Integrating natural history-derived phenomics with comparative genomics to study the genetic architecture of convergent evolution

Sangeet Lamichhaney[1,2], Daren C. Card[1,2,3], Phil Grayson[1,2], João F.R. Tonini[1,2], Gustavo A. Bravo[1,2], Kathrin Näpflin[1,2], Flavia Termignoni-Garcia[1,2], Christopher Torres[4, 5], Frank Burbrink[6], Julia A. Clarke[4,5], Timothy B. Sackton[7] and Scott V. Edwards[1,2*]

Affiliations:

1) Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, USA

2) Museum of Comparative Zoology, Harvard University, Cambridge, USA

3) Department of Biology, University of Texas Arlington, Arlington, USA

4) Department of Biology, The University of Texas at Austin, Austin, USA.

5) Department of Geological Sciences, The University of Texas at Austin, Austin, USA

6) Department of Herpetology, The American Museum of Natural History, New York, NY 10024, USA

7) Informatics Group, Harvard University, Cambridge, USA

* correspondence to: SVE (sedwards@fas.harvard.edu)

Abstract

Evolutionary convergence has been long considered primary evidence of adaptation driven by natural selection and provides opportunities to explore evolutionary repeatability and predictability. In recent years, there has been increased interest in exploring the genetic mechanisms underlying convergent evolution, in part due to the advent of genomic techniques. However, the current 'genomics gold rush' in studies of convergence has overshadowed the reality that most trait classifications are quite broadly defined, resulting in incomplete or potentially biased interpretations of results. Genomic studies of convergence would be greatly improved by integrating deep 'vertical', natural history knowledge with 'horizontal' knowledge focusing on the breadth of taxonomic diversity. Natural history collections have and continue to be best positioned for increasing our comprehensive understanding of phenotypic diversity, with modern practices of digitization and databasing of morphological traits providing exciting improvements in our ability to evaluate the degree of morphological convergence. Combining more detailed phenotypic data with the well-established field of genomics will enable scientists to make progress on an important goal in biology: to understand the degree to which genetic or molecular convergence is associated with phenotypic convergence. Although the fields of comparative biology or comparative genomics alone can separately reveal important insights into convergent evolution, here we suggest that the synergistic and complementary roles of natural history collection-derived phenomic data and comparative genomics methods can be particularly powerful in together elucidating the genomic basis of convergent evolution among higher taxa.

Introduction

Convergent evolution is the independent acquisition of similar features in distantly related lineages [1]. Ever since Darwin suggested that similar traits could arise independently in different organisms [2], understanding the underlying causes and mechanisms of convergence has been one of the fundamental objectives of evolutionary biology. Convergent evolution has been a central component to study evolutionary predictability [3] by integrating phenotypic, phylogenetic, and environmental data [1,4,5]. While convergent evolution is usually presumed to be the result of adaptation [6–8], it is clear that convergent patterns can also result from non-adaptive processes such as exaptation, evolutionary constraints, demographic history [1,5,9,10] or hemiplasy [11]. Conceptual differences in defining phenotypic convergence as process-based (when trait similarity evolves by similar forces of natural selection) versus pattern-based (when lineages independently evolve patterns of similar traits, regardless of mechanism) have practical implications for the adequate identification and measurement of convergent traits [12]. In addition to such challenges for defining convergence at the phenotypic level, additional uncertainties exist for defining the genetic basis of phenotypic convergence [13,14]

Convergent phenotypes may or may not share a genetic basis at many different hierarchical levels (e.g., nucleotide, gene, protein, regulatory networks, function; Fig. 1) [5]. Additionally, high-levels of pleiotropy, already recognized as a likely component of many cases of convergence [citation], means that our definition of “genomic basis” of convergence may require expansion to include the role of individual genes participating in multiple networks as well as functionally overlapping networks that may not share many genes.

Here, we explore how recent advancements in comparative genomics have provided tools to expand genetic studies of convergent phenotypes based on a few candidate genes to entire

genomes, and how such large-scale genomic data are being used to explore the rate and pattern of convergence at different hierarchical levels. In particular, we highlight the need to carefully define the convergent phenotype and utilizing the role of natural history records in aiding this definition. As generating genomic data becomes easier with time, integration of community-wide organismal expertise and natural history collections will remain key to understanding the genomics of convergence. We focus primarily on convergence among distantly related species, largely in animals, and generally do not discuss convergence among close relatives or populations.

Role of organismal expertise in understanding convergence

The resurgence of interest in phenotypic convergence is driven by the desire to add a new layer – genomics - to what has been a long-standing, centuries-old interest in natural history and organismal biology (Fig. 1). Without genomics, studies of phenotypic convergence would no doubt continue as they have for decades, particularly given the firm foundation of comparative biology on which studies of convergence now rest [15,16]. However, the rapidly declining costs of genome sequencing have reinvigorated questions about the degree to which convergent phenotypes share a genetic basis, and generated considerable excitement about using convergence as a means to understand the genetic basis of phenotypes [5]. However, the "genomics gold rush" in studies of convergence has tended to focus on a few easily defined and extensively studied traits, such as the transition to marine [17], subterranean [18], or high-altitude [13] life, loss of flight in birds [19], eusociality in insects [20], social behavior in vertebrates [21,22], vocal learning in birds [23], echolocation in mammals [14,24,25], among others. While these studies have laid the groundwork for the field, organismal and natural history

expertise remains critical for the maturation of studies relating phenotypic convergence and genomics.

Definition and complex nature of convergent phenotypes: Organismal expertise and knowledge of natural history data can inform comparative genomics in several ways. A mechanistic understanding of convergent phenotypes ultimately requires in-depth knowledge of how organisms function in the wild. It is relatively easy to designate a given species as having either a subterranean lifestyle or a lifestyle wholly above ground, but such a simple dichotomy might mask the substantial diversity of ecological and behavioral traits even within the subterranean lifestyle – for example, the diversity of burrow structures and whether or how the work of digging the burrow is shared between the sexes. Similarly, categorization of species as either 'marine' or 'non-marine', or 'volant' and 'flightless' will no doubt capture important components of phenotypic convergence, without necessarily advancing a mechanistic understanding of these phenotypes. Such categorization ignores behavioral, developmental, physiological and ecological complexity that will add nuance to any comparative analysis. For example, birds categorized as "flightless" may exhibit forelimb morphologies varying from complete absence (e.g., moas, *Hesperornis*) to slightly shortened (e.g., ostriches and Galapagos Cormorants) to highly modified forelimbs actively deployed in diving underwater but not in flight (e.g., penguins, Great Auk; Fig. 2a). Similarly, "limblessness" in squamate reptiles can mean complete loss of forelimbs, hindlimbs or both, or partial loss of digits and/or limb long bones (Fig. 2b). While simple binary categorizations of specific character states have proven powerful in guiding comparative genomic analyses [18], finer dissection of convergent phenotypes as a quantitative continuum rather than a binary phenomenon [26] will allow both an adequate testing of the adaptive value of the traits in question and a more detailed categorization

of adaptations themselves. Recent models designed to test the significance of genotype-phenotype associations in a phylogenetic context are an important part of this new framework [27,28]. An important question is how to derive the most statistical power to detect convergent genotype-phenotype associations when phenotypes are defined continuously or with greater than two states. New quantitative methods of assessing convergence in phenotypic traits [29–31], as well as phylogenetic quantitative genetic models [32,33] will both be helpful in accommodating complex characters into genomic studies of convergent evolution.

Diverse types of natural history knowledge can inform comparative genomic studies of convergence: Whereas the above perspective emphasizes 'vertical' knowledge, that is, deep understanding of the natural history of individual species, comprehending the breadth of taxonomic diversity across clades is a second way in which natural history knowledge can inform comparative genomics. Integration of this 'horizontal' knowledge of the total biology of a particular clade of organisms will be important to broaden our perspective on convergent evolution (Fig. 1). The wealth of convergent traits across the Tree of Life is likely to be found not in textbooks but in taxonomic monographs written by naturalists and curators over the last couple of centuries [34]. One example of such a convergent trait is testis color in birds: why are testes in disparate groups of birds black instead of the usual tan? Another example is the evolution of parity mode (oviparous or viviparous), which otherwise being highly conserved trait in amniotes, shows complex mosaic of convergence in squamates. The number of such convergent traits is seemingly limitless, yet we know little about the power of comparative genomics to unravel the molecular basis of these phenotypes, or how our understanding of the link between genotypic and phenotypic convergence will change as the number and type of convergent traits studied from a genomic perspective increase.

Natural history and phenomic knowledge from fossils can inform interpretation of character polarity, diversity and variation by directly informing the number and type of occurrences of convergence in extinct and extant taxa. For example, the number of extant flightless avian taxa is much smaller than the number of flightless avian taxa known from the fossil record, with many convergent instances of flightlessness represented only by extinct taxa (e.g. elephant birds, moa, adzebills, the Atitlán Grebe, the Great Auk, the Kaua'i Mole Duck [35]. The inclusion of fossil taxa near the base of clades can clarify whether traits are derived and potentially convergent or with a single origin and ancestral to a group. This may provide a better estimate of the ancestral phenotype from which the convergent traits evolved. Fossil data from natural history collections are similarly crucial for calibrating phylogenies by time, allowing investigations to assess not only which phenotypes are convergent, but when they arose. In addition, divergence time data among various taxa available in public database (e.g. TimeTree [36]) provide useful information for reconstructing ancestral. The temporal data provided by fossils are important to assessing the viability of potential causal hypotheses of drivers of convergent phenotypes. Recently extinct taxa may also provide genomic data, allowing direct incorporation into molecular phylogenies with extant taxa [37,38].

Museum collections, and the wealth of phenotypic data that they provide, are an excellent source of natural history knowledge [39,40]. Museum specimens are critical for verifying species identities, claims of specific phenotypes published in the scientific literature, are the primary source for scoring characters not yet explored for many species [41], and document diverse aspects of organismal phenotypes including anatomy, environmental context, and various types of nanostructures and chemical profiles. For instance, thorough phenotypic revisions of museum specimens have confirmed the existence of taxonomic misidentifications leading to the

description of new species of Cotinga *(Tijuca condita)*, previously misidentified as *Tijuca atra* [42], or a more comprehensive understanding of phenotypic diversity and conservation needs of endemic Neotropical procyonid genus *Bassaricyon* [43]. Natural history records also provide vital information for interpreting downstream results from genomic analysis. Therefore, it is crucial that published genomes should when possible be based on DNA derived from a traceable, documented source, such as a vouchered museum specimen, known lab variant or strain, captive animal, or lab colony. In cases where this is not possible (e.g., a wild-caught individual that is entirely destroyed in the process of extracting DNA), imaging and documentation of provenance as well as associating as much additional metadata as possible is still crucial.

Many good examples of integrative comparative genomics investigations of convergence come from research teams that includes curators, taxonomists, naturalists or other experts in organismal diversity in morphology, function, and ethology at universities or other research settings [44]. However, a cursory analysis of keywords and author addresses in the Web of Science suggests a paradox: whereas museum scientists have frequently published on the general topic of evolutionary convergence, they now appear underrepresented in the second wave of convergence studies based on comparative genomics (Fig. 3). We recognize that relevant expertise in organismal phenotypes is also housed in great abundance in diverse university settings without affiliated museum collections, and our simple analysis will not capture and likely grossly underestimates these contributions. Nonetheless, we predict that as genomic studies of convergence mature, museum scientists, with their expertise in taxonomy, morphology, ecology, and biogeography will play an increasing role in studies of the genomics of convergence.

Another way in which natural history knowledge can inform comparative genomics is through a relatively recent type of natural history – the natural history of genomes and physiological and biochemical pathways. We regard any deep knowledge of organismal function across diverse clades of organisms as a type of natural history knowledge. A good example of this type of knowledge is our understanding of the taxonomic distribution of the ability to synthesize vitamin C [45]. This case was used to powerfully demonstrate how comparative genomics can help pinpoint the likely genomic basis of convergent traits, in this case the convergent loss of the ability to synthesize vitamin C. Such biochemical knowledge was amassed through measurement by diverse laboratories of vitamin C levels in diverse organisms (in this case, from 20 separate publications spanning 1956 to 2003). Another example is the convergent ability of insects to feed on toxic plants, which is mediated by convergent substitutions, duplications and gene expression changes in a gene called ATPα [46]. Such detailed biochemical knowledge has been a major driver of recent studies of the genomics of convergence. Such an approach is likely to be a powerful method for understanding the genomics of convergence because the association between genotype and phenotype in such adaptations is likely to be tight and involve few genes, and in addition had a clearly defined convergent phenotype. Many studies on the genomics of color in mammals and in squamates provide additional examples of a trait whose molecular basis was aided by knowledge of biochemical pathways across diverse groups of organisms [47–49]. Some molecular and biochemical traits, like genome size [50], proteins and DNA sequences, are organized into well-curated databases, but many such traits, such as vitamin C synthesis, are scattered in the literature. Given what seems like the relative ease of finding the genomic basis for such traits through comparative genomics, it will be important in the future to assemble databases of phenotypic traits that span the gamut from

organismal to biochemical and physiological knowledge e.g. [51–53]. Such databases can rapidly accelerate discovery of the genomic bases of convergent traits.

Genomics of convergence

Outstanding questions in the genomic study of convergence: Apart from the major aim of identifying the genomic or molecular basis of convergent traits, a great diversity of questions also motivate the study of convergent traits. For instance, how does the frequency of convergence change across hierarchical levels and does it differ appreciably at the phenotypic and molecular levels? Does a nonrandom subset of genomic changes explain most instances of convergent evolution, priming convergent evolution to be more likely in certain circumstances? Are such genomic changes more likely to be regulatory or encoded by proteins? Does standing genetic variation or *de novo* mutation account for most examples of convergent evolution? Answers to such questions will not only provide critical information about the genomics of convergence but will also contribute greatly to our understanding of adaptation and evolution in general.

Building comparative genomics resources to study convergence: Once the phenotype of interest is defined, the next most important experimental step in comparative genomics is producing an adequate genomic foundation for downstream work (Fig. 1). While many questions of interest can be investigated with publicly-available data only, if new genome(s) are essential for a study, choosing the optimal methods for generating genomic resources is crucial. Access to high quality samples is often a limiting factor for many nascent projects, a difficulty that can be overcome in part by ensuring high-quality tissue collections are prioritized at natural history repositories. High-molecular-weight DNA is critical for modern genome sequencing

technologies and in many cases, vouchered tissue samples stored in ethanol will not suffice. In addition, proper sampling procedures in the field are the most critical step to ensure high-quality DNA for building genomic resources. As an example, immediately flash-freezing tissues in liquid nitrogen is ideal given that critical molecular information (namely RNA) is quickly degraded at higher temperatures, but in many instances, tissues are transferred to liquid nitrogen after substantial delay or deep cryofreezing is logistically not possible. Though not glamorous, more detailed investigations of tissue preservation practices are vital for enabling genomic investigations, and we advocate that natural history museums should undertake a concerted, transparent effort to create best-practices recommendations for tissue samples that mirror practices already in place for whole-organism preservation [54]. For example, fresh blood stored unfrozen in Queen's lysis buffer [55] at 4°C has provided higher quality DNA from nucleated avian blood cells [56] than museum-grade frozen tissues, and has improved sample collection practices in the Department of Ornithology at Harvard's Museum of Comparative Zoology.

Genome assembly contiguity and gene annotation quality are also critically important for addressing target questions and maximizing the utility and availability of data from rare tissues from natural history collections. [57,58]. Chromosomal-level genome assemblies will allow us to understand the accurate location of genes associated with phenotypic traits across the genome and better understanding of cis- and trans-regulatory factors linked to those phenotypes, as well as ensure near-complete representation of genes in the assembly [59]. For example, a recent study of 78 bird genomes found that approximately 15% of avian genes had been overlooked during genome annotation, mostly due to the effects of GC-biased nucleotide composition [60]. By accounting for these missing genes, the researchers confirmed the expected positive

relationship between rates of protein evolution and life history traits like body mass, longevity, and age of sexual maturity that had been previously missed [61,62].

Molecular convergence in protein evolution: Many recent studies of convergence have focused on protein or codon alignments to identify amino acid positions that have convergently changed in species that share a convergent trait [63–65]. In some circumstances, conflicting placement of convergent phenotypes between gene trees and species tree can be used to identify potential genomic convergence [66, 67]. However, phylogenetic clustering of a particular gene tree can also be a product of other evolutionary or experimental processes, and further analyses are required to confirm that parallel selection in distinct taxa have led to molecular convergence. Indeed, an early study that used phylogenetic signal to identify genes convergently evolving in echolocating mammals [14] was quickly met with sharp criticism [68,69], because such phylogenetic signal could arise stochastically from biased mutational spectra rather than natural selection. Castoe et al. [70], on the other hand, found that a phylogeny based on whole mitochondrial genomes that clustered snakes with agamid lizards, a relationship unsupported by nuclear gene sequences or morphological data, was actually due to strong, convergent protein evolution in just two mitochondrial genes, producing an overwhelming, yet incorrect, phylogenetic signal.

Another approach for investigating convergent protein evolution is examining rates of protein evolution across branches of a species tree and isolating instances where accelerated rates of evolution occur independently on branches leading to organisms with convergent phenotypes [17,18,71]. Such methods utilize amino acid distance trees that are normalized by average divergence rates across the genome for each tree branch and estimate the correlation between relative evolutionary rates of genes and the evolution of a convergent trait across a phylogeny

[17, 72,73]. Such rate estimates as well as ancestral reconstructions have been used to detect classic examples of convergent protein evolution, such as substitutions in Na+, K+-ATPase enzymes of herbivorous insects that mediate resistance to toxic, plant-derived cardenolides [74,75] and substitutions in voltage-gated sodium channel proteins in reptiles, amphibians and fish that mediate resistance to tetrodotoxin [76,77]. Traditional methods of measuring rates of protein evolution, such as those employing the ratio of nonsynonymous to synonymous substitutions per site (dn/ds) [72], are also useful, but care should be taken to ensure that ds is not saturated when comparing distantly related species.

Gene family evolution associated with convergence: Gene duplications can provide raw material for rapid evolutionary innovation [78], hence analyzing the structure of gene families can provide deeper insights into the evolutionary processes underlying convergent traits. [79–81]. Phylogenetic approaches are available to estimate the rate of change in gene family sizes [82,83], and correlated rate shifts in taxa with convergent phenotypes implicate a gene family in the process of convergent evolution. For example, visual opsin and oxygen-binding globin families are known to vary in composition under varying ecological constraints, and convergent patterns of opsin and globin family turnover have occurred in jawed and jawless vertebrates [84,85]. Studies of venom genes across animal kingdom, which form complex protein cocktails used for capturing prey and defense, show that similar protein families are commonly co-opted into hyper-mutable venom gene arrays [86,87].

Relative importance of regulatory regions in convergent evolution: Because many studies on convergent evolution have focused on protein-coding regions, the role of regulatory regions underlying convergent phenotypes is typically only well understood in few model systems, like sticklebacks [88]. It is plausible that regulatory elements are less constrained, and thus able to act

as important drivers of adaptive molecular evolution by altering the timing, location, or level of expression of their target gene. Recent studies have indeed shown that changes in regulatory regions are associated with the origin of key innovations such as feathers and hair [89,90], as well as convergent evolution of traits such as flower pigmentation [91], loss of flight in ratites [19] and ocular degeneration in mammals [18].

Whereas predicting protein-coding genes in genome sequences is made easier both by examining homologous sequence patterns and a wealth of easy to generate functional data (e.g., transcriptomes), *de novo* identification of regulatory regions poses a significant challenge. In the era of comparative genomics, sequence conservation in non-coding regions has served as a useful starting point to identify at least part of the suite of noncoding regulatory regions across the genome [92,93]. Unfortunately, the functional links between regulatory regions and the genes they regulate are often unclear, especially in enhancers that can act over long genomic distances. This uncertainty hinders our understanding of the connections between genotype and phenotype and requires additional approaches discussed below.

Functional characterization of genomic convergence

Given a convergent trait of interest, and candidate loci generated from any number of the genomic investigations described in the sections above, an additional step in understanding the underlying genetic mechanism for a convergent phenotype is functional validation of such genomic loci (Fig. 1). Studies from diverse groups of organisms have indicated that convergence at the genetic level can result from shared regulatory, metabolic, and developmental pathways, protein-coding genes with similar functions, or even identical amino-acid substitutions within the same gene (Table 1). These analyses range in complexity and cost, and include experimental embryology and tissue culture work, and the creation and testing of transgenics.

In recent years, techniques including CRISPR/Cas9 and massively parallelized reporter assays (MPRAs) have been added to the toolkits of those researchers interested in creating transgenic organisms, or testing hundreds or thousands of non-coding variants for enhancer activity [94,95]. Although we are unaware of any studies of genomic convergence utilizing MRPAs at the time of writing, the possibility of functionally assessing thousands of candidate loci in cell lines will almost certainly prove fruitful. CRISPR/Cas9 and other genome editing technologies are considered in many cases to be the gold standard for functional testing, and a recent study on cavefish (*Astyanax mexicanus*) metabolism utilized this technology to demonstrate a convergent insulin resistance phenotype with potential medical relevance across fish and humans [96]. Populations of river-dwelling Mexican tetra have repeatedly become isolated in caves, and the resulting cavefish have convergently evolved pigment, metabolic, and visual adaptations. After analyzing candidate genes within the insulin pathway, researchers uncovered a protein coding change in the insulin receptor gene of two independent populations of cavefish that both show insulin resistance and larger body size compared to their surface-dwelling relatives. This same substitution was also identified in insulin-resistant humans that suffer from Rabson–Mendelhall syndrome, and when placed into a zebrafish background using CRISPR/Cas9, this amino acid change resulted in insulin-resistant zebrafish that were larger than their wild-type siblings [96].

In contrast to this transgenic work, another example of non-model vertebrate convergence in pigeon feather crests was functionally tested with *E. coli* and selective media. Following population genetic analyses that identified a substitution in the gene *EphB2* as a likely candidate for the reversal of feathers on the head of pigeons, researchers discovered a neighboring missense mutation in a species of dove with a convergent crest. In a simple and inexpensive

assay, wild-type and convergently-crested *EphB2* genes were transformed into bacteria and plated; based on the known toxicity of wild-type EphB2 protein to bacteria, it was possible to determine that both crested *EphB2* convergent mutations negatively altered the protein's biochemical function [97].

In the past, there has been considerable debate about whether the link between genotype and phenotype is explained by only few major core genes, or whether it is due to accumulation of small-effect changes at multiple loci across the genome, highlighting the important distinction between polygenic vs Mendelian phenotypes [98–100]. Similarly, there has been growing debate about whether most of the genetic variance is hidden as numerous rare variants of large effect or common variants of small effect [101]. In addition, pleiotropy (involvement of the same genes in multiple traits) poses challenges in associating a particular genetic locus with a phenotypic change [102]. Existence of pleiotropy in complex traits has been widely reported in genome-wide association studies (GWAS) [103], and this observation has been a constant challenge for evolutionary-development (evo-devo) studies [104]. Patterns of pleiotropic variants may confound linking genotypic signatures to a particular trait and systematic approaches are required to identify pleiotropic variants and their associations to infer molecular mechanisms shared by multiple traits [105]. For example, pigmentation has been widely used natural trait to assess the importance of convergent evolution at genetic level with *Agouti* and *MC1R* being identified as obvious candidate genes to have strong effect on pigmentation in vertebrates [106]. But these convergence in pigmentation have been identified at multiple levels of mutations, gene or gene functions [48,107], an example that highlights the underlying challenge of identifying causal genetic architecture associated with phenotypic convergence.

Additionally, questions about the relative roles of regulatory vs. structural protein coding variation as the main drivers of morphological evolution are not new [108,109]. In the past, studies of only few genetic loci did not provide enough resolution to indicate preference for regulatory or protein coding changes for adaptation, but the rise in large scale genomic studies on adaptive evolution in the future will continue to address this debate. Quantitative measures of the contribution of protein-coding versus regulatory to convergent traits are also needed. Even if we had the complete catalogue of mutations underlying a convergent trait, how would we quantify the relative contributions of these two mutational sources to the convergent phenotypes? Phylogenetic analogues to QTL mapping, which could provide estimates of the proportion of trait similarity between species that can be attributed to a given locus, are perhaps a distant goal, but new perspectives on quantitative genetics in a phylogenetic context are already providing glimpses of this future [33,110].

The future of convergent genomic analyses will make use of these complementary functional genomics and analytical techniques for improved resolution of the genetic architecture underlying trait evolution. Thus far, genomic analyses have affirmed that convergence does exist at the phenotypic and molecular levels, with evidence of both protein-coding and regulatory convergence down to the level of single nucleotide mutations (Table 1). A more pressing question is to what extent can functional confirmation of the effect of a given mutation close the explanatory gap between historical scenarios and molecular mechanisms? Does demonstration of a functional effect of a mutation mean that the historical sequence of mutational events has been confirmed? Depending on the experimental and historical context, functional testing of a given mutation today may or may not confirm a specific sequence of mutational events in the past.

Future directions

Renewed interest in the study of evolutionary convergence abounds and is driven in part by the emergence of genomics. As anticipated, genomic data have yielded several examples of convergent genotypic or molecular evolution, many of which are cited in the sections above. However, excitement arising from this area of research has unfortunately overshadowed the reality that most trait classifications are quite broadly defined, resulting in incomplete or potentially biased interpretations of results. Studies of convergence will benefit from having multiple replicates of independent convergences [19] and clear hypotheses and definitions of the phenotypic traits undergoing convergence [66]. It remains challenging to identify instances of convergence for which a genomic perspective will likely lead to significant new insights.

A detailed and nuanced interpretation of phenotypic diversity will be greatly facilitated through the continued support of natural history investigations and extensive and comprehensive digitization and databasing of phenotypic traits. Although the natural history literature serves as an important resource, studies on phenotypic convergence will benefit even more from direct research on natural history collections, particularly given the potential of collections worldwide to house an increasing diversity of specimen types. Emerging databases that catalog the relationships and natural history characteristics of organisms, including the Global Biodiversity Information Facility (GBIF), [111]) and the Encyclopedia of Life [52], are a promising start towards cataloging instances of phenotypic convergence. Moreover, data-rich technologies are emerging that are capable of quickly generating detailed information on natural history characteristics, such as gross organism morphology (e.g., computerized tomography; [112,113]) and environmental preferences (e.g., geographic information systems [114] and thermal imaging [115]). A concerted initiative, such as a broadening of platforms like Phenoscape [53,116], is

needed in order to integrate these data with the skills and knowledge of organismal biologists, physiologists, molecular biologists, geneticists, and other stakeholders to produce detailed, hierarchical, logically coherent and searchable descriptions of organismal phenotypes. Recent efforts in large scale digitization of scientific texts, assembling phylogenomic data matrices [117] and development of automated text mining and natural language processing approaches can also facilitate high-throughput generation of phenomic datasets [118]. In addition, development of ontological phenotypic databases that contains standard terms, definitions and synonyms that can be used to describe a phenotype is also a key in generating such phenomic resources [119]. Such integration, and the computational infrastructure to allow easy access to large data sets [120], will greatly accelerate discovery of the degree, timing, and mechanisms of convergence among taxa.

Large, data-driven and taxon-rich phylogenies with comprehensive metadata attached to each taxon are a prerequisite for scaling up of genomic studies of convergence. A principle use of phylogenies is for testing macroevolutionary models [12,121] to identify whether a trait of interest is statistically associated with other phenotypic traits or with broader ecological variables; such associations can be used to support adaptive scenarios for the evolution of a trait (e.g. [122]). Numerous analytical frameworks have recently been built to address this aim, allowing increasingly complex adaptive landscapes to be modeled and associated with both continuous and discrete phenotypic traits [121]. However, a long-recognized shortcoming of model-testing approaches in this field, and in general, is the possibility that a best-fit model may still poorly reflect empirical evolution of traits across lineages [123]. One potential solution has very recent emerged called phylogenetic natural history, a framework that advocates combining model hypothesis testing with empirically-derived knowledge to better understand

macroevolutionary patterns and associations [123]. Extensions of this and other approaches will be important for the continued improvement of phylogenetic comparative methods.

Phylogenies are also important for inferring evolutionary rates for regions across the genome to identify loci putatively underlying convergent phenotypes. Such analyses are widely used for analyzing protein-coding regions even before the genomic era [124], but analogous methods designed for estimating convergent rate variation in non-coding regions, such as conserved non-exonic elements, are less well developed and therefore require additional attention [28,92,125,126]. In addition, most phylogenies are only represented as purely bifurcating and phylogenetic reticulation are often not considered [127], leaving us unable to discern 'truly convergent' versus 'borrowed' traits. Moreover, our increasing ability to amass evolutionary rate estimates for thousands of genomic loci presents additional challenges of minimizing type II errors [128]. The rapid pace at which quality genome assemblies are being produced will be an important foundation for testing all genomic compartments, both coding and noncoding for a role in convergent phenotypes.

Although new approaches for phenotyping organisms are emerging, new functional genomics approaches have yet to be integrated with comparative genomics approaches [120,129]. A major challenge for the field moving forward will therefore be combining these rich forms of species- or even tissue- or cell- specific data (such as are available for the human genome) with inferences derived from cross-species genomic comparisons to functionally evaluate genomic drivers of in convergent evolution. Integrating genomics data, cutting-edge laboratory and computational techniques, and detailed, multi-level understanding of diverse natural history data will help answer fundamental questions about the propensity for convergent evolution and the genetic and molecular underpinnings of convergent phenotypes.

Data accessibility. This article has no additional data.

Authors' contributions. All authors planned the organization and themes of the paper. SL lead the writing of the paper, with substantial writing from DCC, PG, JT and SVE. All authors edited and approved the final text.

Competing interests. The authors declare no competing interests related to the subject matter.

Funding. This work was supported in part by NSF grant DEB-1355343/EAR-1355292 to SVE, and JAC. SL was supported by a Wenner Gren Postdoctoral Fellowship. DCC was supported by an NSF Postdoctoral Fellowship in Biology (Biological Collections - DBI 1812310). PG was supported by an NSERC PGSD-3 grant. JT was supported by a grant from Lemman Brazil Research Fund at Harvard University to SVE, Naomi Pierce and Cristina Miyaki. KN was supported by a postdoctoral fellowship from the Swiss National Science Foundation. FTG was supported by a Harvard CONACYT (Mexico) Postdoctoral Fellowship.

Acknowledgements. We thank Terry Capellini and Jim Hanken for helpful discussions; Nathan Clarke and one anonymous reviewer for helpful comments on the manuscript; the curators and curatorial associates of the Department of Herpetology of the Museum of Comparative Zoology, Harvard, and of the California Academy of Science, for loaning material; and Lily Lu for the drawings in Fig. 2.

REFERENCES

1. Losos JB. 2011 Convergence, adaptation, and constraint. *Evolution* **65**, 1827–1840. (doi:10.1111/j.1558-5646.2011.01289.x)
2. Darwin 1809-1882 C. 1859 *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life*. London : John Murray, 1859. See https://search.library.wisc.edu/catalog/9934839413602122.
3. Gould SJ. 1992 Wonderful Life; The Burgess Shale and the Nature of History. *J. Gen. Philos. Sci. / Zeitschrift für Allg. Wissenschaftstheorie* **23**.
4. Agrawal AA. 2017 Toward a Predictive Framework for Convergent Evolution: Integrating Natural History, Genetic Mechanisms, and Consequences for the Diversity of Life. *Am. Nat.* **190**, S1–S12.
5. Rosenblum EB, Parent CE, Brandt EE. 2014 The Molecular Basis of Phenotypic Convergence. *Annu. Rev. Ecol. Evol. Syst.* **45**, 203–226. (doi:10.1146/annurev-ecolsys-120213-091851)
6. Schluter D. 2000 *The ecology of adaptive radiation*. Oxford : Oxford University Press, 2000. See https://search.library.wisc.edu/catalog/999914007502121.
7. Mayr E. 1963 *Animal Species and Evolution*. Belknap of Harvard University Press.
8. Simpson 1902- GG. 1953 *The major features of evolution / George Gaylord Simpson*. New York: Columbia University Press.
9. Stayton CT. 2008 Is convergence surprising? An examination of the frequency of convergence in simulated datasets. *J. Theor. Biol.* **252**, 1–14. (doi:https://doi.org/10.1016/j.jtbi.2008.01.008)
10. Wake DB, Wake MH, Specht CD. 2011 Homoplasy: From Detecting Pattern to Determining Process and Mechanism of Evolution. *Science.* **331**, 1032 LP-1035.
11. Guerrero RF, Hahn MW. 2018 Quantifying the risk of hemiplasy in phylogenetic inference. *Proc. Natl. Acad. Sci.* **115**, 12787–12792.
12. Stayton CT. 2015 The definition, recognition, and interpretation of convergent evolution, and two new measures for quantifying and assessing the significance of convergence. *Evolution (N. Y).* **69**, 2140–2153. (doi:10.1111/evo.12729)
13. Natarajan C *et al.* 2015 Convergent Evolution of Hemoglobin Function in High-Altitude Andean Waterfowl Involves Limited Parallelism at the Molecular Sequence Level. *PLOS Genet.* **11**, e1005681.
14. Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ. 2013 Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**, 228–231. (doi:10.1038/nature12511)
15. Harvey PH, Pagel MD. 1991 *The Comparative Method in Evolutionary Biology*. Oxford: Oxford University Press.
16. Martins EP. 2000 Adaptation and the comparative method. *Trends Ecol. Evol.* **15**, 296–299.
17. Chikina M, Robinson JD, Clark NL. 2016 Hundreds of Genes Experienced Convergent Shifts in Selective Pressure in Marine Mammals. *Mol Biol Evol* **33**, 2182–2192. (doi:10.1093/molbev/msw112)
18. Partha R, Chauhan BK, Ferreira Z, Robinson JD, Lathrop K, Nischal KK, Chikina M, Clark NL. 2017 Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *Elife* **6**. (doi:10.7554/eLife.25884)

19. Sackton TB *et al.* 2018 Convergent regulatory evolution and the origin of flightlessness in palaeognathous birds. *bioRxiv*
20. Woodard SH, Fischman BJ, Venkat A, Hudson ME, Varala K, Cameron SA, Clark AG, Robinson GE. 2011 Genes involved in convergent evolution of eusociality in bees. *Proc. Natl. Acad. Sci.* **108**, 7472 LP-7477.
21. Goodson JL, Kingsbury MA. 2013 What's in a name? Considerations of homologies and nomenclature for vertebrate social behavior networks. *Horm. Behav.* **64**, 103–112. (doi:10.1016/j.yhbeh.2013.05.006)
22. O'Connell LA, Hofmann HA. 2012 Evolution of a vertebrate social decision-making network. *Science* **336**, 1154–1157. (doi:10.1126/science.1218889)
23. Pfenning AR *et al.* 2014 Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* **346**, 1256846. (doi:10.1126/science.1256846)
24. Jones G. 2010 Molecular Evolution: Gene Convergence in Echolocating Mammals. *Curr. Biol.* **20**, R62–R64. (doi:10.1016/j.cub.2009.11.059)
25. Lambert MJ, Nevue AA, Portfors C V. 2017 Contrasting patterns of adaptive sequence convergence among echolocating mammals. *Gene* **605**, 1–4. (doi:10.1016/j.gene.2016.12.017)
26. Bolnick DI, Barrett RDH, Oke KB, Rennison DJ, Stuart YE. 2018 (Non)Parallel Evolution. *Annu. Rev. Ecol. Evol. Syst.* **49**, 303–330.
27. Levy Karin E, Wicke S, Pupko T, Mayrose I. 2017 An Integrated Model of Phenotypic Trait Changes and Site-Specific Sequence Evolution. *Syst. Biol.* **66**, 917–933. (doi:10.1093/sysbio/syx032)
28. Hu Z, Sackton TB, Edwards S V., Liu JS. 2018 A hierarchical Bayesian model for detecting convergent rate changes of conserved noncoding elements on phylogenetic trees. *bioRxiv* , 260745.
29. Collyer ML, Sekora DJ, Adams DC. 2015 A method for analysis of phenotypic change for phenotypes described by high-dimensional data. *Heredity (Edinb).* **115**, 357–365. (doi:10.1038/hdy.2014.75)
30. Adams DC, Collyer ML. 2009 A general framework for the analysis of phenotypic trajectories in evolutionary studies. *Evolution (N. Y).* **63**, 1143–1154.
31. Collyer ML, Adams DC. 2007 Analysis of two-state multivariate phenotypic change in ecological studies. *Ecology* **88**, 683–692. (doi:10.1890/06-0727)
32. Hadfield JD, Nakagawa S. 2010 General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *J. Evol. Biol.* **23**, 494–508. (doi:10.1111/j.1420-9101.2009.01915.x)
33. Mendes FK, Fuentes-Gonzalez JA, Schraiber JG, Hahn MW. 2018 A multispecies coalescent model for quantitative traits. *Elife* **7**. (doi:10.7554/eLife.36482)
34. Conway Morris S. 1998 *The crucible of creation: the Burgess shale and the rise of animals*. Oxford, UK: Oxford University Press.
35. IWANIUK N, James HF. 1999 Extraordinary cranial specialization in a new genus of extinct duck (Aves□: Anseriformes) from Kauai , Hawaiian Islands ANDREW.
36. Kumar S, Stecher G, Suleski M, Hedges SB. 2017 TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**, 1812–1819.
37. Heintzman PD, Zazula GD, Cahill JA, Reyes A V, MacPhee RDE, Shapiro B. 2015 Genomic Data from Extinct North American Camelops Revise Camel Evolutionary History. *Mol. Biol. Evol.* **32**, 2433–2440.

38. Qu Q, Haitina T, Zhu M, Ahlberg PE. 2015 New genomic and fossil data illuminate the origin of enamel. *Nature* **526**, 108.
39. Cook JA *et al.* 2014 Natural History Collections as Emerging Resources for Innovative Education. *Bioscience* **64**, 725–734. (doi:10.1093/biosci/biu096)
40. Suarez A V, Tsutsui ND. 2004 The value of museum collections for research and society. *Bioscience* **54**, 66–74.
41. Schmitt CJ, Cook JA, Zamudio KR, Edwards S V. 2018 Museum specimens of terrestrial vertebrates are sensitive indicators of environmental change in the Anthropocene. *Philos. Trans. R. Soc. B-Biological Sci.* **in revisio**.
42. Snow D. 1980 A new species of cotinga from southeastern Brazil. *Bull. Br. Ornithol. Club* **100**, 213–215.
43. Helgen KM, Pinto CM, Kays R, Helgen LE, Tsuchiya MTN, Quinn A, Wilson DE, Maldonado JE. 2013 Taxonomic revision of the olingos (Bassaricyon), with description of a new species, the Olinguito. *Zookeys* , 1–83. (doi:10.3897/zookeys.324.5827)
44. Losos JB *et al.* 2013 Evolutionary Biology for the 21st Century. *PLOS Biol.* **11**, e1001466.
45. Hiller M, Schaar BT, Indjeian VB, Kingsley DM, Hagey LR, Bejerano G. 2012 A 'forward genomics' approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Rep* **2**, 817–823. (doi:10.1016/j.celrep.2012.08.032)
46. Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P. 2012 Parallel molecular evolution in an herbivore community. *Science.* **337**, 1634–1637. (doi:10.1126/science.1226630)
47. Kronforst MR *et al.* 2012 Unraveling the thread of nature's tapestry: the genetics of diversity and convergence in animal pigmentation. *Pigment Cell Melanoma Res* **25**, 411–433. (doi:10.1111/j.1755-148X.2012.01014.x)
48. Manceau M, Domingues VS, Linnen CR, Rosenblum EB, Hoekstra HE. 2010 Convergence in pigmentation at multiple levels: mutations, genes and function. *Philos Trans R Soc L. B Biol Sci* **365**, 2439–2450. (doi:10.1098/rstb.2010.0104)
49. Rosenblum EB, Römpler H, Schöneberg T, Hoekstra HE. 2010 Molecular and functional basis of phenotypic convergence in white lizards at White Sands. *Proc. Natl. Acad. Sci.* **107**, 2113 LP-2117.
50. Gregory TR. 2005 The Animal Genome Size Database. *Anim. Genome Size Database*.
51. MorphoBank. In press. MorphoBank.
52. Parr CS *et al.* 29 ADThe Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth. *Biodivers. Data J.* **2**, e1079.
53. Edmunds RC *et al.* 2016 Phenoscape: Identifying Candidate Genes for Evolutionary Phenotypes. *Mol. Biol. Evol.* **33**, 13–24. (doi:10.1093/molbev/msv223)
54. Global Genome Initiative. Smithsonian National Museum of Natural History. https://ggi.si.edu (accessed 9 January 2019).
55. Seutin G, White BN, Boag PT. 1991 Preservation of avian blood and tissue samples for DNA analyses. *Can. J. Zool.* **69**, 82–90. (doi:10.1139/z91-013)
56. Grayson P, Sin SYW, Sackton T, Edwards S V. 2017 Comparative genomics as a foundation for evo-devo studies in birds. In *Methods in Molecular Biology: Avian and Reptilian Developmental Biology*, New York: Humana Press.
57. Putnam NH *et al.* 2016 Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350.

58. Rice ES *et al.* 2017 Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling. *Genome Res.* **27**, 686–696. (doi:10.1101/gr.213595.116)
59. Koepfli K-P, Paten B, O'Brien SJ. 2015 The Genome 10K Project: a way forward. *Annu. Rev. Anim. Biosci.* **3**, 57–111. (doi:10.1146/annurev-animal-090414-014900)
60. Botero-Castro F, Figuet E, Tilak M-K, Nabholz B, Galtier N. 2017 Avian Genomes Revisited: Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds. *Mol. Biol. Evol.* **34**, 3123–3131.
61. Figuet E, Nabholz B, Bonneau M, Mas Carrio E, Nadachowska-Brzyska K, Ellegren H, Galtier N. 2016 Life History Traits, Protein Evolution, and the Nearly Neutral Theory in Amniotes. *Mol. Biol. Evol.* **33**, 1517–1527.
62. Weber CC, Nabholz B, Romiguier J, Ellegren H. 2014 Kr/Kc but not dN/dS correlates positively with body mass in birds, raising implications for inferring lineage-specific selection. *Genome Biol.* **15**, 542. (doi:10.1186/s13059-014-0542-8)
63. Rey C, Guéguen L, Sémon M, Boussau B. 2018 Accurate detection of convergent amino-acid evolution with PCOC. *Mol. Biol. Evol.* **35**, 2296–2306. (doi:10.1093/molbev/msy114)
64. Xu S, He Z, Guo Z, Zhang Z, Wyckoff GJ, Greenberg A, Wu C-I, Shi S. 2017 Genome-Wide Convergence during Evolution of Mangroves from Woody Plants. *Mol. Biol. Evol.* **34**, 1008–1015.
65. Marcovitz A, Turakhia Y, Gloudemans M, Braun BA, Chen HI, Bejerano G. 2017 A novel unbiased test for molecular convergent evolution and discoveries in echolocating, aquatic and high-altitude mammals. *bioRxiv* , 170985. (doi:10.1101/170985)
66. Pease JB, Haak DC, Hahn MW, Moyle LC. 2016 Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *PLOS Biol.* **14**, e1002379.
67. Muntané G, Farré X, Rodríguez JA, Pegueroles C, Hughes DA, de Magalhães JP, Gabaldón T, Navarro A. 2018 Biological Processes Modulating Longevity across Primates: A Phylogenetic Genome-Phenome Analysis. *Mol. Biol. Evol.* **35**, 1990–2004.
68. Thomas GWC, Hahn MW. 2015 Determining the Null Model for Detecting Adaptive Convergence from Genomic Data: A Case Study using Echolocating Mammals. *Mol. Biol. Evol.* **32**, 1232–1236. (doi:10.1093/molbev/msv013)
69. Zou Z, Zhang J. 2015 No genome-wide protein sequence convergence for echolocation. *Mol. Biol. Evol.* **32**, 1237–1241. (doi:10.1093/molbev/msv014)
70. Castoe TA, Koning APJ de, Kim H-M, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD. 2009 Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc. Natl. Acad. Sci.* **106**, 8986–8991. (doi:10.1073/pnas.0900233106)
71. Kowalczyk A, Meyer WK, Partha R, Mao W, Clark NL, Chikina M. 2018 RERconverge: an R package for associating evolutionary rates with convergent traits. *bioRxiv* , 451138. (doi:10.1101/451138)
72. Yang Z. 2007 PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591. (doi:10.1093/molbev/msm088)
73. Pond SLK, Frost SDW, Muse S V. 2005 HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676–679.
74. Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P. 2012 Parallel Molecular Evolution in an Herbivore Community. *Science.* **337**, 1634–1637. (doi:10.1126/science.1226630)

75. Dobler S, Dalla S, Wagschal V, Agrawal AA. 2012 Community-wide convergent evolution in insect adaptation to toxic cardenolides by substitutions in the Na,K-ATPase. *Proc. Natl. Acad. Sci.* **109**, 13040–13045. (doi:10.1073/pnas.1202111109)
76. Zakon HH. 2012 Adaptive evolution of voltage-gated sodium channels: The first 800 million years. *Proc. Natl. Acad. Sci.* **109**, 10619–10625. (doi:10.1073/pnas.1201884109)
77. McGlothlin JW, Chuckalovcak JP, Janes DE, Edwards S V., Feldman CR, Brodie ED, Pfrender ME, Brodie ED. 2014 Parallel Evolution of Tetrodotoxin Resistance in Three Voltage-Gated Sodium Channel Genes in the Garter Snake Thamnophis sirtalis. *Mol. Biol. Evol.* **31**, 2836–2846. (doi:10.1093/molbev/msu237)
78. Naseeb S, Ames RM, Delneri D, Lovell SC. 2017 Rapid functional and evolutionary changes follow gene duplication in yeast. *Proc. R. Soc. B Biol. Sci.* **284**.
79. Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, Shanmugam D, Roos DS, Stoeckert CJ. 2002 Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups. In *Current Protocols in Bioinformatics*, John Wiley & Sons, Inc.
80. Emms DM, Kelly S. 2015 OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157. (doi:10.1186/s13059-015-0721-2)
81. Li L, Stoeckert CJ, Roos DS. 2003 OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* **13**, 2178–2189. (doi:10.1101/gr.1224503)
82. De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006 CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271. (doi:10.1093/bioinformatics/btl097)
83. Liu L, Yu L, Kalavacharla V, Liu Z. 2011 A Bayesian model for gene family evolution. *BMC Bioinformatics* **12**, 426. (doi:10.1186/1471-2105-12-426)
84. Liegertová M *et al.* 2015 Cubozoan genome illuminates functional diversification of opsins and photoreceptor evolution. *Sci. Rep.* **5**, 11885. (doi:10.1038/srep11885)
85. Hoffmann FG, Opazo JC, Storz JF. 2010 Gene cooption and convergent evolution of oxygen transport hemoglobins in jawed and jawless vertebrates. *Proc. Natl. Acad. Sci.* **107**, 14274–14279. (doi:10.1073/pnas.1006756107)
86. Casewell NR *et al.* 2017 The Evolution of Fangs, Venom, and Mimicry Systems in Blenny Fishes. *Curr. Biol.* **27**, 1184–1191. (doi:10.1016/j.cub.2017.02.067)
87. Whittington CM *et al.* 2010 Novel venom gene discovery in the platypus. *Genome Biol.* **11**, R95. (doi:10.1186/gb-2010-11-9-r95)
88. Xie KT *et al.* 2019 DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science.* **363**, 81 LP-84. (doi:10.1126/science.aan1425)
89. Lowe CB, Clarke JA, Baker AJ, Haussler D, Edwards S V. 2015 Feather development genes and associated regulatory innovation predate the origin of Dinosauria. *Mol. Biol. Evol.* **32**, 23–28. (doi:10.1093/molbev/msu309)
90. Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, Salama SR, Kingsley DM, Lindblad-Toh K, Haussler D. 2011 Three Periods of Regulatory Innovation During Vertebrate Evolution. *Science.* **333**, 1019 LP-1024.
91. Larter M, Dunbar-Wallis A, Berardi AE, Smith SD. 2018 Convergent Evolution at the Pathway Level: Predictable Regulatory Changes during Flower Color Transitions. *Mol. Biol. Evol.* **35**, 2159–2169.
92. Siepel A *et al.* 2005 Evolutionarily conserved elements in vertebrate, insect, worm, and

yeast genomes. *Genome Res.* **15**, 1034–1050. (doi:10.1101/gr.3715005)
93. Woolfe A *et al.* 2004 Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLOS Biol.* **3**, e7. (doi:10.1371/journal.pbio.0030007)
94. Tewhey R *et al.* 2018 Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell*. **172**, 1132–1134. (doi:10.1016/j.cell.2018.02.021)
95. Cong L *et al.* 2013 Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823. (doi:10.1126/science.1231143)
96. Riddle MR *et al.* 2018 Insulin resistance in cavefish as an adaptation to a nutrient-limited environment. *Nature* **555**, 647.
97. Vickrey AI, Domyan ET, Horvath MP, Shapiro MD. 2015 Convergent Evolution of Head Crests in Two Domesticated Columbids Is Associated with Different Missense Mutations in EphB2. *Mol. Biol. Evol.* **32**, 2657–2664. (doi:10.1093/molbev/msv140)
98. Field Y *et al.* 2016 Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764. (doi:10.1126/science.aag0776)
99. Botstein D, Risch N. 2003 Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33 Suppl**, 228–237.
100. Boyle EA, Li YI, Pritchard JK. 2017 An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186. (doi:10.1016/j.cell.2017.05.038)
101. Gibson G. 2012 Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145. (doi:10.1038/nrg3118)
102. Porto A, Schmelter R, VandeBerg JL, Marroig G, Cheverud JM. 2016 Evolution of the Genotype-to-Phenotype Map and the Cost of Pleiotropy in Mammals. *Genetics* **204**, 1601–1612. (doi:10.1534/genetics.116.189431)
103. Yang C, Li C, Wang Q, Chung D, Zhao H. 2015 Implications of pleiotropy: challenges and opportunities for mining Big Data in biomedicine . *Front. Genet.* . **6**, 229.
104. Pavličev M. 2016 Pleiotropy and Its Evolution: Connecting Evo-Devo and Population Genetics BT - Evolutionary Developmental Biology: A Reference Guide. In (eds L Nuno de la Rosa, G Müller), pp. 1–10. Cham: Springer International Publishing. (doi:10.1007/978-3-319-33038-9_52-1)
105. Zhan J, Arking DE, Bader JS. 2018 Discovering patterns of pleiotropy in genome-wide association studies. *bioRxiv* , 273540. (doi:10.1101/273540)
106. Hubbard JK, Uy JAC, Hauber ME, Hoekstra HE, Safran RJ. 2010 Vertebrate pigmentation: from underlying genes to adaptive function. *Trends Genet.* **26**, 231–239. (doi:10.1016/j.tig.2010.02.002)
107. Kratochwil CF, Liang Y, Gerwin J, Woltering JM, Urban S, Henning F, Machado-Schiaffino G, Hulsey CD, Meyer A. 2018 Agouti-related peptide 2 facilitates convergent evolution of stripe patterns across cichlid fish radiations. *Science.* **362**, 457 LP-460. (doi:10.1126/science.aao6809)
108. Carroll SB. 2000 Endless forms: the evolution of gene regulation and morphological diversity. *Cell* **101**, 577–580.
109. Hoekstra HE, Coyne JA, Rausher M. 2007 The locus of evolution: Evo-Devo and the genetics of adaptation. *Evolution (N. Y).* **61**, 995–1016. (doi:10.1111/j.1558-5646.2007.00105.x)
110. Broman KW, Kim S, Sen S, Ané C, Payseur BA. 2012 Mapping quantitative trait loci

onto a phylogenetic tree. *Genetics* **192**, 267–279. (doi:10.1534/genetics.112.142448)
111. Samy G, Chavan V, Ariño AH, Otegui J, Hobern D, Sood R, Robles E. 2013 Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials. *Biodivers. Informatics; Vol 8, No 2 (2013)DO - 10.17161/bi.v8i2.4124*
112. Metscher BD. 2009 MicroCT for comparative morphology: simple staining methods allow high-contrast 3D imaging of diverse non-mineralized animal tissues. *BMC Physiol.* **9**, 11. (doi:10.1186/1472-6793-9-11)
113. Hörnschemeyer T, Beutel RG, Pasop F. 2002 Head structures of Priacma serrata leconte (coleptera, archostemata) inferred from X-ray tomography. *J. Morphol.* **252**, 298–314. (doi:10.1002/jmor.1107)
114. Kozak KH, Graham CH, Wiens JJ. 2008 Integrating GIS-based environmental data into evolutionary biology. *Trends Ecol. Evol.* **23**, 141–148. (doi:10.1016/j.tree.2008.02.001)
115. Goller M, Goller F, French SS. 2014 A heterogeneous thermal environment enables remarkable behavioral thermoregulation in Uta stansburiana. *Ecol. Evol.* **4**, 3319–3329. (doi:10.1002/ece3.1141)
116. Deans AR *et al.* 2015 Finding Our Way through Phenotypes. *PLOS Biol.* **13**, e1002033.
117. O'Leary MA *et al.* 2013 The Placental Mammal Ancestor and the Post–K-Pg Radiation of Placentals. *Science.* **339**, 662 LP-667. (doi:10.1126/science.1229237)
118. Burleigh JG *et al.* 2013 Next-generation phenomics for the Tree of Life. *PLoS Curr.* **5**. (doi:10.1371/currents.tol.085c713acafc8711b2ff7010a4b03733)
119. Smith CL, Eppig JT. 2009 The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **1**, 390–399. (doi:10.1002/wsbm.44)
120. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. 2015 Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.* **16**, 85.
121. Mahler DL, Weber MG, Wagner CE, Ingram T. 2017 Pattern and Process in the Comparative Study of Convergent Evolution. *Am. Nat.* **190**, S13–S28. (doi:10.1086/692648)
122. Mahler DL, Ingram T, Revell LJ, Losos JB. 2013 Exceptional Convergence on the Macroevolutionary Landscape in Island Lizard Radiations. *Science.* **341**, 292 LP-295.
123. Uyeda JC, Zenil-Ferguson R, Pennell MW. 2018 Rethinking phylogenetic comparative methods. *Syst. Biol.*
124. Storz JF. 2016 Causes of molecular convergence and parallelism in protein evolution. *Nat. Rev. Genet.* **17**, 239.
125. Kostka D, Holloway AK, Pollard KS. 2018 Developmental Loci Harbor Clusters of Accelerated Regions That Evolved Independently in Ape Lineages. *Mol. Biol. Evol.* **35**, 2034–2045.
126. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010 Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121. (doi:10.1101/gr.097857.109)
127. Burbrink FT, Gehara M. 2018 The Biogeography of Deep Time Phylogenetic Reticulation. *Syst. Biol.* **67**, 743–755.
128. Storey JD, Tibshirani R. 2003 Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100**, 9440 LP-9445.
129. Lappalainen T. 2015 Functional genomics bridges the gap between quantitative genetics

and molecular biology. *Genome Res.* **25**, 1427–1431. (doi:10.1101/gr.190983.115)
130. Projecto-Garcia J *et al.* 2013 Repeated elevational transitions in hemoglobin function during the evolution of Andean hummingbirds. *Proc. Natl. Acad. Sci.* **110**, 20669 LP-20674. (doi:10.1073/pnas.1315456110)
131. Hu Y *et al.* 2017 Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. *Proc. Natl. Acad. Sci.* **114**, 1081 LP-1086.
132. Zhang G *et al.* 2014 Comparative genomics reveals insights into avian genome evolution and adaptation. *Science.* **346**, 1311–1320.
133. Kapheim KM *et al.* 2015 Genomic signatures of evolutionary transitions from solitary to group living. *Science.*
134. Gallant JR, Losilla M, Tomlinson C, Warren WC. 2017 The Genome and Adult Somatic Transcriptome of the Mormyrid Electric Fish Paramormyrops kingsleyae. *Genome Biol. Evol.* **9**, 3525–3530. (doi:10.1093/gbe/evx265)
135. Berens AJ, Hunt JH, Toth AL. 2015 Comparative transcriptomics of convergent evolution: different genes but conserved pathways underlie caste phenotypes across lineages of eusocial insects. *Mol. Biol. Evol.* **32**, 690–703.

Figure legends

Fig 1. Conceptual framework for studies on the genomics of phenotypic convergence. Starting from the top, organismal expertise and knowledge of natural history are the starting points for such studies. Environmental gradients and constraints in physiology, biochemical and developmental pathways may limit or direct trait evolution, potentially driving phenotypic convergence. Phylogenetic comparative methods can be used to test, quantify and visualize instances of convergence. Finally, comparative genomics methods can be used to test whether convergent phenotypes have common underlying genomic mechanisms at various hierarchical level of individual genetic loci or regulatory networks. Functional validations of genes or pathways identified from genome-wide scans provide means to test the role of specific genomic regions in producing a given convergent phenotype and to attempt historical reconstruction of evolutionary events.

Fig. 2. Illustration of the continuous nature of limb states often categorized as binary in studies of evolutionary convergence. In both panels, taxa were chosen to illustrate a range of forelimb character states and for which we had easily verified details from microscopy, photographs or specimens. X indicates complete absence of fore- or hindlimb elements. Drawings not to scale. Top, tree for palaeognathous birds (topology after [18]), with representative drawings of forelimbs for taxa casually deemed volant or flightless. Taxa and sources as follows: Elegant Crested Tinamou (*Eudromia elegans*, MCZ343064 and 340325); Little Bush Moa (*Anomalopteryx didiformis*; all forelimb elements absent); Emu (*Dromaius novaehollandiae*; after photo by JAC from Muséum National d'Histoire Naturelle [MHNH]); Southern Cassowary (Casuarius casuarius, JAC MHNH photo, MCZ364589); Little Spotted Kiwi (*Apteryx owenii*,

MCZ340308); Greater Rhea (*Rhea americana*, JAC MHNH photo, MCZ341488); Common Ostrich (*Struthio camelus*, JAC MHN photo, MCZ341420); Chicken (*Gallus gallus*, online sources). Bottom, examples of limbed and limbless squamates. Relationships after [104]. Common Crag Lizard (*Pseudocordylus melanotus*, CAS173019); Cape Grass Lizard (*Chamaesaura anguina*, MCZ R-173157); European Legless Lizard (*Pseudopus apodus*, CAS-184449), body cavity shown in light gray to provide positional context for hindlimb; Mexican Mole Lizard (*Bipes biporus*, CAS-142262, hindlimb absent); New Guinea Blind Lizard (*Dibamus novaeguineae,* CAS-SU 27070). MCZ, Museum of Comparative Zoology, Harvard; CAS, California Academy of Sciences. All limb drawings by Lily Lu.

Fig. 3. Numbers of papers on different kinds of evolutionary convergence by authors with or without a museum address (Search 1, see below). The goal of the searches was to determine if scientists with extensive natural history knowledge of organisms were participating in the second wave of studies on convergence informed by genomics (see text). We reasoned that museum specialists would constitute an important component of this community of researchers. We therefore conducted two searches on the Web of Science Core Collection on September 23, 2018 (Searches 1 and 2) and two searches on Pubmed (searches 3 and 4), on the same date. For each database we conducted two searches: Searches 1 and 3: Topic: “convergen*” AND “evolution” without or with, respectively, “genom*” as an additional topic keyword. To determine which papers had authors with museum addresses, we included “museum” OR “musee” or “museo” as part of the author address. For searches 2 and 4, we used “convergent evolution” OR “parallel evolution” as topic keywords, again, without or with, respectively, “genom*” as an additional topic keyword. Museum addresses were determined as in searches 1 and 3. The graph and

associated data (Supplementary Table 1) suggests that researchers with a museum address publish extensively on general evolutionary convergence, appearing on between 7.0% and 24.2% of papers in this literature depending on the search terms. However, researchers with a museum address appear on only 4.9% to 7.0% of papers on genomics of convergence (Supplementary Table 1). These addresses are underrepresented on papers on genomics of convergence by 29%-71%, depending on the analysis. We recognize that our search is likely to miss many individuals with extensive knowledge of organismal diversity that do not work in museums or have a museum address on their publications. Additionally, our search terms are likely to detect many papers that are tangential to this analysis (see Suppementary Table 1). Nonetheless, we suspect that the trends indicated reflect the coarse-grained approach to phenotypes that have partly characterized the second wave of studies on evolutionary convergence informed by genomics. We predict that the actual numbers of authors with extensive natural history knowledge, irrespective of their work addresses, and who have participated in studies of the genomics of convergence, would change the slopes but not relative magnitude of the trends seen here.

Table 1: Examples of studies identifying genomic signals of convergence at different hierarchical levels.

Convergent phenotypes studied	Methods used	Main findings	Level of convergence identified	Reference
High altitude adaptation in hummingbirds	Comparative genetics	Convergent amino acid substitutions	Amino acid	[130]
Pseudothumb and bamboo diet in Giant Panda	Comparative genomics	Convergent amino acid substitutions	Amino acid	[131]
Skin coloration in lizards	Cell-based assays	Convergent amino acid substitutions	Amino acid	[49]
Vocal learning in birds	Comparative genomics	Convergent accelerations in genes	Gene	[132]
Eusociality in bees	Comparative transcriptomics	Convergent accelerated evolution in genes	Gene expression	[20]
Vitamin C synthesis in mammals	Comparative genomics	Gene duplication/loss	Gene	[45]
Loss of flight in birds	Comparative genomics	Convergent rate shifts in noncoding DNA	Regulatory	[19]
Transitions from solitary to group living	Comparative genomics	Increase in potential for gene regulation and decrease in diversity and abundance of transposable elements	Regulatory	[133]
Electric organ in fish	Comparative transcriptomics	Convergence in similar transcription factors, developmental and cellular pathways	Regulatory	[134]
Eusociality in insects	Comparative transcriptomics	Convergent expression in biological pathways	Regulatory	[135]
Evolution of stripe patterns across cichlid fish radiations	CRISPR-Cas9 Genome editing	Regulatory changes of the gene act as molecular switches	Regulatory	[107]

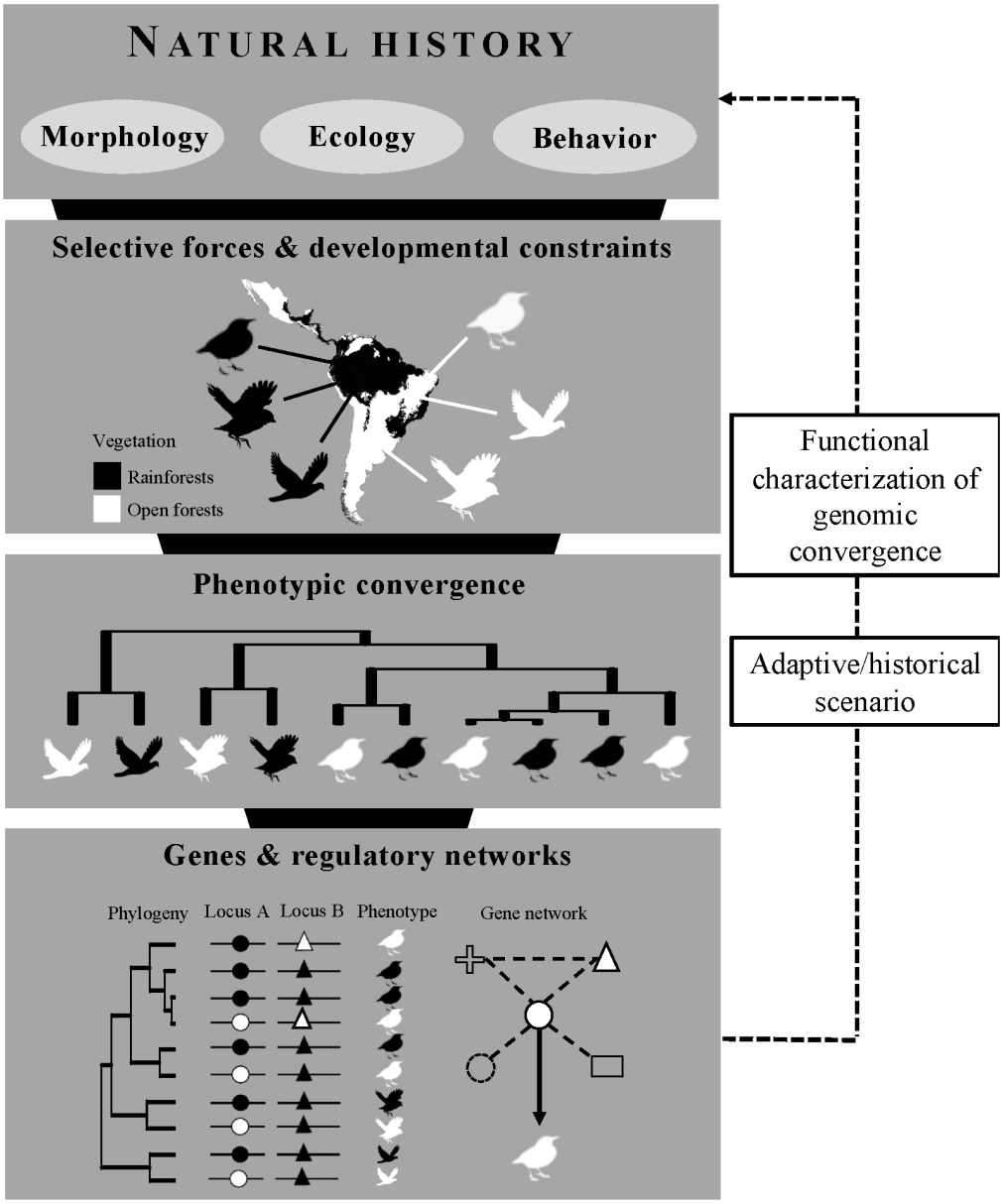

Fig. 1

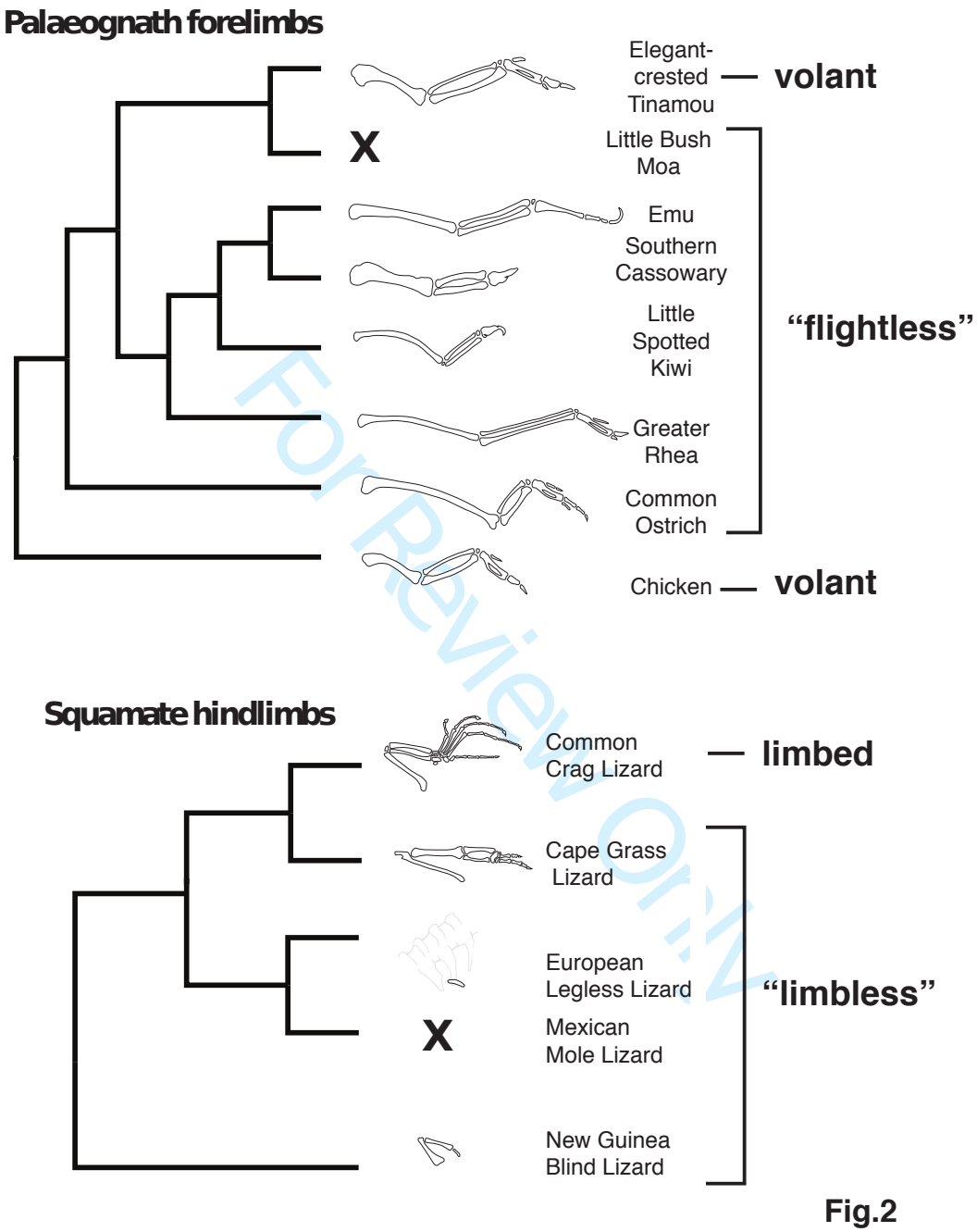

Fig.2

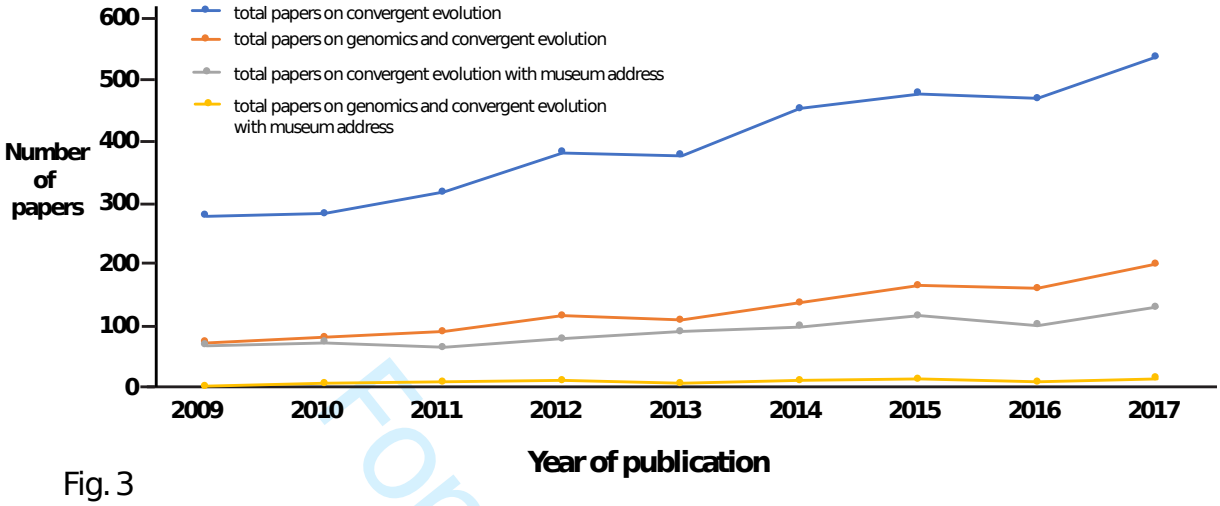

Fig. 3