








Evolutionary sparse learning reveals the shared genetic basis of convergent traits

Received: 30 March 2023

Accepted: 18 March 2025

Published online: 04 April 2025



John B. Allard ^{1,2}, Sudip Sharma ^{1,2}, Ravi Patel ^{1,2}, Maxwell Sanderford^{1,2}, Koichiro Tamura ^{3,4}, Slobodan Vucetic⁵, Glenn S. Gerhard⁶  & Sudhir Kumar ^{1,2} 

Cases abound in which nearly identical traits have appeared in distant species facing similar environments. These unmistakable examples of adaptive evolution offer opportunities to gain insight into their genetic origins and mechanisms through comparative analyses. Here, we present an approach to build genetic models that underlie the independent origins of convergent traits using evolutionary sparse learning with paired species contrast (ESL-PSC). We tested the hypothesis that common genes and sites are involved in the convergent evolution of two key traits: C4 photosynthesis in grasses and echolocation in mammals. Genetic models were highly predictive of independent cases of convergent evolution of C4 photosynthesis. Genes contributing to genetic models for echolocation were highly enriched for functional categories related to hearing, sound perception, and deafness, a pattern that has eluded previous efforts applying standard molecular evolutionary approaches. These results support the involvement of sequence substitutions at common genetic loci in the evolution of convergent traits. Benchmarking on empirical and simulated datasets showed that ESL-PSC could be more sensitive in proteome-scale analyses to detect genes with convergent molecular evolution associated with the acquisition of convergent traits. We conclude that phylogeny-informed machine learning naturally excludes apparent molecular convergences due to shared species history, enhances the signal-to-noise ratio for detecting molecular convergence, and empowers the discovery of common genetic bases of trait convergences.

Organisms continuously adapt to their natural environment. Under similar environmental conditions, the same adaptations may evolve independently in clades across the tree of life. For example, the convergent evolution of the ability to echolocate in some bats and toothed whales is an example of adaptation brought on by transitions to new environments requiring similar physiological innovations. Evolutionary biologists have long sought the common genetic bases of these

convergent adaptations under the hypothesis that they may share the same pathways, genes, and/or base substitutions. However, “the extent to which convergent traits evolve by similar genetic and molecular pathways is not clear”¹. Despite many molecular evolutionary investigations, the strongest evidence for molecular convergence thus far appears to be a marginally significant (FDR-corrected $P = 0.0486$) enrichment of sound perception genes in which convergent and

¹Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA. ²Department of Biology, Temple University, Philadelphia, PA, USA. ³Department of Biological Sciences, Tokyo Metropolitan University, Tokyo, Japan. ⁴Research Center for Genomics and Bioinformatics, Tokyo Metropolitan University, Tokyo, Japan. ⁵Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. ⁶Lewis Katz School of Medicine at Temple University, Philadelphia, PA, USA. ✉e-mail: gsgershard@temple.edu; s.kumar@temple.edu

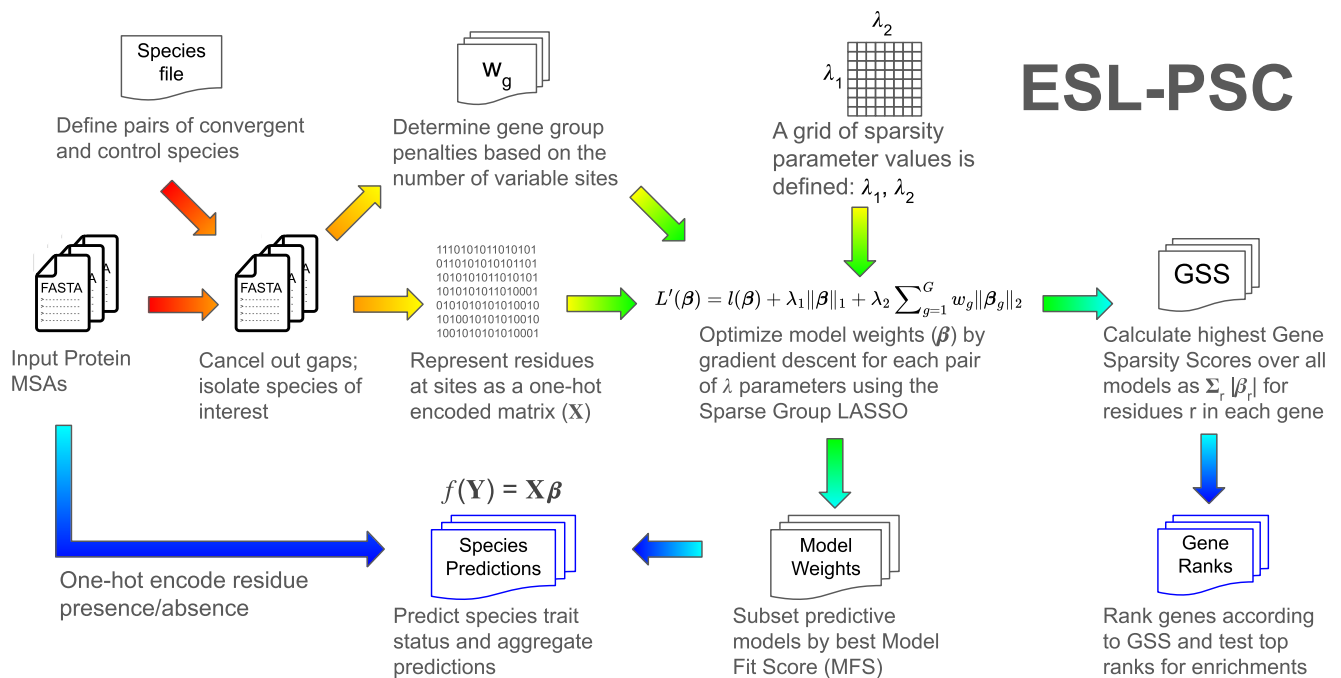


Fig. 1 | The ESL-PSC procedure. The inputs are a set of orthologous protein multiple sequence alignments (MSAs) and a file indicating a set of paired convergent and control species. A single set of species pairs can be given, or alternate species from the same convergent and control clades may be provided, which will result in an ensemble of models using all possible combinations. The outputs (blue) are Species Trait Predictions and Gene Ranks. Species trait predictions include a predicted phenotype in the form of a Sequence Prediction Score for each species in the

input MSA that was not used to build the given model. The Gene Ranks output consists of an ordered list of the input genes (MSAs) according to their Group (gene) Sparsity Scores (GSSs)⁸ that measure the degree to which they inform ESL models of the genetic distinction between the convergent and ancestral trait (see the Methods section). The highest-ranking genes can then be tested for ontology enrichments in order to detect relevant pathways and biological categories that show an abundance of evidence of convergent molecular evolution.

parallel amino acid substitutions were observed^{2–4}. Although these results hint at the possible presence of some shared genetic basis for the evolution of echolocation in independent clades, some studies could not detect such an enrichment³, casting doubt on the robustness of the results, the general applicability of the methodology, and even the presence of a common genetic basis.

The lack of consistent and statistically significant results may be due to insufficient commonality in the genetic bases of these traits, i.e., different genes and different sites may perform similar functions in independent clades. Alternatively, the lack of sufficient statistical power or inability to fully exclude non-adaptive convergence may be hampering efforts to detect genes and sites associated with the evolution of convergent traits^{5–7}. Furthermore, current state-of-the-art approaches primarily reveal retrospective patterns, but they do not produce a quantitative model of genetic changes in convergent trait evolution to make statistical predictions of the presence or absence of the convergent trait.

We have addressed these challenges by building predictive genetic models of convergent trait evolution using evolutionary sparse learning (ESL⁸). ESL is supervised machine learning in which genomic components (e.g., genes and sites) are model parameters and sequence variation among species in multiple sequence alignments are observations⁸. ESL analyses do not estimate evolutionary parameters, such as rates of substitutions between amino acids, variable rates among positions, or branch lengths of the phylogeny. Instead, ESL directly employs the concordance of variation across sequences in an alignment with the presence/absence of the evolutionary trait of interest. ESL model building is actually model selection that incorporates genes and sites with the highest ability to distinguish between two classes of species: those with and without the trait of interest. In ESL, the presence of each possible residue is assigned a weight in the

model. In the optimized model, most sites will have weights of zero, meaning that the solution is sparse. The genes whose sites contribute to the model are considered selected genes. Most genes will have no residues with non-zero weights, so the model will be sparse at the gene level as well. Importantly, ESL simultaneously considers all the genes and sites and their sequence differences among species during computational analysis, obviating the need to set arbitrary evolutionary conservation thresholds and convergent substitution cut-offs necessary in some other approaches^{2,3,7,9,10}.

We developed a paired species contrast (PSC) design to select the data used for ESL with the goal of automatically masking neutral (background) sequence convergence that can lead to spurious inferences and reduce the power to detect the genetic basis of convergence^{5,6,11}. ESL-PSC produces a genetic model to predict the presence/absence of a convergent trait in any species based on its sequence (Fig. 1). This predictive model can test the biological hypothesis of commonality of genetic basis in the independent evolution of the same trait. Lists of loci included in the genetic model can be subjected to additional analyses to test for enrichment of functional categories relevant to the trait^{12,13}, which is a common protocol to establish the biological relevance of candidate loci derived from large-scale scans for molecular convergence in the absence of alternatives^{2–4,11,14–16}.

Here we show that the ESL-PSC approach distinguishes species with and without convergent traits by identifying selected genes and sites that effectively predict traits such as C4 photosynthesis in grasses and echolocation in mammals^{4,17–23}. We demonstrate that this method not only enriches for functional categories relevant to these traits but also outperforms classical approaches when tested on both empirical and simulated datasets. These findings provide compelling evidence that supervised machine learning can quantitatively capture the genetic basis of convergent evolution.

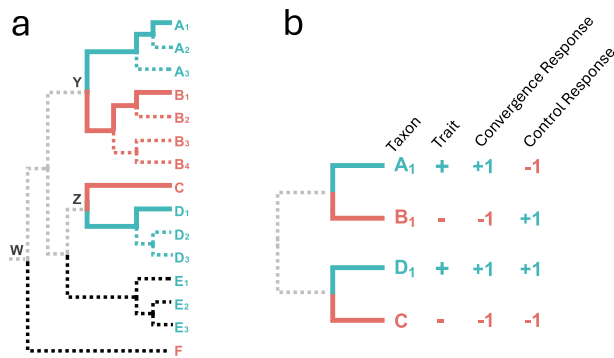


Fig. 2 | The paired species contrast (PSC) design. **a** An example phylogeny with one set of selected species (solid blue and red lines). Extraneous lineages (black dotted lines) and shared evolutionary history (gray dotted lines). **b** A schematic depiction of the four species selected for ESL-PSC analysis. In ESL, the response variable refers to the binary trait, where +1 represents the convergent trait, and -1 represents the ancestral trait.

Results

ESL-PSC for building genetic models of convergent traits

ESL-PSC builds a genetic model to predict trait-positive and trait-negative species, which can be numerically encoded as +1 and -1, respectively^{8,24}. In this analysis, the Least Absolute Shrinkage and Selection Operator (LASSO) compares alternative genetic models by imposing penalties for including amino acid positions and genes in the model while seeking high trait prediction accuracy for the input species. ESL-PSC produces models that incorporate only those proteins whose member sites make a significant contribution to the ability of the genetic model to classify species according to their traits rather than their ancestry.

To build the ESL model, PSC selects a balanced input dataset comprising equal numbers of trait-positive and trait-negative species. Every trait-positive species is paired with a closely related trait-negative species. In PSC, species pairs are required to be from evolutionarily independent clades to avoid introducing correlations among pairs due to shared evolutionary history, which would cause spurious associations^{5,6,11}. As an example, we could select trait-positive species A₁ and D₁ and trait-negative species B₁ and C, respectively, to satisfy the above condition (Fig. 2a).

PSC ensures that the most recent common ancestor (MRCA) of each trait-positive and trait-negative species pair selected will not be an ancestor of any other species in the analysis, nor will it have descended from the MRCA of any other pair. Thus, each pair is evolutionarily independent of all other pairs. In the above example, neither the MRCA of A₁ and B₁ (Y) nor that of C and D₁ (Z) is a descent of the other. However, if C were replaced with F in the second pair, the PSC would not be satisfied because their MRCA (W) is an ancestor of A₁ and B₁. Also, ESL-PSC automatically excludes all branches in the phylogeny that are unrelated to the evolution of the convergent trait (dotted branches in Fig. 2a). This means that model building is focused on the molecular evolutionary changes between trait-positive and trait-negative species, which are solid blue and red branches, respectively. It is also possible to employ the criteria used for pair selection in the phylogenetic independent contrast method, which may allow for more comparisons²⁵ (see also Supplementary Text 1). If there are multiple species in some trait-positive and trait-negative clades (dotted red and blue branches), different combinations of species sets may be used to build multiple genetic models followed by model ensembling (see Methods). ESL-PSC analysis produces a list of proteins included in the genetic models, ranked by their estimated relative importance, and an equation to predict the presence/absence of the trait in a species based on its genetic sequences (Fig. 1). Species not used for building the

genetic model can be utilized for testing evolutionary hypotheses, which we discuss below.

Genetic models for convergent acquisition of C4 photosynthesis

We applied ESL-PSC to build genetic models of photosynthesis evolution using a 64-species alignment of 67 chloroplast proteins²² (see Methods). Many of these grass species have convergently evolved the C4 photosynthetic pathway for carbon concentration^{26,27}, while others have retained the ancestral C3 photosynthetic pathway. Previous studies of convergence in C4 have focused primarily on Ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCo), the most abundant enzyme, which has multiple sites with convergent amino acid substitutions in multiple lineages of plants^{20–22,28,29}. Others^{22,23} have suggested the involvement of additional chloroplast proteins as well. However, the extent to which chloroplast proteins other than RuBisCo represent a predictable and common genetic basis of C4 evolution remain uncertain.

Six clades in the molecular phylogeny contain sibling species of both C4 and C3 phenotypes (Fig. 3), which yielded six species pairs satisfying the PSC design. Each pair contained a species with C4 photosynthesis and a closely related species with C3 photosynthesis. Because some clades contain multiple candidate trait-positive (C4) and trait-negative (C3) species, we selected the species with the least missing data in the sequence alignment in our first analysis (lineages with solid lines in Fig. 3). The lengths of individual protein sequence alignments varied from 30 to 1,528 amino acids, with a total of 16,362 positions in 67 chloroplast proteins²².

A key feature of ESL analysis is the use of sparsity penalties to control the inclusion of sites and proteins in the genetic model built using LASSO⁸. Penalizing the inclusion of genes and sites in the LASSO model is analogous to maximizing the product of the likelihood and a prior on gene and site penalties, making the importance estimates of genes and sites based on regression coefficients comparable to the maximum *a posteriori* estimate considering a Gaussian likelihood function and a Laplacian prior⁸. Statistically, LASSO with bilevel sparsity in ESL's Sparse Group LASSO is desirable because solutions are invariant under group-wise orthogonal reparameterizations and statistically consistent when a sparse solution is expected^{30,31}. The sparse solution is biologically realistic since only a subset of sites and genes are expected to be associated with trait convergence⁸.

We considered a series of penalties and compared resulting genetic models by using a newly developed Model Fit Score (MFS) that is analogous to the Brier score in logistic regression (see Methods). The genetic model with the best MFS contained five proteins. The two largest contributors were RuBisCo and NADH dehydrogenase-like complex subunit I, which is consistent with previous experimental and analytical reports^{20,22,28,32}. The *clpP*, *petA* and *rpt8* were minor components and have not been highlighted previously. This model correctly predicted 97% (36 of 37) of C4 species not used for model building and 100% (15 of 15) of C3 species not used for model building, achieving a balanced accuracy of 98.5%. An ensemble of genetic models with similar MFS scores (best 5%; Supplementary Fig. 1) also performed equally well (Fig. 4a).

The genetic model with the best MFS was accurate (95.5%) in predicting the trait status of C4 species that are siblings of those used for model building in Fig. 3. This suggests that multiple C4 species within a clade likely inherited the trait from a common ancestor, consistent with the parsimonious reconstruction of independent C4 trait evolution³³. Consequently, genetic models built using different species combinations were also highly accurate (96%, Fig. 5b). The best MFS models were also highly predictive (100% accuracy) of the C4/C3 status of species from evolutionarily independent clades (black dotted branches in Fig. 3) that did not contribute any species for building the model (Fig. 4b). These results suggest that many of the

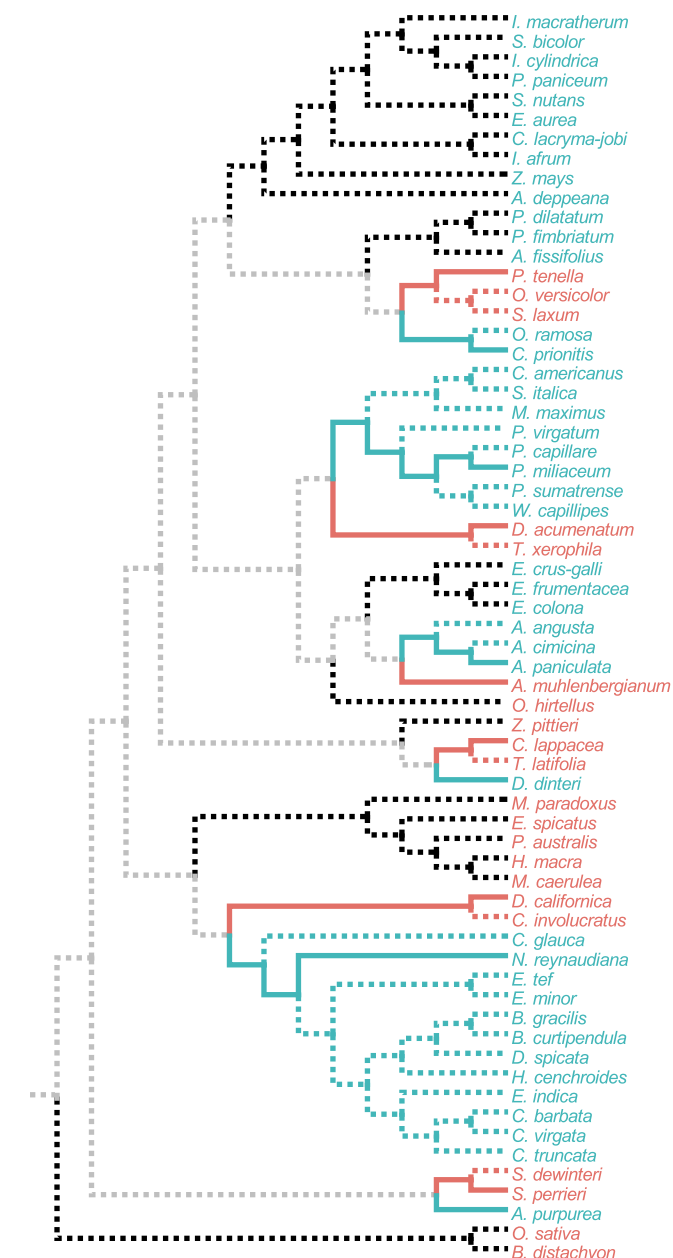


Fig. 3 | Experimental design for ESL-PSC modeling of convergent acquisition of C4 photosynthesis. Evolutionary relationships of 64 grass species based on the phylogeny in Casola and Li²². From the 64 available species, six pairs of trait-positive (C4) and trait-negative (C3) species were chosen according to the PSC approach. Where multiple species met the topological requirements for a contrast pair, we selected the two species that were closest in the evolutionary distance and that had the fewest gaps in the alignments. Evolutionary lineages for the selected species are shown with solid lines, and all other branches are depicted as dotted lines. Lineages begin at the internal node that represents the common ancestor of each pair. Thus, substitutions on lineages with solid lines will be included in ESL-PSC modeling. Blue (C4) and red (C3) dotted lines represent alternative sibling species of the selected species. Black dotted branches represent clades that are evolutionarily independent of the contrast pairs. These include both C4 and C3 species. Gray branches represent the evolutionary history that is shared equally by selected C4 and C3 species. Substitutions that occurred on these branches are expected to cancel out automatically in the modeling process, that is, they will appear equally in both C3 and C4 species used to build the model.

same substitutions have contributed to the acquisition of C4 photosynthesis independently.

By contrast, we found that evolutionarily naive model building, which did not use the PSC design, could only achieve 64% accuracy in identifying C4 species in independent clades (black branches in Fig. 3). In this experiment, we conducted a direct comparison of the PSC and naive approaches by selecting 100 input sets of six C4 and six C3 species from among the siblings of the PSC species, but without respecting the PSC design. The average true positive rate (TPR), a measure of the ability of the model to recognize C4 species on the basis of information in convergent sites, was only 64% over all of these ensembles compared with a TPR of 94% for the ensembles built using the PSC approach (Fig. 5b). This reduction in accuracy reflects the fact that non-PSC models incorporate not only sites whose residues are correlated with the phenotype due to convergent evolution but also sites correlated with the trait purely due to shared ancestry within the inputs. The latter type of site carries no information relevant to the prediction of phenotype in clades whose trait-positive species have acquired the trait independently. This result establishes that our PSC design can produce more effective genotype-phenotype models than naive machine learning.

ESL-PSC provides us with an opportunity to test the hypothesis that other chloroplast proteins have contributed to C4 evolution. For this testing, we built ESL-PSC models without RuBisCo and examined their ability to predict the presence of C4. These RuBisCo-free models achieved modestly high (89%) accuracy, suggesting that the convergent basis of the C4 trait extends to other chloroplast genes (Fig. 5a). Interestingly, these models correctly predicted C4 photosynthesis in *Alloteropsis angusta*, which was the only error (false negative) for the ESL model containing RuBisCo. *A. angusta* is believed to have undergone a C3 to C4 transition independently from the other members of its own genus, including *A. paniculata*³⁴. Interestingly, *A. angusta*'s RuBisCo protein lacks key amino acid substitutions in RuBisCo that are highly diagnostic of other C4 species (Supplementary Fig. 2b).

Therefore, chloroplast proteins other than RuBisCo have likely contributed significantly to C4 evolution in *A. angusta* and other species. While the findings of Casola and Li²² hinted at such a possibility, their statistical analyses using a convergence counting approach did not find a significant excess of convergent substitutions in C4 species as compared to the background C3 species. Also, phylogenetic relatedness likely explains a report²³ that accurate C3/C4 classifiers could be trained using residues in genes other than RuBisCo because species were divided into training and testing sets without regard to phylogeny. This allows the same-trait evolutionary siblings of test species to be present in the training sets, which means that the estimated accuracy in predicting C3/C4 could be due to substitutions shared with sibling species rather than convergent substitutions. The PSC approach is specifically designed to prevent this. Therefore, the ESL-PSC framework provided a new way to investigate the genetics of convergent traits and test hypotheses that have been intractable until now.

We also tested ESL-PSC on multiple synthetic datasets that were generated based on the example of the empirical chloroplast protein dataset for C4/C3 photosynthesis (Supplementary Fig. 3a). For each dataset, we simulated 100 protein-coding MSAs of 500 residues each, along a bifurcating tree of 64 tips. The tree was divided into two independent clades to exclude the impact of phylogenetic dependence on the accuracy measurement. One clade (32 tips) was used for building the ESL-PSC model, and the other (32 tips) for testing its predictions (Supplementary Fig. 3a). The true convergent sites were included in only 10 MSAs out of 100, and the strength of simulated convergent selection was varied by a scaling factor for branch lengths leading to the convergence³⁵. With limited convergence allowed by a scaling factor of 2, with 5, 10, or 15 sites per convergence-affected MSA,

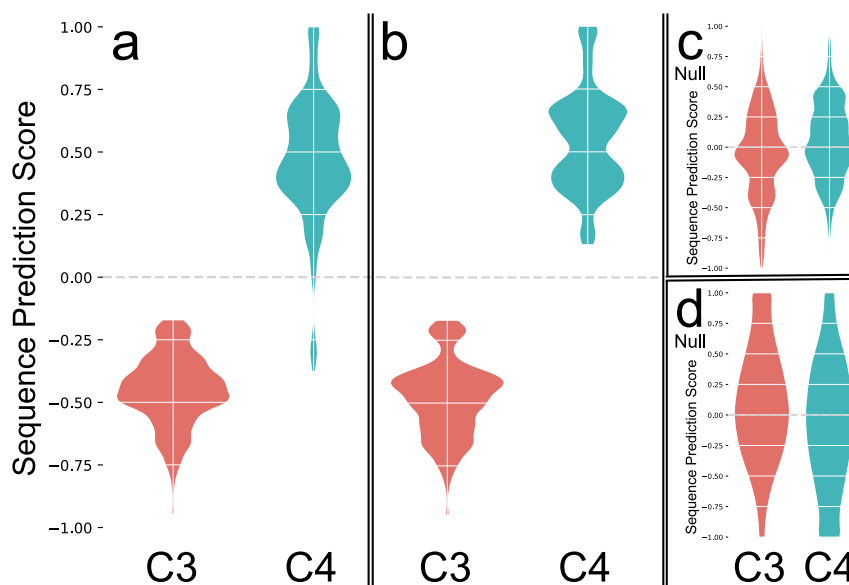


Fig. 4 | Predictive ability of ESL-PSC genetic models of C4/C3 photosynthesis. Sequence (trait) prediction scores (SPS, see Methods) from model ensembles are shown for known C4 (blue) and C3 (red) species in kernel density estimation plots. Negative SPS indicates a prediction of the C3 (trait-negative), and positive SPS indicates a prediction of the C4. Predictions shown are for: **a** all species, **b** species in clades independent of the clades contributing species for model building, **c** response-flipped null ESL-PSC models of C4/C3 photosynthesis. Null models were constructed by flipping the trait

response values of 3 out of 6 of the input contrast pairs. This was done for all 10 distinct combinations of 3 out of 6 contrast pairs, and all model predictions were aggregated. **d** Pair-randomized null ESL-PSC models of C4/C3 photosynthesis. Null models were constructed by randomly flipping or not flipping the residues between each species contrast pair at every variable residue in the MSA. For each of the 25 alternative PSC input species combinations, randomized pair-flipped alignments were generated, and model ensembles were produced for each. Aggregated predictions are shown.

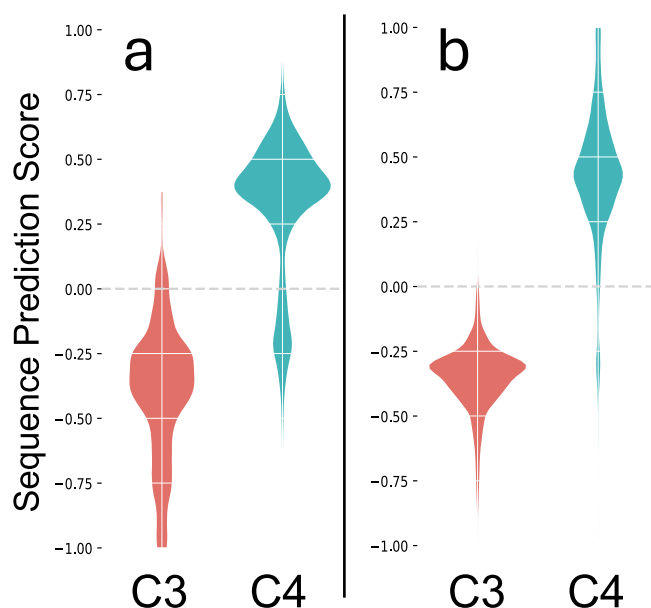


Fig. 5 | Alternative C4/C3 models. **a** Predictions from models developed without the inclusion of RuBisCo are shown for independent species. **b** Alternative PSC combinations. 100 alternative species combinations of PSC pairs were generated, and ensemble models were constructed as above. Predictions were aggregated for only the independent clades (black branches in Fig. 3). SPSs from the best 5% of models by MFS are shown from the aggregate of all ensemble models.

models achieved accuracies of 72%, 86%, and 91%, respectively (Supplementary Fig. 3b). The accuracy reached 100% for a scaling factor of 5 with at least 10 sites affected per convergent alignment. This result confirms that ESL-PSC can accurately model convergent evolutionary change.

Convergent evolution of echolocation

The independent acquisition of echolocation in bats and whales is among the most well-studied cases of convergent molecular and trait evolution. Still, as noted above, attempts to reveal significant evidence of trait-relevant convergent amino acid evolution in proteome-wide scans have produced limited success^{2–4,6,11}. We selected the little brown bat (*Myotis lucifugus*) and the bottlenose dolphin (*Tursiops truncatus*) as trait-positive species (echolocators) because previous studies involving exome-scale searches for convergence in echolocating mammals have often focused on the comparison of microbats and toothed whales^{2,3,11,36}. In the PSC design, we selected a non-echolocating sister species, the large flying fox (*Pteropus vampyrus*), for the little brown bat, and the sheep (*Ovis aries*), for the bottlenose dolphin (Fig. 6; Methods). We retrieved 14,509 protein alignments from the OrthoMAM database of orthologous protein-coding sequences for mammalian genomes³⁷.

Because there were only two clades with trait convergence and, thus, only two species pairs for the PSC design, we made inferences from a collection of ESL models obtained using a range of species combinations within the available clades (Fig. 6, see Methods). The collection of ESL models was then used to generate a ranked list of candidate proteins associated with convergent evolution (Supplementary Data 1). As expected, among the highest-ranked proteins were several that have been previously characterized to have signatures of molecular convergence in echolocators, including Prestin (SLC26a5), TMC1, PJKV (DFNB59), CDH23, CASQ1, and CABP2^{3,17,18,38–40}. In some cases, specific amino acid sites within these proteins have been implicated in conferring the functional changes necessary for the echolocation phenotype, revealed by laboratory assays where mutations to residues found in echolocating species were observed to alter protein function in a manner consistent with echolocation^{3,19}. Therefore, the presence of such proteins in the results serves as a validation of the efficacy of ESL-PSC. In addition, several other hearing-related proteins are included in the top-100 ranking results that have not been previously connected with the convergent evolution of echolocation, including CHRNA9, GOLGA1, GRXCR2, and MREG.

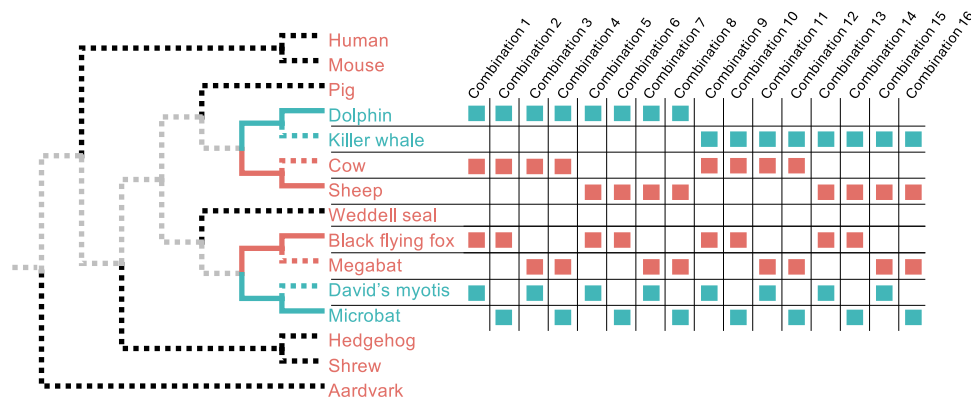


Fig. 6 | Echolocation analysis. The topology of species used in the echolocation analysis is depicted in a cladogram, which is derived from the consensus species tree of mammals³⁹. Echolocation evolved twice in mammals in this phylogeny. Therefore, two contrast pairs can be constructed (e.g., solid blue branches, echolocating; solid red branches, non-echolocating). A series of 15 comparable sets

of input pairs can be constructed by including alternative sibling species (dotted blue and red species) in all possible combinations. Species not included in the contrast pairs do not affect the analysis (black dotted branches). Shared ancestry is canceled out (gray dotted branches). Only a subset of the species not used in the analysis are depicted here to highlight the phylogenetic context.

We generated Benjamini-Hochberg adjusted P -values to gauge the functional enrichment in the top-ranking proteins included in the genetic models. We tested for ~20,000 biological processes and phenotypes (see “Methods”), which revealed the top 100 proteins to be highly enriched for the sensory perception of sound genes (GO:0007605) with an adjusted P -value $< 10^{-4}$ (Fig. 7, intersection of vertical and horizontal boxes). The enrichment adjusted P -value was significant, even for 50, 150, and 200 top proteins identified by ESL-PSC (adjusted $P < 10^{-3}$), suggesting that the results are robust to the size of the gene list analyzed. This top-100 gene list was also significantly enriched for many Phenotype Ontology (PO) terms directly related to hearing and sound perception such as cochlear inner hair cell degeneration (MP:0004398), increased or absent threshold for auditory brainstem response (MP:0011967), cochlear ganglion degeneration (MP:0002857), and organ of Corti degeneration (MP:0000043) (Fig. 7). We also found a highly significant enrichment (adjusted $P = 4.5 \times 10^{-3}$) for the top-level mammalian PO term hearing/vestibular/ear phenotype (MP:0005377).

The top 100 proteins produced by null genetic models did not show significant enrichment for any ontology term. In these models, one of the two contrast pairs had its trait designations reversed, such that the echolocating dolphin and non-echolocating large flying fox were treated as sharing a convergent trait, while the other two species were treated as paired contrast partners. This configuration had the property of canceling out both the shared phylogenetic signal and any shared convergent trait signal from the genuine trait of echolocation. Therefore, biases in enrichment tests and ESL-PSC analyses cannot explain the significant P -values observed (Fig. 7).

We also developed and conducted another null test of ESL-PSC by analyzing only fourfold degenerate sites expected to evolve largely neutrally in mammals because data contamination and other types of errors can drive inferences of amino acid convergence³⁵. No significant enrichment was found for any of the relevant ontology categories. To rule out false positives due to nonspecific effects, we generated simulated alignments mimicking the lengths, amino acid equilibrium frequencies, and other evolutionary parameters for each of the 14,509 OrthoMaM protein alignments and ran ESL-PSC as above (see Supplementary Methods). These experiments did not produce gene lists with significant enrichments for any ontology terms.

Comparison with other methods

We implemented and applied six existing convergence detection methods to scan for convergence-containing proteins in the empirical mammalian echolocation dataset (see Supplementary Methods and

Supplementary Text 2 for details on the methods). Ontology enrichment results based on the output of each method are shown in Fig. 7. Only 3 of the tested methods obtained significant ontology term enrichments in these analyses: Ancestral State Reconstruction using codon data (ASR-Codon), Convergence at Conservative Sites (CCS⁷), and Convergent SUBSTITUTION (CSUBST³⁵). Of these, only CCS succeeded in finding significant enrichments for some hearing-related ontology terms. The GO Biological Process term Cilium Movement (GO:0003341) was significantly enriched in the results produced by both CCS and ASR-Codon, and its enrichment in the CCS results was the most highly significant of the ontology terms enrichments found by methods other than ESL-PSC.

CCS uses a symmetric phylogeny similar to the PSC approach; it reconstructs ancestral states by assuming that the residue present in the outgroup species is the ancestral state. In contrast, ESL-PSC does not use an outgroup sequence or reconstruct ancestral states. In CCS, a site is inferred to have undergone convergent evolution when at least two convergent lineages have substituted to the same residue that differs from the ancestral residue. Only three proteins (PER1, CDH23, and SLC26A5) overlapped between the top 100 ESL-PSC results and the top 100 CCS results, of which the latter two are known to be involved in echolocation from previous studies^{17,18,39}. No significant enrichments were found from the results of other methods: ASR-AA, Branch Site Unrestricted Statistical Test of Episodic Diversification (BUSTED)⁴¹ or Target species-specific Amino Acid Substitution (TAAS)⁴².

ESL-PSC for echolocation with 16 species combinations took only approximately 30 min on a desktop computer using 1 CPU core to complete all of the steps in Fig. 1. In contrast, methods that rely on evolutionary model-based ancestral state inference (ASR, BUSTED, and CSUBST) took up to 2000-times longer than ESL-PSC.

Trait predictions to test functional hypotheses

The existence of significant enrichments of ontology terms among the top-ranking proteins found by ESL-PSC and other methods serves as evidence that these methods have detected non-random signals of convergent amino acid evolution among certain biological categories, which in turn implicates these ontology categories in the biology of the trait. However, an alternative explanation is that some methods are biased toward selecting proteins that share certain characteristics unrelated to convergent evolution. This could result in spurious ontology term enrichments because some evolutionary attributes, such as evolutionary rates, can covary among proteins belonging to the same functional categories^{43,44}. In order to further assess whether

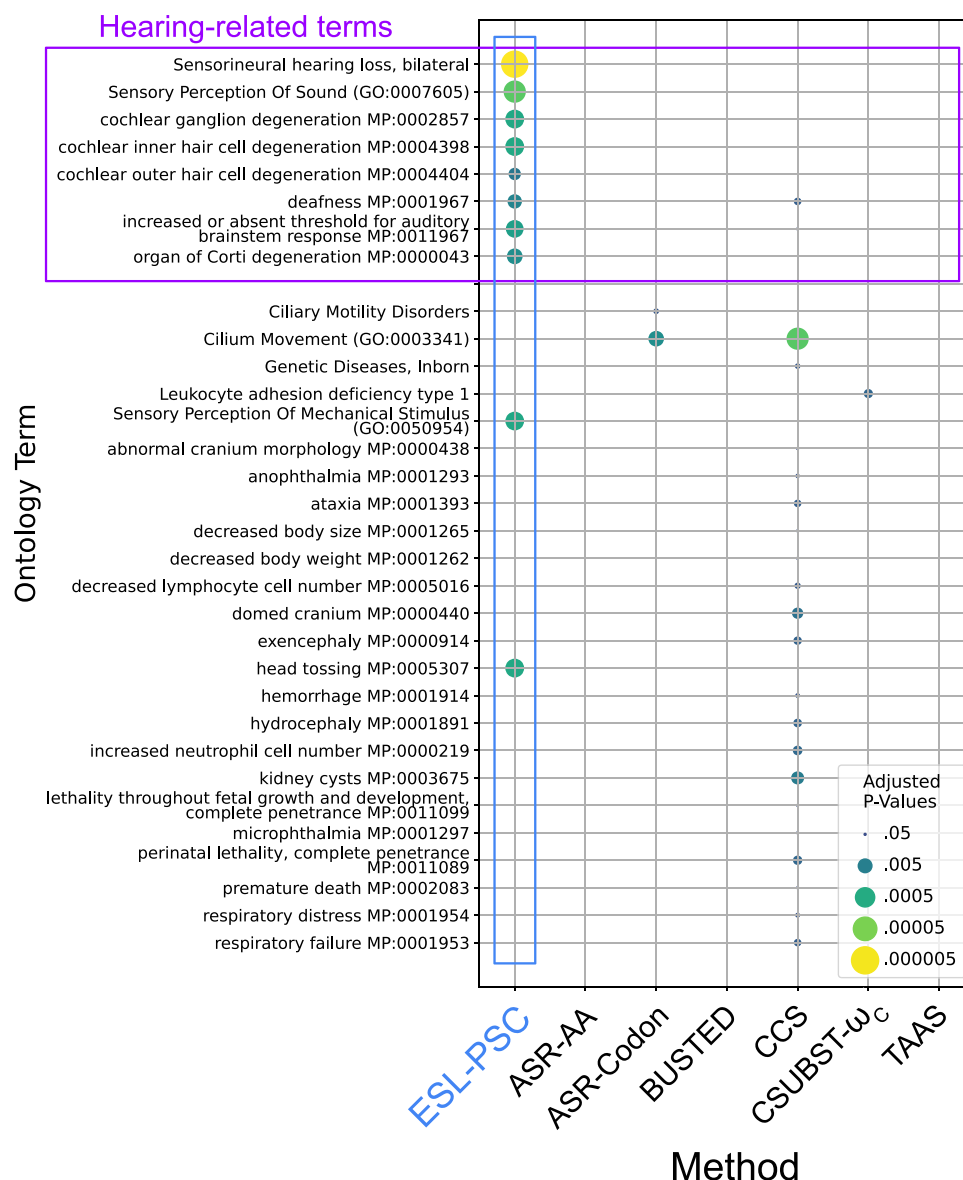


Fig. 7 | Ontology term enrichments. Enrichment tests (Fisher's exact test) were performed for Gene, Phenotype, and Disease ontology terms (see Methods) for the top 100 highest-ranking trait proteins in our echolocation multiple species combination ensemble model integration analysis and in the highest ranking 100 proteins from the other tested methods. *P*-values are adjusted by the Benjamini-

Hochberg method to account for multiple tests (see Supplementary Methods). ASR: Ancestral Sequence Reconstruction; BUSTED: Branch Site Unrestricted Statistical Test of Episodic Diversification⁴¹; CCS: Convergence at Conservative Sites⁷; CSUBST- ω_c : Convergent SUBSTITution ω_c ³⁵; TAAS: Target species specific Amino Acid Substitution⁴².

the proteins and evolutionary substitutions contributing to an ontology term enrichment could be functionally linked to the trait, we developed a simple approach using the predictive modeling capability of ESL-PSC framework. If a category of proteins contributing to an enrichment has, in fact, undergone convergent amino acid evolution that enabled the trait, then their sequences should carry information allowing the accurate prediction of the trait status of other species that were not used in model building.

ESL-PSC models built using the 136 proteins bearing the Cilium Movement (GO:0003341) GO term, which were significantly enriched in the top-ranked CCS and ASR-Codon results, had a 75% balanced accuracy (Supplementary Fig. 4d). This is much lower than the 96% balanced accuracy achieved by models built using 144 proteins bearing the Sensory Perception of Sound (GO:0007605) term that was found to be enriched only in the top-100 ESL-PSC proteins (Supplementary Fig. 4c). Models built using 39 proteins, contributing to all of the numerous non-hearing-related significant ontology enrichments of the

CCS method, produced models with a balanced accuracy of only 62%, suggesting that the non-hearing related enrichments may be spurious and caused by unshared neutral background convergence (Supplementary Fig. 4f).

Enrichment of convergent residues in echolocating species

Adaptive convergent residues that are functionally important would be expected to be more highly correlated with the trait across all species in the dataset compared with neutral background convergences. As an independent approach to assess the trait-relevance of the top-ranking proteins found by each method, we identified the convergent residues found by the CCS method in the top 100 CCS proteins (526 total) and the top-100 ESL-PSC proteins (170 total). For each convergent site, we used Fisher's exact test to evaluate the enrichment of the convergent residue at that site in all of the echolocating species in the data set relative to the non-echolocating species. We also performed this test on all CCS-

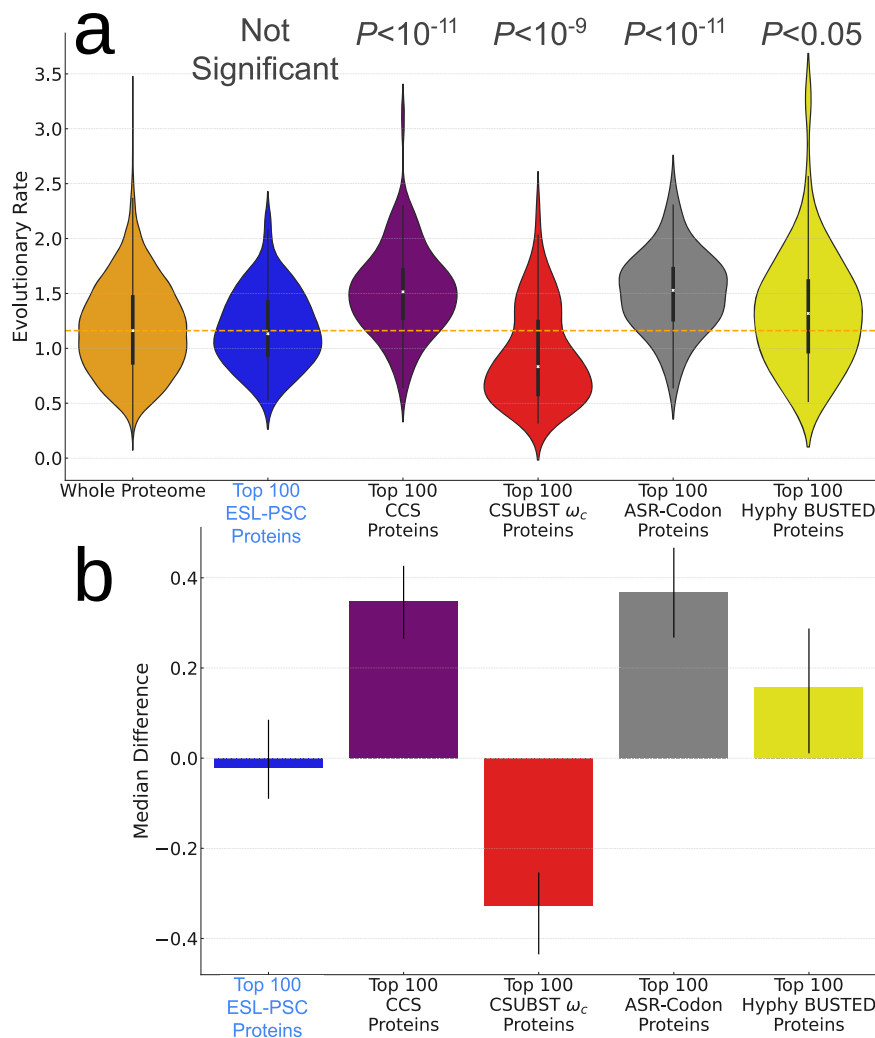


Fig. 8 | Comparison of evolutionary rate distributions for top-ranking proteins from each method with the whole proteome. **a** The distributions of mammalian evolutionary rates are shown for the top 100 ranking convergent proteins for ESL-PSC, CCS, CSUBST ω_c , ASR-Codon, and Hyphy BUSTED, along with the distribution for all 14,509 orthologous protein alignments. The orange dashed line shows the median evolutionary rate for the whole proteome dataset. Differences between each top 100-protein distribution and the whole-proteome distribution were assessed by a two-sided two-sample Kolmogorov–Smirnov test. The exact P -values for the comparisons are as follows: ESL-PSC, $P = 0.4789346$; CCS, $P = 3.320747 \times 10^{-12}$; CSUBST ω_c , $P = 6.009956 \times 10^{-10}$; ASR-Codon, $P = 1.241988 \times 10^{-12}$; Hyphy BUSTED, $P = 0.01697650$. Differences between the distributions for each top 100 protein list and the whole proteome were assessed by two-sample Kolmogorov–Smirnov tests with P -values shown above each method. The distribution for the top 100 ESL-PSC proteins is not significantly different than that of the whole proteome (2 sample K-S $P = 0.479$), indicating that ESL-PSC results were not biased toward proteins exhibiting pervasively

slower or faster evolution across the mammalian tree. The Hyphy BUSTED method is included for comparison, but as a method that detects diversifying selection and not convergent substitutions, it is expected to find somewhat faster-evolving proteins. Evolutionary rates are measured in substitutions per site per billion years and are calculated for each gene alignment as the total branch length of the maximum likelihood gene tree as reported by the OrthoMaM database³⁷ divided by the total time in the consensus species timetree⁵⁹ pruned for taxon sampling variation across gene alignments. Violin plots show the kernel density estimate of the evolutionary rate distributions. Embedded box plots indicate the median (center line), interquartile range (bounds of the box), and 1.5 times the interquartile range (whiskers).

b Differences in the median evolutionary rate between the top 100 proteins of each method and that of the whole proteome. Error bars indicate 95% bootstrap confidence intervals of the median difference for each method (see Supplementary Methods).

detected convergent sites throughout the proteome and compared the distributions of enrichment P -values for each method's top-100 subset to the overall proteome-wide distribution with the top ESL-PSC proteins removed. The top-100 ESL-PSC proteins had a significantly lower (more highly enriched) distribution of enrichment P -values (Mann–Whitney U test $P = 2.88 \times 10^{-6}$), but the top-100 CCS distribution was not significantly different from the whole proteome, indicating that the convergent residues found in the top-100 CCS proteins are on average no more preferentially found in echolocators compared with non-echolocators than the average for all sites found by the CCS method (Supplementary Fig. 5, Supplementary methods). This finding accords with the greater predictive

power of the top-100 ESL-PSC proteins compared with other methods.

Evolutionary rate differences in candidate convergent proteins

We examined the distributions of evolutionary rates of the top 100 proteins found by ESL-PSC and other methods (Fig. 8). There was no significant difference between the distribution of top ESL-PSC proteins and the whole proteome, suggesting that ESL-PSC results are not biased toward faster or slower evolving proteins. In contrast, CCS and ASR-based methods, which use counts of sites with convergent change, had significantly higher mammalian evolutionary rates in their top-100 proteins. This result could be explained by the fact that neutral

background convergence will tend to be higher in faster-evolving proteins. CSUBST ω_c , which quantifies excess convergence over neutral expectations based on an analysis of nonsynonymous and synonymous substitutions (dN/dS) and corrects for synonymous convergence rate, produced genes with significantly lower evolutionary rates. The top-100 proteins by ω_c included many that had low Observed Convergent Nonsynonymous substitutions (OCN) (Supplementary Fig. 6b), but these had high values of ω_c due to very small denominators in the dN/dS ratios (Supplementary Fig. 6c, d). To assess whether these proteins were responsible for the low evolutionary rates of the top-100 subsets, we divided the top-100 proteins for ω_c into the upper half and lower half by OCN. In each case, the lower half had significantly lower evolutionary rates than the whole proteome (Supplementary Fig. 6a). In view of this, we filtered the output of CSUBST using a cutoff of 1 for OCN before re-ranking the proteins, but the resulting top-100 subsets were still not enriched for any ontology terms. Notably, with many proteins receiving ω_c values in the thousands due in part to near-zero denominators, some truly convergent proteins could not reach the top-100 results. For instance, SLC26A5 (prestin) had a highly elevated ω_c in our analysis (918.5) but was ranked 133rd when ranking by ω_c . Thus, in some cases, CSUBST and other methods based on classical approaches may include more background and fewer true convergent proteins in their highest-ranking results because they tend to be predisposed to including either fast-evolving or slow-evolving proteins.

Analysis of synthetic datasets

We also evaluated the performance of ESL-PSC for detecting convergence in simulated datasets inspired by the example of echolocation. In this empirical-synthetic study, partitions of 5, 10, or 15 sites generated by simulating convergent evolution were added to 25 randomly selected empirical protein alignments that we notionally considered to belong to a hypothetical functional category (Supplementary Methods, Supplementary Fig. 8). ESL-PSC, CCS and CSUBST were run on each full proteome dataset containing these modified alignments (see Supplementary text 3 for information regarding the choice of these methods for comparison). The objective of each test analysis was to detect an enrichment of the hypothetical category within the top-ranked proteins. We conducted 100 replicate analyses for each of 20 different strengths of convergent selection and compared the results from each method. ESL-PSC performed the best in the majority of conditions in which any method achieved the notional significance threshold of 5 proteins in the top-ranked results (Supplementary Fig. 9).

We further evaluated the performance of ESL-PSC and CCS on empirical-synthetic datasets for larger numbers of convergent species, as these two produced significant gene enrichments for echolocation and a larger-scale evaluation of CSUBST was computationally prohibitive. Randomly selected combinations of 2, 3, 5 or 10 mammalian species were designated as having undergone simulated convergent evolution (Supplementary Fig. 7). Species combinations were either drawn from *Laurasiatheria* (2, 3, and 5 species combinations) or the broader set of all *Theria* (5 and 10 species combinations). A range of the number of sites affected and foreground branch scaling factors were explored to vary the strength of molecular signatures of convergent evolution (see Supplementary Methods). We performed ESL-PSC and CCS analyses on 52,500 whole proteome datasets on a High-Performance Computing (HPC) cluster. The number of simulated convergence-affected alignments appearing in the top 100 results was recorded for each dataset, which is shown in Fig. 9. ESL-PSC performed largely identically on combinations of 5 convergent species regardless of whether they were chosen from within *Laurasiatheria* or all of *Theria*, but CCS improved its performance slightly on combinations chosen from *Theria* (Fig. 9).

ESL-PSC exhibited substantially greater sensitivity than CCS across all tests, especially with smaller numbers of sites in each protein subjected to simulated convergent evolution and when larger numbers of convergent species are analyzed (Fig. 9). Notably, with 5 or 10 convergent species, ESL-PSC could identify alignments containing only a single simulated site per convergent protein, while CCS could not detect any of these alignments, even at the highest strength of convergent evolution. The increasingly better performance of ESL-PSC with larger numbers of convergent species in the dataset shows that it benefits with larger datasets, as would be expected of a machine learning method.

Discussion

The discovery of genotype-phenotype relationships is of central importance in functional genomics and evolution. Repeated evolution of the same trait in species of independent clades offers an opportunity to reveal the genetic architecture shared by these independent trait acquisitions. We have presented a comparative genomics approach using machine learning (ESL-PSC), informed by molecular phylogenies, to build quantitative genetic models of trait convergences. The application of ESL-PSC to two distinct, previously investigated examples establishes that there can be a significant commonality in the genetic basis of trait evolution among species in independent lineages.

ESL-PSC models correctly predict the presence of C4 photosynthesis with high accuracy in grass clades not involved in building the model (Fig. 4a, b). The same was true for echolocation when building a predictive model using genes involved in hearing (Supplementary Fig. 4b, c). Classical molecular evolutionary methods do not commonly afford this type of quantitative prediction. The high accuracy of genetic models of C4 trait evolution in which the well-studied convergent protein RuBisCo was excluded is suggestive of the potential role of additional chloroplast proteins in the convergent gain of C4 photosynthesis. These analyses also showed that not all species with convergent traits harbor the same substitutions in the sites included in genetic modes. In fact, no more than four out of six C4 species shared the same amino acid residue in the sites selected during ESL model building. Therefore, ESL model building can automatically harness relevant information from incomplete molecular convergence correlation with the trait convergence, obviating the need to use ad hoc cut-offs and subset the data by evolutionary conservation^{2,3,5,7,45}. This makes ESL-PSC convergent evolution analyses more objective and reproducible.

ESL-PSC also identified genes involved in the convergent acquisition of echolocation in mammals. The list of top genes in ESL models was found to be highly enriched for ontology categories involved in auditory processes at FDR-corrected *P*-values that were more significant than previously reported, implying that the machine learning approach to building genetic models can be more effective than previous approaches in some situations, as seen in the analysis of synthetic datasets (Fig. 9). The validation of ESL-PSC derived from the enrichment of functional categories is arguably circumstantial, but direct experimental approaches are beyond the scope of this investigation. However, further support may be found by assessing the potential functional relevance of the selected genes to determine whether mutations in these genes are known to be associated with diseases due to relevant functional disruptions. In the analysis of Disease Ontology categories, we found the Sensorineural hearing loss, bilateral term to be highly enriched in the top genes in ESL-PSC models (adjusted $P < 10^{-5}$; Fig. 7). No previous study has reported such an enrichment, and no other methods tested produced gene lists with highly significant enrichments (Fig. 7; Supplementary Data 2).

The ESL approach differs conceptually from existing statistical methods designed to detect molecular signatures for convergent trait evolution. Unlike classical methods, ESL does not model the process of base substitution using sophisticated mathematical models or

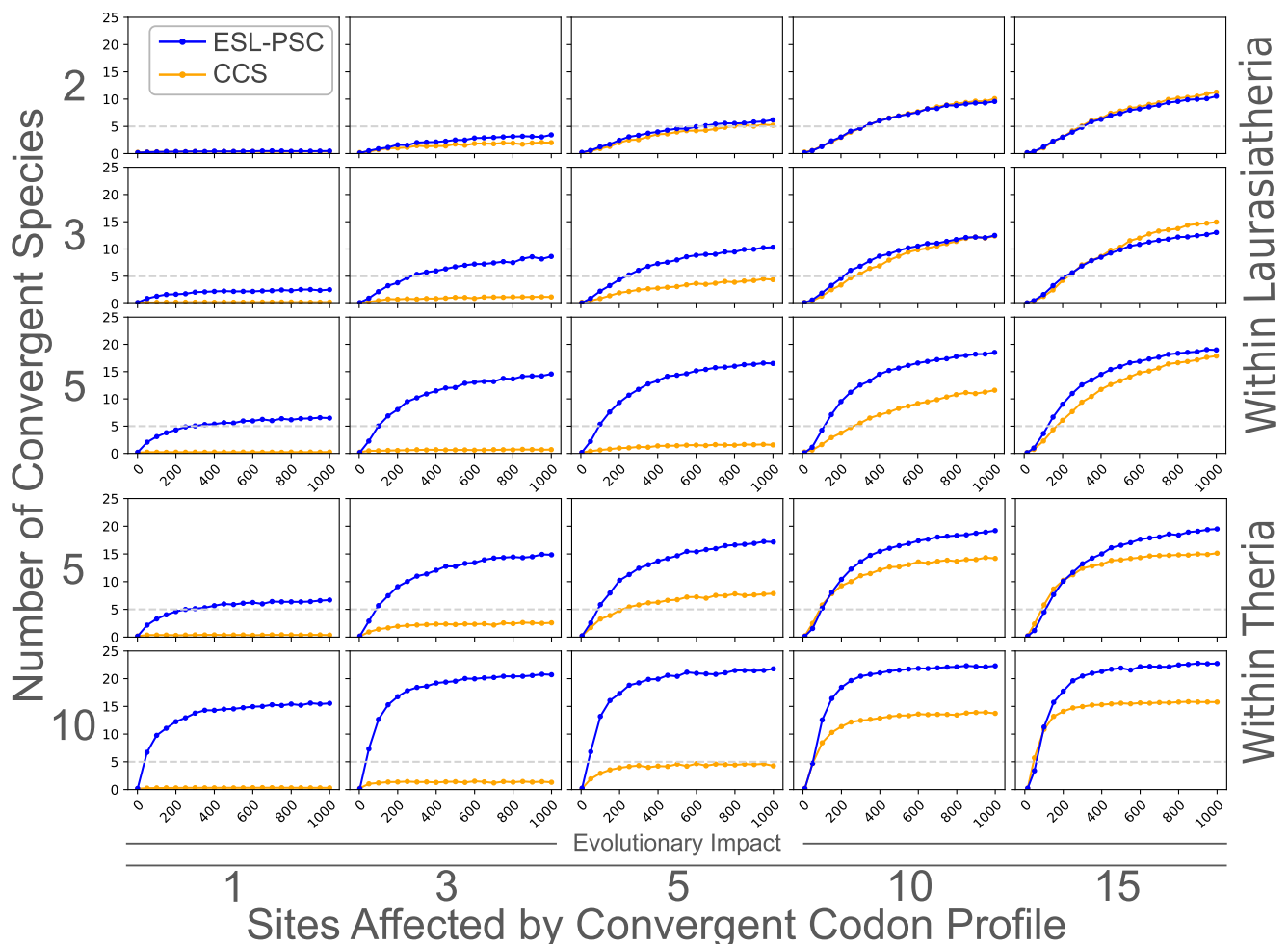


Fig. 9 | Results of simulation studies with randomized species combinations.

The Y axis of each panel shows the number of simulated convergent alignments that were in the top 100 ranks of all proteins for each of the methods, averaged over 100 species combinations and sets of seed alignments for each point. For each row of panels, 100 species combinations were chosen at random consisting of the number of simulated convergent species shown at the left along with a control sibling for each one that met the topological requirements for ESL-PSC and CCS (see Supplementary Methods). For the upper three rows of panels, species combinations were chosen from Laurasiatheria (50 species in the dataset). For the lower two rows, combinations were chosen from all of Theria. For each column of panels, the number of amino acid sites shown at the bottom was simulated using the CSUBST simulate function³⁵ for each of 25 randomly chosen alignments for each species combination, with simulations repeated for each of

20 foreground scaling factors that are adjusted for each column of panels such that the product of the number of sites and the scaling factors produces the same range, which we define as the “Evolutionary Impact” of convergence in each panel. For each panel, an additional run was conducted with a scaling factor of 1 and the convergent codon profile turned off as a negative control. For each species combination, for each number of sites for each scaling factor, ESL-PSC and CCS were run on the full proteome data set with the 25 chosen alignments having the given number of sites replaced by those simulated as described above (see Supplementary Methods). This required a total of 52,500 runs of ESL-PSC and CCS and ~1.3 million simulated alignment partitions. Dashed lines at $Y=5$ represent a notional detection threshold of 5 genes that may be required to accept a significant enrichment, as we and others have used in order to reduce type 1 error due to many small ontology categories.

estimate evolutionary parameters⁸. Instead, ESL seeks models based on patterns of sequence variation that are the most concordant with the presence or absence of the convergent trait in the species. Thus, ESL analysis avoids the assumption of identical substitution patterns across lineages and sites, along with the assumptions of stationarity and reversibility, which are needed for analytical tractability in standard evolutionary genetics methods but are known to be violated often (see Supplementary text 1)^{8,46}. Sites and lineages experiencing convergent evolution are unlikely to exhibit the same patterns of substitutions as neutrally evolving sites, and the choice of substitution model can have a large impact on inferences of convergence^{35,47,48}. ESL also does not require the branch lengths of the phylogeny, which need to be inferred from substitutions aggregated across all sites. Such branch length estimates may not apply to sites in lineages experiencing adaptive evolution, leading to biased inferences.

Interestingly, ESL-PSC can detect the genetic basis of convergent traits using very few instances of evolutionary innovations in the datasets analyzed. This success is confirmed in the analysis of simulated datasets evolved with as few as two instances of convergent evolution. This performance of ESL-PSC with few samples is amenable to the paradigm of few-shot learning^{49,50}, which is an essential development for studying convergent evolution in which there are rarely large samples of convergent species. The success of ESL-PSC in this regard is attributable in part to the PSC design, which uses the organismal phylogeny to select equal numbers of trait-positive and trait-negative species (Fig. 2), maximizing the information relevant to convergence in the data. In addition, the preferential occurrence of convergence in only a small subset of genes relevant to a trait allows the sparse group LASSO to greatly mitigate the curse of dimensionality.

In ESL-PSC, some lineages, represented by dotted lines in Figs. 3 and 6, and their substitutions are not used in building a specific ESL model, unlike classical statistical methods that use such substitutions for establishing background patterns in detecting convergent evolution^{35,47,48,51}. However, the benefits of using such substitutions are subject to assumptions about their rates and patterns being applicable to those sites and positions involved in adaptive evolution. In contrast, clades and species not used for model building in ESL-PSC are still valuable because they provide an opportunity to quantitatively test whether the same genetic changes were involved in the evolution of convergent traits in independent lineages, as reported here (Supplementary Figs. 4, 5).

Therefore, ESL-PSC complements existing statistical approaches of molecular evolution and addresses challenges researchers face in detecting and testing molecular signatures of convergent trait evolution. We expect ESL-PSC to be useful as a comparative genomics tool for uncovering the common genetic foundations of the evolution of traits shared between species. We envision that ESL-PSC will be applied to generate a candidate gene list, which may be followed by a series of hypothesis tests addressing the commonality of the genetic basis of trait convergences and in-depth analyses of candidate genes using traditional molecular evolutionary analyses to gain additional insights regarding individual sites and selective processes at play.

Methods

Genomic alignment data retrieval and processing

Alignments of chloroplast genes were retrieved from the supplemental data in ref. 22. We generated translated amino acid sequences from the provided nucleic acid alignments for ESL-PSC analyses. The OrthoMaM data set³⁷ of mammalian one-to-one orthologous protein sequence alignments was downloaded from <https://orthomam.mbb.cnrs.fr/>. Following previous studies in which exome-scale scans for convergence in echolocating mammals were performed, we analyzed echolocation in microbats and toothed whales^{2–4,11,36} and used megabats and non-cetacean artiodactyls as non-echolocating sister taxa^{2,5,6,11}. In ESL-PSC analyses, we excluded sites containing missing data or alignment gaps in individual data sets. All multiple sequence alignments (MSAs) were one-hot encoded⁸, which transforms them into a numerical format that is required by the model-building algorithm. The presence of the convergent trait was represented numerically by +1 and its absence by -1.

Selection of contrast pairs

For both the Chloroplast C4/C3 photosynthesis and mammalian echolocation datasets, we selected the largest number of possible contrast pair clades that met the criteria for evolutionary independence for all contrast pairs, namely that the MRCA of any pair cannot be ancestral to any other pair or member of a pair. The procedure we used was as follows: We first selected the most recent bifurcating node whose daughter clades consisted entirely of species of opposite phenotypes (ancestral and convergent). A single contrast pair of species is chosen from the two clades by taking the two species with the longest overlapping genomic sequences. All other descendants and ancestors of their MRCA were then removed from consideration and the procedure repeats until no further valid clades remain. For the mammalian echolocation analyses, this procedure was applied for a set of species used in previous studies, as described in the main text.

Building genetic models

ESL-PSC uses the Least Absolute Shrinkage and Selection Operator (LASSO)²⁴ logistic regression, in which coefficients were chosen to minimize a combination of the difference between observed and predicted response values of the input species (the logistic loss). It uses an inclusion penalty term that scales with the sum of the absolute

values of the model coefficients and, therefore, induces sparsity⁸. We used bilevel sparsity in which separate penalties are applied for the inclusion of sites and groups of sites (e.g., proteins). The loss function is given by:

$$L(\beta) = l(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \sum_{g=1}^G w_g \|\beta_g\|_2 \quad (1)$$

where β is the vector of model weights, $l(\beta)$ is the logistic loss, λ_1, λ_2 are site and group inclusion penalty parameters respectively, w_g is the group penalty for group g , and G is the number of groups in the dataset (e.g., proteins in the proteome). The λ_1, λ_2 values are expressed as a proportion of the maximum value that will generate a nontrivial solution. The loss function is minimized by gradient descent⁵², which was re-implemented in the myESL software package used for ESL-PSC implementation⁵³. We estimated a new Model Fit Score (MFS) for a given genetic model, which is the root mean squared difference between the input trait value (+1 and -1) and predicted trait values for all species used for building the model.

$$\text{MFS} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

Where y_i is the input trait response value for species i . The best-fit genetic models have the lowest MFS value, i.e., the input and output of the genetic model are the most concordant. This is needed because optimal inclusion penalties are not known beforehand in LASSO. So, the genetic model with the best MFS is chosen.

In our analysis, the size of the penalty for each protein (group of sites) was globally controlled by the inclusion penalties, but can also vary for each individual group depending on its composition. Group penalties in applications of the LASSO method are typically based on the square root of the number of columns belonging to the group in dataset³¹. Applying this system produced models in which proteins with fewer variable sites and lower total entropy were penalized more than those with many variable sites, in the exome-wide analysis. However, highly conserved proteins containing even a few variable sites can be important. Therefore, we devised a penalty function in which the group penalty scales linearly with the number of variable sites plus a constant equal to the median number of variable sites across the proteins in the dataset (excluding fully invariant proteins). This function was effective for both small-scale (chloroplast exome) and large-scale (mammalian proteome) analyses.

Predictive model ensembles

Models with similar MFS scores were combined to form ensembles of models for predictions. For all model ensembles, we used a range of group and site inclusion penalty values from 1%-99% of the maximum penalty that can be applied before a trivial solution in which all model feature weights are set to 0 is obtained. The inclusion penalty values were taken from a logspace over this range. Unless specified, we selected genetic models with the best MFS or those with the top-5% MFS values.

Building the candidate protein list

We estimate the Group Sparsity Score (GSS) for every selected protein in every model overall inclusion penalty combinations. GSS is the sum of absolute values of regression coefficients for all the selected positions in the given protein⁸:

$$\text{GSS}_g = \sum_{r \in R_g} |\beta_r| \quad (3)$$

where R_g is the set of binary encoded variable columns belonging to protein g . The higher the GSS, the greater their importance. Proteins

not included in the genetic model receive $GSS = 0$. Each protein g is ranked within each model by its GSS value, denoted as $GSS_g^{(j)}$ for model j . The best rank, $rank_{g,best}$, for protein g across all models is the smallest rank it achieves in any model:

$$rank_{g,best} = \min_j (rank_g^{(j)}) \quad (4)$$

where j indexes over all models. To obtain the overall ranking, proteins are ordered by $rank_{g,best}$ with equally ranked proteins being further ordered according to the maximum GSS_g they attained in any model $GSS_{g,best}$. This yields an ordered list of proteins whose convergent sites stand out compared with the rest of the proteome in number, proportion, and strength of the concordance of their convergent site patterns with the species phenotypes, without privileging any one of those considerations.

When each of the input species has at least one sibling species that share its phenotype for the trait being studied, then different combinations of these allowable input species can be used interchangeably, and models over all inclusion penalty combinations can be built for each of the species combinations. $GSS_{g,best}$ is obtained for each protein for each species combination and proteins are then ranked according to the number of combinations in which $GSS_{g,best}$ is non-zero:

$$C_g = \sum_{k=1}^K \mathbb{1}(\exists_j GSS_g^{(j,k)} > 0) \quad (5)$$

where $\mathbb{1}$ is the indicator function and K the total number of species combinations. Additionally, the number of species combinations where the protein g is ranked in the top positions is counted:

$$T_g = \sum_{k=1}^K \mathbb{1}(rank_g^{(j,k)} \leq \alpha \cdot N) \quad (6)$$

where α is the top proportion (0.01 in our analyses), and N is the total number of proteins. The final ranking is then obtained as follows:

$$Rank(g) = \text{sort}(\{g\}, (-C_g, -T_g, rank_{g,best}, -GSS_{g,best})) \quad (7)$$

Ontology analysis

Ontology enrichment testing was performed using Enrichr⁵⁴, and P -values were adjusted for multiple testing. We considered only three specific ontology libraries consistently throughout the study in order to avoid issues with multiple testing across many libraries. Gene ontologies were obtained from GO⁵⁵. We tested for the biological process GO ontologies using the GO_Biological_Process_2021 library in Enrichr (6036 terms). Phenotype ontologies were derived from MGI⁵⁶. Enrichr provides PO testing using a trimmed version of the MGI phenotype vocabulary, which excludes the top three levels of PO terms (4601 terms). Disease ontologies were derived from DisGeNet (9828 terms)⁵⁷. To determine enrichment and overlapping genes for the top-level PO term hearing/ vestibular/ ear phenotype (MP:0005377), we used the MouseMine⁵⁸ ontology testing tool and the Benjamini-Hochberg adjustment to obtain a multiple testing adjusted P -value. By common convention, enrichments were only considered valid if accounted for by an overlap of at least 5 genes in the sample set. Phenotype ontology terms were retrieved from the Mouse Genome Informatics mammalian phenotype vocabulary, and gene lists associated with phenotype ontology terms were generated from the Mouse/Human Orthology with Phenotype Annotations (downloaded from <http://www.informatics.jax.org/downloads/reports/index.html#pheno>). For gene enrichment analyses, we found it unnecessary to use ensembles of 400 models (20 values for each inclusion penalty) because the gene ranks are based on the maximum model weights which do not change significantly when using a

denser grid search over the space of inclusion penalties. Results shown here were based on ensembles using 4 values of each inclusion penalty (16 models) in each ensemble for each species combination.

Null genetic model ensembles

There are a number of different ways to test the genetic models produced by machine learning. We built null genetic models by reversing trait designations of a subset of datasets such that both the shared evolutionary history and shared basis of the convergent trait between trait-positive species were canceled out (Fig. 3c). For an even number $2n$ of input species contrast pairs, the largest scrambling of the input phenotype designations is achieved by flipping n pairs. There are $1/2^{2n}C_n$ possible distinct null configurations. For a small n , it is possible to generate and combine all null predictions, but a random subset of possible null configurations can be sampled when n is large. Another type of null model can be constructed by randomly flipping (or not flipping) the residues between the two members of each contrasting pair at each site (Fig. 3d). This preserves any phylogenetic relationships present in the alignment but, when averaging over a large number of such pair-randomized alignments, destroys the correlations that are due to convergence. Both of these null model experiments are expected to produce models whose prediction accuracy on test species not used in model building is comparable to random chance. Protein lists developed by using null genetic models are not expected to be enriched in any functional ontology terms beyond that expected by random chance alone.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data generated in this study are provided in the Supplementary Data and Source Data files. Grass and mammalian protein sequence alignments required to reproduce the ESL-PSC analyses in this article can be found at: <https://github.com/kumarlabgit/ESL-PSC>. Source data are provided with this paper.

Code availability

A GitHub repository containing scripts and software used to perform the ESL-PSC analyses in this study is available at <https://github.com/kumarlabgit/ESL-PSC>. Code used to implement comparison methods and simulations is also included in the repository.

References

1. Sackton, T. B. & Clark, N. Convergent evolution in the genomics era: new insights and directions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20190102 (2019).
2. Marcovitz, A. et al. A functional enrichment test for molecular convergent evolution finds a clear protein-coding signal in echolocating bats and whales. *Proc. Natl Acad. Sci. USA* **116**, 21094–21103 (2019).
3. Lee, J.-H. et al. Molecular parallelism in fast-twitch muscle proteins in echolocating mammals. *Sci. Adv.* **4**, eaat9660 (2018).
4. Liu, Z., Qi, F.-Y., Xu, D.-M., Zhou, X. & Shi, P. Genomic and functional evidence reveals molecular insights into the origin of echolocation in whales. *Sci. Adv.* **4**, eaat8821 (2018).
5. Zou, Z. & Zhang, J. No genome-wide protein sequence convergence for echolocation. *Mol. Biol. Evol.* **32**, 1237–1241 (2015).
6. Thomas, G. W. C. & Hahn, M. W. Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. *Mol. Biol. Evol.* **32**, 1232–1236 (2015).
7. Xu, S. et al. Genome-wide convergence during evolution of mangroves from woody plants. *Mol. Biol. Evol.* **34**, 1008–1015 (2017).

8. Kumar, S. & Sharma, S. Evolutionary sparse learning for phylogenomics. *Mol. Biol. Evol.* **38**, 4674–4682 (2021).
9. Yuan, Y. et al. Comparative genomics provides insights into the aquatic adaptations of mammals. *Proc. Natl. Acad. Sci. USA*. **118**, e2106080118 (2021).
10. He, Z. et al. Convergent adaptation of the genomes of woody plants at the land-sea interface. *Natl. Sci. Rev.* **7**, 978–993 (2020).
11. Parker, J. et al. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**, 228–231 (2013).
12. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*. **102**, 15545–15550 (2005).
13. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
14. Partha, R., Kowalczyk, A., Clark, N. L. & Chikina, M. Robust method for detecting convergent shifts in evolutionary rates. *Mol. Biol. Evol.* **36**, 1817–1830 (2019).
15. Kowalczyk, A., Partha, R., Clark, N. L. & Chikina, M. Pan-mammalian analysis of molecular constraints underlying extended lifespan. *Elife* **9**, e51089 (2020).
16. Farré, X. et al. Comparative analysis of mammal genomes unveils key genomic variability for human life span. *Mol. Biol. Evol.* **38**, 4948–4961 (2021).
17. Liu, Y. et al. Convergent sequence evolution between echolocating bats and dolphins. *Curr. Biol.* **20**, R53–R54 (2010).
18. Li, Y., Liu, Z., Shi, P. & Zhang, J. The hearing gene Prestin unites echolocating bats and whales. *Curr. Biol.* **20**, R55–56 (2010).
19. Liu, Z., Qi, F.-Y., Zhou, X., Ren, H.-Q. & Shi, P. Parallel sites implicate functional convergence of the hearing gene prestin among echolocating mammals. *Mol. Biol. Evol.* **31**, 2415–2424 (2014).
20. Christin, P.-A. et al. Evolutionary switch and genetic convergence on rbcL following the evolution of C4 photosynthesis. *Mol. Biol. Evol.* **25**, 2361–2368 (2008).
21. Parto, S. & Lartillot, N. Molecular adaptation in Rubisco: Discriminating between convergent evolution and positive selection using mechanistic and classical codon models. *PLoS One* **13**, e0192697 (2018).
22. Casola, C. & Li, J. Beyond RuBisCO: convergent molecular evolution of multiple chloroplast genes in C4 plants. *PeerJ* **10**, e12791 (2022).
23. Yogadasan, N., Doxey, A. C. & Chuong, S. D. X. A machine learning framework identifies plastid-encoded proteins harboring C3 and C4 distinguishing sequence information. *Genome Biol. Evol.* **15**, evad129 (2023).
24. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.* **58**, 267–288 (1996).
25. Maddison, W. P. Testing character correlation using pairwise comparisons on a phylogeny. *J. Theor. Biol.* **202**, 195–204 (2000).
26. Heyduk, K., Moreno-Villena, J. J., Gilman, I. S., Christin, P.-A. & Edwards, E. J. The genetics of convergent evolution: insights from plant photosynthesis. *Nat. Rev. Genet.* **20**, 485–493 (2019).
27. Gowik, U. & Westhoff, P. The path from C3 to C4 photosynthesis. *Plant Physiol.* **155**, 56–63 (2011).
28. Kapralov, M. V., Smith, J. A. C. & Filatov, D. A. Rubisco evolution in C₄ eudicots: an analysis of Amaranthaceae sensu lato. *PLoS One* **7**, e52974 (2012).
29. Besnard, G. et al. Phylogenomics of C4 photosynthesis in sedges (Cyperaceae): multiple appearances and genetic convergence. *Mol. Biol. Evol.* **26**, 1909–1919 (2009).
30. Meier, L., Van De Geer, S. & Bühlmann, P. The group lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70**, 53–71 (2008).
31. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. A Sparse-Group Lasso. *J. Comput. Graph. Stat.* **22**, 231–245 (2013).
32. Parto, S. & Lartillot, N. Correction: Molecular adaptation in Rubisco: discriminating between convergent evolution and positive selection using mechanistic and classical codon models. *PLoS One* **13**, e0196267 (2018).
33. Grass Phylogeny Working Group II. New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. *N. Phytol.* **193**, 304–312 (2012).
34. Dunning, L. T. et al. Introgression and repeated co-option facilitated the recurrent emergence of C4 photosynthesis among close relatives. *Evolution* **71**, 1541–1555 (2017).
35. Fukushima, K. & Pollock, D. D. Detecting macroevolutionary genotype–phenotype associations using error-corrected rates of protein convergence. *Nat. Ecol. Evol.* **7**, 155–170 (2023).
36. Chabrol, O., Royer-Carenzi, M., Pontarotti, P. & Didier, G. Detecting the molecular basis of phenotypic convergence. *Methods Ecol. Evol.* **9**, 2170–2180 (2018).
37. Scornavacca, C. et al. OrthoMaM v10: scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Mol. Biol. Evol.* **36**, 861–862 (2019).
38. Davies, K. T. J., Cotton, J. A., Kirwan, J. D., Teeling, E. C. & Rossiter, S. J. Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence. *Heredity* **108**, 480–489 (2012).
39. Shen, Y.-Y., Liang, L., Li, G.-S., Murphy, R. W. & Zhang, Y.-P. Parallel evolution of auditory genes for echolocation in bats and toothed whales. *PLoS Genet.* **8**, e1002788 (2012).
40. Li, G., Wang, J., Rossiter, S. J., Jones, G. & Zhang, S. Accelerated FoxP2 evolution in echolocating bats. *PLoS One* **2**, e900 (2007).
41. Murrell, B. et al. Gene-wide identification of episodic selection. *Mol. Biol. Evol.* **32**, 1365–1371 (2015).
42. Zhang, G. et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
43. Clark, N. L., Alani, E. & Aquadro, C. F. Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Res.* **22**, 714–720 (2012).
44. Little, J., Chikina, M. & Clark, N. L. Evolutionary rate covariation is a reliable predictor of co-functional interactions but not necessarily physical interactions. *Elife* **12**, RP93333 (2024).
45. Thomas, G. W. C., Hahn, M. W. & Hahn, Y. The effects of increasing the number of taxa on inferences of molecular convergence. *Genome Biol. Evol.* **9**, 213–221 (2017).
46. Naser-Khdour, S., Minh, B. Q., Zhang, W., Stone, E. A. & Lanfear, R. The prevalence and impact of model violations in phylogenetic analysis. *Genome Biol. Evol.* **11**, 3341–3352 (2019).
47. Zhang, J. & Kumar, S. Detection of convergent and parallel evolution at the amino acid sequence level. *Mol. Biol. Evol.* **14**, 527–536 (1997).
48. Zou, Z. & Zhang, J. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Mol. Biol. Evol.* **32**, 2085–2096 (2015).
49. Wang, Y., Yao, Q., Kwok, J. T. & Ni, L. M. Generalizing from a few examples: a survey on few-shot learning. *ACM Comput. Surv.* **53**, 1–34 (2021).
50. Song, Y., Wang, T., Cai, P., Mondal, S. K. & Sahoo, J. P. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Comput. Surv.* **55**, 1–40 (2023).
51. Castoe, T. A. et al. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc. Natl. Acad. Sci. USA* **106**, 8986–8991 (2009).
52. Liu, J., Ji, S. & Ye, J. SLEP: Sparse Learning with Efficient Projections. *Arizona State University, Tempe, AZ, USA* (2009).
53. Sanderford, M., Sharma, S., Stecher, G., Liu, J., Ye, J. & Kumar, S. MyESL: Sparse learning in molecular evolution and phylogenetic analysis. *arXiv preprint arXiv:2501.04941*. (2025).
54. Xie, Z. et al. Gene set knowledge discovery with enrichr. *Curr. Protoc.* **1**, e90 (2021).

55. Gene Ontology Consortium. The Gene Ontology resource: enriching a GO mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).
56. Smith, C. L. & Eppig, J. T. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **1**, 390–399 (2009).
57. Piñero, J. et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855 (2020).
58. Motenko, H., Neuhauser, S. B., O’Keefe, M. & Richardson, J. E. MouseMine: a new data warehouse for MGI. *Mamm. Genome* **26**, 325–330 (2015).
59. Kumar, S. et al. TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174 (2022).

Acknowledgements

The authors would like to thank Drs. Alessandra Lamarca, Jack Craig, and Sayaka Miura for reading the manuscript and providing many helpful suggestions; and Avery Selberg for technical assistance with Hyphy. This work was supported by research grants from the National Institutes of Health to S.K. (R35GM139540-05), a fellowship to J.A. from Temple University, a National Academy of Medicine Catalyst Award to G.G. and S.K., and Temple University OVPR Catalyst Award to G.G. and S.K. This research includes calculations carried out on HPC resources supported in part by the National Science Foundation through major research instrumentation grant number 1625061 and by the US Army Research Laboratory under contract number W911NF-16-2-0189.

Author contributions

S.K. conceived the idea and developed the initial method; M.S., S.S., and R.P. implemented the initial method and the underlying MyESL software; J.A. and S.K. designed the experiments. J.A. extended the method, implemented the ESL-PSC pipeline, conducted the experiments, and designed and implemented simulations and benchmarking analyses; S.S. contributed ideas to improve the technique and ran CSUBST analyses for benchmarking; J.A., S.K., and G.G. wrote the manuscript; K.T., S.V., and G.G. contributed conceptual advice and feedback; and all authors contributed to intellectual discussions about the method and results.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-58428-8>.

Correspondence and requests for materials should be addressed to Glenn S. Gerhard or Sudhir Kumar.

Peer review information *Nature Communications* thanks László Nagy and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025