nature reviews genetics

https://doi.org/10.1038/s41576-026-00933-7

Review article

Check for updates

The genetic foundations of convergent traits

John B. Allard[1,2] & Sudhir Kumar[1,2] ✉

Abstract

Convergent phenotypic evolution, the independent acquisition of similar or nearly identical traits in multiple species, is widespread throughout the tree of life. These cases of repeated evolution offer an opportunity to investigate shared genetic changes underlying shared traits, thereby linking genotypes to phenotypes. Genetic convergence can take many forms: identical amino acid or nucleotide substitutions; non-identical changes in orthologous genes or other elements; losses or gains of the same genetic elements; or convergent shifts in molecular evolutionary characteristics, such as substitution rates, amino acid preferences and selection strength. However, identifying adaptive genetic convergence, whereby evolved traits provide a fitness advantage, is challenging due to a pervasive background of random convergence that causes low signal-to-noise ratios. Numerous computational methods, including machine learning and artificial intelligence approaches, have been developed to detect, interpret and predict molecular convergence across multiple levels of genetic organization in multicellular organisms. These emerging approaches offer novel avenues to uncover the genetic foundations of complex and biomedically important traits.

Sections

[1]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA. [2]Department of Biology, Temple University, Philadelphia, PA, USA. ✉e-mail: s.kumar@temple.edu

Introduction

The convergent evolution of similar traits in independent lineages is a pervasive phenomenon across the tree of life[1]. Examples include major morphological and physiological adaptations such as the camera eye of vertebrates and invertebrates, powered flight in birds, mammals and insects, and various body plan optimizations such as streamlining for aquatic living[2]. Trait convergence is also commonly seen at finer biological scales, encompassing metabolism, diet, sensory capabilities, resistance to environmental stress, and other characteristics that may not be visible or easily detected[2]. Convergent phenotypic evolution demonstrates repeatability and determinism in evolution[3] and provides compelling evidence that adaptive evolution, rather than fitness-neutral genetic drift, has driven phenotypic change in such cases.

The presence of convergent traits in organisms belonging to distinct evolutionary lineages offers the potential to identify changes at the genetic level whose molecular mechanistic effects underlie a phenotypic trait[4]. Molecular evolutionary convergence associated with trait convergence can manifest in various ways and at multiple genetic organizational levels[5–10], visible at different degrees of 'magnification' (Fig. 1). At the most granular scale are identical nucleotide substitutions at alignable sites, or identical amino acid substitutions at the same sites in homologous proteins[11–14]. Other forms of molecular convergence include different substitutions at the same sites[15], substitutions at different sites within the same protein[5], convergent structural variants such as insertion and deletions (indels)[16] and copy number changes[17], and gains and losses of entire genes[18]. Beyond coding sequences, convergence can arise from modifications in regulatory elements that alter the amplitude, timing or spatial patterning of gene expression[19,20], or from changes in different genes with related functional roles[21,22]. Truly unique genetic solutions might also exist for the evolution of convergent traits. The most expansive definition of genetic convergence would encompass diverse changes across genetic elements within similar biological categories and pathways. For example, phenotypic convergence may be driven by distinct genetic changes that trigger similar molecular mechanistic effects, such as gene copy number expansion in some lineages and amino acid substitutions in others, as in the case of glyphosate resistance across plant species[23]. This phenomenon should also be considered genetic convergence. Thus, at the highest levels of genetic organization, the line between convergence and unique solutions can become a continuum.

Traditional comparative genomics approaches can reveal functional elements by using extraordinary evolutionary conservation across distant taxa as an indication of strong purifying selection (that is, the selective removal of alleles that are deleterious; also known as negative selection) to preserve function[24]. Within a given clade, some genomic elements are uniquely constrained in this way compared with related clades, which could be due to the preservation of clade-specific innovations[25] (Fig. 2a). This basic approach has been a powerful tool for discovering and annotating important regions of genomes[26]. However, most clades share many traits, making it difficult to link specific constrained elements to specific phenotypes without additional evidence. In cases of convergent evolution, where traits have arisen independently in distant organisms (Fig. 2b), comparative scans for shared genetic signatures of convergence can enable more precise mapping of genotype to phenotype. However, the task of mapping molecular convergence to trait convergence is complicated by several factors, including the nature of genetic innovations, the choice of computational methods, the need for high-quality genomes and alignments, and the difficulty of establishing evolutionary orthology of genetic loci, particularly for non-coding elements. Additionally, adaptive molecular convergence can involve a small number of changes in one or a few genomic segments, making it difficult to establish statistically against a large neutral background[27]. With the advent of new computational techniques and more affordable sequencing, genotype–phenotype links can now be pursued systematically to uncover the genetic basis of complex traits. In numerous cases, genetic convergence has been experimentally confirmed to confer the molecular properties enabling the corresponding convergent phenotype[28–37]. Many well-documented examples exist that suggest convergent genomic evolution can, in fact, explain convergent traits, sometimes through only a handful of key genetic changes, such as the amino acid substitutions that underlie the evolution of red visual pigments from green pigments in humans and fish[38–40].

In this Review, we synthesize current knowledge on the spectrum of molecular convergence in multicellular organisms at various levels of genetic organization (Fig. 1). We discuss the strengths and weaknesses of available classes of methods in the context of cases in which comparative genomics of convergent phenotypes is shedding light on complex traits and human health (see ref. 41 for additional methodological details). Finally, we examine new artificial intelligence-based methods that show promise in detecting previously elusive molecular convergent patterns and advancing beyond retrospective detection towards predictive models.

Convergent amino acid substitutions

This section traces the field's historical trajectory, starting from the first examples of genetic convergence noted in the literature at the amino acid level to the recent trends towards proteome-scale scans for convergence in protein sequences. As a note on nomenclature, convergent and parallel evolutionary substitutions differ in that the ancestral state (residue) in the former is different in the species considered, whereas they are the same in the latter. For simplicity, we consider both together and refer to them as convergent substitutions throughout this Review.

Observations in individual proteins

Early on, distantly related taxa bearing the same convergent trait were found to cluster together in phylogenies reconstructed using protein sequence alignments, such as the harbour seal and the whales and dolphins (aquatic mammals) in the case of vertebrate myoglobin, and the langur monkey and the ruminants (foregut fermenters) in the case of lysozyme[11–13]. A preponderance of convergent residue substitutions misled phylogenetic methods and produced incorrect species relationships[11–13]. However, claims of sequence convergence based on conflicts between true and reconstructed trees were challenged[42] because distortion of species groupings in a molecular phylogeny (incongruence) is neither necessary nor sufficient to conclude that molecular sequences have converged[42,43]. In 1997, a statistical test was introduced that estimates the expected number of convergent amino acid substitutions under neutral evolution and calculates the probability of observed convergence using species phylogeny, branch lengths and amino acid frequencies; this study found statistically significant traces of molecular convergence in the lysozyme proteins of foregut fermenters[44].

At the turn of the twenty-first century, as sequencing technologies and volumes advanced, studies began to find new, clear examples of sequence convergence. Convergent evolution was documented in genes underlying C4 photosynthesis, a highly efficient carbon fixation pathway that evolved independently multiple times in grasses[45,46]. Phosphoenolpyruvate carboxylase (PEPC) and RuBisCo were found to have undergone convergent amino acid changes in multiple C4 lineages, often via identical nucleotide substitutions. Widespread convergent evolution of resistance to tetrodotoxin through convergent amino acid substitutions in many paralogous sodium channel genes was reported across diverse species of pufferfishes[47]. The *SLC26A5* gene, which encodes the motor protein prestin that is essential for high-frequency hearing in mammals[48], was reported to have experienced convergent amino acid substitutions in echolocating bats and toothed whales[49–51]. Similar patterns of convergence were later identified in additional hearing-related genes[52,53]. These examples, spanning traits as diverse as carbon metabolism, toxin resistance and sensory adaptation, demonstrated that amino acid convergence could be repeatedly detected and linked to iconic phenotypes.

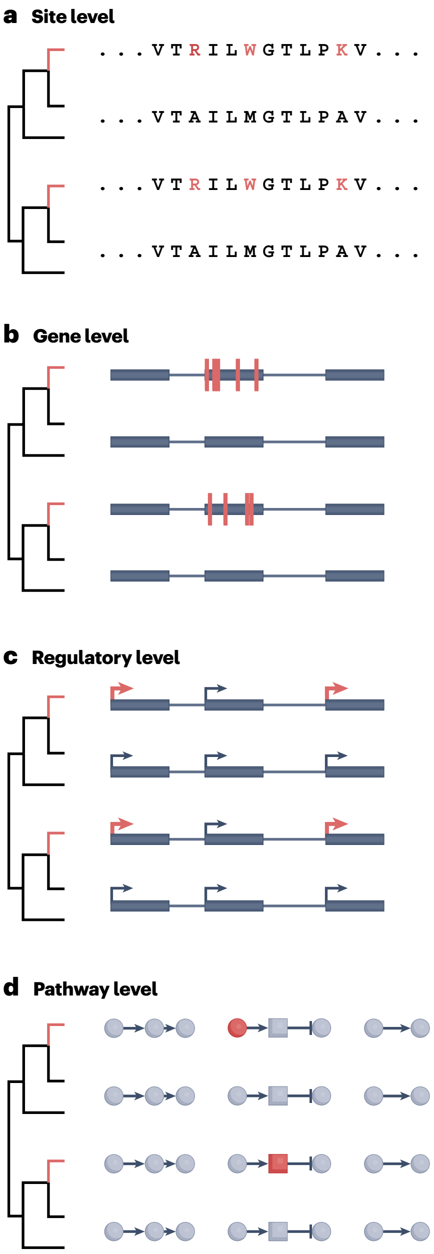

Fig. 1 | Convergent molecular evolution at different genetic levels of organization. a–**d**, The trees shown at the left contain two convergent species (red branches) and all other tips and ancestral nodes (black branches) exhibit the ancestral phenotype and genotype. Different forms of molecular convergence that occurred independently on branches leading to convergent species are shown to the right of the convergent species in each panel, and the ancestral genetic states are shown to the right of the non-convergent sibling species. **a**, Site-level convergence. The most granular level of molecular convergence is an identical substitution. Three identical amino acid substitutions are shown (red residues). Identical nucleotide substitutions may also be involved in convergence in regulatory regions, but this is much less studied. **b**, Gene-level convergence. Three coding sequences are depicted (dark grey rectangles). In convergent species, red vertical markers indicate the positions of derived residues, which in this case are not shared between the convergent species. However, the gene in question has had many changes in the two convergent species and not in the siblings that maintain the ancestral phenotype. This gene may thus be detected in a scan for positively selected genes or evolutionary rate shifts. **c**, Regulatory and expression-level convergence. A genomic region containing three genes (dark grey rectangles) is shown. Angled arrows depict expression levels, with larger arrows indicating greater expression. Expression levels have increased convergently in these genes in the convergent species. **d**, Pathway-level convergence. Three pathways are shown for each convergent and ancestral lineage. The affected pathway component genes in the convergent species are highlighted (red). Convergent changes affected the same pathway, but not the same components in both convergent species.

Whole-proteome scans for convergent substitutions

As genome-scale data became available, researchers sought to extend the detection of convergence from isolated genes to the full proteome. This shift promised to uncover not only known cases of convergence at scale but also novel loci associated with convergent traits whose molecular basis remained poorly understood. For example, a dramatic case of convergence across mitochondrial proteins was reported between snakes and agamid lizards[54]; phylogenies reconstructed from reptile mitochondrial genome sequences did not recover some of the species relationships evident in phylogenies based on nuclear genes due to an excess of convergent substitutions. As another example, a nuclear proteome-scale scan focused on echolocation in mammals used a maximum likelihood approach to contrast the site-specific likelihood support for the known species phylogeny with that of a phylogeny that placed echolocating species in the same group[55]. Convergent molecular signatures across more than 200 loci were reported, including genes previously linked to high-frequency hearing and echolocation; however, there was no significant enrichment for genes related to hearing among these loci[55]. Soon after, reports emerged that the site-specific likelihood support approach could produce similar levels of convergence signatures even when clustering species with no shared trait, for example, grouping non-echolocating species such as cows with echolocating bats[56]. Additionally, there was no excess of convergent substitutions compared with the quantity of divergent substitutions on the phylogeny branches on which echolocation is believed to have evolved[57]. These studies highlighted how the prevalence of background neutral convergence makes it difficult to robustly detect adaptive molecular convergence[54,58–61].

Since then, a series of studies have continued the quest to find signals of convergence across entire proteomes, often using techniques based on ancestral sequence reconstruction. Such studies were

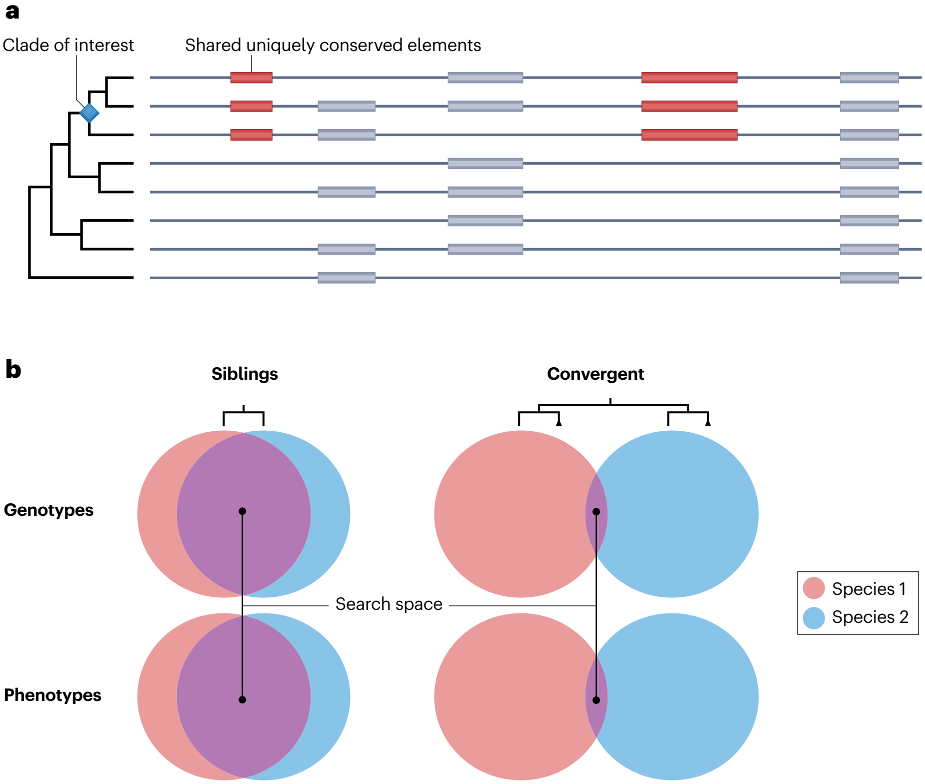

Fig. 2 | A comparison of conservation-based and convergent evolution-based comparative genomics. a, In traditional comparative genomics, the shared genetic elements that are uniquely conserved within a clade are linked to the shared phenotypes of that clade. **b**, Among sibling species in a clade, the number of uniquely shared phenotypes and genotypes can be large because they are mainly inherited from a common ancestor, leading to a large search space for the link between any one phenotype and its genotypic correlates. This increases the difficulty of linking specific traits to specific genetic elements. When distant species convergently acquired a certain trait, the search space for linked genotypic correlates is much smaller because neutral evolution tends to create separation over time in the absence of convergent selective pressure.

successful in detecting some signs of convergence in genes related to echolocation, the aquatic transition of marine mammals, intertidal adaptation of mangroves and vocal learning in birds[34,60,62–68]. To reduce the noise of apparent convergence due to background neutral sequence change, some studies applied filters to exclude non-conserved sites on the assumption that convergent substitutions at otherwise conserved sites are more likely to represent adaptive evolution[34,60,62,65,67]. This approach does result in an improvement in the signal-to-noise ratio. However, useful genetic information can be lost when ~90% of sites are excluded. For example, one study conducted ontology enrichment analysis on genes with radical convergent substitutions at conserved sites, and found strong enrichments for fast-twitch muscle-related genes but no significant enrichments for hearing-related genes[62], which other studies applying different methods have found[34,67,69]. To what extent diffuse convergence at somewhat substitution-tolerant sites may be important for the fine-tuning of protein functions that underlie convergent traits remains unclear.

A recent method, CSUBST, applied to coding DNA rather than amino acid sequences, estimates rates of both non-synonymous and synonymous convergent substitutions, whose ratio (ω_c) represents an estimate of the non-neutral convergence rate[70]. This rate is typically increased only in cases of convergent amino acid substitutions. However, similar to most methods that require statistical models of sequence evolution using maximum likelihood approaches, CSUBST is computationally demanding for genome-scale analyses with many species, but has excelled in studies using focused candidate sets[71,72]. Other methods have been developed that use evolutionary models to detect levels of sequence convergence above neutral expectations while incorporating additional approaches, such as the use of computer simulations to produce synthetic datasets to derive expectations or the application of site-specific amino acid profiles[73–77]. Because of the computational costs of fitting evolutionary models across genome-scale or proteome-scale datasets, a different class of strategies has been developed to scan the extant sequences for patterns that strictly reflect convergence to a derived residue in species with convergent traits[60,65,66,78,79].

Despite this methodological diversity, how to distinguish true adaptive convergence from the background noise of neutral or nearly neutral evolution, with statistical confidence, remains a key challenge. The choice of amino acid substitution model can greatly affect the statistical significance of observed differences[59]. Additionally, the need for multiple test corrections for gene-by-gene *P* values greatly reduces the statistical power of analyses. Consequently, studies relying on tests for enrichments of gene ontology categories and pathways among the set of inferred convergent genes have had mixed success. The case of echolocation is an excellent example, where despite numerous findings of convergence in experimental studies of individual high-frequency hearing-related genes, enrichments for hearing-related ontology terms were often only marginal[34,67]. Some machine learning-based methods have found significant enrichments of hearing-related genes among convergent echolocation genes, which we discuss below[69].

A growing collection of reports establish that convergent traits can evolve by convergent amino acid substitutions, which suggests that a narrow set of viable molecular solutions exist for the evolution of some traits. Across diverse traits and taxa, selection has repeatedly targeted the same residues, underscoring the strong biochemical and structural constraints that shape protein function. Although it remains uncertain how commonly convergent amino acid substitutions underlie convergent evolution, these repeated solutions suggest that the path to adaptation in certain functional contexts is not only reproducible but, to some extent, predictable.

Convergent evolution at the gene level

Gene-level convergence encompasses shared genetic solutions that do not require the same amino acid changes at the same sites, ranging from distinct function-altering substitutions to gene copy number and family size changes, or the de novo acquisition of similar genes[21,80–85]. Convergent shifts in evolutionary rates and other characteristics may reflect similar changes in selective pressures across lineages and can indicate gene-level convergence.

Same genes, different sites

In some cases, different amino acid substitutions at multiple sites in the same (homologous) genes across independently evolving lineages can yield similar functional outcomes. This type of gene-level convergence will not be detected by methods focused on finding identical substitutions. For example, adaptation to hypoxia in many species of high-altitude birds involved mostly unique substitutions in haemoglobin, relative to their lower-altitude siblings, that produced similar increases in O_2 affinity[36,37,86]. Digestive RNases in ruminants and foregut-fermenting colobine monkeys also evolved similar functions through entirely different substitutions[87].

The likelihood of repeated selection on the same gene without identical substitutions depends on structural and functional constraints (Fig. 3a,b). In electric fish, different amino acid substitutions in similar regions of the *scn4aa* gene, which encodes the voltage-gated sodium channel subunit expressed in the electric organ, contribute to species-specific communication signals[88]. By contrast, in many species, the homologous voltage-gated sodium channel expressed in muscle and targeted by tetrodotoxin has acquired identical convergent amino acid substitutions conferring toxin resistance[28,47]. Toxin molecules can impose strong constraints requiring specific conserved amino acid residues to change[89]. Loss-of-function mutations targeting conserved sites, disruption of protein interactions or changes affecting protein stability can each be achieved via multiple distinct substitutions. Some traits depend on changes in an aggregate property of a protein that can involve many different sites, such as the convergent increase in surface charge of myoglobin in diving aquatic mammals, which enables it to be present in very high concentrations without aggregating[90]. Conversely, extreme functional tuning in highly conserved molecular machinery – such as prestin in echolocating mammals[33,49], which must oscillate at ultra-high frequencies, and voltage-gated potassium channels in electric fish, which require extreme voltage sensitivity[40] – is often limited to a few viable substitutions due to stringent constraints (Fig. 3b).

Computational detection of gene-level convergence can be performed using various methods, which have often been study-specific. In many cases, candidate genes or gene families have been compared, rather than whole-genome scans[80,82]. Genome-wide selection scans sometimes make use of population data within or across species[91–93]. Scans for positively selected genes, as detected by an elevated ratio of non-synonymous to synonymous substitution rates aggregated over sites, can also reveal gene-level convergence[94–96]. However, a lack of statistical signals of positive selection for a gene does not necessarily rule out the presence of site-level amino acid convergence (Box 1). For convergent positive selection scans, the BUSTED-PH method (branch-site unrestricted statistical test for episodic diversification and association with phenotype)[96] improves on standard designs by testing not only for positive selection on the foreground evolutionary lineages but also a lack thereof on the background lineages, and a difference in selection strength between the two[97–100]. When selective pressure elevates evolutionary rates in the same genes across species, these changes may also be detected by evolutionary rate-convergence methods as discussed below.

Gene gains and losses

Another form of gene-level convergence involves signatures of independent gene gains or losses associated with convergent trait acquisition. Convergent gene gains often arise via gene duplications, exemplified by amylase copy number expansions in multiple human populations with starch-heavy diets[17] and gene family expansions linked to multicellularity and metabolic traits in diverse fungi[21,101]. Similarly, the acquisition of eusociality in insects is associated with expanding gene families[102,103]. Convergent whole-genome duplications can facilitate convergent genetic evolution, as in the case of crop domestication, where pre-existing duplications seem to have enhanced genomic evolvability, enabling the acquisition of advantageous traits[104]. Lineage-specific gene duplications can provide material for neofunctionalization, whereby one gene copy retains its original function and the other copy acquires a new function through mutation and selection[105]. Therefore, a search for convergent genetic signatures using single copy orthologues will miss cases of convergence in which a gene is modified or upregulated in one lineage and duplicated and neofunctionalized in another.

Convergence can also result from de novo gene evolution – the emergence of new genes from non-coding DNA, regions of existing genes or both. For example, structurally distinct anti-freeze proteins (AFPs) in fish independently arose in various lineages[83,106], such as nearly identical type I AFPs in flounders and Arctic sculpins, structurally distinct type II AFPs in sea raven and herring[106], and nearly identical anti-freeze glycoproteins in Antarctic notothenioid fish and Arctic cod[107]. These cases illustrate the capacity of selection to independently repurpose ancestral genetic elements to achieve convergent outcomes[83].

Convergent gene losses are also common, particularly where ancestral traits have independently regressed[18]. Such losses tend to occur when ecological changes render certain functions unnecessary, allowing them to decay. For example, across independent clades of obligate herbivores and obligate carnivores, many genes related to digestion and energy homeostasis were repeatedly lost in the course of dietary specialization[108]. Similar patterns occur when sensory or organ functions are no longer needed, such as the repeated loss of vision and pigmentation genes in subterranean animals or the coordinated decay of gastric acid genes in stomachless fishes[109–112].

Streamlining of gene content is adaptive when the cost of maintaining a superfluous trait is high, as in the loss of nitrogen-fixing symbioses in plants[113]. In extreme cases of adaptive streamlining, thousands of genes can be lost, as seen after the repeated losses of multicellularity in fungi[114]. Although gene losses can provide strong evidence linking genes to a convergent trait, new biochemical functions or increases in the functional capabilities of an existing modality, such as improved sensory acuity, are more likely to involve gene gains or sequence changes.

There are three primary approaches to find gene gains and losses that may underlie convergent traits. One approach is to count gene copy numbers and then test the relationship between counts for different gene families and the trait in question using phylogenetic regression and related methods[101,115]. Another approach is to cluster gene

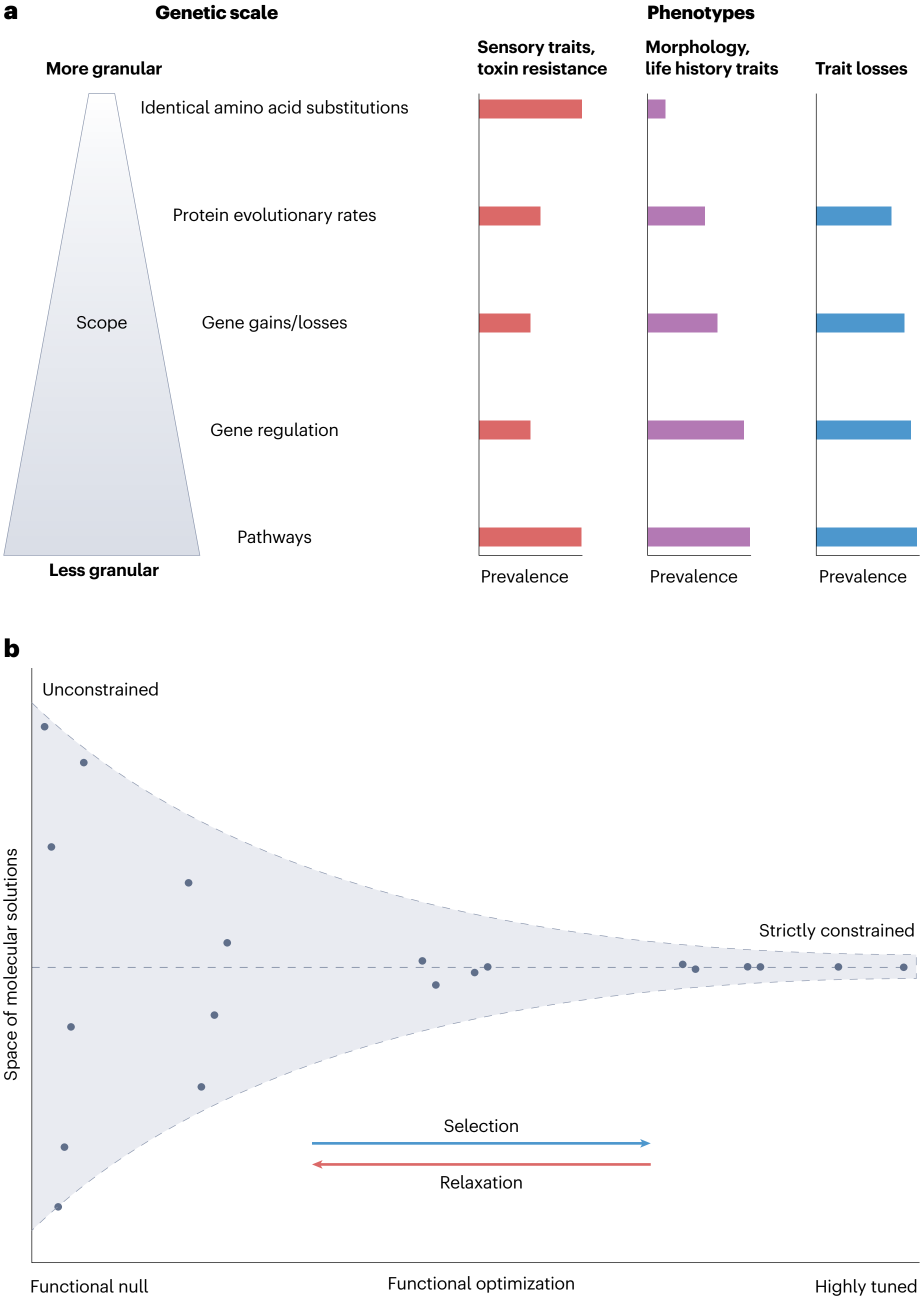

Fig. 3 | Prevalence of molecular convergence at different genetic organizational levels in different categories of traits. a, Hypothetical distributions of convergent changes across genetic scales from more to less granular. Traits whose underlying biological mechanisms are tied closely to highly tuned biochemical properties of proteins, such as many sensory traits and resistance to toxins, may be more likely to depend on convergent amino acid substitutions. Conversely, traits that are farther removed from the molecular underpinnings and more dependent on emergent interactions of many components at a macroscopic level, such as many morphological traits, may depend more on convergence at higher levels of genetic organization. Convergent losses of specific traits are more likely to occur at higher levels of organization as well. **b**, Functional optimization as a driver of granular convergence. The shaded area represents the 'space of molecular solutions', that is, the volume of genetic variation capable of producing a given phenotype. Although the absence of a trait ('functional null') permits a near-infinite number of non-functional sequence states, highly optimized traits may be supported by only a small number of molecular configurations. This canalization implies that moving 'up' the gradient of optimization is more likely to require identical residue substitutions, whereas trait losses can occur by many different paths when selection is relaxed.

sequences across species and then infer gene trees, reconcile them with species trees, and map duplications and losses onto branches. Rates of loss and gain can then be inferred on branches and compared across foreground and background branches[21,114,116]. Finally, a third category involves fitting birth–death models of gene family evolution[117,118], in which the observed copy numbers at the tips are modelled as the outcome of gene duplication and loss events along a time tree, and branches with the convergent trait are allowed to have their own duplication and loss rates. Convergence is inferred when the same gene families show significantly increased expansion or contraction on multiple independent trait-bearing lineages[119–122]. Convergent gene losses may also be detected by convergent evolutionary rate acceleration due to relaxation of negative selection, as we discuss below.

Evolutionary rate convergence

Alterations in selective pressures on a gene or regulatory element can manifest as changes in the evolutionary rates of molecular substitutions, without necessarily involving the same sites. Increased constraint leads to slower rates, whereas relaxed constraint and positive selection cause accelerated rates. Several computational methods with a similar conceptual basis have been developed to search for genes or genomic elements whose evolutionary rates show convergent shifts in connection with convergent trait acquisitions, including RERConverge[123–127] (Table 1). In addition to gene-level convergence, rate convergence has also been applied to study gene regulatory convergence (discussed below)[125,128–132]. One advantage of these methods is that they do not rely on site-by-site tests, making them well suited to studying traits shaped by distributed genomic changes. However, they generally cannot distinguish between rate acceleration due to positive selection and relaxation of purifying selection (for example, when a gene becomes a pseudogene)[123,128]. Nevertheless, these methods have excelled at finding the genetic basis of trait losses because loss of purifying selection can lead to rapid evolutionary change in formerly conserved elements[133], such as genes associated with vision in subterranean mammals[109,125], gustatory and olfactory genes in marine mammals[134] and hair-related genes in hairless mammals[132], among others[133].

The convergent transition to marine aquatic life in mammals, which occurred independently in cetaceans, pinnipeds and sirenians, is a particularly interesting example for contrasting site-based and rate-based methods. It has been the focus of multiple studies that sought convergent amino acid substitutions[63,64,68], and has been investigated using multiple methods to search for evolutionary rate convergence in genes[127,128,133–135]. Amino acid convergence methods found relatively few genes with convergent substitutions and only weak pathway or category enrichments that were difficult to interpret biologically[63,64,68]. By contrast, rate-based methods revealed hundreds of genes with convergent rate shifts, strongly enriched for skin components, sensory systems and muscle or metabolic adaptations clearly relevant to marine life[127,134]. The strength of rate-based methods, in this case, may stem from the fact that what we observe as a single phenotype is often a multiplicity of phenotypes that coevolved with marine transitions, in many cases through the regression of ancestral terrestrial adaptations[133,136–138]. Rate convergence has thus emerged as a useful approach for traits that manifest through changes in constraint across multiple dispersed genomic loci[139].

Higher-level convergence

A higher level of genetic convergence includes cases in which the shared genetic solutions affect elements beyond homologous gene sequences, including regulatory DNA and non-homologous genes that belong to the same pathways or functional categories.

Regulatory and expression-level convergence

A substantial amount of evolutionary change between species can be explained by changes in spatial, temporal and conditional gene expression patterns resulting from non-coding regulatory DNA changes[140–143]. Convergent evolution similarly can arise through parallel alterations in gene regulation leading to parallel changes in gene expression, which underlies traits including social insect castes[144], flightlessness and tarsus length in birds[145,146], transitions to asexuality in stick insects[147], electric organ development in electric fish[105,148], viviparity in cyprinodontiform fish[99], floral coloration and pollinator type[149,150], hibernation and body size in mammals[151,152], and venom production in various animal lineages[153]. A change in gene regulation may be less prone to deleterious pleiotropic effects compared with amino acid substitutions because it might affect expression quantity, timing or tissue specificity only mildly or conditionally, whereas a change in the amino acid sequence will affect the protein in all settings[140].

Convergence in gene regulation can be detected, even when the underlying genomic changes are not known, by experimentally measuring expression levels in one or more focal tissue types and comparing between species with and without the convergent trait. For traits whose mechanism is known to involve a certain anatomical feature, such as the electric organ in electric fish[148], this organ or tissue provides a natural focus for expression studies. Expression can be quantified using RNA sequencing, reverse transcription quantitative PCR or microarray analyses, and convergence is inferred by testing whether orthologous genes show consistent direction (and often magnitude) of expression change across independent origins using a variety of statistical methods, including replicated paired designs or phylogenetically aware

Box 1 | The relationship between positive selection and amino acid convergence

Adaptive evolution of amino acid sequences under positive selection is often detected using the non-synonymous/synonymous rate ratio (ω or dN/dS), which is expected to be elevated (ω>1) relative to the (strict) neutral expectation (ω=1)[205,206]. Because convergent evolution is an adaptive process, it has often been assumed that convergent protein sequence evolution should result in an elevated ω value. Many studies of the genetics of convergent traits have begun with a search for genes with evidence of positive selection using ω-based methods[63,64,95,207,208]. Positive selection that occurred in the same genes or sites across multiple independent clades (and not outside them) can certainly be evidence of convergent molecular evolution[45,46,209]. However, statistical evidence of positive selection in genes and/or sites is neither necessary nor sufficient to demonstrate adaptive convergence in amino acid sequences.

Convergent amino acid substitutions can result from directional selection due to a change in amino acid preferences at a site, but the affected site will experience purifying selection for most of its history, which would make ω<1 on that evolutionary lineage[15,73,210]. It also can be difficult to infer positive selection at the gene level when the proportion of codons affected is small[27,211]. Even a single change can be enough to alter a phenotype at the molecular level[39,40,171], and the most thoroughly investigated convergent genes found to date only have a handful of verified convergent substitutions in most cases[33,45,52,212]. For example, no hearing-related genes were found to be positively selected on the stem branch of the echolocating toothed whales, despite many classic convergent echolocation genes present in this clade[213]. Studies that have searched for both positively selected genes and genes with convergent substitutions have, in most cases, found relatively small overlaps between the two categories[63,64,214–216]. It is likely that scans for convergence using ω-based methods detect cases of genes that have undergone diversifying selection, which tends to produce much stronger elevations in ω compared with directional selection[217]. Therefore, although convergent amino acid substitutions are sometimes found at sites and in genes with elevated ω values, it is best to consider these as distinct forms of molecular convergence and, when possible, use specialized amino acid convergence methods along with ω-based methods.

Table 1 | Major families of methods for detecting genetic convergence underlying convergent trait evolution

Method family	Subcategory	Primary signal / when to use	Main strengths	Main caveats	Representative tools and examples
Amino acid convergence		Recurrent amino acid changes in orthologous proteins on convergent evolutionary lineages	Direct link to protein sites and biochemistry	Generally sensitive to alignment/tree error; could be low power for polygenic traits	
	ASR-based inference of convergent substitutions	Identical amino acid changes reconstructed on foreground versus background lineages	Very intuitive; high specificity for 'same site, same substitution' scenario	Needs substitution model choices and assumptions; rather stringent; many real cases may use different amino acids; can be computationally costly	CSUBST[70]; ASR-based counting tests and their genome-wide extensions[34,44,62,63,67]
	Pattern-based amino acid convergence (no ASR)	Extant amino acid patterns enriched in trait species (without reconstructing ancestors)	Avoids ASR; can use tip sequences directly	Still sensitive to alignment; may pick up clade-specific patterns unrelated to trait	CCS[60]; and related pattern-based approaches
	Codon/AA profile-shift methods	Foreground lineages share an altered amino acid preference/profile at sites	More flexible than identical-substitution tests; detects shared biochemical preferences	Model-dependent; heavier computation; interpretation at 'profile' level, not single amino acids	PCOC[75]; TDG09[203]; Pelicon and related profile-change codon/amino acid models[74]
Selection and rate convergence in coding genes		Convergent changes in selection strength or evolutionary rate across genes	Good for distributed, polygenic signals and subtle constraint changes	Hard to distinguish positive selection versus relaxed constraint; trait-correlated confounders matter	
	Branch-wise ω ($\mathrm{d}N/\mathrm{d}S$) and selection-intensity models	Elevated or relaxed ω or selection intensity on convergent branches	Highlights genes under altered selection during trait evolution; works well at the gene level	ω can change for many reasons; limited power with few convergent lineages; can miss directional selection at a few sites	Branch/branch-site/clade models (HyPhy and codeml); BUSTED-PH[96]
	Gene-wise relative rate convergence	Correlation between relative gene rates and trait across the phylogeny	Powerful and fast for genome-wide scans; does not need specific substitutions	Sensitive to branch-length error and confounders (e.g., life history); multiple-testing burden	RERConverge[123]; Forward Genomics[125]; and related rate-based methods
Gene content (gain or loss and family size)		Convergent expansions, contractions or losses of gene families in trait lineages	Natural for gene dosage, innovation and regressive traits; models duplication/loss explicitly	Depends on good annotation and orthogroups; whole-genome events can obscure trait-specific changes	Gene family birth–death models[117,118]; gene tree-based duplication mapping
Regulatory sequence convergence		Convergent acceleration, deceleration or activity shifts of CNEs/enhancers/TFBSs	Captures non-coding mechanisms when coding signal is weak; many traits may evolve primarily through regulatory changes	Needs good multi-species genomes/alignments; mapping elements to target genes is uncertain	RERConverge[123]; PhyloAcc[145], REforge[204]
Expression-level convergence		Convergent gene expression patterns (differentially expressed genes, modules, principle components) in trait species	Shows functional read-out; does not require genome alignments	Strongly affected by tissue matching, environment, batch effects; orthology mapping limits power	RNA-seq and differential expression analyses[153,156]
Machine learning-based methods		High-dimensional patterns in sequences or regulatory data that predict trait versus non-trait	Integrates many weak signals; can use rich representations (embeddings, CNN features)	Risk of overfitting; interpretation can be challenging in some cases; needs careful validation and good labels	
	Sequence-based sparse or penalized models	Trait prediction from aligned sequences (positions or groups as features)	Transparent and interpretable feature weights; highlights key genes/sites; does not need branch lengths	Orthology assignments and design of contrasts are key; traits may evolve without any identical amino acid convergence	ESL-PSC[69]
	Regulatory machine learning models	Trait prediction from regulatory intervals (e.g., OCRs, enhancers) with CNNs	Captures complex motifs and combinatorial regulatory patterns; links enhancers to traits	Requires many species with high-quality full genome alignments; model more 'black-box'	TACIT[174]
	pLM-based convergence	Detection of convergence in protein structural/functional features beyond identical amino acid substitutions	Uses information in embeddings generated by pretrained models; may detect cryptic similarity beyond amino acid identity	Very new; best practices and benchmarks still emerging; embeddings are high dimensional	ACEP[198]

ACEP, adaptive convergence by embedding of protein; ASR, ancestral state reconstruction; BUSTED-PH, branch-site unrestricted statistical test for episodic diversification and association with phenotype; CCS, convergence at conservative sites; CNE, conserved non-coding element; CNN, convolutional neural network; ESL-PSC, evolutionary sparse learning with paired species Contrast; HyPhy, hypothesis testing using phylogenies; OCR, open chromatin region; PCOC, profile change with one change; pLM, protein language model; RNA-seq, RNA sequencing; TACIT, tissue-aware conservation inference toolkit; TDG09, Tamuri–dos Reis–Goldman 2009 profile-change codon model; TFBS, transcription factor binding site; ω ($\mathrm{d}N/\mathrm{d}S$), non-synonymous/synonymous rate ratio.

regression models[147–149,153–156]. Convergent changes in expression can result from shared *trans*-regulatory mechanisms[157] or from primarily parallel *cis*-regulatory alterations[158].

At the regulatory DNA level, computational methods that search for evolutionary rate acceleration in conserved non-coding elements (CNEs), including RERConverge and PhyloACC[125,128,159–161] (Table 1), have revealed thousands of CNEs whose accelerated substitution patterns track convergent phenotypes, such as wing loss in ratite birds[145] and seasonal hibernation in mammals[151], among others[20,99,131]. In a small number of cases, the chain of causality from convergent nucleotide-level changes to expression changes and, finally, to the phenotypic change has been identified[19,145,162]. For example, to examine three convergent origins of gliding membranes in marsupials, scanning for accelerated non-coding regions in one or more gliders revealed a gene, *Emx2*, that was expressed in the developing gliding membrane and was enriched for nearby accelerated regions[20]. Reporter assays showed that accelerated DNA regions near *Emx2* acted as enhancers, boosting *Emx2* expression, and in vivo *Emx2* knockdown in developing tissue reduced membrane outgrowth, supporting a causal role for EXM2 (ref. 20).

Where scans for regulatory convergence and other forms of molecular convergence have been performed on the same genomes, there is generally little overlap between genes affected by each type of convergence[99,145]. In a joint scan for expression and amino acid convergence (no particular target trait) across thousands of lineage pairs on a tree of vertebrates and >16,000 orthology groups, only 33 pairs were found to have convergence of both types[70]. Differences in methodological power and data abundance make it difficult to interpret the scarcity of such cases. Whether molecular convergence appears predominantly in a single level of genetic organization for a given trait or whether multiple levels are commonly involved remains an open question, and more studies addressing convergence across multiple genetic organizational levels are warranted.

Convergence at the level of pathways

Even when the specific genetic changes responsible for a convergent trait differ across lineages, the repeated involvement of a shared pathway or functional category of genes represents evidence of convergent phenotypic evolution. We can think of this as the highest level of genetic organizational convergence (Fig. 1d).

Across diverse systems, convergent phenotypes repeatedly arise from changes that primarily affect different genes within the same pathways or gene networks. High-altitude adaptations have repeatedly involved modifications to different genes within the hypoxia-inducible factor (HIF) pathway to similarly enhance oxygen transport in low-oxygen environments[82,92]. Distinct but functionally similar genes facilitated independent foraging adaptations in Myotis bats[22]. Multiple lineages of sulfide-tolerant fishes independently adapted through convergent expression changes in genes within mitochondrial pathways involved in oxidative phosphorylation and sulfide detoxification, although the specific affected genes differed among species[163]. Across independently evolved eusocial insects, caste-associated transcriptomes show only partial reuse of the same genes but consistent enrichment of overlapping reproductive, developmental and metabolic pathways, indicating that convergent caste phenotypes mainly arise by repeatedly tapping into similar gene networks rather than identical loci[144,164]. This aligns with a proposed framework in which eusociality repeatedly recruits conserved nutrition–growth–reproduction regulatory systems[165]. In plants, repeated genetic co-option within the cytokinin biosynthesis pathway has driven the independent evolution and loss of prickles across diverse species[80]; arctic Brassicaceae species exhibit similar physiological adaptations to extreme cold and drought, using different genes within shared stress response pathways[166]; and different genes within anthocyanin biosynthesis pathways repeatedly drove floral pigmentation changes in Iochrominae[149]. Collectively, these findings underscore how adaptive evolution can convergently harness the same biological pathways, illustrating functional predictability at the pathway level despite variability in the specific genetic details.

Some evidence from microbial experimental evolution studies suggests that convergent change becomes progressively more likely at higher levels of genetic organization[5,167–169], although the extent to which this is true may depend on the specific traits and environmental adaptations under consideration. Higher-level convergence might be more common in traits that are farthest from molecular phenotypes or depend on a vast concert of different gene functions, such as adaptation to high temperature[5] or life history traits[167]. Gene- and amino acid-level convergence may be required for traits that are enabled directly by molecular biochemical mechanisms in specific proteins, such as sensory abilities, toxin or drug resistance and electrical properties, especially when extremely high molecular optimization is required[33,40,47,53,62,89,170–173] (Fig. 3a,b).

Morphological and other developmentally determined traits that result from slight differences in the timing or localization of morphogenic signalling proteins might depend more on regulatory and gene regulatory convergence[146]. Trait losses are more likely to involve gene losses, whereas great increases in complexity or wholly new functionalities could involve gene family expansions or de novo gene births respectively. It is difficult to draw strong conclusions given the limited catalogue of verified cases of molecular convergence, but this hypothesis deserves further consideration and future studies should determine the frequencies of different modes of molecular convergence for different types of traits.

Machine learning and artificial intelligence for convergent comparative genomics

Traditional methods for detecting genetic convergence typically rely on explicit evolutionary models to reconstruct ancestral states, estimate substitution rates or test for shifts in selection strength associated with a trait. These approaches have yielded many insights, but they require strong assumptions about how sequences evolve and often analyse sites, genes or regulatory elements in isolation, which can limit power and scalability when convergence is sparse, polygenic or distributed across molecular levels. Machine learning frameworks provide a complementary strategy: they learn patterns that distinguish convergent from non-convergent lineages directly from high-dimensional genomic data, can jointly consider many loci and features at once, and reduce dependence on any single prespecified evolutionary model. As a result, they offer a flexible route to detect genetic signatures that underlie convergent phenotypes across all genetic levels.

The ESL-PSC (evolutionary sparse learning with paired species contrast) method departs from the standard molecular phylogenetic protocols by building a genetic model of trait convergence using all relevant loci simultaneously[69]. The experimental design relies on paired comparisons between closely related species, where one member of each pair exhibits the convergent trait and the other does not (Fig. 4a). This pairing allows the neutral and ancestral phylogenetic background to be implicitly subtracted. The inferred ESL-PSC models can be used to predict the trait states in species that are evolutionarily independent of the species used in building the model, and the genes can be scored and ranked based on

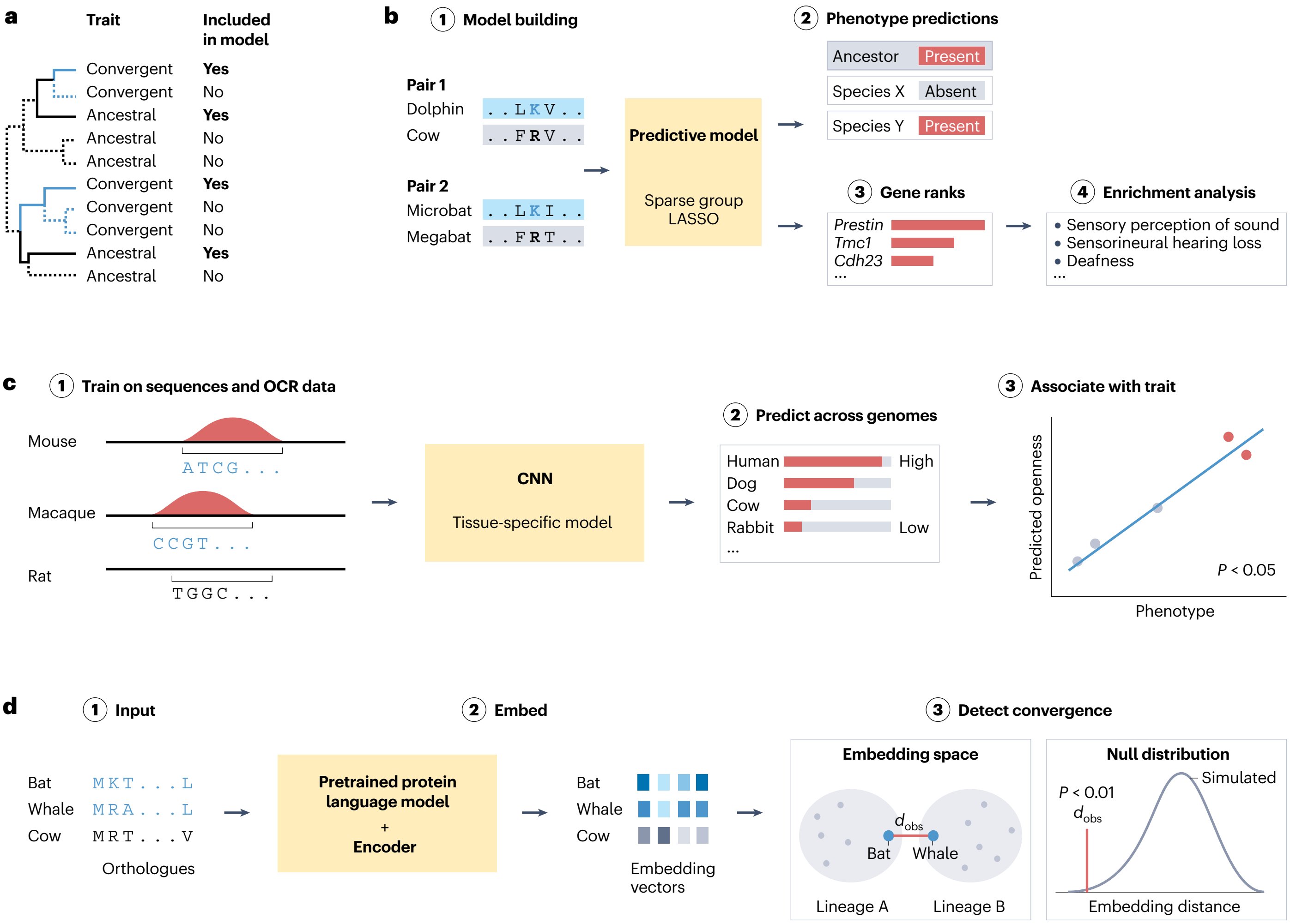

Fig. 4 | Machine learning-based methods for detecting genetic convergence. **a**, The paired species contrast (PSC) approach. A phylogeny is shown containing two convergent clades (blue branches) that are separated by lineages bearing the ancestral phenotype (black branches). A single evolutionary sparse learning (ESL) model includes pairs of convergent and ancestral (control) species (solid branches) and the shared ancestry of each pair, and other lineages (dashed branches) are excluded. Sibling species in each clade of convergent and control species can be exchanged to create an ensemble of models using many combinations of species. **b**, The eESL-PSC method. The echolocation trait is used as an example: orthologous protein sequences (numerically encoded) are used as the inputs to build an ensemble of sparse group LASSO (least absolute shrinkage and selection operator) models (**1**); these models can be used to generate phenotype predictions for other species that can be aligned with those used to build the models, including evolutionarily independent species (**2**); model contribution scores for each sequence can be used to generate a ranking of genes with the strongest evidence of convergence (**3**); and these can then be analysed for gene ontology enrichments (**4**). **c**, The tissue-aware conservation inference toolkit (TACIT) links genotype to phenotype by predicting regulatory activity: a convolutional neural network (CNN) is trained to distinguish open chromatin regions (OCRs) from closed regions using DNA sequences from a relevant tissue in a few representative species (for example, mouse and macaque) (**1**); the trained model is then applied to orthologous sequences across a large phylogeny (for example, 222 mammals) to generate quantitative predictions of chromatin openness for every species, regardless of whether experimental data exist for them (**2**); and these predicted openness scores are tested for association with the convergent phenotype using phylogenetic regression (**3**). **d**, The adaptive convergence by embedding of protein (ACEP) method uses artificial intelligence to detect high-order convergence beyond site-specific identity: orthologous protein sequences from convergent lineages (for example, bats and whales) and control lineages are processed by a pretrained protein language model (pLM) and a bottleneck encoder to generate fixed-length numerical embeddings that capture latent structural and physico-chemical features (**1**); in the high-dimensional embedding space, proteins with convergent adaptive functions may cluster closer together (d_{obs}) than expected from their phylogenetic distance (**2**); and statistical significance is determined by comparing the observed embedding distance between convergent taxa against a null distribution of distances derived from simulations of neutral evolution (**3**).

their contribution to many models (Fig. 4b). Applied to a proteome-scale search for convergence underlying echolocation in mammals[69], ESL-PSC revealed with high statistical significance that high-scoring genes were strongly enriched for many hearing-related ontology terms. Benchmarking on empirical and simulated datasets found that ESL-PSC outperformed a range of methods in many situations[69].

Another machine learning approach, the tissue-aware conservation inference toolkit (TACIT), seeks links between gene regulatory convergence and convergent traits[174]. Open chromatin regions (OCRs) detected by methods such as ATAC-seq (assay for transposase-accessible chromatin using sequencing) tend to coincide with gene regulatory regions. The TACIT trains convolutional neural networks (CNNs) to predict chromatin openness based on genomic DNA sequence intervals centred on OCRs (that is, ATAC-seq peaks) as positive cases and intervals sampled from closed chromatin regions as negative cases. Once trained in a small number of reference mammalian species with high-quality OCR data, these models can predict the likelihood of OCRs across other genomes (Fig. 4c). These probabilities were used to identify candidate regions linked to a certain trait of interest (that is, predicted as OCR openness was associated with the trait). For instance, the TACIT was applied to OCR profiles from a bat vocal-motor brain region, predicting enhancer openness across 222 mammalian genomes and pinpointing 48 candidate enhancers whose predicted activity convergently associates with vocal learning behaviour across species[98]. By linking regulatory sequence activity to trait status in a lineage-aware manner, the TACIT provides new means to detect enhancer-level convergence.

In the near future, artificial intelligence technologies that power large language models[175–177] hold potential for detecting cryptic signatures of molecular convergence. Given that proteins can converge on similar structures and functional motifs[84,178,179] without identical amino acid substitutions or changes in standard evolutionary measurements, such as evolutionary rates, these cases are invisible to current approaches. However, finding rule-based patterns and long-range interdependencies in sequence data is a task at which large language models excel, and biological sequences contain such latent patterns and are amenable to this approach[180–182]. Protein language models (pLMs), which are pretrained on millions of protein sequences, have been shown to facilitate prediction of the structure, function and fitness of proteins from their sequences alone[183–187]. During training, pLMs gain the ability to generate high-dimensional numerical representations of residues in proteins, known as embeddings. Each residue's embedding encodes key aspects of its molecular context in the sequence that determine its properties and behaviour[183,188]. These embedding vectors can be extracted and used effectively as inputs for downstream prediction tasks[189–194], which opens new avenues for linking genotypes to phenotypes. For example, unexpected similarity of two residues' embedding vectors in a protein in species from distinct evolutionary lineages can signify a similar structural or functional role, analogous to the original quest for convergent substitutions. For example, an analysis of influenza virus nucleoprotein sequences revealed that pLM embeddings reveal patterns consistent with convergent evolution[195]. Using evolutionary velocity analysis, a framework that estimates the directional flow of evolution using pLM-derived embeddings, it was reported that two avian influenza strains became more similar to human H1N1 influenza nucleoprotein sequences over time[195]. In a study of venom proteins, residue embeddings revealed unexpected clustering of centipede toxins with snake three-finger toxins, despite their distant taxonomic origins[196]. Structural analyses confirmed that these proteins share a remarkably similar fold, suggesting an instance of convergent evolution. Convergent motifs were also detected, in some cases, in pLM-based tools designed to search for sequences with remote homology[197].

A method using pLM embeddings to test for adaptive convergence, termed adaptive onvergence by embedding of protein (ACEP)[198], takes a multiple-sequence alignment and a fixed species-tree topology for each gene, and infers conventional branch lengths, site-wise rates, equilibrium frequencies and ancestral states. Using this information, ACEP simulates molecular evolution to build a neutral reference and then generates embeddings for real and simulated sequences using a pretrained pLM and a trained bottleneck encoder to obtain fixed-length per-protein embeddings. The mean embedding distance (cosine or Euclidean) between convergent lineages is then compared with the simulated neutral reference, with the expectation that proteins harbouring convergent genetic changes will show a smaller embedding distance than the neutral reference (Fig. 4d). Applied to echolocating mammals, ACEP recovered classic candidate proteins, such as prestin (encoded by *SLC26A5*), and a significant enrichment of 'sensory perception' genes in bats and toothed whales, whereas a control pairing (bats versus bovids) lacked such an enrichment. However, enrichments for terms related specifically to hearing were not found, unlike previous studies that searched for identical amino acid convergences[34,67,69]. This trait may simply be one for which identical amino acid convergence is more pronounced due to the high level of functional tuning it requires (Fig. 3). Notably, embedding-level convergence detection persisted after masking known convergent sites in one protein, suggesting that ACEP can capture high-order physico-chemical or structural features beyond identical amino acid convergences[198]. These emerging trends suggest that machine learning and artificial intelligence approaches hold considerable promise for detecting hidden signatures of genetic convergence underlying the convergent evolution of traits.

Challenges and future directions

Despite recent progress in computational methods, several challenges continue to hinder efforts to uncover the genetic foundations of convergent traits. The first problem is the plethora of competing approaches and software tools. We (Table 1) and others[41,139] have provided guidance on which method to choose, but a daunting array of choices nevertheless remains. Very few, if any, truly independent benchmarking studies have compared the performance of existing methods, potentially because benchmarking requires a source of ground truth, that is, cases where the underlying genetic basis of complex convergent traits is fully known and understood. However, even for the most widely studied traits, such as echolocation, our understanding remains limited and provisional. Simulated data are often used in benchmarking studies for other kinds of computational methods (that is, phylogeny reconstruction), but confidently simulating adaptive convergent evolution would require an understanding of the idiosyncratic selective processes at play in these sporadic cases, which eludes us. One can only simulate using a wide range of plausible parameters and conditions and demonstrate robust capabilities of detection approaches[69,70,74].

Another problem is that the many different modes of convergent genetic evolution that can occur at different organizational levels, each requiring dedicated computational methods. For studies of traits whose genetic basis is uncertain, a full unbiased search requires a large array of methods to be used. Within each class, each method has different sensitivities, which are generally not known or intuitive. A related problem is that we lack a strong understanding of the factors that make a certain mode of genetic convergence more likely for a given trait. We have discussed some hypotheses here (Fig. 3), but these remain to be rigorously investigated. In addition, it is uncertain how frequently genetic convergence occurs at all, at any given level. Many cases exist in the literature, as we have reviewed, but a publication bias likely inflates the perception that most traits have an underlying shared genetic basis. That said, when convergence at the level of pathways, shared functions and gene families is accounted for, there is relatively little scope for genuinely unique genetic solutions.

Future developments in artificial intelligence-based methods present a possible path to solving many of the above problems. An ideal approach would involve a model pretrained on a vast array of genomes, population variation and experimental measurements, such that the model learns patterns of convergent similarities across levels of organization, which could be harnessed to detect generalized genetic convergence with one analysis. Models such as DeepMind's AlphaGenome[199] already show that one model can take ~1 Mb of DNA as input, predict thousands of genome tracks, such as contact maps and tissue-specific expression at base-pair resolution across modalities, and diagnose function-altering variants in one pass. Similarly, whole-genome language models trained across diverse species, such as Evo 2 (ref. 200), now learn long-context, interpretable representations of coding and regulatory sequence that unify splicing, expression and chromatin features. In the not too distant future, representations from a deep learning model of biological sequences, variation and experimental read-outs may be able to detect multiple levels of convergent genetic evolution in a single analysis. If realized, such foundational models, and the analytic methods that leverage them, using principles similar to methods such as ESL-PSC, the TACIT and ACEP, would make it possible to unlock the hidden structure of evolutionary genotype to phenotype relationships across the tree of life.

We envision that convergent genomics research will advance from detection to prediction by building genetic models of convergent traits. Whereas most traditional methods for detecting molecular convergence have focused on identifying sites, genes and pathways whose evolutionary changes are associated with a shared trait, machine learning approaches enable a more ambitious goal of building predictive genetic models of convergent traits[69,201,202]. These models can infer a minimal set of genetic features that jointly and quantitatively predict the presence or absence of a trait across lineages. By applying the model to ancestral sequence reconstructions or taxa at intermediate stages of evolution, one can estimate the trajectory or likelihood of acquiring the convergent trait. This ability to retrospectively and prospectively infer trait states represents a notable advance over descriptive models of convergence and may transform the study of convergent evolution from an explanatory discipline into a predictive science, one that can inform trait annotations and their evolutionary reconstruction, among other applications.

Conclusions

Convergent evolution provides powerful insights into evolutionary determinism, adaptation and the genetic basis of complex traits,

Box 2 | Convergence as a tool for biomedical discovery

Convergence-based methods can shed light on many complex health-related traits whose genetic basis may be largely fixed within the human species but variable across species. These include traits such as longevity, cancer resistance, metabolic adaptation and sensory decline, some of which may not be amenable to study using traditional genome-wide association studies or single-species models because they have limited variability within the species[218]. Because convergent evolution represents independent 'natural experiments' in solving similar physiological problems, it enables robust inference of functionally important genes and genomic regions. Convergent patterns indicate that selection has repeatedly acted on the same systems, suggesting that the corresponding genes are under meaningful functional constraint or selection.

One convergent trait with substantial biomedical relevance is that of prolonged healthy longevity, which has appeared in many independent lineages[219]. Maximum lifespan varies from around 2 years to more than 200 years across mammals[220], underscoring the importance of evolutionary comparative studies to understand the genetic basis of dramatic differences in healthy ageing between species. Several studies have leveraged comparative genomic analyses to reveal genetic pathways consistently associated with increased lifespan across distinct mammalian lineages[127,129,221,222]. An analysis of mammals showed that genes involved in DNA repair, NF-κB signalling, cell cycle regulation and immune functions exhibit increased evolutionary constraint in long-lived mammals[221]. This approach, which quantitatively correlated evolutionary rate shifts with lifespan across mammals, robustly demonstrated that genes maintaining genomic integrity and controlling inflammation are critical for longevity. Similarly, another study identified strong signatures of positive selection in DNA repair genes, immune modulation and insulin signalling pathways in exceptionally long-lived rockfish species, including species documented to live for more than two centuries, suggesting common mechanisms driving lifespan extension across distant taxa[223]. These findings were subsequently confirmed by targeted genomic sequencing, highlighting nucleotide excision repair and chromatoid body regulation, pathways related to transposon suppression and genome stability[224]. Future work leveraging new methods and high-quality genomes will continue to shed light on many open questions related to the mechanistic basis of longevity.

Across mammals, cancer mortality is only weakly related to body size and lifespan, a phenomenon known as Peto's paradox, consistent with repeated evolution of cancer suppression mechanisms. Pan-mammalian associations between evolutionary rate constraints and genes involved in cancer resistance have been identified[221], notably those regulating cell cycle, DNA repair and immune responses. This finding aligns with single-species studies showing convergent adaptive changes in tumour suppressor genes, such as *ADAMTS9* in microbats and the naked mole-rat[225]. Similarly, convergent evolutionary rate shifts in cancer-related pathways were identified among birds and bats[215], reinforcing the concept of conserved evolutionary pressures favouring cancer resistance mechanisms. The process of tumorigenesis itself can also be considered as an evolutionary process in which malignancy is acquired convergently, and cancer genomics studies identify driver mutations along with the genes and pathways in which they occur based on their independent acquisitions across tumours and individuals[226]. Together, these studies underscore how computational approaches to identifying convergent genetic changes can illuminate the fundamental biological pathways underpinning complex, medically important traits.

revealing how similar phenotypes arise through repeatable molecular changes across diverse lineages. From shared amino acid substitutions to regulatory element reuse, molecular convergence highlights nature's recurring solutions to selective pressures. It serves not only as a retrospective lens but also as a predictive framework for linking genotype to phenotype. This principle applies both across and within species, including in cancer genomics, where repeated occurrence of somatic mutations in tumours is of great functional importance (Box 2). The convergence of molecular mechanisms across distantly related lineages thus not only deepens our understanding of evolution but also holds promise for identifying key adaptive variants in human traits and offers a powerful framework for identifying genes and pathways underlying complex traits of biomedical relevance. Advanced computational tools now enable scalable, interpretable detection of molecular convergence across genomes. These approaches are driving progress in trait prediction, ancestral reconstruction and understanding of evolutionary innovation, with implications across evolutionary biology and genetics.

Published online: 02 February 2026

References

1. Stayton, C. T. The definition, recognition, and interpretation of convergent evolution, and two new measures for quantifying and assessing the significance of convergence. *Evolution* **69**, 2140–2153 (2015).
2. McGhee, G. *Convergent Evolution: Limited Forms Most Beautiful* (MIT Press, 2011).
3. Blount, Z. D., Lenski, R. E. & Losos, J. B. Contingency and determinism in evolution: replaying life's tape. *Science* **362**, eaam5979 (2018).
4. Wake, D. B., Wake, M. H. & Specht, C. D. Homoplasy: from detecting pattern to determining process and mechanism of evolution. *Science* **331**, 1032–1035 (2011).
5. Tenaillon, O. et al. The molecular diversity of adaptive convergence. *Science* **335**, 457–461 (2012).
6. Rosenblum, E. B., Parent, C. E. & Brandt, E. E. The molecular basis of phenotypic convergence. *Annu. Rev. Ecol. Evol. Syst.* **45**, 203–226 (2014).
7. Stern, D. L. The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013).
8. Cerca, J. Understanding natural selection and similarity: convergent, parallel and repeated evolution. *Mol. Ecol.* **32**, 5451–5462 (2023).
9. Xu, S., Wang, J., Guo, Z., He, Z. & Shi, S. Genomic convergence in the adaptation to extreme environments. *Plant. Commun.* **1**, 100117 (2020).
10. Martin, A. & Orgogozo, V. The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution* **67**, 1235–1250 (2013).
11. Romero-Herrera, A. E., Lehmann, H., Joysey, K. A. & Friday, A. E. On the evolution of myoglobin. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **283**, 61–163 (1978).
12. Stewart, C. B., Schilling, J. W. & Wilson, A. C. Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* **330**, 401–404 (1987).
13. Swanson, K. W., Irwin, D. M. & Wilson, A. C. Stomach lysozyme gene of the langur monkey: tests for convergence and positive selection. *J. Mol. Evol.* **33**, 418–425 (1991).
14. Storz, J. F. Causes of molecular convergence and parallelism in protein evolution. *Nat. Rev. Genet.* **17**, 239–250 (2016).
15. Rey, C. et al. Detecting adaptive convergent amino acid evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20180234 (2019).
16. Schenk, M. F. et al. Population size mediates the contribution of high-rate and large-benefit mutations to parallel evolution. *Nat. Ecol. Evol.* **6**, 439–447 (2022).
17. Bolognini, D. et al. Recurrent evolution and selection shape structural diversity at the amylase locus. *Nature* **634**, 617–625 (2024).
18. Albalat, R. & Cañestro, C. Evolution by gene loss. *Nat. Rev. Genet.* **17**, 379–391 (2016).
19. Chan, Y. F. et al. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *Pitx1* enhancer. *Science* **327**, 302–305 (2010).
20. Moreno, J. A. et al. Emx2 underlies the development and evolution of marsupial gliding membranes. *Nature* **629**, 127–135 (2024).
21. Merényi, Z. et al. Unmatched level of molecular convergence among deeply divergent complex multicellular fungi. *Mol. Biol. Evol.* **37**, 2228–2240 (2020).
22. Morales, A. E. et al. Distinct genes with similar functions underlie convergent evolution in Myotis bat ecomorphs. *Mol. Biol. Evol.* **41**, msae165 (2024).
23. Gaines, T. A., Patterson, E. L. & Neve, P. Molecular mechanisms of adaptive evolution revealed by global selection for glyphosate resistance. *N. Phytol.* **223**, 1770–1775 (2019).
24. Hardison, R. C. Comparative genomics. *PLoS Biol.* **1**, E58 (2003).
25. Alföldi, J. & Lindblad-Toh, K. Comparative genomics as a tool to understand evolution and disease. *Genome Res.* **23**, 1063–1068 (2013).
26. Hilgers, L. & Hiller, M. Linking phenotype to genotype using comprehensive genomic comparisons. *Curr. Opin. Genet. Dev.* **94**, 102384 (2025).
27. Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky Pond, S. L. & Tamura, K. Statistics and truth in phylogenomics. *Mol. Biol. Evol.* **29**, 457–472 (2012).
28. Feldman, C. R., Brodie, E. D. Jr, Brodie, E. D. 3rd & Pfrender, M. E. Constraint shapes convergence in tetrodotoxin-resistant sodium channels of snakes. *Proc. Natl. Acad. Sci. USA.* **109**, 4556–4561 (2012).
29. Dobler, S., Dalla, S., Wagschal, V. & Agrawal, A. A. Community-wide convergent evolution in insect adaptation to toxic cardenolides by substitutions in the Na,K-ATPase. *Proc. Natl. Acad. Sci. USA.* **109**, 13040–13045 (2012).
30. Karageorgi, M. et al. Genome editing retraces the evolution of toxin resistance in the monarch butterfly. *Nature* **574**, 409–412 (2019).
31. Zhen, Y., Aardema, M. L., Medina, E. M., Schumer, M. & Andolfatto, P. Parallel molecular evolution in an herbivore community. *Science* **337**, 1634–1637 (2012).
32. Zhang, J. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nat. Genet.* **38**, 819–823 (2006).
33. 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).
34. 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).
35. Projecto-Garcia, J. et al. Repeated elevational transitions in hemoglobin function during the evolution of Andean hummingbirds. *Proc. Natl Acad. Sci. USA* **110**, 20669–20674 (2013).
36. Natarajan, C. et al. Convergent evolution of hemoglobin function in high-altitude Andean waterfowl involves limited parallelism at the molecular sequence level. *PLoS Genet.* **11**, e1005681 (2015).
37. Zhu, X. et al. Divergent and parallel routes of biochemical adaptation in high-altitude passerine birds from the Qinghai–Tibet Plateau. *Proc. Natl Acad. Sci. USA* **115**, 1865–1870 (2018).
38. Yokoyama, R. & Yokoyama, S. Convergent evolution of the red- and green-like visual pigment genes in fish, *Astyanax fasciatus*, and human. *Proc. Natl. Acad. Sci. USA.* **87**, 9315–9318 (1990).
39. Bricelj, V. M. et al. Sodium channel mutation leading to saxitoxin resistance in clams increases risk of PSP. *Nature* **434**, 763–767 (2005).
40. Todorovic, J., Swapna, I., Suma, A., Carnevale, V. & Zakon, H. Dual mechanisms contribute to enhanced voltage dependence of an electric fish potassium channel. *Biophys. J.* **123**, 2097–2109 (2024).
41. Macdonald, A. R., James, M. E., Mitchell, J. D. & Holland, B. R. From trees to traits: a review of advances in PhyloG2P methods and future directions. *Genome Biol. Evol.* **17**, evaf150 (2025).
42. Doolittle, R. F. Convergent evolution: the need to be explicit. *Trends Biochem. Sci.* **19**, 15–18 (1994).
43. Steenwyk, J. L., Li, Y., Zhou, X., Shen, X.-X. & Rokas, A. Incongruence in the phylogenomics era. *Nat. Rev. Genet.* **24**, 834–850 (2023).
44. Zhang, J. & Kumar, S. Detection of convergent and parallel evolution at the amino acid sequence level. *Mol. Biol. Evol.* **14**, 527–536 (1997).
45. Christin, P.-A. et al. Evolutionary switch and genetic convergence on *rbcL* following the evolution of C_4 photosynthesis. *Mol. Biol. Evol.* **25**, 2361–2368 (2008).
46. Christin, P.-A., Salamin, N., Savolainen, V., Duvall, M. R. & Besnard, G. C4 photosynthesis evolved in grasses via parallel adaptive genetic changes. *Curr. Biol.* **17**, 1241–1247 (2007).
47. Jost, M. C. et al. Toxin-resistant sodium channels: parallel adaptive evolution across a complete gene family. *Mol. Biol. Evol.* **25**, 1016–1024 (2008).
48. Dallos, P. Cochlear amplification, outer hair cells and prestin. *Curr. Opin. Neurobiol.* **18**, 370–376 (2008).
49. Li, Y., Liu, Z., Shi, P. & Zhang, J. The hearing gene prestin unites echolocating bats and whales. *Curr. Biol.* **20**, R55–R56 (2010).
50. Liu, Y. et al. Convergent sequence evolution between echolocating bats and dolphins. *Curr. Biol.* **20**, R53–R54 (2010).
51. Li, G. et al. The hearing gene prestin reunites echolocating bats. *Proc. Natl. Acad. Sci. USA.* **105**, 13959–13964 (2008).
52. 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).
53. 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).
54. Castoe, T. A. et al. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc. Natl. Acad. Sci. USA.* **106**, 8986–8991 (2009).
55. Parker, J. et al. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**, 228–231 (2013).
56. Zou, Z. & Zhang, J. No genome-wide protein sequence convergence for echolocation. *Mol. Biol. Evol.* **32**, 1237–1241 (2015).
57. 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).
58. Rokas, A. & Carroll, S. B. Frequent and widespread parallel evolution of protein sequences. *Mol. Biol. Evol.* **25**, 1943–1953 (2008).
59. 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).
60. Xu, S. et al. Genome-wide convergence during evolution of mangroves from woody plants. *Mol. Biol. Evol.* **34**, 1008–1015 (2017).

61. Goldstein, R. A., Pollard, S. T., Shah, S. D. & Pollock, D. D. Nonadaptive amino acid convergence rates decrease over time. *Mol. Biol. Evol.* **32**, 1373–1381 (2015).
62. Lee, J.-H. et al. Molecular parallelism in fast-twitch muscle proteins in echolocating mammals. *Sci. Adv.* **4**, eaat9660 (2018).
63. Foote, A. D. et al. Convergent evolution of the genomes of marine mammals. *Nat. Genet.* **47**, 272–275 (2015).
64. Zhou, X., Seim, I. & Gladyshev, V. N. Convergent evolution of marine mammals is associated with distinct substitutions in common genes. *Sci. Rep.* **5**, 16550 (2015).
65. He, Z. et al. Convergent adaptation of the genomes of woody plants at the land–sea interface. *Natl Sci. Rev.* **7**, 978–993 (2020).
66. Zhang, G. et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
67. 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).
68. Yuan, Y. et al. Comparative genomics provides insights into the aquatic adaptations of mammals. *Proc. Natl. Acad. Sci. USA.* **118**, e2106080118 (2021).
69. Allard, J. B. et al. Evolutionary sparse learning reveals the shared genetic basis of convergent traits. *Nat. Commun.* **16**, 3217 (2025).
70. Fukushima, K. & Pollock, D. D. Detecting macroevolutionary genotype–phenotype associations using error-corrected rates of protein convergence. *Nat. Ecol. Evolution* **7**, 155–170 (2023).
71. Okuyama, Y. et al. Convergent acquisition of disulfide-forming enzymes in malodorous flowers. *Science* **388**, 656–661 (2025).
72. Sadanandan, K. R. et al. Convergence in hearing-related genes between echolocating birds and mammals. *Proc. Natl. Acad. Sci. USA.* **120**, e2307340120 (2023).
73. Parto, S. & Lartillot, N. Detecting consistent patterns of directional adaptation using differential selection codon models. *BMC Evol. Biol.* **17**, 147 (2017).
74. Duchemin, L., Lanore, V., Veber, P. & Boussau, B. Evaluation of methods to detect shifts in directional selection at the genome scale. *Mol. Biol. Evol.* **40**, msac247 (2022).
75. Rey, C., Guéguen, L., Sémon, M. & Boussau, B. Accurate detection of convergent amino-acid evolution with PCOC. *Mol. Biol. Evol.* **35**, 2296–2306 (2018).
76. Chabrol, O., Royer-Carenzi, M., Pontarotti, P. & Didier, G. Detecting the molecular basis of phenotypic convergence. *Methods Ecol. Evol.* **9**, 2170–2180 (2018).
77. Morel, M., Zhukova, A., Lemoine, F. & Gascuel, O. Accurate detection of convergent mutations in large protein alignments with ConDor. *Genome Biol. Evol.* **16**, evae040 (2024).
78. Pease, J. B., Haak, D. C., Hahn, M. W. & Moyle, L. C. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biol.* **14**, e1002379 (2016).
79. Barteri, F. et al. CAAStools: a toolbox to identify and test convergent amino acid substitutions. *Bioinformatics* **39**, btad623 (2023).
80. Satterlee, J. W. et al. Convergent evolution of plant prickles by repeated gene co-option over deep time. *Science* **385**, eado1663 (2024).
81. Bohutínská, M. & Peichel, C. L. Divergence time shapes gene reuse during repeated adaptation. *Trends Ecol. Evol.* **39**, 396–407 (2024).
82. Graham, A. M. & McCracken, K. G. Convergent evolution on the hypoxia-inducible factor (HIF) pathway genes *EGLN1* and *EPAS1* in high-altitude ducks. *Heredity* **122**, 819–832 (2019).
83. Rives, N., Lamba, V., Cheng, C. H. C. & Zhuang, X. Diverse origins of near-identical antifreeze proteins in unrelated fish lineages provide insights into evolutionary mechanisms of new gene birth and protein sequence convergence. *Mol. Biol. Evol.* **41**, msae182 (2024).
84. Gherardini, P. F., Wass, M. N., Helmer-Citterich, M. & Sternberg, M. J. E. Convergent evolution of enzyme active sites is not a rare phenomenon. *J. Mol. Biol.* **372**, 817–845 (2007).
85. Casewell, N. R. et al. Solenodon genome reveals convergent evolution of venom in eulipotyphlan mammals. *Proc. Natl. Acad. Sci. USA.* **116**, 25745–25755 (2019).
86. Natarajan, C. et al. Predictable convergence in hemoglobin function has unpredictable molecular underpinnings. *Science* **354**, 336–339 (2016).
87. Zhang, J. Parallel functional changes in the digestive RNases of ruminants and colobines by divergent amino acid substitutions. *Mol. Biol. Evol.* **20**, 1310–1317 (2003).
88. Zakon, H. H., Lu, Y., Zwickl, D. J. & Hillis, D. M. Sodium channel genes and the evolution of diversity in communication signals of electric fishes: convergent molecular evolution. *Proc. Natl. Acad. Sci. USA.* **103**, 3675–3680 (2006).
89. Ujvari, B. et al. Widespread convergence in toxin resistance by predictable molecular evolution. *Proc. Natl. Acad. Sci. USA.* **112**, 11911–11916 (2015).
90. Mirceta, S. et al. Evolution of mammalian diving capacity traced by myoglobin net surface charge. *Science* **340**, 1234192 (2013).
91. Witt, K. E. & Huerta-Sánchez, E. Convergent evolution in human and domesticate adaptation to high-altitude environments. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20180235 (2019).
92. Wu, D.-D. et al. Convergent genomic signatures of high-altitude adaptation among domestic mammals. *Natl. Sci. Rev.* **7**, 952–963 (2020).
93. Reid, N. M. et al. The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science* **354**, 1305–1308 (2016).
94. Wu, B. et al. The genomes of two billfishes provide insights into the evolution of endothermy in teleosts. *Mol. Biol. Evol.* **38**, 2413–2427 (2021).
95. Sun, Y.-B. et al. Species groups distributed across elevational gradients reveal convergent and continuous genetic adaptation to high elevations. *Proc. Natl. Acad. Sci. USA.* **115**, E10634–E10641 (2018).
96. Murrell, B. et al. Gene-wide identification of episodic selection. *Mol. Biol. Evol.* **32**, 1365–1371 (2015).
97. Melendez-Vazquez, F. et al. Ecological interactions and genomic innovation fueled the evolution of ray-finned fish endothermy. *Sci. Adv.* **11**, eads8488 (2025).
98. Wirthlin, M. E. et al. Vocal learning-associated convergent evolution in mammalian proteins and regulatory elements. *Science* **383**, eabn3263 (2024).
99. Yusuf, L. H., Saldívar Lemus, Y., Thorpe, P., Macías Garcia, C. & Ritchie, M. G. Genomic signatures associated with transitions to viviparity in Cyprinodontiformes. *Mol. Biol. Evol.* **40**, msad208 (2023).
100. Cicconardi, F., McLellan, C. F., Seguret, A., McMillan, W. O. & Montgomery, S. H. Convergent molecular evolution associated with repeated transitions to gregarious larval behavior in Heliconiini. *Mol. Biol. Evol.* **42**, msaf179 (2025).
101. David, K. T. et al. Convergent expansions of keystone gene families drive metabolic innovation in Saccharomycotina yeasts. *Proc. Natl. Acad. Sci. USA.* **122**, e2500165122 (2025).
102. Simola, D. F. et al. Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome Res.* **23**, 1235–1247 (2013).
103. Shell, W. A. et al. Sociality sculpts similar patterns of molecular evolution in two independently evolved lineages of eusocial bees. *Commun. Biol.* **4**, 253 (2021).
104. Salman-Minkov, A., Sabath, N. & Mayrose, I. Whole-genome duplication as a key factor in crop domestication. *Nat. Plants* **2**, 16115 (2016).
105. Thompson, A., Vo, D., Comfort, C. & Zakon, H. H. Expression evolution facilitated the convergent neofunctionalization of a sodium channel gene. *Mol. Biol. Evol.* **31**, 1941–1955 (2014).
106. Fletcher, G. L., Hew, C. L. & Davies, P. L. Antifreeze proteins of teleost fishes. *Annu. Rev. Physiol.* **63**, 359–390 (2001).
107. Chen, L., DeVries, A. L. & Cheng, C. H. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proc. Natl. Acad. Sci. USA.* **94**, 3817–3822 (1997).
108. Hecker, N., Sharma, V. & Hiller, M. Convergent gene losses illuminate metabolic and physiological changes in herbivores and carnivores. *Proc. Natl. Acad. Sci. USA.* **116**, 3036–3041 (2019).
109. Partha, R. et al. Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *eLife* **6**, e25884 (2017).
110. Protas, M. E. et al. Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nat. Genet.* **38**, 107–111 (2006).
111. Espinasa, L., Robinson, J. & Espinasa, M. Mc1r gene in *Astroblepus pholeter* and *Astyanax mexicanus*: convergent regressive evolution of pigmentation across cavefish species. *Dev. Biol.* **441**, 305–310 (2018).
112. Kato, A. et al. Convergent gene losses and pseudogenizations in multiple lineages of stomachless fishes. *Commun. Biol.* **7**, 408 (2024).
113. Griesmann, M. et al. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science* **361**, eaat1743 (2018).
114. Nagy, L. G. et al. Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. *Nat. Commun.* **5**, 4471 (2014).
115. Wilhoit, K. et al. Convergent evolution and predictability of gene copy numbers associated with diets in mammals. *Genome Biol. Evol.* **17**, evaf008 (2025).
116. Nagy, L. G. et al. Genetic bases of fungal white rot wood decay predicted by phylogenomic analysis of correlated gene–phenotype evolution. *Mol. Biol. Evol.* **34**, 35–44 (2017).
117. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2021).
118. Librado, P. & Rozas, J. Reconstructing gene gains and losses with BadiRate. *Methods Mol. Biol.* **2569**, 213–232 (2022).
119. Aristide, L. & Fernández, R. Genomic insights into mollusk terrestrialization: parallel and convergent gene family expansions as key facilitators in out-of-the-sea transitions. *Genome Biol. Evol.* **15**, evad176 (2023).
120. Balart-García, P. et al. Parallel and convergent genomic changes underlie independent subterranean colonization across beetles. *Nat. Commun.* **14**, 3842 (2023).
121. Freitas, L. & Nery, M. F. Expansions and contractions in gene families of independently-evolved blood-feeding insects. *BMC Evol. Biol.* **20**, 87 (2020).
122. Zhang, X. et al. Genomic convergence underlying high-altitude adaptation in alpine plants. *J. Integr. Plant. Biol.* **65**, 1620–1635 (2023).
123. 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).
124. Hiller, M. et al. A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Rep.* **2**, 817–823 (2012).
125. Prudent, X., Parra, G., Schwede, P., Roscito, J. G. & Hiller, M. Controlling for phylogenetic relatedness and evolutionary rates improves the discovery of associations between species' phenotypic and genomic differences. *Mol. Biol. Evol.* **33**, 2135–2150 (2016).
126. Redlich, R. et al. RERconverge expansion: using relative evolutionary rates to study complex categorical trait evolution. *Mol. Biol. Evol.* **41**, msae210 (2024).
127. Treaster, S., Daane, J. M. & Harris, M. P. Refining convergent rate analysis with topology in mammalian longevity and marine transitions. *Mol. Biol. Evol.* **38**, 5190–5203 (2021).
128. Hu, Z., Sackton, T. B., Edwards, S. V. & Liu, J. S. Bayesian detection of convergent rate changes of conserved noncoding elements on phylogenetic trees. *Mol. Biol. Evol.* **36**, 1086–1100 (2019).
129. Gemmell, P., Sackton, T. B., Edwards, S. V. & Liu, J. S. A phylogenetic method linking nucleotide substitution rates to rates of continuous trait evolution. *PLoS Comput. Biol.* **20**, e1011995 (2024).

130. Yan, H. et al. PhyloAcc-GT: a Bayesian method for inferring patterns of substitution rate shifts on targeted lineages accounting for gene tree discordance. *Mol. Biol. Evol.* **40**, msad195 (2023).
131. Roscito, J. G. et al. Convergent and lineage-specific genomic differences in limb regulatory elements in limbless reptile lineages. *Cell Rep.* **38**, 110280 (2022).
132. Kowalczyk, A., Chikina, M. & Clark, N. Complementary evolution of coding and noncoding sequence underlies mammalian hairlessness. *eLife* **11**, e76911 (2022).
133. Sharma, V. et al. A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nat. Commun.* **9**, 1215 (2018).
134. Chikina, M., Robinson, J. D. & Clark, N. L. Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. *Mol. Biol. Evol.* **33**, 2182–2192 (2016).
135. Saputra, E., Kowalczyk, A., Cusick, L., Clark, N. & Chikina, M. Phylogenetic permulations: a statistically rigorous approach to measure confidence in associations in a phylogenetic context. *Mol. Biol. Evol.* **38**, 3004–3021 (2021).
136. Meyer, W. K. et al. Ancient convergent losses of paraoxonase 1 yield potential risks for modern marine mammals. *Science* **361**, 591–594 (2018).
137. Lopes-Marques, M. et al. Complete inactivation of sebum-producing genes parallels the loss of sebaceous glands in Cetacea. *Mol. Biol. Evol.* **36**, 1270–1280 (2019).
138. Huelsmann, M. et al. Genes lost during the transition from land to water in cetaceans highlight genomic changes associated with aquatic adaptations. *Sci. Adv.* **5**, eaaw6671 (2019).
139. Clark, N. L., Kowalczyk, A., Kopania, E. E. K. & Chikina, M. Phylogenomic approaches to study adaptive evolution in mammals: from aging to aquatic lifestyles. *Annu. Rev. Genet.* **59**, 461–483 (2025).
140. Hill, M. S., Vande Zande, P. & Wittkopp, P. J. Molecular and evolutionary processes generating variation in gene expression. *Nat. Rev. Genet.* **22**, 203–215 (2021).
141. Levine, M. Transcriptional enhancers in animal development and evolution. *Curr. Biol.* **20**, R754–R763 (2010).
142. Rebeiz, M. & Tsiantis, M. Enhancer evolution and the origins of morphological novelty. *Curr. Opin. Genet. Dev.* **45**, 115–123 (2017).
143. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell* **167**, 1170–1187 (2016).
144. Berens, A., Hunt, J. H. & Toth, A. Comparative transcriptomics of convergent evolution: different genes but conserved pathways underlie caste phenotypes across lineages of eusocial insects. *Mol. Biol. Evol.* **32**, 690–703 (2015).
145. Sackton, T. B. et al. Convergent regulatory evolution and loss of flight in paleognathous birds. *Science* **364**, 74–78 (2019).
146. Shakya, S. B., Edwards, S. V. & Sackton, T. B. Convergent evolution of noncoding elements associated with short tarsus length in birds. *BMC Biol.* **23**, 52 (2025).
147. Parker, D. J. et al. Repeated evolution of asexuality involves convergent gene expression changes. *Mol. Biol. Evol.* **36**, 350–364 (2019).
148. Gallant, J. R. et al. Nonhuman genetics. Genomic basis for the convergent evolution of electric organs. *Science* **344**, 1522–1525 (2014).
149. Larter, M., Dunbar-Wallis, A., Berardi, A. E. & Smith, S. D. Convergent evolution at the pathway level: predictable regulatory changes during flower color transitions. *Mol. Biol. Evol.* **35**, 2159–2169 (2018).
150. Darragh, K., Kay, K. M. & Ramírez, S. R. The convergent evolution of hummingbird pollination results in repeated floral scent loss through gene downregulation. *Mol. Biol. Evol.* **42**, msaf027 (2025).
151. Nakayama, D. & Makino, T. Convergent accelerated evolution of mammal-specific conserved non-coding elements in hibernators. *Sci. Rep.* **14**, 11754 (2024).
152. Okamoto, A. S. & Capellini, T. D. Parallel evolution at the regulatory base-pair level contributes to mammalian interspecific differences in polygenic traits. *Mol. Biol. Evol.* **41**, msae157 (2024).
153. Zancolli, G., Reijnders, M., Waterhouse, R. M. & Robinson-Rechavi, M. Convergent evolution of venom gland transcriptomes across Metazoa. *Proc. Natl. Acad. Sci. USA.* **119**, e2111392119 (2022).
154. Pfenning, A. R. et al. Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* **346**, 1256846 (2014).
155. Barts, N. et al. Molecular evolution and expression of oxygen transport genes in livebearing fishes (Poeciliidae) from hydrogen sulfide rich springs. *Genome* **61**, 273–286 (2018).
156. Lipshutz, S. E. et al. Repeated behavioural evolution is associated with convergence of gene expression in cavity-nesting songbirds. *Nat. Ecol. Evol.* **9**, 845–856 (2025).
157. Hart, J. C., Ellis, N. A., Eisen, M. B. & Miller, C. T. Convergent evolution of gene expression in two high-toothed stickleback populations. *PLoS Genet.* **14**, e1007443 (2018).
158. Durkin, S. M., Ballinger, M. A. & Nachman, M. W. Tissue-specific and *cis*-regulatory changes underlie parallel, adaptive gene expression evolution in house mice. *PLoS Genet.* **20**, e1010892 (2024).
159. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
160. Siepel, A., Pollard, K. S. & Haussler, D. in *Lecture Notes in Computer Science* (ed. Apostolico, A. et al.) 190–205 (Springer, 2006).
161. Thomas, G. W. C. et al. Practical guidance and workflows for identifying fast evolving non-coding genomic elements using PhyloAcc. *Integr. Comp. Biol.* **64**, 1513–1525 (2024).
162. Van Belleghem, S. M. et al. High level of novelty under the hood of convergent evolution. *Science* **379**, 1043–1049 (2023).
163. Greenway, R. et al. Convergent evolution of conserved mitochondrial pathways underlies repeated adaptation to extreme environments. *Proc. Natl. Acad. Sci. USA.* **117**, 16424–16430 (2020).
164. Warner, M. R., Qiu, L., Holmes, M. J., Mikheyev, A. S. & Linksvayer, T. A. Convergent eusocial evolution is based on a shared reproductive groundplan plus lineage-specific plastic genes. *Nat. Commun.* **10**, 2651 (2019).
165. Kay, T., Piekarski, P. K. & Kronauer, D. J. C. Convergent evolution of a conserved molecular network underlies parenting and sociality. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-025-00903-5 (2025).
166. Birkeland, S., Gustafsson, A. L. S., Brysting, A. K., Brochmann, C. & Nowak, M. D. Multiple genetic trajectories to extreme abiotic stress adaptation in Arctic Brassicaceae. *Mol. Biol. Evol.* **37**, 2052–2068 (2020).
167. Spor, A. et al. Phenotypic and genotypic convergences are influenced by historical contingency and environment in yeast: multiple life-history traits convergence in yeast. *Evolution* **68**, 772–790 (2014).
168. Bailey, S. F., Rodrigue, N. & Kassen, R. The effect of selection environment on the probability of parallel evolution. *Mol. Biol. Evol.* **32**, 1436–1448 (2015).
169. Woods, R., Schneider, D., Winkworth, C. L., Riley, M. A. & Lenski, R. E. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proc. Natl. Acad. Sci. USA.* **103**, 9107–9112 (2006).
170. Murrell, B. et al. Modeling HIV-1 drug resistance as episodic directional selection. *PLoS Comput. Biol.* **8**, e1002507 (2012).
171. Hill, J. et al. Recurrent convergent evolution at amino acid residue 261 in fish rhodopsin. *Proc. Natl. Acad. Sci. USA.* **116**, 18473–18478 (2019).
172. Hagen, J. F. D., Roberts, N. S. & Johnston, R. J. Jr. The evolutionary history and spectral tuning of vertebrate visual opsins. *Dev. Biol.* **493**, 40–66 (2023).
173. Zou, D. et al. Comparative genomics sheds new light on the convergent evolution of infrared vision in snakes. *Proc. Biol. Sci.* **291**, 20240818 (2024).
174. Kaplow, I. M. et al. Relating enhancer genetic variation across mammals to complex phenotypes using machine learning. *Science* **380**, eabm7993 (2023).
175. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. *OpenAI* https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (2018).
176. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* (eds Guyon, I. et al.) 30 (NIPS, 2017).
177. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1 (Long and Short Papers) (eds Burstein, J. et al.) 4171–4186 (ACL, 2019).
178. Riziotis, I. G. et al. Paradigms of convergent evolution in enzymes. *FEBS J.* **292**, 537–555 (2025).
179. Wright, E. Tandem repeats provide evidence for convergent evolution to similar protein structures. *Genome Biol. Evol.* **17**, evaf013 (2025).
180. Vu, M. H. et al. Linguistically inspired roadmap for building biologically reliable protein language models. *Nat. Mach. Intell.* **5**, 485–496 (2023).
181. Ferruz, N. & Höcker, B. Controllable protein design with language models. *Nat. Mach. Intell.* **4**, 521–532 (2022).
182. Simon, E., Swanson, K. & Zou, J. Language models for biological research: a primer. *Nat. Methods* **21**, 1422–1429 (2024).
183. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA.* **118**, e2016239118 (2021).
184. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
185. Unsal, S. et al. Learning functional properties of proteins with language models. *Nat. Mach. Intell.* **4**, 227–245 (2022).
186. Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. In *35th Conference on Neural Information Processing Systems* (NeurIPS, 2021).
187. Yu, T. et al. Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023).
188. Heinzinger, M. et al. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinforma.* **20**, 723 (2019).
189. Bepler, T. & Berger, B. Learning the protein language: evolution, structure, and function. *Cell Syst.* **12**, 654–669.e3 (2021).
190. Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T. & Rost, B. Embeddings from deep learning transfer GO annotations beyond homology. *Sci. Rep.* **11**, 1160 (2020).
191. Bernhofer, M. & Rost, B. TMbed: transmembrane proteins predicted through language model embeddings. *BMC Bioinforma.* **23**, 326 (2022).
192. Yeung, W., Zhou, Z., Li, S. & Kannan, N. Alignment-free estimation of sequence conservation for identifying functional sites using protein sequence embeddings. *Brief. Bioinforma.* **24**, bbac599 (2023).
193. Pokharel, S., Pratyush, P., Heinzinger, M., Newman, R. & Kc, D. B. Improving protein succinylation sites prediction using embeddings from protein language model. *Sci. Rep.* **12**, 16933 (2022).
194. Rosen, Y. et al. Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN. *Nat. Methods* **21**, 1492–1500 (2024).
195. Hie, B. L., Yang, K. K. & Kim, P. S. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Syst.* **13**, 274–285.e6 (2022).
196. Senoner, T. et al. ProtSpace: a tool for visualizing protein space. *J. Mol. Biol.* **437**, 168940 (2025).

197. Kaminski, K., Ludwiczak, J., Pawlicki, K., Alva, V. & Dunin-Horkawicz, S. pLM-BLAST: distant homology detection based on direct comparison of sequence representations from protein language models. *Bioinformatics* **39**, btad579 (2023).
198. Cao, Z., Zhang, H. & Zou, Z. Language models reveal a complex sequence basis for adaptive convergent evolution of protein functions. *Proc. Natl. Acad. Sci. USA.* **122**, e2418254122 (2025).
199. Avsec, Ž. et al. AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model. Preprint at *bioRxiv* https://doi.org/10.1101/2025.06.25.661532 (2025).
200. Brixi, G. et al. Genome modeling and design across all domains of life with Evo 2. Preprint at *bioRxiv* https://doi.org/10.1101/2025.02.18.638918 (2025).
201. 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).
202. Gonçalves, C. et al. Diverse signatures of convergent evolution in cactus-associated yeasts. *PLoS Biol.* **22**, e3002832 (2024).
203. Tamuri, A. U., Dos Reis, M., Hay, A. J. & Goldstein, R. A. Identifying changes in selective constraints: host shifts in influenza. *PLoS Comput. Biol.* **5**, e1000564 (2009).
204. Langer, B. E., Roscito, J. G. & Hiller, M. REforge associates transcription factor binding site divergence in regulatory elements with phenotypic differences between species. *Mol. Biol. Evol.* **35**, 3027–3040 (2018).
205. Anisimova, M. & Kosiol, C. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol. Biol. Evol.* **26**, 255–271 (2009).
206. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170 (1988).
207. Zhang, Z. et al. The parallel molecular adaptations to the Antarctic cold environment in two psychrophilic green algae. *Genome Biol. Evol.* **11**, 1897–1908 (2019).
208. Hao, Y., Qu, Y., Song, G. & Lei, F. Genomic insights into the adaptive convergent evolution. *Curr. Genomics* **20**, 81–89 (2019).
209. Martin, D. P. et al. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* **184**, 5189–5200.e7 (2021).
210. Thiltgen, G., Dos Reis, M. & Goldstein, R. A. Finding direction in the search for selection. *J. Mol. Evol.* **84**, 39–50 (2017).
211. Murrell, B. et al. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **8**, e1002764 (2012).
212. Casola, C. & Li, J. Beyond RuBisCO: convergent molecular evolution of multiple chloroplast genes in C4 plants. *PeerJ* **10**, e12791 (2022).
213. McGowen, M. R., Tsagkogeorga, G., Williamson, J., Morin, P. A. & Rossiter, A. S. J. Positive selection and inactivation in the vision and hearing genes of cetaceans. *Mol. Biol. Evol.* **37**, 2069–2083 (2020).
214. Lu, B., Jin, H. & Fu, J. Molecular convergent and parallel evolution among four high-elevation anuran species from the Tibetan region. *BMC Genomics* **21**, 839 (2020).
215. Matsuda, Y. & Makino, T. Comparative genomics reveals convergent signals associated with the high metabolism and longevity in birds and bats. *Proc. Biol. Sci.* **291**, 20241068 (2024).
216. Wang, H. et al. Evolutionary basis of high-frequency hearing in the cochleae of echolocators revealed by comparative genomics. *Genome Biol. Evol.* **12**, 3740–3753 (2020).
217. Hughes, A. L. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* **99**, 364–373 (2007).
218. Smith, S. D., Pennell, M. W., Dunn, C. W. & Edwards, S. V. Phylogenetics is the new genetics (for most of biodiversity). *Trends Ecol. Evol.* **35**, 415–425 (2020).
219. Rechsteiner, C., Morandini, F., Kim, S. J., Seluanov, A. & Gorbunova, V. Unlocking longevity through the comparative biology of aging. *Nat. Aging* **5**, 1686–1703 (2025).
220. de Magalhães, J. P. & Costa, J. A database of vertebrate longevity records and their relation to other life-history traits. *J. Evol. Biol.* **22**, 1770–1774 (2009).
221. Kowalczyk, A., Partha, R., Clark, N. L. & Chikina, M. Pan-mammalian analysis of molecular constraints underlying extended lifespan. *eLife* **9**, e51089 (2020).
222. Liu, W. et al. Large-scale across species transcriptomic analysis identifies genetic selection signatures associated with longevity in mammals. *EMBO J.* **42**, e112740 (2023).
223. Kolora, S. R. R. et al. Origins and evolution of extreme life span in Pacific Ocean rockfishes. *Science* **374**, 842–847 (2021).
224. Treaster, S. et al. Convergent genomics of longevity in rockfishes highlights the genetics of human life span variation. *Sci. Adv.* **9**, eadd2743 (2023).
225. Lambert, M. J. & Portfors, C. V. Adaptive sequence convergence of the tumor suppressor ADAMTS9 between small-bodied mammals displaying exceptional longevity. *Aging* **9**, 573–582 (2017).
226. Pienta, K. J., Hammarlund, E. U., Axelrod, R., Amend, S. R. & Brown, J. S. Convergent evolution, evolving evolvability, and the origins of lethal cancer. *Mol. Cancer Res.* **18**, 801–810 (2020).

Acknowledgements

The authors thank J. M. Craig and S. Sharma for their many constructive comments on this manuscript. The authors' work was supported by a fellowship to J.B.A. from Temple University and a grant from the National Institutes of Health (NIH) to S.K.

Author contributions

J.B.A. researched the literature, wrote the initial draft, and created the figures. Both authors contributed equally to all other aspects of the article.

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Genetics* thanks Bastien Boussau, who co-reviewed with Pierre Gérenton; Nathan Clark; and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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