colin Barnstable, and Josephine Hoh. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science (New York, N.Y.)*, 308(5720):385, April 2005.

[125] Gladys Y.-P. Ko. Circadian regulation in the retina: From molecules to network. *Eur. J. Neurosci.*, 51(1):194–216, January 2020.

[126] Dennis Kostka, Tara Friedrich, Alisha K. Holloway, and Katherine S. Pollard. motifDiverge: a model for assessing the statistical significance of gene regulatory motif divergence between two DNA sequences. *Stat. Interface*, 8(4):463–476, 2015.

[127] Dennis Kostka, Melissa J. Hubisz, Adam Siepel, and Katherine S. Pollard. The Role of GC-Biased Gene Conversion in Shaping the Fastest Evolving Regions of the Human Genome. *Mol. Biol. Evol.*, 29(3):1047, March 2012.

[128] Amanda Kowalczyk, Wynn K. Meyer, Raghavendran Partha, Weiguang Mao, Nathan L. Clark, and Maria Chikina. RERconverge: an R package for associating evolutionary rates with convergent traits. *Bioinformatics*, 35(22):4815–4817, November 2019.

[129] Amanda Kowalczyk, Raghavendran Partha, Nathan L Clark, and Maria Chikina. Pan-mammalian analysis of molecular constraints underlying extended lifespan. *eLife*, 9:e51089, feb 2020.

[130] Gregory V. Kryukov, Len A. Pennacchio, and Shamil R. Sunyaev. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.*, 80(4):727–739, April 2007.

[131] Robert M. Kuhn, David Haussler, and W. James Kent. The UCSC genome browser and associated tools. *Briefings Bioinf.*, 14(2):144–161, March 2013.

[132] Ivan V. Kulakovskiy, Ilya E. Vorontsov, Ivan S. Yevshin, Ruslan N. Sharipov, Alla D. Fedorova, Eugene I. Rumynskiy, Yulia A. Medvedeva, Arturo Magana-Mora, Vladimir B. Bajic, Dmitry A. Papatsenko, Fedor A. Kolpakov, and Vsevolod J. Makeev. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, 46(D1):252–259, January 2018.

[133] Elena Kulinskaya. On two-sided p-values for non-symmetric distributions. *arXiv*, October 2008.

[134] Ian Laiker and Nicolás Frankel. Pleiotropic Enhancers are Ubiquitous Regulatory Elements in the Human Genome. *Genome Biol. Evol.*, 14(6):evac071, June 2022.

[135] Björn E Langer, Juliana G Roscito, and Michael Hiller. REforge Associates Transcription Factor Binding Site Divergence in Regulatory Elements with Phenotypic Differences between Species. *Molecular Biology and Evolution*, 35(12):3027–3040, 09 2018.

[136] Alfonso Lavado and Guillermo Oliver. Prox1 expression patterns in the developing and adult murine brain. *Dev. Dyn.*, 236(2):518–524, February 2007.

[137] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.*, 9(8):e1003118, August 2013.

[138] Jiwon Lee, Yool Lee, Min Joo Lee, Eonyoung Park, Sung Hwan Kang, Chin Ha Chung, Kun Ho Lee, and Kyungjin Kim. Dual modification of BMAL1 by SUMO2/3 and ubiquitin promotes circadian activation of the CLOCK/BMAL1 complex. *Mol. Cell. Biol.*, 28(19):6056–6065, October 2008.

[139] Sei-Jung Lee, Yi Ran No, Duyen T. Dang, Long H. Dang, Vincent W. Yang, Hyunsuk Shim, and C. Chris Yun. Regulation of hypoxia-inducible factor 1α (HIF-1α) by lysophosphatidic acid is dependent on interplay between p53 and Krüppel-like factor 5. *J. Biol. Chem.*, 288(35):25244–25253, August 2013.

[140] Timothy M. Lenton, Tais W. Dahl, Stuart J. Daines, Benjamin J. W. Mills, Kazumi Ozaki, Matthew R. Saltzman, and Philipp Porada. Earliest land plants created modern levels of atmospheric oxygen. *Proc. Natl. Acad. Sci. U.S.A.*, 113(35):9704–9709, August 2016.

[141] Mingzhou Li, Shilin Tian, Long Jin, Guangyu Zhou, Ying Li, Yuan Zhang, Tao Wang, Carol K. L. Yeung, Lei Chen, Jideng Ma, Jinbo Zhang, Anan Jiang, Ji Li, Chaowei Zhou, Jie Zhang, Yingkai Liu, Xiaoqing Sun, Hongwei Zhao, Zexiong Niu, Pinger Lou, Lingjin Xian, Xiaoyong Shen, Shaoqing Liu, Shunhua Zhang, Mingwang Zhang, Li Zhu, Surong Shuai, Lin Bai, Guoqing Tang, Haifeng Liu, Yanzhi Jiang, Miaomiao Mai, Jian Xiao, Xun Wang, Qi Zhou, Zhiquan Wang, Paul Stothard, Ming Xue, Xiaolian Gao, Zonggang Luo, Yiren Gu, Hongmei Zhu, Xiaoxiang Hu, Yaofeng Zhao, Graham S. Plastow, Jinyong Wang, Zhi Jiang, Kui Li, Ning Li, Xuewei Li, and Ruiqiang Li. Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat. Genet.*, 45(12):1431–1438, December 2013.

[142] Yan Li, Dong-Dong Wu, Adam R. Boyko, Guo-Dong Wang, Shi-Fang Wu, David M. Irwin, and Ya-Ping Zhang. Population Variation Revealed High-Altitude Adaptation of Tibetan Mastiffs. *Mol. Biol. Evol.*, 31(5):1200–1205, May 2014.

[143] Chang Liu, Siming Li, Tiecheng Liu, Jimo Borjigin, and Jiandie D. Lin. Transcriptional coactivator PGC-1alpha integrates the mammalian clock and energy metabolism. *Nature*, 447(7143):477–481, May 2007.

[144] Chuanyu Liu, Mingyue Wang, Xiaoyu Wei, Liang Wu, Jiangshan Xu, Xi Dai, Jun Xia, Mengnan Cheng, Yue Yuan, Pengfan Zhang, Jiguang Li, Taiqing Feng, Ao Chen, Wenwei Zhang, Fang Chen, Zhouchun Shang, Xiuqing Zhang, Brock A. Peters, and Longqi Liu. An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Sci. Data*, 6(65):1–10, May 2019.

[145] Hannah K. Long, Sara L. Prescott, and Joanna Wysocka. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell*, 167(5):1170–1187, November 2016.

[146] Ari Löytynoja. Phylogeny-aware alignment with PRANK. In *Multiple Sequence Alignment Methods*, pages 155–170. Humana Press, 2014.

[147] J. Yuyang Lu, Matthew Simon, Yang Zhao, Julia Ablaeva, Nancy Corson, Yongwook Choi, KayLene Y.H. Yamada, Nicholas J. Schork, Wendy R. Hood, Geoffrey E. Hill, Richard A. Miller, Andrei Seluanov, and Vera Gorbunova. Comparative transcriptomics reveals circadian and pluripotency networks as two pillars of longevity regulation. *Cell Metabolism*, 34(6):836–856.e5, 2022.

[148] Ying Lu, Jin Liu, Yang Liu, Yaru Qin, Qun Luo, Quanli Wang, and Haifeng Duan. TLR4 plays a crucial role in MSC-induced inhibition of NK cell function. *Biochem. Biophys. Res. Commun.*, 464(2):541–547, August 2015.

[149] Dandan Luo, Zhao He, Chunxiao Yu, and Qingbo Guan. Role of p38 MAPK Signalling in Testis Development and Male Fertility. *Oxid. Med. Cell. Longevity*, 2022, 2022.

[150] Michael Lynch. Methods for the Analysis of Comparative Data in Evolutionary Biology on JSTOR. *Evolution*, 45(5):1065–1080, August 1991.

[151] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, May 2015.

[152] Ian J. Majewski, Matthew E. Ritchie, Belinda Phipson, Jason Corbin, Miha Pakusch, Anja Ebert, Meinrad Busslinger, Haruhiko Koseki, Yifang Hu, Gordon K. Smyth, Warren S. Alexander, Douglas J. Hilton, and Marnie E. Blewitt. Opposing roles of polycomb repressive complexes in hematopoietic stem and progenitor cells. *Blood*, 116(5):731–739, August 2010.

[153] Saba Manshaei, Thea Willis, Dominic Withers, Jesus Gil, Cynthia Lilian Andoniadou, and Juan Pedro Martinez-Barbera. Paracrine Signalling From SOX2-Expressing Pituitary Embryonic Cells Is Required for Terminal Differentiation of Hormone-Producing Cells. *J. Endocr. Soc.*, 5(Supplement_1):A547–A548, May 2021.

[154] Saba Manshaei, Thea L. Willis, Virinder Reen, Husayn Pallikonda, Jodie Birch, Dominic J. Withers, Jesus Gil, Cynthia L. Andoniadou, and Juan Pedro Martinez-Barbera. RF13 | PMON143 BRF1-Mediated Paracrine Signalling by a Subset of SOX2-Expressing Stem Cells is Required for Normal Development of the Stem Cell Compartment and Terminal Differentiation of Pituitary Committed Progenitors. *J. Endocr. Soc.*, 6(Supplement_1):A580–A581, December 2022.

[155] Amir Marcovitz, Robin Jia, and Gill Bejerano. "Reverse Genomics" Predicts Function of Human Conserved Noncoding Elements. *Mol. Biol. Evol.*, 33(5):1358–1369, May 2016.

[156] El Martins. Adaptation and the comparative method. *Trends Ecol. Evol.*, 15(7):296–299, July 2000.

[157] Hassan Marzban, Nathan Hoy, Tooka Aavani, Diana K. Sarko, Kenneth C. Catania, and Richard Hawkes. Compartmentation of the cerebellar cortex in the naked mole-rat (Heterocephalus glaber). *Cerebellum*, 10(3):435–448, September 2011.

[158] Benjamin Mayne, Oliver Berry, Campbell Davies, Jessica Farley, and Simon Jarman. A genomic predictor of lifespan in vertebrates. *Sci. Rep.*, 9(17866):1–10, December 2019.

[159] E. V. Maytin and J. F. Habener. Transcription factors C/EBP alpha, C/EBP beta, and CHOP (Gadd153) expressed during the differentiation program of keratinocytes in vitro and in vivo. *J. Invest. Dermatol.*, 110(3):238–246, March 1998.

[160] Elizabeth S. Maywood, Johanna E. Chesham, John A. O'Brien, and Michael H. Hastings. A diversity of paracrine signals sustains molecular circadian cycling in suprachiasmatic nucleus circuits. *Proc. Natl. Acad. Sci. U.S.A.*, 108(34):14306–14311, August 2011.

[161] Adam T. McLain and Christopher Faulk. The evolution of CpG density and lifespan in conserved primate and mammalian promoters. *Aging (Albany NY)*, 10(4):561, April 2018.

[162] Cory Y. McLean, Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, 28:495–501, May 2010.

[163] Eóin N. McNamee, Joanne C. Masterson, Paul Jedlicka, Martine McManus, Almut Grenz, Colm B. Collins, Marcel F. Nold, Claudia Nold-Petry, Philip Bufler, Charles A. Dinarello, and Jesús Rivera-Nieves. Interleukin 37 expression protects mice from colitis. *Proc. Natl. Acad. Sci. U.S.A.*, 108(40):16711–16716, October 2011.

[164] Lin Mei and Klaus-Armin Nave. Neuregulin-ERBB signaling in the nervous system and neuropsychiatric diseases. *Neuron*, 83(1):27–49, July 2014.

[165] Robert W. Meredith, Jan E. Janečka, John Gatesy, Oliver A. Ryder, Colleen A. Fisher, Emma C. Teeling, Alisha Goodbla, Eduardo Eizirik, Taiz L. L. Simão, Tanja Stadler, Daniel L. Rabosky, Rodney L. Honeycutt, John J. Flynn, Colleen M. Ingram, Cynthia Steiner, Tiffani L. Williams, Terence J. Robinson, Angela Burk-Herrick, Michael Westerman, Nadia A. Ayoub, Mark S. Springer, and William J. Murphy. Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. *Science*, 334(6055):521–524, October 2011.

[166] Wynn K. Meyer, Jerrica Jamison, Rebecca Richter, Stacy E. Woods, Raghavendran Partha, Amanda Kowalczyk, Charles Kronk, Maria Chikina, Robert K. Bonde, Daniel E. Crocker, Joseph Gaspard, Janet M. Lanyon, Judit Marsillach, Clement E. Furlong, and Nathan L. Clark. Ancient convergent losses of Paraoxonase 1 yield potential risks for modern marine mammals. *Science*, 361(6402):591–594, August 2018.

[167] Aashiq H. Mirza, Simranjeet Kaur, Caroline A. Brorsson, and Flemming Pociot. Effects of GWAS-Associated Genetic Variants on lncRNAs within IBD and T1D Candidate Loci. *PLoS One*, 9(8):e105723, August 2014.

[168] Germán Montoya-Sanhueza and Anusuya Chinsamy. Cortical bone adaptation and mineral mobilization in the subterranean mammal Bathyergus suillus (Rodentia: Bathyergidae): effects of age and sex. *PeerJ*, 6:e4944, June 2018.

[169] L. G. Moore, M. Shriver, L. Bemis, B. Hickler, M. Wilson, T. Brutsaert, E. Parra, and E. Vargas. Maternal Adaptation to High-altitude Pregnancy: An Experiment of Nature—A Review. *Placenta*, 25:S60–S71, April 2004.

[170] Damian Moran, Rowan Softley, and Eric J. Warrant. The energetic cost of vision and the evolution of eyeless Mexican cavefish. *Sci. Adv.*, 1(8):e1500363., September 2015.

[171] Alan M. Moses, Daniel A. Pollard, David A. Nix, Venky N. Iyer, Xiao-Yong Li, Mark D. Biggin, and Michael B. Eisen. Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput. Biol.*, 2(10):e130., October 2006.

[172] Gerard Muntané, Xavier Farré, Juan Antonio Rodríguez, Cinta Pegueroles, David A. Hughes, João Pedro de Magalhães, Toni Gabaldón, and Arcadi Navarro. Biological Processes Modulating Longevity across Primates: A Phylogenetic Genome-Phenome Analysis. *Mol. Biol. Evol.*, 35(8):1990–2004, August 2018.

[173] Miklós Müller, Marek Mentel, Jaap J. van Hellemond, Katrin Henze, Christian Woehle, Sven B. Gould, Re-Young Yu, Mark van der Giezen, Aloysius G. M. Tielens,

and William F. Martin. Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol. Mol. Biol. Rev.*, 76(2):444–495, June 2012.

[174] Chandrasekhar Natarajan, Noriko Inoguchi, Roy E. Weber, Angela Fago, Hideaki Moriyama, and Jay F. Storz. Epistasis among adaptive mutations in deer mouse hemoglobin. *Science*, 340(6138):1324–1327, June 2013.

[175] Matthew R. Nelson, Hannah Tipney, Jeffery L. Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, Pak Chung Sham, Mulin Jun Li, Junwen Wang, Lon R. Cardon, John C. Whittaker, and Philippe Sanseau. The support of human genetic evidence for approved drug indications. *Nat. Genet.*, 47(8):856–860, August 2015.

[176] Ella Preger-Ben Noon, Gonzalo Sabarís, Daniela M. Ortiz, Jonathan Sager, Anna Liebowitz, David L. Stern, and Nicolás Frankel. Comprehensive Analysis of a cis-Regulatory Region Reveals Pleiotropy in Enhancer Function. *Cell Rep.*, 22(11):3021–3031, March 2018.

[177] Jackie L. Norrie, Marybeth S. Lupo, Beisi Xu, Issam Al Diri, Marc Valentine, Daniel Putnam, Lyra Griffiths, Jiakun Zhang, Dianna Johnson, John Easton, Ying Shao, Victoria Honnell, Sharon Frase, Shondra Miller, Valerie Stewart, Xin Zhou, Xiang Chen, and Michael A. Dyer. Nucleome Dynamics during Retinal Development. *Neuron*, 104(3):512–52811, November 2019.

[178] Gherman Novakovsky, Oriol Fornes, Manu Saraswat, Sara Mostafavi, and Wyeth W. Wasserman. ExplaiNN: interpretable and transparent neural networks for genomics. *Genome Biol.*, 24(1):1–24, December 2023.

[179] David Ochoa and Florencio Pazos. Practical aspects of protein co-evolution. *Front. Cell Dev. Biol.*, 2:87573, April 2014.

[180] Emily A. O'Connor and Charlie K. Cornwallis. Immunity and lifespan: answering long-standing questions with comparative genomics. *Trends Genet.*, 38(7):650–661, July 2022.

[181] G. Oliver, B. Sosa-Pineda, S. Geisendorf, E. P. Spana, C. Q. Doe, and P. Gruss. Prox 1, a prospero-related homeobox gene expressed during mouse development. *Mech. Dev.*, 44(1):3–16, November 1993.

[182] Benson Otarigho and Alejandro Aballay. Immunity-longevity tradeoff neurally controlled by GABAergic transcription factor PITX1/UNC-30. *Cell Rep.*, 35(8):109187, May 2021.

[183] Ping Ouyang, Kun Wu, Liudan Su, Weifang An, Yanhong Bie, He Zhang, Haixian Kang, Enping Jiang, Wei Zhu, Yunhong Yao, Xinrong Hu, Zhangquan Chen, and Sen Wang. Inhibition of human cervical cancer cell invasion by IL-37 involving runt related transcription factor 2 suppression. *Annals of Translational Medicine*, 7(20), October 2019.

[184] Mark Pagel and Andrew Meade. Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo. *Am. Nat.*, June 2006.

[185] Lavisha Parab, Sampriti Pal, and Riddhiman Dhar. Transcription factor binding process is the primary driver of noise in gene expression. *PLos Genet.*, 18(12):e1010535, December 2022.

[186] Mathilde Paris, Tommy Kaplan, Xiao Yong Li, Jacqueline E. Villalta, Susan E. Lott, and Michael B. Eisen. Extensive Divergence of Transcription Factor Binding in Drosophila Embryos with Highly Conserved Gene Expression. *PLos Genet.*, 9(9):e1003748, September 2013.

[187] Raghavendran Partha, Bharesh K. Chauhan, Zelia Ferreira, Joseph D. Robinson, Kira Lathrop, Ken K. Nischal, Maria Chikina, and Nathan L. Clark. Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *eLife*, October 2017.

[188] Raghavendran Partha, Amanda Kowalczyk, Nathan L. Clark, and Maria Chikina. Robust Method for Detecting Convergent Shifts in Evolutionary Rates. *Mol. Biol. Evol.*, 36(8):1817–1830, August 2019.

[189] Rupali P. Patwardhan, Joseph B. Hiatt, Daniela M. Witten, Mee J. Kim, Robin P. Smith, Dalit May, Choli Lee, Jennifer M. Andrie, Su-In Lee, Gregory M. Cooper, Nadav Ahituv, Len A. Pennacchio, and Jay Shendure. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.*, 30:265–270, March 2012.

[190] Katherine S. Pollard, Melissa J. Hubisz, Kate R. Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, 20(1):110–121, January 2010.

[191] Katherine S. Pollard, Sofie R. Salama, Bryan King, Andrew D. Kern, Tim Dreszer, Sol Katzman, Adam Siepel, Jakob S. Pedersen, Gill Bejerano, Robert Baertsch, Kate R. Rosenbloom, Jim Kent, and David Haussler. Forces Shaping the Fastest Evolving Regions in the Human Genome. *PLos Genet.*, 2(10), October 2006.

[192] Christine D. Pozniak, Abraham J. Langseth, Gerrit J. P. Dijkgraaf, Youngshik Choe, Zena Werb, and Samuel J. Pleasure. Sox10 directs neural stem cells toward the oligodendrocyte lineage by decreasing Suppressor of Fused expression. *Proc. Natl. Acad. Sci. U.S.A.*, 107(50):21795, December 2010.

[193] Daniela Praher, Bob Zimmermann, Rohit Dnyansagar, David J. Miller, Aurelie Moya, Vengamanaidu Modepalli, Arie Fridrich, Daniel Sher, Lene Friis-Møller, Per Sundberg, Sylvain Fôret, Regan Ashby, Yehu Moran, and Ulrich Technau. Conservation and turnover of miRNAs and their highly complementary targets in early branching animals. *Proceedings of the Royal Society B: Biological Sciences*, 288(1945), February 2021.

[194] Joana Projecto-Garcia, Chandrasekhar Natarajan, Hideaki Moriyama, Roy E. Weber, Angela Fago, Zachary A. Cheviron, Robert Dudley, Jimmy A. McGuire, Christopher C. Witt, and Jay F. Storz. Repeated elevational transitions in hemoglobin function during the evolution of Andean hummingbirds. *Proc. Natl. Acad. Sci. U.S.A.*, 110(51):20669–20674, December 2013.

[195] Xavier Prudent, Genis Parra, Peter Schwede, Juliana G. Roscito, and Michael Hiller. Controlling for Phylogenetic Relatedness and Evolutionary Rates Improves the Discovery of Associations Between Species' Phenotypic and Genomic Differences. *Mol. Biol. Evol.*, 33(8):2135–2150, August 2016.

[196] Xuebin Qi, Qu Zhang, Yaoxi He, Lixin Yang, Xiaoming Zhang, Peng Shi, Linping Yang, Zhengheng Liu, Fuheng Zhang, Fengyun Liu, Shiming Liu, Tianyi Wu, Chaoying Cui, Ouzhuluobu, Caijuan Bai, Baimakangzhuo, Jianlin Han, Shengguo Zhao, Chunnian Liang, and Bing Su. The Transcriptomic Landscape of Yaks Reveals Molecular Pathways for High Altitude Adaptation. *Genome Biol. Evol.*, 11(1):72–85, January 2019.

[197] Qiang Qiu, Guojie Zhang, Tao Ma, Wubin Qian, Junyi Wang, Zhiqiang Ye, Changchang Cao, Quanjun Hu, Jaebum Kim, Denis M. Larkin, Loretta Auvil, Boris Capitanu, Jian Ma, Harris A. Lewin, Xiaoju Qian, Yongshan Lang, Ran Zhou, Lizhong Wang, Kun Wang, Jinquan Xia, Shengguang Liao, Shengkai Pan, Xu Lu, Haolong Hou, Yan Wang, Xuetao Zang, Ye Yin, Hui Ma, Jian Zhang, Zhaofeng Wang, Yingmei Zhang, Dawei Zhang, Takahiro Yonezawa, Masami Hasegawa, Yang Zhong, Wenbin Liu, Yan Zhang, Zhiyong Huang, Shengxiang Zhang, Ruijun Long, Huanming Yang,

Jian Wang, Johannes A. Lenstra, David N. Cooper, Yi Wu, Jun Wang, Peng Shi, Jian Wang, and Jianquan Liu. The yak genome and adaptation to life at high altitude. *Nat. Genet.*, 44:946–949, August 2012.

[198] Víctor Quesada, Sandra Freitas-Rodríguez, Joshua Miller, José G. Pérez-Silva, Zi-Feng Jiang, Washington Tapia, Olaya Santiago-Fernández, Diana Campos-Iglesias, Lukas F. K. Kuderna, Maud Quinzin, Miguel G. Álvarez, Dido Carrero, Luciano B. Beheregaray, James P. Gibbs, Ylenia Chiari, Scott Glaberman, Claudio Ciofi, Miguel Araujo-Voces, Pablo Mayoral, Javier R. Arango, Isaac Tamargo-Gómez, David Roiz-Valle, María Pascual-Torner, Benjamin R. Evans, Danielle L. Edwards, Ryan C. Garrick, Michael A. Russello, Nikos Poulakakis, Stephen J. Gaughran, Danny O. Rueda, Gabriel Bretones, Tomàs Marquès-Bonet, Kevin P. White, Adalgisa Caccone, and Carlos López-Otín. Giant tortoise genomes provide insights into longevity and age-related disease. *Nat. Ecol. Evol.*, 3:87–95, January 2019.

[199] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010.

[200] Elaine W. Raines. The extracellular matrix can regulate vascular cell migration, proliferation, and survival: relationships to vascular disease. *Int. J. Exp. Path.*, 81(3):173, June 2000.

[201] Andrea I. Ramos and Scott Barolo. Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1632), December 2013.

[202] Jason Raymond and Daniel Segreé. The Effect of Oxygen on Biochemical Networks and the Evolution of Complex Life. *Science*, 311(5768):1764–1767, March 2006.

[203] Stacey L. Reeber, Tom S. Otis, and Roy Vincent Sillitoe. New roles for the cerebellum in health and disease. *Front. Syst. Neurosci.*, 7:66786, November 2013.

[204] Peter R. Rich and Amandine Maréchal. The mitochondrial respiratory chain. *Essays Biochem.*, 47(1-23.):;, 2010.

[205] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, 43(7):e47., April 2015.

[206] Cosette M. Rivera-Cruz, Joseph J. Shearer, Manoel Figueiredo Neto, and Marxa L. Figueiredo. The Immunomodulatory Effects of Mesenchymal Stem Cell Polarization within the Tumor Microenvironment Niche. *Stem Cells International*, 2017, 2017.

[207] Jonathan Romiguier and Camille Roux. Analytical Biases Associated with GC-Content in Molecular Evolution. *Front. Genet.*, 8:246001, February 2017.

[208] Juliana G. Roscito, Katrin Sameith, Genis Parra, Bjoern E. Langer, Andreas Petzold, Claudia Moebius, Marc Bickle, Miguel Trefaut Rodrigues, and Michael Hiller. Phenotype loss is associated with widespread divergence of the gene regulatory landscape in evolution. *Nat. Commun.*, 9(4737):1–15, November 2018.

[209] Magali Saint-Geniez, Aihua Jiang, Stephanie Abend, Laura Liu, Harry Sweigard, Kip M. Connor, and Zoltan Arany. PGC-1α regulates normal and pathological angiogenesis in the retina. *Am. J. Pathol.*, 182(1):255–265, January 2013.

[210] Manabu Sakamoto and Chris Venditti. Phylogenetic non-independence in rates of trait evolution. *Biol. Lett.*, 14(10):20180502, October 2018.

[211] S. Sanyal, H. G. Jansen, W. J. de Grip, E. Nevo, and W. W. de Jong. The eye of the blind mole rat, Spalax ehrenbergi. Rudiment with hidden function? *Invest. Ophthalmol. Visual Sci.*, 31(7):1398–1404, July 1990.

[212] Elysia Saputra, Amanda Kowalczyk, Luisa Cusick, Nathan Clark, and Maria Chikina. Phylogenetic Permulations: A Statistically Rigorous Approach to Measure Confidence in Associations in a Phylogenetic Context. *Molecular Biology and Evolution*, 38(7):3004–3021, 03 2021.

[213] Elysia Saputra, Weiguang Mao, Nathan Clark, and Maria Chikina. Prediction of local convergent shifts in evolutionary rates with phyloConverge characterizes the phenotypic associations and modularity of regulatory elements. *bioRxiv*, page 2022.05.02.490345, May 2022.

[214] Diana K. Sarko, Duncan B. Leitch, and Kenneth C. Catania. Cutaneous and periodontal inputs to the cerebellum of the naked mole-rat (Heterocephalus glaber). *Front. Neuroanat.*, 7:64791, November 2013.

[215] Daniel J. Schaid, Wenan Chen, and Nicholas B. Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.*, 19(8):491, August 2018.

[216] Laura B. Scheinfeldt, Sameer Soi, Simon Thompson, Alessia Ranciaro, Dawit Woldemeskel, William Beggs, Charla Lambert, Joseph P. Jarvis, Dawit Abate, Gurja Belay, and Sarah A. Tishkoff. Genetic adaptation to high altitude in the Ethiopian highlands. *Genome Biol.*, 13(1):1–9, January 2012.

[217] Schizophrenia Working Group of the Psychiatric Genomics Consortium, Stephan Ripke, Benjamin M. Neale, Aiden Corvin, James T. R. Walters, Kai-How Farh, Peter A. Holmans, Phil Lee, Brendan Bulik-Sullivan, David A. Collier, Hailiang Huang, Tune H. Pers, Ingrid Agartz, Esben Agerbo, Margot Albus, Madeline Alexander, Farooq Amin, Silviu A. Bacanu, Martin Begemann, Richard A. Belliveau, Jr., Judit Bene, Sarah E. Bergen, Elizabeth Bevilacqua, Tim B. Bigdeli, Donald W. Black, Richard Bruggeman, Nancy G. Buccola, Randy L. Buckner, William Byerley, Wiepke Cahn, Guiqing Cai, Dominique Campion, Rita M. Cantor, Vaughan J. Carr, Noa Carrera, Stanley V. Catts, Kimberley D. Chambert, Raymond C. K. Chan, Ronald Y. L. Chan, Eric Y. H. Chen, Wei Cheng, Eric F. C. Cheung, Siow Ann Chong, C. Robert Cloninger, David Cohen, Nadine Cohen, Paul Cormican, Nick Craddock, James J. Crowley, David Curtis, Michael Davidson, Kenneth L. Davis, Franziska Degenhardt, Jurgen Del Favero, Ditte Demontis, Dimitris Dikeos, Timothy Dinan, Srdjan Djurovic, Gary Donohoe, Elodie Drapeau, Jubao Duan, Frank Dudbridge, Naser Durmishi, Peter Eichhammer, Johan Eriksson, Valentina Escott-Price, Laurent Essioux, Ayman H. Fanous, Martilias S. Farrell, Josef Frank, Lude Franke, Robert Freedman, Nelson B. Freimer, Marion Friedl, Joseph I. Friedman, Menachem Fromer, Giulio Genovese, Lyudmila Georgieva, Ina Giegling, Paola Giusti-Rodríguez, Stephanie Godard, Jacqueline I. Goldstein, Vera Golimbet, Srihari Gopal, Jacob Gratten, Lieuwe de Haan, Christian Hammer, Marian L. Hamshere, Mark Hansen, Thomas Hansen, Vahram Haroutunian, Annette M. Hartmann, Frans A. Henskens, Stefan Herms, Joel N. Hirschhorn, Per Hoffmann, Andrea Hofman, Mads V. Hollegaard, David M. Hougaard, Masashi Ikeda, Inge Joa, Antonio Julià, René S. Kahn, Luba Kalaydjieva, Sena Karachanak-Yankova, Juha Karjalainen, David Kavanagh, Matthew C. Keller, James L. Kennedy, Andrey Khrunin, Yunjung Kim, Janis Klovins, James A. Knowles, Bettina Konte, Vaidutis Kucinskas, Zita Ausrele Kucinskiene, Hana Kuzelova-Ptackova, Anna K. Kähler, Claudine Laurent, Jimmy Lee, S. Hong Lee, Sophie E. Legge, Bernard Lerer, Miaoxin Li, Tao Li, Kung-Yee Liang, Jeffrey Lieberman, Svetlana Limborska, Carmel M. Loughland, Jan Lubinski, Jouko Lönnqvist, Milan Macek, Patrik K. E. Magnusson, Brion S. Maher, Wolfgang Maier, Jacques Mallet, Sara Marsal, Manuel Mattheisen, Morten Mattingsdal, Robert W. McCarley, Colm McDonald, Andrew M. McIntosh, Sandra Meier, Carin J. Meijer, Bela Melegh, Ingrid Melle, Raquelle I. Mesholam-Gately, Andres Metspalu, Patricia T. Michie, Lili Milani, Vihra Milanova, Younes Mokrab, Derek W. Morris, Ole Mors, Kieran C. Murphy, Robin M. Murray, Inez Myin-Germeys, Bertram Müller-Myhsok, Mari Nelis, Igor Nenadic, Deborah A. Nertney, Gerald Nestadt, Kristin K. Nicodemus, Liene Nikitina-Zake, Laura Nisenbaum, Annelie Nordin, Eadbhard O'Callaghan, Colm O'Dushlaine, F. Anthony O'Neill, Sang-Yun Oh, Ann Olincy, Line Olsen, Jim Van Os, Psychosis Endophenotypes International Consortium, Chris-

tos Pantelis, George N. Papadimitriou, Sergi Papiol, Elena Parkhomenko, Michele T. Pato, Tiina Paunio, Milica Pejovic-Milovancevic, Diana O. Perkins, Olli Pietiläinen, Jonathan Pimm, Andrew J. Pocklington, John Powell, Alkes Price, Ann E. Pulver, Shaun M. Purcell, Digby Quested, Henrik B. Rasmussen, Abraham Reichenberg, Mark A. Reimers, Alexander L. Richards, Joshua L. Roffman, Panos Roussos, Douglas M. Ruderfer, Veikko Salomaa, Alan R. Sanders, Ulrich Schall, Christian R. Schubert, Thomas G. Schulze, Sibylle G. Schwab, Edward M. Scolnick, Rodney J. Scott, Larry J. Seidman, Jianxin Shi, Engilbert Sigurdsson, Teimuraz Silagadze, Jeremy M. Silverman, Kang Sim, Petr Slominsky, Jordan W. Smoller, Hon-Cheong So, Chris C. A. Spencer, Eli A. Stahl, Hreinn Stefansson, Stacy Steinberg, Elisabeth Stogmann, Richard E. Straub, Eric Strengman, Jana Strohmaier, T. Scott Stroup, Mythily Subramaniam, Jaana Suvisaari, Dragan M. Svrakic, Jin P. Szatkiewicz, Erik Söderman, Srinivas Thirumalai, Draga Toncheva, Sarah Tosato, Juha Veijola, John Waddington, Dermot Walsh, Dai Wang, Qiang Wang, Bradley T. Webb, Mark Weiser, Dieter B. Wildenauer, Nigel M. Williams, Stephanie Williams, Stephanie H. Witt, Aaron R. Wolen, Emily H. M. Wong, Brandon K. Wormley, Hualin Simon Xi, Clement C. Zai, Xuebin Zheng, Fritz Zimprich, Naomi R. Wray, Kari Stefansson, Peter M. Visscher, Wellcome Trust Case-Control Consortium 2, Rolf Adolfsson, Ole A. Andreassen, Douglas H. R. Blackwood, Elvira Bramon, Joseph D. Buxbaum, Anders D. Børglum, Sven Cichon, Ariel Darvasi, Enrico Domenici, Hannelore Ehrenreich, Tõnu Esko, Pablo V. Gejman, Michael Gill, Hugh Gurling, Christina M. Hultman, Nakao Iwata, Assen V. Jablensky, Erik G. Jönsson, Kenneth S. Kendler, George Kirov, Jo Knight, Todd Lencz, Douglas F. Levinson, Qingqin S. Li, Jianjun Liu, Anil K. Malhotra, Steven A. McCarroll, Andrew McQuillin, Jennifer L. Moran, Preben B. Mortensen, Bryan J. Mowry, Markus M. Nöthen, Roel A. Ophoff, Michael J. Owen, Aarno Palotie, Carlos N. Pato, Tracey L. Petryshen, Danielle Posthuma, Marcella Rietschel, Brien P. Riley, Dan Rujescu, Pak C. Sham, Pamela Sklar, David St Clair, Daniel R. Weinberger, Jens R. Wendland, Thomas Werge, Mark J. Daly, Patrick F. Sullivan, and Michael C. O'Donovan. Biological Insights From 108 Schizophrenia-Associated Genetic Loci. *Nature*, 511(7510):421, July 2014.

[218] Klaus Peter Schliep. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, February 2011.

[219] Jeremy D. Schmahmann. The role of the cerebellum in cognition and emotion: personal reflections since 1982 on the dysmetria of thought hypothesis, and its historical evolution from theory to therapy. *Neuropsychol. Rev.*, 20(3):236–260, September 2010.

[220] Dominic Schmidt, Michael D. Wilson, Benoit Ballester, Petra C. Schwalie, Gordon D. Brown, Aileen Marshall, Claudia Kutter, Stephen Watt, Celia P. Martinez-Jimenez, Sarah Mackay, Iannis Talianidis, Paul Flicek, and Duncan T. Odom. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, 328(5981):1036–1040, May 2010.

[221] H. Scholz, H. J. Schurek, K. U. Eckardt, and C. Bauer. Role of erythropoietin in adaptation to hypoxia. *Experientia*, 46(11-12):1197–1201, December 1990.

[222] I. N. Sergeev, R. Buffenstein, and J. M. Pettifor. Vitamin D receptors in a naturally vitamin D-deficient subterranean mammal, the naked mole rat (Heterocephalus glaber): biochemical characterization. *Gen. Comp. Endocrinol.*, 90(3):338–345, June 1993.

[223] Esther Serrano-Saiz, Burcu Gulez, Laura Pereira, Marie Gendrel, Sze Yen Kerk, Berta Vidal, Weidong Feng, Chen Wang, Paschalis Kratsios, James B. Rand, and Oliver Hobert. Modular Organization of Cis-regulatory Control Information of Neurotransmitter Pathway Genes in Caenorhabditis elegans. *Genetics*, 215(3):665–681, July 2020.

[224] Dror Sharon, Hiroyuki Yamamoto, Terri L. McGee, Vivian Rabe, Robert T. Szerencsei, Robert J. Winkfein, Clemens F. M. Prinsen, Claire S. Barnes, Sten Andreasson, Gerald A. Fishman, Paul P. M. Schnetkamp, Eliot L. Berson, and Thaddeus P. Dryja. Mutated alleles of the rod and cone Na-Ca+K-exchanger genes in patients with retinal diseases. *Invest. Ophthalmol. Visual Sci.*, 43(6):1971–1979, June 2002.

[225] Karthik Shekhar, Sylvain W. Lapan, Irene E. Whitney, Nicholas M. Tran, Evan Z. Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z. Levin, James Nemesh, Melissa Goldman, Steven A. McCarroll, Constance L. Cepko, Aviv Regev, and Joshua R. Sanes. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell*, 166(5):1308–132330, August 2016.

[226] Xing-Xing Shen, Jacob L. Steenwyk, Abigail L. LaBella, Dana A. Opulente, Xiaofan Zhou, Jacek Kominek, Yuanning Li, Marizeth Groenewald, Chris T. Hittinger, and Antonis Rokas. Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum Ascomycota. *Sci. Adv.*, 6(45):eabd0079, November 2020.

[227] Adam Siepel, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W. Hillier, Stephen Richards, George M. Weinstock, Richard K. Wilson, Richard A. Gibbs, W. James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15(8):1034–1050, August 2005.

[228] Adam Siepel, Katherine S. Pollard, and David Haussler. New Methods for Detecting Lineage-Specific Selection. In *Research in Computational Molecular Biology*, pages 190–205. Springer, Berlin, Germany, 2006.

[229] Tatum S. Simonson, Yingzhong Yang, Chad D. Huff, Haixia Yun, Ga Qin, David J. Witherspoon, Zhenzhong Bai, Felipe R. Lorenzo, Jinchuan Xing, Lynn B. Jorde, Josef T. Prchal, and RiLi Ge. Genetic Evidence for High-Altitude Adaptation in Tibet. *Science*, 329(5987):72–75, July 2010.

[230] Brian Tilston Smith, Iii William M. Mauck, Brett W. Benz, and Michael J. Andersen. Uneven Missing Data Skew Phylogenomic Relationships within the Lories and Lorikeets. *Genome Biol. Evol.*, 12(7):1131, July 2020.

[231] James A. Smythies, Min Sun, Norma Masson, Rafik Salama, Peter D. Simpson, Elizabeth Murray, Viviana Neumann, Matthew E. Cockman, Hani Choudhry, Peter J. Ratcliffe, and David R. Mole. Inherent DNA-binding specificities of the HIF-1α and HIF-2α transcription factors in chromatin. *EMBO Rep.*, 20(1):e46401., January 2019.

[232] Valentina Snetkova, Athena R. Ypsilanti, Jennifer A. Akiyama, Brandon J. Mannion, Ingrid Plajzer-Frick, Catherine S. Novak, Anne N. Harrington, Quan T. Pham, Momoe Kato, Yiwen Zhu, Janeth Godoy, Eman Meky, Riana D. Hunter, Marie Shi, Evgeny Z. Kvon, Veena Afzal, Stella Tran, John L. R. Rubenstein, Axel Visel, Len A. Pennacchio, and Diane E. Dickel. Ultraconserved enhancer function does not require perfect sequence conservation. *Nat. Genet.*, 53:521–528, April 2021.

[233] Shen Song, Na Yao, Min Yang, Xuexue Liu, Kunzhe Dong, Qianjun Zhao, Yabin Pu, Xiaohong He, Weijun Guan, Ning Yang, Yuehui Ma, and Lin Jiang. Exome sequencing reveals genetic differentiation due to high-altitude adaptation in the Tibetan cashmere goat (Capra hircus). *BMC Genomics*, 17(1):1–12, December 2016.

[234] B. Sosa-Pineda, J. T. Wigle, and G. Oliver. Hepatocyte migration during liver development requires Prox1. *Nat. Genet.*, 25(3):254–255, July 2000.

[235] Mikhail Spivakov. Spurious transcription factor binding: non-functional or genetically redundant? *Bioessays*, 36(8):798–806, August 2014.

[236] Graham N. Stone, Sean Nee, and Joseph Felsenstein. Controlling for non-independence in comparative analysis of patterns across populations within species. *Phil. Trans. R. Soc. B*, 366(1569):1410–1424, May 2011.

[237] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.*, 100(16):9440–9445, August 2003.

[238] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck Rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–190221, June 2019.

[239] Tim Stuart, Avi Srivastava, Shaista Madad, Caleb A. Lareau, and Rahul Satija. Single-cell chromatin state analysis with Signac. *Nat. Methods*, 18:1333–1341, November 2021.

[240] A. Subramanian, P. Tamayo, Mootha V. K., S. Mukherjee, Ebert B. L., Gillette M. A., A. Paulovich, Pomeroy S. L., Golub T. R., Lander E. S., and Mesirov J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 102(43):15545–15550, September 2005.

[241] Kazumitsu Sugiura, Yoshinao Muro, Kyoko Futamura, Kenji Matsumoto, Noriko Hashimoto, Yuji Nishizawa, Tetsuro Nagasaka, Hirohisa Saito, Yasushi Tomita, and Jiro Usukura. The unfolded protein response is activated in differentiating epidermal keratinocytes. *J. Invest. Dermatol.*, 129(9):2126–2135, September 2009.

[242] Georgina Sweet. The Eyes of Chrysochloris hottentota and C. asiatica. *J. Cell Sci.*, s2-53(210):327–338, January 1909.

[243] Robi Tacutu, Daniel Thornton, Emily Johnson, Arie Budovsky, Diogo Barardo, Thomas Craig, Eugene Diana, Gilad Lehmann, Dmitri Toren, Jingwei Wang, Vadim E Fraifeld, and João P de Magalhães. Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Research*, 46(D1):D1083–D1090, 11 2017.

[244] W. R. Taylor and R. G. Smith. The role of starburst amacrine cells in visual signal processing. *Visual Neurosci.*, 29(1):73–81, January 2012.

[245] María Touceda-Suárez, Elizabeth M. Kita, Rafael D. Acemel, Panos N. Firbas, Marta S. Magri, Silvia Naranjo, Juan J. Tena, Jose Luis Gómez-Skarmeta, Ignacio Maeso, and Manuel Irimia. Ancient Genomic Regulatory Blocks Are a Source for Regulatory Gene Deserts in Vertebrates after Whole-Genome Duplications. *Mol. Biol. Evol.*, 37(10):2857–2864, October 2020.

[246] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nat. Rev. Methods Primers*, 1(59):1–21, August 2021.

[247] Miguel A'ngel Vargas, Mathieu St-Louis, Luc Desgroseillers, Jean-Louis Charli, and Guy Boileau. Parathyroid Hormone-Related Protein(1–34) Regulates Phex Expression in Osteoblasts through the Protein Kinase A Pathway. *Endocrinology*, 144(11):4876–4885, November 2003.

[248] Jeff Vierstra, John Lazar, Richard Sandstrom, Jessica Halow, Kristen Lee, Daniel Bates, Morgan Diegel, Douglas Dunn, Fidencio Neri, Eric Haugen, Eric Rynes, Alex Reynolds, Jemma Nelson, Audra Johnson, Mark Frerker, Michael Buckley, Rajinder Kaul, Wouter Meuleman, and John A. Stamatoyannopoulos. Global reference mapping of human transcription factor footprints. *Nature*, 583(7818):729–736, Jul 2020.

[249] Diego Villar, Camille Berthelot, Sarah Aldridge, Tim F. Rayner, Margus Lukk, Miguel Pignatelli, Thomas J. Park, Robert Deaville, Jonathan T. Erichsen, Anna J. Jasinska, James M. A. Turner, Mads F. Bertelsen, Elizabeth P. Murchison, Paul Flicek, and Duncan T. Odom. Enhancer Evolution across 20 Mammalian Species. *Cell*, 160(3):554–566, January 2015.

[250] Bajic Vladan, Spremo-Potparevic Biljana, Vesna Mandusic, Milicevic Zorana, and Lada Zivkovic. Instability in X chromosome inactivation patterns in AMD: a new risk factor? *Medical Hypothesis, Discovery and Innovation in Ophthalmology*, 2(3):74, 2013.

[251] Heinz Wässle, Liane Heinze, Elena Ivanova, Sriparna Majumdar, Jan Weiss, Robert J. Harvey, and Silke Haverkamp. Glycinergic transmission in the mammalian retina. *Front. Mol. Neurosci.*, 2:702, July 2009.

[252] Sen Wang, Weifang An, Yunhong Yao, Renhuai Chen, Xiaoxuan Zheng, Wanyong Yang, Yi Zhao, Xinrong Hu, Enping Jiang, Yanhong Bie, Zhangquan Chen, Ping Ouyang, He Zhang, and Hui Xiong. Interleukin 37 Expression Inhibits STAT3 to Suppress the Proliferation and Invasion of Human Cervical Cancer Cells. *J. Cancer*, 6(10):962–969, August 2015.

[253] Tao Wang, Qidi Peng, Bo Liu, Xiaoli Liu, Yongzhuang Liu, Jiajie Peng, and Yadong Wang. eQTLMAPT: Fast and Accurate eQTL Mediation Analysis With Efficient Permutation Testing Approaches. *Front. Genet.*, 10:1309., January 2020.

[254] Wei Wang and Matthew Stephens. Empirical Bayes Matrix Factorization. *arXiv*, February 2018.

[255] Ashley E. Webb, Elizabeth A. Pollina, Thomas Vierbuchen, Noelia Urbán, Duygu Ucar, Dena S. Leeman, Ben Martynoga, Madhavi Sewak, Thomas A. Rando, François

Guillemot, Marius Wernig, and Anne Brunet. FOXO3 shares common targets with ASCL1 genome-wide and inhibits ASCL1-dependent neurogenesis. *Cell Rep.*, 4(3), August 2013.

[256] Joel O. Wertheim, Ben Murrell, Martin D. Smith, Sergei L. Kosakovsky Pond, and Konrad Scheffler. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. *Mol. Biol. Evol.*, 32(3):820–832, March 2015.

[257] J. T. Wigle, K. Chowdhury, P. Gruss, and G. Oliver. Prox1 function is crucial for mouse lens-fibre elongation. *Nat. Genet.*, 21(3):318–322, March 1999.

[258] Emily S. Wong, Dawei Zheng, Siew Z. Tan, Neil I. Bower, Victoria Garside, Gilles Vanwalleghem, Federico Gaiti, Ethan Scott, Benjamin M. Hogan, Kazu Kikuchi, Edwina McGlinn, Mathias Francois, and Bernard M. Degnan. Deep conservation of the enhancer regulatory code in animals. *Science*, 370(6517):eaax8137, November 2020.

[259] Gregory A. Wray. Do Convergent Developmental Mechanisms Underlie Convergent Phenotypes? *Brain Behav. Evol.*, 59(5-6):327–336, July 2002.

[260] Gregory A. Wray. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.*, 8:206–216, March 2007.

[261] Di Wu and Gordon K. Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.*, 40(17):e133, September 2012.

[262] Fuguo Wu, Tadeusz J. Kaczynski, Santhosh Sethuramanujam, Renzhong Li, Varsha Jain, Malcolm Slaughter, and Xiuqian Mu. Two transcription factors, Pou4f2 and Isl1, are sufficient to specify the retinal ganglion cell fate. *Proc. Natl. Acad. Sci. U.S.A.*, 112(13):1559–1568, March 2015.

[263] Fu-Hui Xiao, Qin Yu, Zhi-Li Deng, Ke Yang, Yunshuang Ye, Ming-Xia Ge, Dongjing Yan, Hao-Tian Wang, Xiao-Qiong Chen, Li-Qin Yang, Bin-Yu Yang, Rong Lin, Wen Zhang, Xing-Li Yang, Lei Dong, Yonghan He, Jumin Zhou, Wang-Wei Cai, Ji Li, and Qing-Peng Kong. ETS1 acts as a regulator of human healthy aging via decreasing ribosomal activity. *Sci. Adv.*, 8(17):eabf2017, April 2022.

[264] Shuhua Xu, Shilin Li, Yajun Yang, Jingze Tan, Haiyi Lou, Wenfei Jin, Ling Yang, Xuedong Pan, Jiucun Wang, Yiping Shen, Bailin Wu, Hongyan Wang, and Li Jin. A Genome-Wide Search for Signals of High-Altitude Adaptation in Tibetans. *Mol. Biol. Evol.*, 28(2):1003–1011, February 2011.

[265] Hiroshi Yajima and Kiyoshi Kawakami. Low Six4 and Six5 gene dosage improves dystrophic phenotype and prolongs life span of mdx mice. *Dev. Growth Differ.*, 58(6):546–561, August 2016.

[266] Jian Yang, Zi-Bing Jin, Jie Chen, Xiu-Feng Huang, Xiao-Man Li, Yuan-Bo Liang, Jian-Yang Mao, Xin Chen, Zhili Zheng, Andrew Bakshi, Dong-Dong Zheng, Mei-Qin Zheng, Naomi R. Wray, Peter M. Visscher, Fan Lu, and Jia Qu. Genetic signatures of high-altitude adaptation in Tibetans. *Proc. Natl. Acad. Sci. U.S.A.*, 114(16):4189–4194, April 2017.

[267] Tzu-Hsien Yang. Transcription factor regulatory modules provide the molecular mechanisms for functional redundancy observed among transcription factors in yeast. *BMC Bioinf.*, 20(23):1–16, December 2019.

[268] Ziheng Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, 24(8):1586–1591, August 2007.

[269] Alexander I. Young, Stefania Benonisdottir, Molly Przeworski, and Augustine Kong. Deconstructing the sources of genotype-phenotype associations in humans. *Science (New York, N.Y.)*, 365(6460):1396, September 2019.

[270] Ling Yu, Hongbing Liu, Mingquan Yan, Jing Yang, Fanxin Long, Ken Muneoka, and YiPing Chen. Shox2 is Required for Chondrocyte Proliferation and Maturation in Proximal Limb Skeleton. *Dev. Biol.*, 306(2):549, June 2007.

[271] Yang Yu, Qingyun Zhang, Qinggui Meng, Chen Zong, Lei Liang, Xue Yang, Rui Lin, Yan Liu, Yang Zhou, Hongxiang Zhang, Xiaojuan Hou, Zhipeng Han, and Jiwen Cheng. Mesenchymal stem cells overexpressing Sirt1 inhibit prostate cancer growth by recruiting natural killer cells and macrophages. *Oncotarget*, 7(44):71112–71122, November 2016.

[272] Dinghong Zhang, Bin Yu, Jing Liu, Weiqian Jiang, Taorong Xie, Ran Zhang, Dali Tong, Zilong Qiu, and Haishan Yao. Altered visual cortical processing in a mouse model of MECP2 duplication syndrome. *Sci. Rep.*, 7, 2017.

[273] Hong-Yong Zhang, Jian Li, Ying-Chun Ouyang, Tie-Gang Meng, Chun-Hui Zhang, Wei Yue, Qing-Yuan Sun, and Wei-Ping Qian. Cell Division Cycle 5-Like Regulates Metaphase-to-Anaphase Transition in Meiotic Oocyte. *Front. Cell Dev. Biol.*, 0, 2021.

[274] Xiaoyu Zhang, Irene M. Kaplow, Morgan Wirthlin, Tae Yoon Park, and Andreas R. Pfenning. HALPER facilitates the identification of regulatory element orthologs across species. *Bioinformatics*, 36(15):4339–4340, August 2020.