Estimates of molecular convergence reveal pleiotropic genes underlying adaptive variation across teleost fish

Agneesh Barua[1,2*], Brice Beinsteiner[3], Vincent Laudet[4,5], Marc Robinson-Rechavi[1,2]

[1]Department of Ecology and Evolution, University of Lausanne
[2]Swiss Institute of Bioinformatics
[3]Helmholtz Pioneer Campus, Helmholtz Zentrum München, Neuherberg, Germany
[4]Marine Eco-Evo-Devo Unit, Okinawa Institute of Science and Technology Graduate University, Japan
[5]Marine Research Station, Institute of Cellular and Organismic Biology (ICOB), Academia Sinica, 23-10, Dah-Uen Rd, Jiau Shi, I-Lan 262, Taiwan

*Corresponding author

Abstract

Teleosts are the most diverse group of vertebrates on earth. Their diversity is a testament to the combined effects of genetic, developmental, and evolutionary forces. However, disentangling the interactions between these forces is challenging due to the complexity of the genotype-phenotype relationship and the masking of adaptive genetic signals by genetic noise. Estimates of molecular convergence where changes in the sequence of protein-coding genes lead to identical amino acid substitution across multiple lineages provide strong evidence of adaptive evolution. In this study, we estimated signals for molecular convergence in protein-coding genes across 143 teleost genomes to identify genes and processes that experienced adaptive changes. We find that genes with signals of molecular convergence are implicated in diverse

processes ranging from embryonic development, tissue morphogenesis, metabolism, to hormone and heat response. Some convergent substitutions are located on functionally important sites on proteins potentially providing the molecular basis for adaptations to hypoxia, salinity fluctuations, and varying skeletal morphologies. Additionally, single-cell RNA sequencing data from zebrafish showed that the convergent genes have dynamic expression across various cell types during embryonic development. These results highlight the functional importance of the convergent genes as well as their pleiotropic nature. Although traditionally considered a source of genetic constraint, we argue that adaptation via changes in pleiotropic genes are particularly advantageous during periods of ecological shifts. We present the pleiotropic release model which describes how adaptive variation on pleiotropic genes can have large fitness effects, allowing organisms to overcome selective pressures during periods of ecological shifts.

Introduction

Teleosts fishes represent the most specious and diverse groups of vertebrates on Earth, encompassing a wide array of phenotypes, morphologies, habitats, and life histories [1,2]. This diversity positions teleosts as an ideal group for exploring the broad evolutionary processes giving rise to species diversity and richness. Furthermore, the availability of high-quality genomes for numerous species, coupled with single-cell sequencing and functional data from zebrafish (*Danio rerio*), offers an excellent opportunity to investigate the potential effects variations in coding sequence might have on tissue functionality and animal phenotypes [3,4]. While research has shown that genetic variation resulting from whole genome duplications has an important role in the evolution of teleosts [5], disentangling the combined effects of evolution, genetics, and development requires identifying truly adaptive genetic changes from noise is a challenge.

An especially promising avenue involves performing genome-wide scans of selection to determine the genetic basis behind the origin of convergent traits [10–14]. Evolutionary convergence, characterised by different lineages independently acquiring similar character states, in response to similar ecological challenges, is pervasive in nature [15]. While convergence typically invokes images of recurring phenotypic traits, convergence can also occur at the molecular level where the same amino acid substitution can occur at the same site in orthologous proteins [16–18]. Identifying

47 molecular signatures of convergence is integral to understanding the genetic basis
48 behind the evolution of phenotypes, as well as aiding in recognising adaptive
49 developmental processes across diverse groups of animals [19–21]. Indeed, studies of
50 molecular convergence have provided insights into the origin of traits like
51 echolocation in birds and mammals, toxin resistance in beetles, and carnivory in
52 plants, just to name a few [22–24].

53 As methodologies for detecting molecular convergence events have improved,
54 scientists have been able to re-examine and enhance prior studies to identify or
55 confirm general trends of convergent evolution associated with phenotypes [18,24–27]. A
56 notable recent development is the CSUBST (Combinatorial SUBSTitutions) algorithm
57 introduced by Fukushima and Pollock [28]. This algorithm stands out for its ability to
58 explore a multi-dimensional sequence space comprising extant and ancestral
59 sequences, to identify non-synonymous amino acid substitutions at homologous sites
60 across multiple lineages. It combines a pairwise exhaustive search across all possible
61 branch combinations, followed by a heuristic step that examines branch combinations
62 in more than two independent lineages, without predefining specific foreground
63 branches [28]. This approach enables a data-driven exploration of genomic datasets,
64 uncovering patterns of repeated molecular convergence.

65 In this study, we use this algorithm at a genome-scale to identify patterns of repeated
66 molecular convergence in protein-coding orthologs across 143 teleost fish species. We

67 set out to answer three major questions. First, which genes have evolved convergent
68 substitutions in the sampled teleost species? Second, which processes are these genes
69 involved in? Lastly, what can these genes tell us about the evolution of diversity in
70 form and function in teleosts? We show that throughout the evolutionary history of
71 teleosts, convergent substitutions have occurred in genes that have a role in
72 development and maintaining tissue function. These genes are multi-functional and
73 expressed in multiple cell types across embryonic development in zebrafish. We
74 propose a model which describes how natural selection can favour changes in
75 pleiotropic genes during periods of ecological shifts, leading to molecular
76 convergence.
77

78 Results

79 Data collection and orthology analysis

80 We selected teleost fish species that had chromosomal-level annotated genomes from
81 the National Center for Biotechnology Information (NCBI) [29] and Ensembl genome
82 databases [30]. Our final list of genomes comprised 143 species from 40 orders (most
83 sampled were *Perciformes* with 17 species) covering 85 families (most sampled were
84 *Salmonidae* with 10 species) (Fig 1a, Table S1). We use the spotted gar, *Lepisosteus*
85 *oculatus*, a non-teleost that did not undergo the teleost-specific whole genome
86 duplication [31] as our outgroup. Using OrthoFinder [32] we constructed orthogroups for
87 all species, which resulted in 40,940 orthogroups containing over 4 million protein-

88 coding orthologs. 1723 orthogroups contained all the species. To ensure that we focus
89 on genes which reflect the evolution of teleost fish diversity, we selected orthogroups
90 with at least one species from each of the 40 orders in our dataset. Additionally, to
91 keep computational times reasonable we restricted our analysis to orthogroups
92 containing a maximum of 1500 genes. This resulted in a final dataset of 9273
93 orthogroups with around 2 million orthologous genes (Dataset S1), i.e. more than half
94 of all the genes in the original 40,940 orthogroups. We annotated and functionally
95 characterised the excluded orthogroups, and found that they were mostly associated
96 with immune processes and ion transport, whereas the gene sets we used in our
97 analysis were involved in many diverse processes (Fig S1 and S4). Although processes
98 related to ion transport and immunity are important, our sample encompassed a
99 diverse array of processes, and provided a powerful window into teleost evolution.
100 We processed the coding sequences of these 2 million orthologous genes based on the
101 pre-processing steps recommended by CSUBST (more details in the methods section).
102 We ran the analysis using a rooted time-calibrated species tree obtained from
103 timetree.org [33].
104

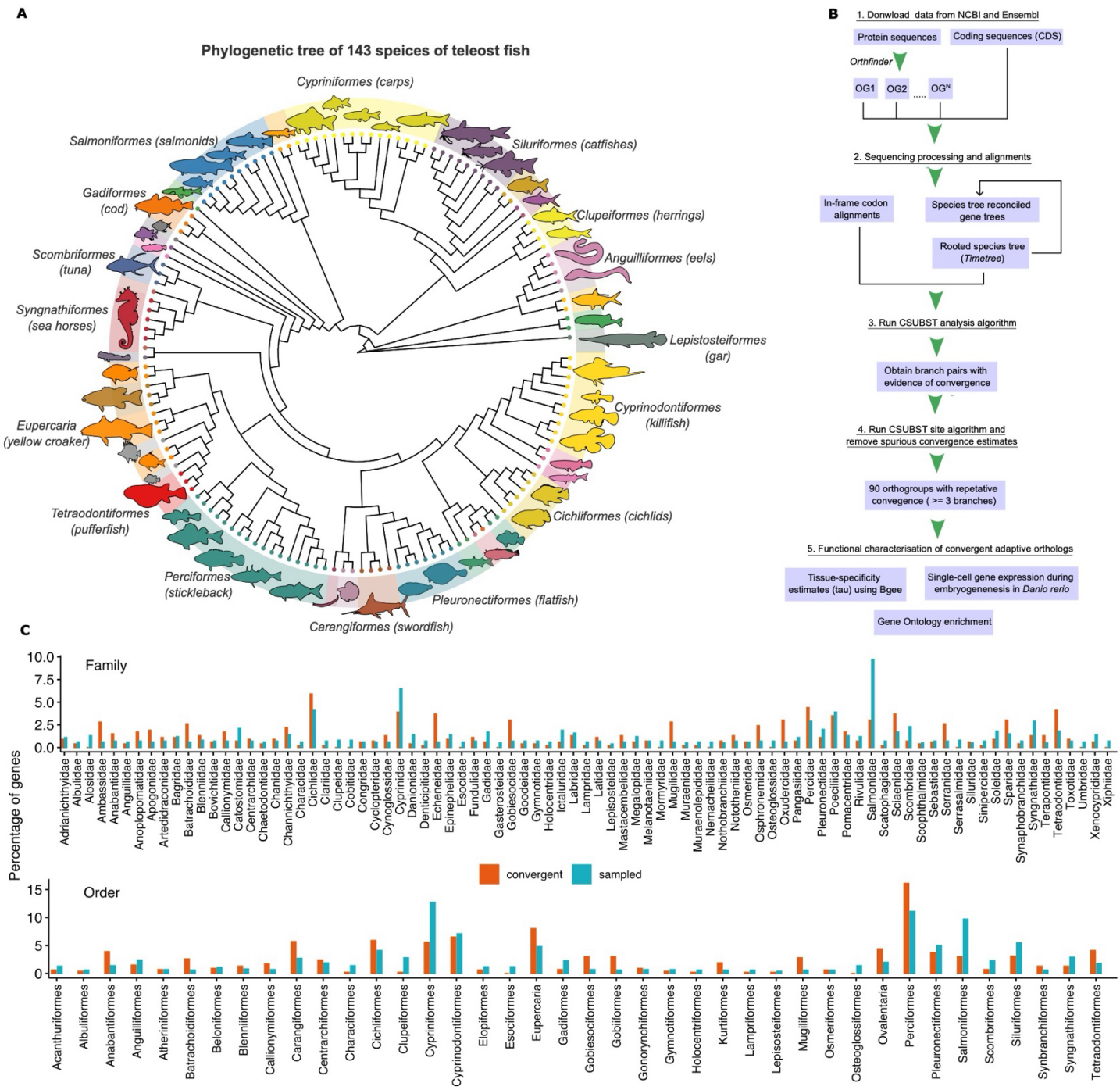

Fig 1: Phylogenetic tree of species, and taxonomic distribution of convergence. (A) Phylogenetic tree of 143 teleost fish species used in the study. The silhouettes are not to scale and are representative of each order. (B) Schematic describing the workflow of this study. (C) Distribution plots showing the proportion of genes that are sampled, (i.e. the genes used as inputs) versus the proportion that are estimated to have convergence substitutions. The top panel shows the distribution grouped by family, while the bottom panel shows the distribution grouped by order. The panels tell us whether some groups of fish are overrepresented in the convergent set than expected by sampling (e.g. Perciformes).

114
115

116 Which genes have evolved convergent substitutions in teleosts?

117 CSUBST draws inspiration from the widely used *dN*/*dS* ratio (parameterised as ω)
118 which measures the rate of protein evolution as a ratio between non-synonymous and
119 synonymous substitution rates [34]. Fukushima and Pollock [28] developed a similar
120 metric, ω_c, which applies to substitutions repeatedly occurring on combinations of
121 separate phylogenetic branches. This metric is designed to assess the relative rates of
122 convergent evolution by comparing the rates of non-synonymous convergence and
123 synonymous convergence across a range of distinct phylogenetic branches. The
124 rationale behind this approach is that truly adaptive functional changes to proteins
125 are most frequently caused by non-synonymous codon substitutions that alter the
126 amino acid sequence. If distinct (distantly related) lineages exhibit the same amino
127 acid change at the same position this is a strong signal for adaptive evolution as the
128 emergence of such repeated substitution by neutral evolution alone is extremely low.
129 The ω_c metric has high sensitivity and error tolerance while suppressing false
130 positives [28]. CSUBST generates estimates for various types of combinatorial
131 substitutions. These include divergent changes, characterised by substitutions from
132 any ancestral amino acid to a different extant amino acid, and convergent changes,
133 which involve substitutions from any ancestral amino acid to a specific amino acid

134 occurring at the same position across multiple lineages. Our study specifically
135 concentrates on convergent changes.
136 We estimated ω_c for all independent branch pairs across 9273 orthogroups and
137 classified branch pairings that jointly satisfied minimum thresholds of 5 for ω_c and for
138 observed non-synonymous convergence (OCN) ($\omega_c \geq 5$ & OCN ≥ 5) as evidence for
139 convergence. We only focussed on convergence that was observed in three or more
140 branch combinations. The OCN values are calculated using an empirical Bayesian
141 method where probabilities of codon substitutions across each branch of a tree are
142 compared, resulting in estimates across all branch combinations. Results could be
143 biassed by the higher power of the OCN convergence estimate for longer genes or
144 orthogroups containing more genes simply because the sequence space is larger. We
145 used a Pearson correlation test to check whether the length or the number of genes in
146 each orthogroup introduced any bias. We found no significant relationship between
147 gene number or gene length and the OCN metric, suggesting that the OCN value
148 strictly depends on sequence variation (Fig S2), and that such bias is not a concern in
149 our data.
150 Next, we used the CSUBST *site* analysis to map convergent amino acid substitutions
151 on predicted protein structures. There can be certain instances of spurious
152 convergence events even when stringent convergence metrics are used. This can often
153 occur due to misalignment of sequences, due to splice variants represented
154 inconsistently among different species [28]. An ad-hoc filtering step was employed to

155 remove spurious convergence estimates (Fig S3). The final selection comprised 89
156 orthogroups, each showing evidence for repetitive convergence in multiple
157 independent branches across teleosts.
158 Most convergence involved three branches, with a few cases involving four, five, or
159 seven branches. The orders *Perciformes*, *Eupercaria*, and *Cyprinodontiformes* had the
160 most convergent genes, as did the families *Cichlidae*, *Percidae*, and *Tetraodontidae* (Fig
161 1C). The taxonomic distribution of convergent genes was different from that simply
162 expected by sampling (Kolmogorov-Smirnov test; *p-value* < 0.05) (FigS4). In terms of
163 ecological characteristics (based on available data from FishBase), the highest
164 proportion of convergent genes were estimated in tropical sea fish (Fig S5). However,
165 we could not infer any overall significant relationship between convergence in a
166 specific orthogroups and ecological features such as climate, ecosystem type, or food
167 items (chi-square test > 0.05).
168
169

Convergence statistics

arity	n	OmegaC	OCN	dNc	dSc
3	85	129.6	9.9	52.1	0.33
4	2	Inf	10.0	3677.3	0
4	1	Inf	5.3	20230.8	0
7	1	1.46e+10	5.7	427e+5	0.29

arity = Number of combinatorial branches

n = Number of Orthogroups

OmegaC = Median ωc value

OCN = Median observed convergence rate

dNc = Median value for the number of non-synonymous convergent substitutions

dSc = Median value for the number of synonymous convergent substitutions

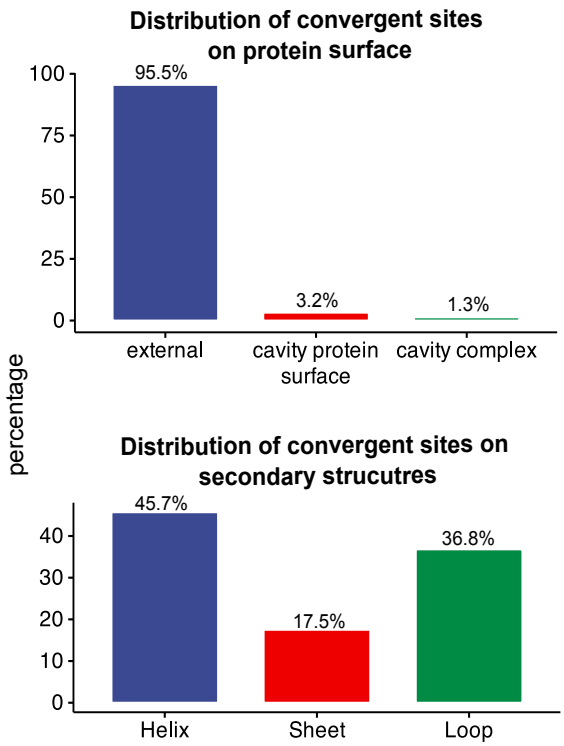

Fig 2: Convergent statistics and structural features of convergent proteins. The table shows the different convergence statistics estimate by CSUBST. We can see that for each branch combination (arity) the values of ω_c are high. This is due to the high number of non-synonymous convergent substitutions (dN_c) compared to synonymous amino acid substitutions (dS_c). The pie charts provide a visualisation of the distribution of convergent sites along the protein. Qualitative estimates show that most of the convergent substitutions are on the external site of the protein suggesting they may play a role in mediating protein-protein interactions.

181

182 Although we mapped convergent substitutions onto the protein structure, we could
183 not perform statistical tests to assess the distribution of the convergent sites due to the
184 heterogeneous nature of the data (see methods). However, qualitatively we observed
185 that most convergent sites were located on the surface, potentially enabling interaction
186 with other proteins (Fig 2). Additionally, in cases where it is possible to identify site-
187 specific functions, we found convergent substitutions at critical functional sites. For
188 instance, in embryonic haemoglobin alpha 5 protein (*hbae5*; OG0000107), we observed
189 both convergent and divergent substitutions located around the protoporphyrin IX
190 containing Fe ligand binding site (Fig 3A). *Hbae5* expression has been linked to
191 increased tolerance to acute hypoxia, therefore, alterations in the binding affinity of
192 the Fe ligand, which plays a role in oxygen transport, might impart increased
193 resistance to hypoxic conditions caused due to environmental change [35]. The
194 convergent substitutions were observed in *Electrophorus electricus* (electric eel),
195 *Larimichthys crocea* (yellow croaker), and *Oreochromis niloticus* (Nile tilapia). Electric eel
196 and Nile tilapia can both resort to air-breathing during times of hypoxia [36,37], while
197 the yellow croaker, which is often found in muddy estuaries and coastal waters, has
198 evolved protective, haemoglobin-containing mucous that aids in oxygen transport
199 during hypoxia and air exposure [38]. The convergent substitutions in *Hbae5* could be
200 linked to these adaptations.

201 We also detected convergent substitutions within the interior of the channel of the
202 barrel-like structure of the solute-carrier protein *slc22a21* (OG0002545), near the 1-
203 methyl-4-phenylpyridinium ligand binding site (Fig 3B). Given the role of *slc22a21* in
204 maintaining the homeostasis of various organic ions, these substitutions could impact
205 processes related to the absorption, distribution, metabolism, and excretion of
206 biomolecules [39]. Members of the *Salmonidae* family, as well as *Periophthalmus*
207 *magnuspinnatus* (mudskippers), and *Hypomesus transpacificus* (Delta smelt), which
208 harbour the substitutions, all experience varying levels of salinity throughout their life
209 cycles [40–42]. The convergent substitutions could have helped these species adapt to
210 changes in salinity.
211 The orthogroup OG0003815 encodes for N-acetylneuraminic acid synthase (*nansa*
212 gene) and plays a crucial role in skeletal development, particularly in the development
213 of the head [43]. The convergent substitutions are present at multiple sites on the protein
214 surface, including a substitution within the protein's interior at a functional ligand-
215 binding site (Fig 3C). Additionally, the orthogroup exhibits a high degree of species
216 diversity, comprising 21 species across 9 orders that exhibit these convergent
217 substitutions. The species within this orthogroup display distinct morphologies,
218 ranging from the fusiform bodies shapes of *Orechromis niloticus*, *Astatotilapia calliptera*
219 (eastern river bream), and *Toxotes jaculatrix* (banded archerfish), to the elongated body
220 shapes of *Salaris fasciatus* (jewelled blenny), *Muraenoplis orangiensis* (Patagonian
221 morray cod), and *Gouania willdenowi* (blunt-snouted clingfish), as well as the flat body

222 shapes of *Soela solea* (common sole) and *Hippoglossus Stenolepis* (Pacific halibut).
223 Interestingly, the remora, which has evolved a flat oval sucking disk (modified dorsal
224 fin) used to attach to larger marine animals like sharks, also harbours substitutions in
225 the *nansa* gene. The occurrence of convergent substitutions appearing in organisms
226 with divergent phenotypes has been observed in marine mammals as well; for
227 instance, substitutions in the *S100A9* and *MGP* genes were linked to both high and
228 low bone density occurring in shallow diving manatees and deep diving dolphins
229 respectively [14]. Therefore, the convergent substitutions in the *nansa* gene likely
230 influenced the sialylation of glycoproteins which altered downstream pathways
231 affecting head and skeletal morphology.
232 ADP-ribosylarginine hydrolase (*adprhl1*; OG0001509) exhibited convergent
233 substitutions at the adenosine-5′-diphosphate ligand site (Fig 3D). This active site is
234 important for the post-translational modifications of proteins [44]. Although the ADP-
235 ribosylarginine hydrolase gene family affects multiple processes, its link to
236 tumorigenesis might be particularly relevant for the *Xiphophorus* (platyfish) species
237 which readily form skin melanomas [44,45].

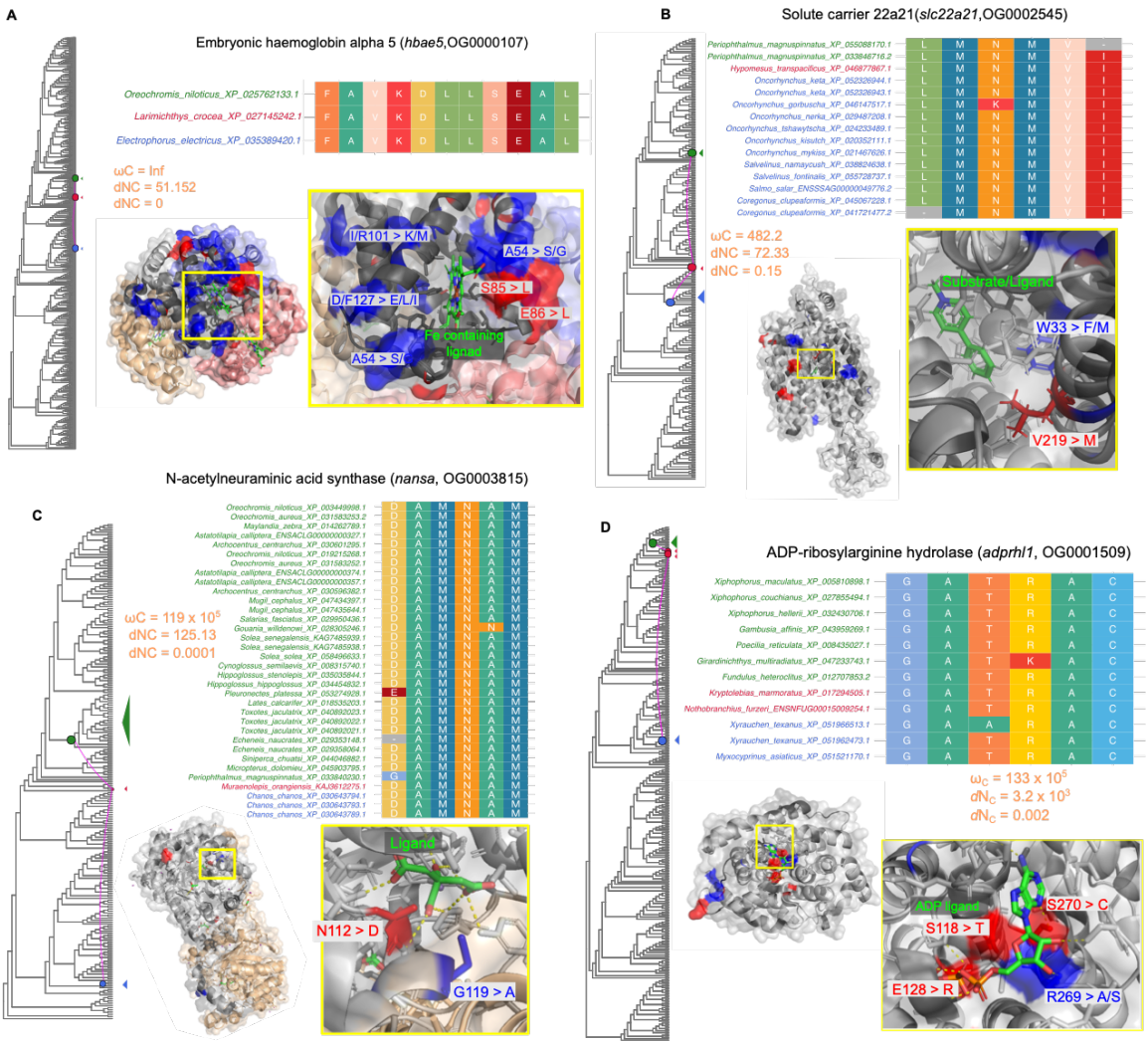

Fig 3: Convergent substitutions mapped onto protein structures. The red sites represent convergent sites while the blue represent divergent sites. The amino acid letter and site followed by the arrow represent the change in sequence at that particular position on the protein structure. For example, in *hbae5*, S85>L means a serine at position 85 on the protein structure changed to a leucine. These sites are localised in important ligand (green) binding sites. The phylogeny is a species-reconciled gene tree. The coloured dots represent the branches in which convergence was found. The species names and colours correspond to the extant species descended from these branches. The alignment shows each of the aligned sites in the species with convergent substitutions. These sites are different from those in other species (full alignment in electronic supplementary material). Convergence metrics for each branch combination are also shown.

Which biological processes are the convergent genes involved in?

We performed a Gene Ontology (GO) enrichment to get a general overview of the biological processes associated with these convergent genes. The analysis revealed a

254 significant enrichment (False Discovery Rate, FDR < 0.05) for terms predominantly
255 linked to biomolecule metabolism, response to hormones or stimuli such as heat, and
256 processes of embryonic development and tissue morphogenesis (Fig S6). To better
257 resolve the function of these convergent genes, particularly in the context of
258 development and tissue formation, we carried out a comprehensive analysis of gene
259 expression across species and developmental transitions.

260 Using gene expression data from the Bgee database [46], we calculated tissue specificity
261 of the convergent orthologs in multiple fish species and across eight different tissue
262 types (brain, eye, heart, liver, muscle, ovary, skin, and testis). Approximately one-
263 third (28 out of 89) of the convergent genes were tissue-specific, with the highest
264 number in the liver (Fig S7). We observed that the proportion of convergent genes that
265 are tissue specific is no different from the proportion of tissue specific genes in our
266 dataset as a whole (Wilcoxon signed rank test *p-value* = 1). Additionally, using Fisher's
267 exact test we found no significant difference in the convergent/non-convergent ratio
268 between tissue-specific and non-tissue-specific genes; in other words, tissue-specific
269 genes aren't more likely to be convergent. This observation was surprising
270 considering that genes that are tissue-specific, i.e. less pleiotropic, are expected to be
271 better targets for adaptation by changes in protein sequence. Of note, our estimates of
272 tissue specificity were limited to adult fish.

273

274 To get an insight into the expression dynamics of the convergent orthologs during
275 development, we analysed single-cell RNA sequencing data for embryonic

276 development in zebrafish (*Danio rerio*) from the Zebrahub atlas [4]. Based on their
277 expression across embryonic development, we observed three groups of convergent
278 genes. First, genes with ubiquitous expression throughout embryonic development
279 and across multiple cell types (Fig 4). These included genes such as *adam17a*, *eef2b*, and
280 *eef1a1|1* (complete list in Table S1). While these genes had an overall ubiquitous
281 expression, variations did exist. For instance, *adam17a* (ADAM metallopeptidase 17a;
282 OG0003642), which is involved in the Notch signalling pathway, showed stable
283 expression in the early stages at 10hpf (hours post-fertilisation), 16hpf, and 2dpf (days
284 post-fertilisation) but declined at 10dpf. For *eef2b* (eukaryotic translation elongation
285 factor 2b; OG0001781) and *eef1a1|1* (eukaryotic translation elongation factor 1 alpha 1;
286 OG0000135) high expression persisted across development, with a slight reduction in
287 eye-related cell types at 10dpf.
288

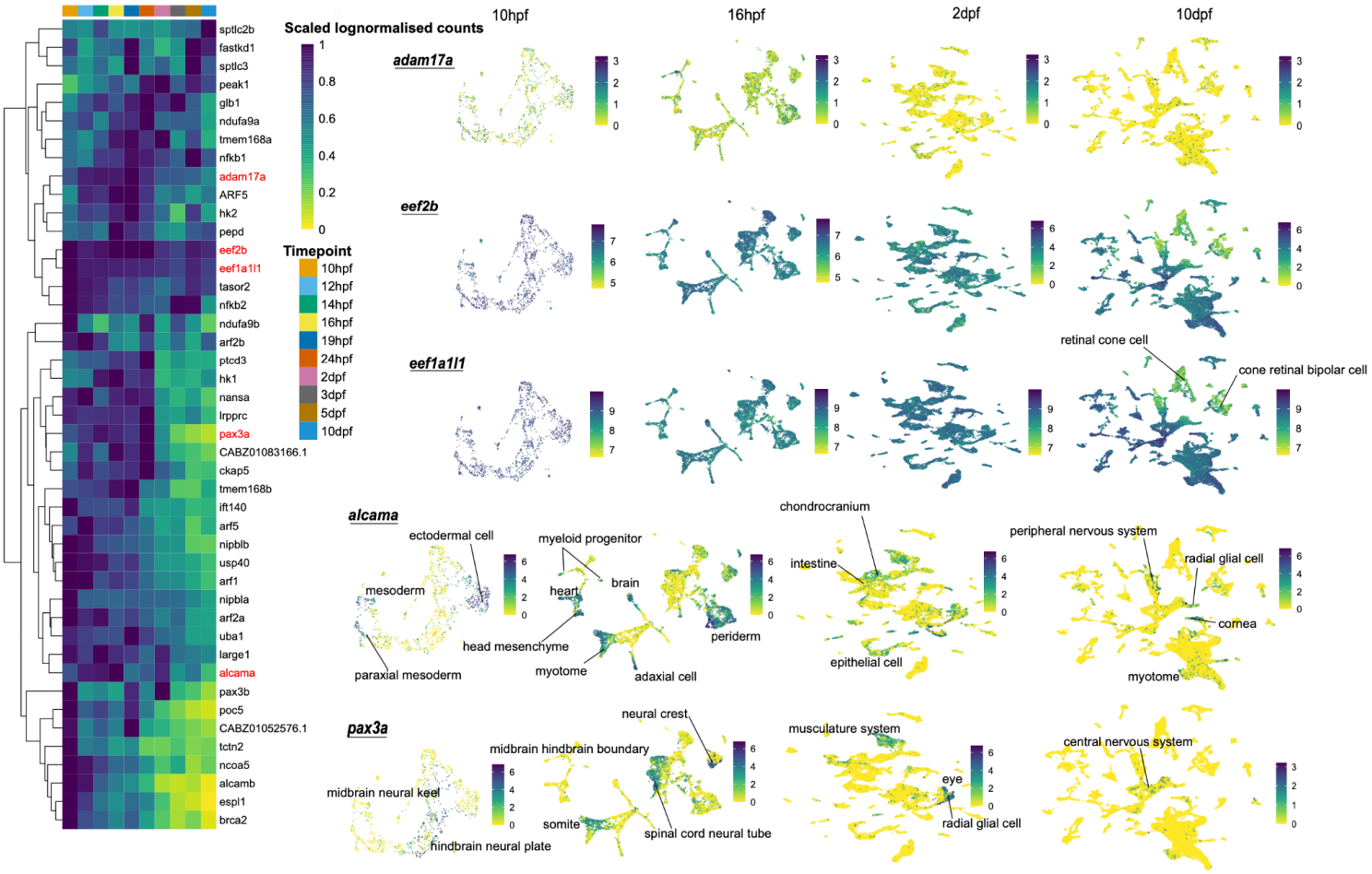

Fig 4: Convergent genes expressed in multiple cell types throughout embryonic development in zebrafish. Single-cell RNA sequencing shows that genes expressed throughout embryonic development can be divided into two groups. Genes like *adam17a*, *eef2b*, and *eef1a1l1* have expression in almost all cell types. While genes like *alcama* and *pax3a* show more restricted expression in specific cell types. A key feature of both these groups is that the genes are expressed in multiple tissue systems (e.g. heart, musculature, nervous system) suggesting that they are pleiotropic.

296

297 The second group comprised genes expressed throughout embryonic development
298 but in specific cell types. Examples include *alcam-a* (activated leukocyte cell adhesion
299 molecule a; OG0001823) and *pax3a* (paired box 3a; OG0000460) (Fig 4). *Alcam-a*,
300 primarily involved in retinal ganglion differentiation, exhibited high expression in cell
301 types related to lens placode, neural crest, radial glial cell, and cornea (Fig 3 10dpf) [47].
302 Between 10hpf and 2dpf, we found that *alcam-a* was also expressed in cell types related
303 to somites, musculature, periderm, heart, and head mesenchyme, corroborating a
304 recent study showing that *alcam-a* is essential for the production of mesenchymal
305 stromal cells from the somites [48]. With high expression in cell types related to somites,
306 spinal cord neural tubes, radial glial cells, neural crest cells, and the central nervous
307 system, the expression of *pax3a* aligned with its known role in eye development and
308 pigmentation [49].
309 The third group of genes were predominantly expressed late in embryonic
310 development between 24hpf and 10dpf (Fig 5). These genes showed more restricted
311 expression patterns than those expressed throughout development. For example,
312 *slc22a21* (OG0002545) was expressed in cells related to the vasculature and excretory
313 systems. In addition to maintaining the homeostasis of various organic ions, the *slc22a*
314 family also shows a strong association with modulating steroid hormone levels,
315 highlighting its potential impact on multiple biological processes [50]. *Hbae5*
316 (OG0000107) showed restricted expression in the hematopoietic system and nucleated

317 erythrocyte cells (Fig 5). In contrast to *pax3a*, *pax7a* (OG0000460) was expressed later
318 in development in multiple cell types related to the neural crest, the brain, and muscles
319 (myoblast and myotome). Finally, genes like *postnb* (periostin, osteoblast specific
320 factor b; OG0002859) and *gstt1b* (glutathione S-transferase theta b, OG0004038) both
321 showed expression in cell types related to muscles, the skin, and blood vasculature.
322 *Postnb* is important for the function of fibroblasts and plays a role in the formation of
323 the extracellular matrix of the skin [51]. *Gstt1b* is regarded as an important player in the
324 detoxification and clearance of intracellular metabolites, which helps the liver detoxify
325 new metabolites during a shift in diets [52].

326 To summarise, we identified genes with convergent substitutions in multiple branches
327 across the phylogeny of teleosts. Most of the convergent substitutions appear to be on
328 the external surface of the proteins while at least some are in critical functional sites of
329 proteins. The convergent genes are also involved in multiple processes ranging from
330 development to metabolism and function in multiple tissue systems. Additionally, the
331 convergent genes as a whole are not tissue specific and are instead expressed in
332 multiple cell types during embryonic development, or in multiple tissues in adults,
333 suggesting that many of these convergent genes are pleiotropic.

334
335

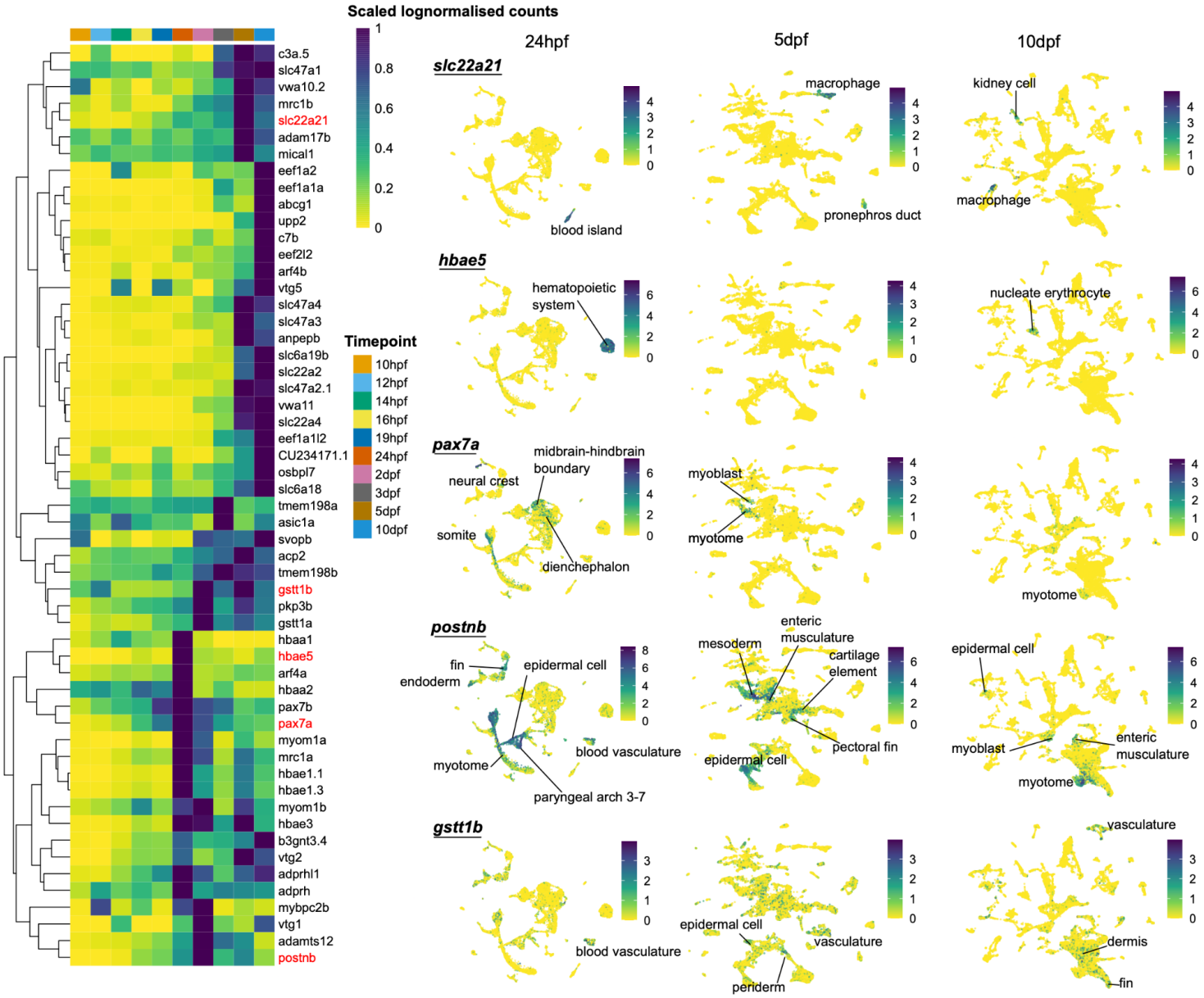

Fig 5: Convergent genes expressed in multiple cell types at specific stages of embryonic development in zebrafish. Compared to the genes expressed throughout embryonic development, the orthologs of these genes also show highly restricted expression patterns in multiple different cell types.

341
342

343 # Discussion

344

345 Convergent amino acid substitutions represent evolutionary endpoints of natural
346 selection. These endpoints in sequence space can be considered as peaks reached by
347 extant species after traversing an adaptive landscape of multidimensional sequence
348 space [53–55]. Since these peaks were reached repeatedly, they can be used to generate
349 hypotheses regarding the mechanisms behind the evolution of phenotypes. In this
350 study, we carried out a genome-scale analysis of convergence across 143 teleost fish
351 species and identified convergent genes representing adaptation throughout
352 macroevolutionary time.

353 **Molecular convergence and evolution in teleosts**

354 One unexpected observation from our study was that most genes with convergent
355 substitutions were expressed in multiple cell types or tissues. Functional data from
356 the ZFIN database [3] also show that mutations in these convergent genes can affect
357 several phenotypes, with a majority of the effects showing a decrease or reduction in
358 affected structures or processes, while a few show an increase (Table S2)(Fig S8). These
359 observations suggest that the convergent genes are mostly pleiotropic. Pleiotropic
360 genes are usually considered poor candidates for the evolution of new functions since
361 their multiple functions are expected to increase genetic constraints [56]. However,

362 studies in several model systems as well as natural populations have shown that the
363 idea of pleiotropy constraining adaptation needs to be reconsidered [57,58]. For example,
364 a vital gene responsible for the adaptation to freshwater habitats in sticklebacks,
365 Ectodysplasin (*Eda*), is highly pleiotropic and responsible for ectodermal appendage
366 formation, sensory system patterning, and schooling behaviour [59,60]. Intermediate
367 levels of pleiotropy might even enhance adaptive potential by causing greater fitness
368 gain and increasing the fixation of beneficial alleles [61].

369 We describe a model which we call the 'pleiotropic release' model to illustrate how
370 pleiotropic genes underpin adaptation and phenotypic convergence across diverse
371 taxa. Consider a multidimensional phenotypic adaptive landscape where two species
372 reside at two distinct adaptive peaks (Fig 6). An environmental shift introduces a new
373 common adaptive optimum such that both species now occupy regions of low-fitness.
374 To survive, the species must navigate out of these low-fitness regions and move
375 towards the new optimum. There are two possible ways to achieve this. Adaptive
376 variation can be fixed in genes that affect one trait at a time causing incremental
377 changes in phenotype towards the optima (orange arrows). These phenotypic changes
378 will be of low magnitude because allometry generally prevents large changes in one
379 trait without proportional changes in another (e.g. height cannot change without
380 affecting weight) [62]. In the other scenario, adaptive variation can accumulate on
381 pleiotropic genes which have a large cumulative effect on both traits enabling the
382 traversal of large distances on the adaptive landscape. If selection favours changes in

the pleiotropic genes, the species can escape (be released) from the low fitness regions faster and move towards the new shared optima. Since the environmental shift produced a single, shared optima the species which started with different phenotypes will converge to a similar one. And if a few pleiotropic genes were selected, they have more chances of having fixed convergent substitutions. It is important to note that the pleiotropic release model does not suggest a lack of cost on multiple traits. Adaptation along the axes of pleiotropic genes would likely have costs on multiple traits, but the relative increase in fitness outweighs these costs.

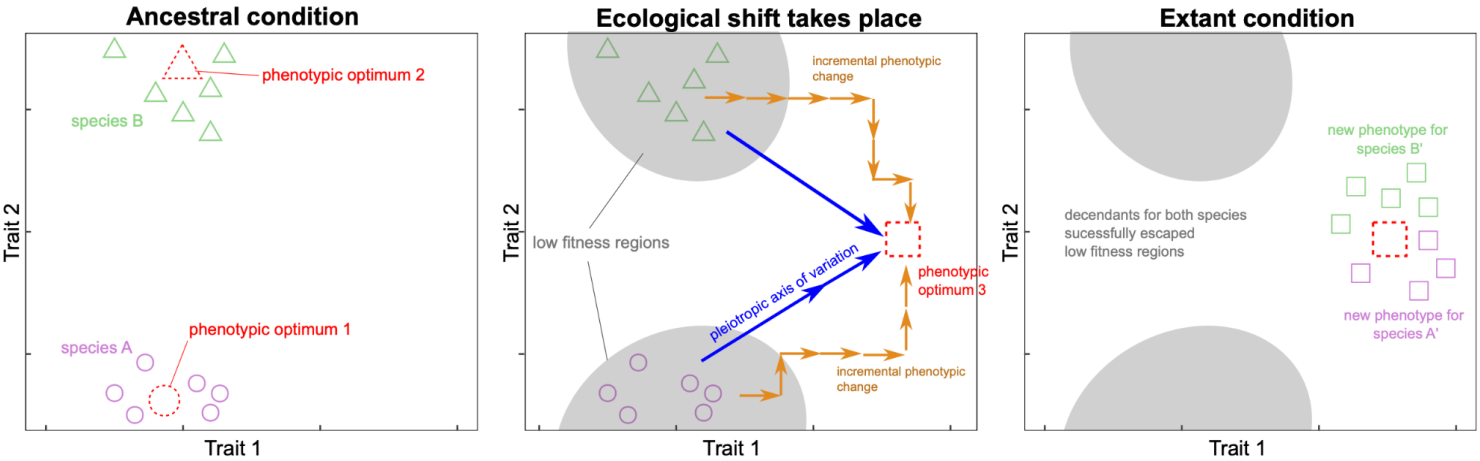

Fig 6: The pleiotropic release model. In the ancestral condition two species occupy separate phenotypic optima in an adaptive landscape. Following an ecological shift, the optima changes and the species now occupy regions of low fitness (grey). Escape from the low fitness regions can happen through either incremental change in low, single effect genes (orange) or through large effect pleiotropic genes (blue). Changes along the pleiotropic axis of variation will be favoured as they will enable the species to escape the low fitness regions faster. Overtime, species will evolve a convergent phenotype and occupy the new optima in the extant condition.

403

404 We can use this general model to explain the patterns that we observe in our study.
405 For example, an environmental shift might have resulted in the ancestors of
406 salmonids, mudskippers, and delta smelt to each experience varying levels of salinity.
407 The varying levels of salinity could also impact the available food resources as food
408 webs differ between marine and freshwater habitats. To adapt to the new
409 environmental conditions changes in the *slc22a21* gene might prove particularly
410 beneficial. Along with maintaining the homeostasis of organic ions between tissues
411 and the interfacing body fluids, this gene family can impact tissue systems such as the
412 liver, kidneys, skeletal system, and central nervous system [63]. Additionally, they also
413 facilitate inter-organism communication between the gut microbiome and the other
414 tissue systems [63]. Therefore, adaptive variations in *slc22a21* gene could cumulatively
415 enhance liver detoxification, improve ion reabsorption by the kidneys, and strengthen
416 signalling between the gut microbiome and the central nervous system, all of which
417 would be particularly beneficial during ecological shifts involving changes in salinity
418 and dietary changes as a result of altered nutrient availability. Considering that
419 teleosts occupy diverse niches in different environments, evolution through changes
420 in pleiotropic genes might have been particularly relevant during the initial periods
421 of colonisation into new niches.

422
423

424 **Evolution using pleiotropic genes**

425 In a recent experimental evolution study in yeast, Kinsler *et. al.* [64] showed that large
426 effect adaptive mutations are generally pleiotropic, improving both respiration and
427 fermentation, and tend to occur during the early stages of evolution. Although they
428 observed a general shift from pleiotropic to modular adaptation, they did observe
429 some strongly adaptive yeast clones that continued to improve in both respiration and
430 fermentation causing them to achieve high levels of fitness. This highlights that long
431 term evolution via pleiotropic adaptation, while rare, is possible. The authors also
432 discuss observations in multiple studies which show that when yeast are evolved in a
433 glucose-limited environment, the first adaptive mutations are always in pleiotropic
434 genes that enhance both respiration and fermentation. Therefore, pleiotropic release
435 via evolution along the axis of pleiotropic genes might be a general feature of
436 evolution during the initial periods following ecological shifts.

437 Genetic redundancy brought about by single gene or genome duplications, and
438 modularity, are two possible mechanisms that can overcome the adaptation-impeding
439 effects of pleiotropy [58]. In our analysis, the number of gene copies is generally higher
440 in the convergent orthogroups than in the non-convergent ones (one sided
441 Kolmogorov-Smirnov test *p-value* < 0.05) (Fig S9). Additionally, some genes show clear
442 differences in the expression of their paralogs (Fig S9). The presence of multiple copies
443 coupled with the divergence in expression patterns between paralogs could enable
444 partitioning of effects, allowing for variation among paralogs to persist non-

445 adaptively as cryptic divergent variations (hereafter cryptic divergence). Cryptic
446 divergence is the genetic difference between paralogs variations that have negligible
447 or no phenotypic effects, but during periods of environmental shifts, produce
448 heritable phenotypic variation. It is analogous to cryptic variation among alleles [65].
449 The developmental impacts of cryptic variation in pleiotropic genes have been
450 experimentally demonstrated in models like *Drosophila melanogaster*, *C. elegans*,
451 *Arabidopsis thaliana*, and zebrafish [66–69]. *In vivo* studies in natural environments, such
452 as those in sticklebacks and cavefish, have highlighted the critical role of ancestral
453 cryptic variation in facilitating the adaptation to ecological changes [70,71]. Additionally,
454 experimental evidence indicates that mutations in protein-coding sequences of
455 enzymes can generate cryptic variation which maintains the enzyme's native function
456 on its ancestral substrate while enhancing its adaptive potential in a novel
457 substrate [72,73]. Therefore, a change in environment can cause the cryptic variation to
458 manifest, leading to phenotypes with enhanced fitness.
459 The functional implications of the convergent substitutions can also be seen at the level
460 of the protein structure. We observed a qualitative trend that a high proportion of
461 convergent substitution were present on protein surfaces. The accumulation of
462 substitutions on protein surfaces could influence protein-protein interactions,
463 enhancing specificity and efficiency of biological processes. Overtime, the effect of
464 these substitutions can further optimise protein-protein interactions, fine tuning their
465 activity and making them highly specialised for different cellular contexts. In addition

466 to enhancing protein-protein interactions, changes at the amino acid level can also
467 cause changes in protein conformation and tertiary structure, enabling proteins to
468 acquire new functions without affecting existing catalytic sites [74]. This can allow
469 proteins to have multifaceted roles, affecting different processes in different cell types
470 without cross disruption across cell types. These features highlight the importance of
471 considering the cellular basis of traits along with the genetic basis.
472

473 # Conclusion

474 In our study, we employed a phylogenetic approach to obtain data-driven estimates
475 of molecular convergence depicting the axes of adaptation in the teleost lineage. Our
476 findings reveal that genes with convergent adaptive substitutions have a role in the
477 development and functioning of a diverse array of tissue systems. A notable
478 characteristic of many of these genes is their versatile and dynamic expression across
479 diverse cell types, signifying their pleiotropic nature. We discuss the pleiotropic
480 release model which describes how adaptive variations accumuling in pleiotropic
481 genes can cause large phenotypic changes enabling species to escape regions of low
482 fitness during periods of ecological shifts. Although the model establishes a theoretical
483 link between pleiotropy and convergence, the extent to which this model represents a
484 general evolutionary mechanism still needs to be tested, opening avenues for future
485 research.
486
487

489 # Materials and Methods

490 ## Data collection and estimating orthologs

491 In our study, we obtained chromosomal-level annotated teleost fish genomes from the
492 NCBI genomes [29] and Ensembl genome databases [30] (Table S3). Using an initial list of
493 species we obtained taxonomic information from FishBase (Froese, R. and D. Pauly.
494 Editors. 2023. www.fishbase.org,) and a time-calibrated, rooted species tree from
495 timetree.org [33]. We only included species that were present in the species tree and
496 downloaded their protein and coding sequences (CDS files). Using the protein
497 sequences from a final set of 143 species, we constructed orthologs using OrthoFinder
498 [32], resulting in 40,940 orthogroups containing 98.2% of all the genes (Table S4). We
499 focussed on orthogroups with at least one species from each of the 40 orders in our
500 dataset. Additionally, to keep computational times reasonable we restricted our
501 analysis to orthogroups containing a maximum of 1000 genes. This resulted in a final
502 dataset of 9224 orthogroups with around 2 million orthologous genes (Dataset S1 in
503 electronic supplementary material).

504
505

506 ## Processing of the orthologs

507 Prior to running the CSUBST algorithm, we executed a series of preprocessing steps,
508 closely following the guidelines outlined in the CSUBST wiki
509 (https://github.com/kfuku52/csubst/wiki/Preparing-CSUBST-inputs). Briefly, we

510 carried out the following steps. First, we used the CDSKIT v0.91 tool
511 (https://github.com/kfuku52/cdskit) to isolate the longest isoform for each gene and
512 make the sequences in-frame. Next, using TRANSeq v6.6.0 [79] we translated the coding
513 sequences into protein sequences followed by alignment using MAFFT v7.481 with
514 the 'auto' option [80]. We then performed a translation alignment using TRANALIGN
515 v6.6.0 [79] and trimmed poorly aligned sequences using TRIMAL v1.4.1[81] with the
516 '*ignorestopcodon*' and '*automated1*' options. Finally, we masked and removed
517 ambiguous sites in the alignment using the '*mask*' and '*hammer*' functions of CDSKIT
518 v0.91.
519 We used IQ-TREE v2.0.3 [82] with the MPF model, and 1000 ultra-fast optimized
520 bootstrapping replicates ('*bnni*' option), to construct gene trees. We rooted and
521 reconciled the gene trees with the species trees using a combination of GeneRax v1.1.0
522 [83] and NWKIT v0.10.0 (https://github.com/kfuku52/nwkit).
523 Detailed scripts and commands utilised throughout these steps can be found within
524 the 'Scripts' folder in the electronic supplementary material. We also provide a conda
525 environment YAML file that can be used to recreate the environment for the current
526 pre-processing step as well as the following analysis step.
527

528 Convergence analysis using CSUBST

529 CSUSBT employs a metric similar to the popular *dN/dS* ratio (parameterised as ω)
530 which represents the rate of protein evolution as a ratio between non-synonymous
531 and synonymous substitution rates [34]. Distinct from ω, the ωc parameter captures
532 repeated substitutions across various combinations of phylogenetic branches. These
533 combinations range from a minimum of two branches, indicating pairwise
534 substitutions, to a maximum of ten branches. The software quantifies these
535 combinations as *arity (K),* where *K*=2 represents pairwise comparisons, and *K*=3
536 signifies a three-way branch comparison. By estimating all possible combinations of
537 amino acid substitution (across multiple branch combinations) occurring in a multi-
538 dimensional sequence space, ωc measures how much proteins in different lineages
539 have moved towards the same evolutionary endpoints in this sequence space.
540 CSUBST reports ωc statistics for nine types of combinatorial substitution [28]. We focus
541 primarily on *omegaCany2spe* - denoting combinatorial substitution from any ancestral
542 codon to a specific descendant codon, namely, convergent substitutions - and
543 *OCNany2spe*, which denotes the observed rate for non-synonymous convergence. We
544 ran the CSUSBT algorithm using the 'CSUBST *analyze*' command with five threads
545 and conducted an exhaustive search for a max *K*=10. The output of this command can
546 be found in Dataset S3. Orthologs were categorised based on the presence of
547 convergence across these branch combinations and further analyzed using the
548 '*CSUBST site*' command (input files in Dataset S4) to map convergent amino acids onto

549 protein structures. This mapping facilitates the visualisation of convergent amino acid
550 sites, particularly their proximity to functional domains within the protein. After
551 excluding cases of spurious convergence (Fig S3) we identified 90 orthogroups that
552 showed evidence for reliable convergence. Output in Dataset S5, with examples of
553 spurious convergence in Dataset S6.

554

555 Protein structure analysis

556 The '*CSUBST site*' command obtains protein structures by searching the RCSB PDB
557 database (RCSB.org) [84]. If no experimentally determined structures are found,
558 CSUBST uses NCBI's QBLAST [29] search against the UNIPROT database [85] to obtain a
559 .pdb file from the AlphaFold Protein Structure Database [86]. As a result, the structures
560 obtained are often heterogeneous, comprising X-ray crystallography structures of
561 protein in complex with a ligand, non-complexed protein in crystal, or
562 computationally estimated structures from AlphaFold. This heterogeneity makes
563 statistical comparisons of protein sites difficult. Nonetheless, we applied a systematic
564 selection method to obtain as objective a comparison as possible. We used Pymol [87]
565 version 2.4.5 to systematically classify the residues. For each Pymol session,
566 convergent residues were selected based on their colour (red), surface residues were
567 determined using a cutoff of 1.4 $Å^2$ to determine their exposure to the solvent.

568 Residues with a distance equal to or less than 4 Å from the ligand were classified as in
569 the vicinity of the ligand. Buried residues were those that did not have access to the
570 surface or internal cavity of the protein. The secondary structures of helices, sheets,
571 and loops are selected directly using the keywords provided in Pymol. For all these
572 selections, a second b-factor parameter > 0 was added to eliminate residues that were
573 present in the sequence but not built into the structure. From this set, we can then
574 cross-reference the selections to determine the desired list and number of residues.
575 These results are in Table S5.

576 GO term annotations and enrichment

577 We used DeepGOPlus v1.0.15 [88] to annotate all the genes (~4 million) in our dataset
578 with gene ontology (GO) terms. DeepGOPlus integrates a convoluted neural network
579 along with sequence similarity scores of protein sequences to annotate GO terms. In
580 our analysis, we use a prediction threshold of 0.3 [88]. Using the GOstats and GO.db
581 Bioconductor [89] packages we performed GO enrichment for GO terms related to
582 Biological Processes. In this analysis, convergent orthologs constituted the test set,
583 while the remainder served as the background universe. Enriched terms were
584 identified using a false discovery rate (FDR) cut-off of <0.05.

585
586

587 Bgee data and estimating Tau

588 To gain deeper insight into the functional relevance of the convergent genes we used
589 data from the Bgee database [46] and studied expression across eleven species
590 (*Lepisosteus oculatus*, *Danio rerio*, *Astyanax mexicanus*, *Esox lucius*, *Salmo salar*, *Gadus*
591 *morhua*, *Oryzias latipes*, *Astatotilapia calliptera*, *Neolamprologus brichardi*, *Scophthalmus*
592 *maximus*, and *Nothobranchius furzeri*) and eight tissues. Bgee is based exclusively on
593 curated healthy wild-type expression data, which is processed to produce calls of
594 presence/absence of expression which is integrated with other information like gene
595 orthology and organ homology allowing for comprehensive comparison of gene
596 expression across species and tissues. We collected data for the brain, eye, heart, liver,
597 muscle tissue, ovary, and testis. The Bgee data sometimes includes expression levels
598 of tissues from different experiments. We checked for potential batch effects in tissue
599 expression between different experiments and found evidence only in *Lepisosteus*
600 *oculatus* (Fig S10). We corrected this using the ComBat_seq function from the sva R
601 package [90] followed by quantile normalisation, an approach which was successfully
602 used previously [91]. To calculate tissue specificity, we used the Tau metric because of
603 its efficiency in recovering biological signals [92]. We classified a gene as tissue-specific
604 if it had a tau > 0.8 and its expression in the target tissue was greater than the sum of
605 its expression in other tissues. Since we are comparing tissue specificity across
606 different species, we identified genes that are tissue-specific in the same tissue across

607 all the species sampled. These stringent criteria ensured we captured robust signals
608 for tissue specificity. Around one-third of these genes had signals of convergent
609 evolution with the largest being in the liver (Fig S4).

610 Analysis of single-cell RNA-seq from Zebrahub

611 We integrated several types of data to obtain greater functional insights into the
612 convergent genes. We used single-cell RNA-seq (scRNAseq) from the Zebrahub
613 consortium [4]. This data comprises a high-resolution scRNAseq atlas of zebrafish
614 development achieved by sequencing individual embryos across ten developmental
615 stages [4]. The dataset contained 120,444 cells and a total of 529 cell clusters which were
616 annotated based on published literature and the ZFIN database [3]. The entire dataset
617 is available at https://zebrahub.ds.czbiohub.org/. The consortium provides h5ad
618 objects with quality control, clustering, and uniform manifold projections (UMAP)
619 (https://github.com/czbiohub-sf/zebrahub_analysis). We log-normalise this processed
620 data using the scater R package [93] and use it for our study. To visualise the expression
621 of the convergent genes across embryonic development we calculated the mean
622 expression of each gene at each time point and constructed a scaled (max scaled to 1)
623 heatmap using the scuttle [93] and dittoSeq [94] R packages. We visualised heatmaps using
624 the plotReducedDim function from the scater R package [93]. All UMAP plots with cell
625 type and gene expression at each time point are available in Dataset S9.

Data availability

An electronic report containing output of R code and major analyses in this study can be found at:

https://agneeshbarua.github.io/Teleost_convergence/

References

1. Streelman, J. T., Peichel, C. L. & Parichy, D. M. Developmental Genetics of Adaptation in Fishes: The Case for Novelty. *Annu. Rev. Ecol. Evol. Syst.* **38**, 655–681 (2007).
2. McMenamin, S. K. & Parichy, D. M. Metamorphosis in teleosts. *Curr. Top. Dev. Biol.* **103**, 127–165 (2013).
3. Bradford, Y. M. *et al.* Zebrafish information network, the knowledgebase for Danio rerio research. *Genetics* **220**, (2022).
4. Lange, M. *et al.* Zebrahub – Multimodal Zebrafish Developmental Atlas Reveals the State-Transition Dynamics of Late-Vertebrate Pluripotent Axial Progenitors. *bioRxiv* 2023.03.06.531398 (2023) doi:10.1101/2023.03.06.531398.
5. Ravi, V. & Venkatesh, B. The Divergent Genomes of Teleosts. *Annu Rev Anim Biosci* **6**, 47–68 (2018).
6. Gagnaire, P.-A. Comparative genomics approach to evolutionary process connectivity. *Evol. Appl.* **13**, 1320–1334 (2020).
7. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).
8. Lewin, H. A. *et al.* The Earth BioGenome Project 2020: Starting the clock. *Proc. Natl. Acad. Sci. U. S. A.* **119**, (2022).
9. Christmas, M. J. *et al.* Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**, eabn3943 (2023).
10. Martin, A. & Orgogozo, V. The Loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution* **67**, 1235–1250 (2013).
11. Zhang, G. *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
12. Partha, R. *et al.* Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *Elife* **6**, (2017).
13. Allio, R. *et al.* Genome-wide macroevolutionary signatures of key innovations in

butterflies colonizing new host plants. *Nat. Commun.* **12**, 354 (2021).
14. Foote, A. D. *et al.* Convergent evolution of the genomes of marine mammals. *Nat. Genet.* **47**, 272–275 (2015).
15. Rosenblum, E. B., Parent, C. E. & Brandt, E. E. The molecular basis of phenotypic convergence. *Annu. Rev. Ecol. Evol. Syst.* **45**, 203–226 (2014).
16. Christin, P.-A., Weinreich, D. M. & Besnard, G. Causes and evolutionary significance of genetic convergence. *Trends Genet.* **26**, 400–405 (2010).
17. Storz, J. F. Causes of molecular convergence and parallelism in protein evolution. *Nat. Rev. Genet.* **17**, 239–250 (2016).
18. Rey, C. *et al.* Detecting adaptive convergent amino acid evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20180234 (2019).
19. Wake, D. B. Homoplasy: The Result of Natural Selection, or Evidence of Design Limitations? *Am. Nat.* **138**, 543–567 (1991).
20. Steppan, S. J., Phillips, P. C. & Houle, D. Comparative quantitative genetics: evolution of the G matrix. *Trends Ecol. Evol.* **17**, 320–327 (2002).
21. Losos, J. B. Convergence, adaptation, and constraint. *Evolution* **65**, 1827–1840 (2011).
22. 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. U. S. A.* **109**, 13040–13045 (2012).
23. Fukushima, K. *et al.* Genome of the pitcher plant Cephalotus reveals genetic changes associated with carnivory. *Nat Ecol Evol* **1**, 59 (2017).
24. Sadanandan, K. R. *et al.* Convergence in hearing-related genes between echolocating birds and mammals. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2307340120 (2023).
25. 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).
26. Mendes, F. K., Hahn, Y. & Hahn, M. W. Gene Tree Discordance Can Generate Patterns of Diminishing Convergence over Time. *Mol. Biol. Evol.* **33**, 3299–3307 (2016).
27. Cerca, J. Understanding natural selection and similarity: Convergent, parallel and repeated evolution. *Mol. Ecol.* **32**, 5451–5462 (2023).
28. Fukushima, K. & Pollock, D. D. Detecting macroevolutionary genotype-phenotype associations using error-corrected rates of protein convergence_SUPP. *Nat Ecol Evol* **7**, 155–170 (2023).
29. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).
30. Martin, F. J. *et al.* Ensembl 2023. *Nucleic Acids Res.* **51**, D933–D941 (2023).
31. Braasch, I. *et al.* The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat. Genet.* **48**, 427–437 (2016).
32. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
33. Kumar, S. *et al.* TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol. Biol. Evol.* **39**, (2022).
34. Yang, Z. *Molecular Evolution: A Statistical Approach*. (OUP Oxford, 2014).
35. Ragsdale, A. *et al.* Paternal hypoxia exposure primes offspring for increased hypoxia resistance. *BMC Biol.* **20**, 185 (2022).
36. Farber, J. & Rahn, H. Gas exchange between air and water and the ventilation pattern in the electric eel. *Respir. Physiol.* **9**, 151–161 (1970).

37. Obirikorang, K. A., Acheampong, J. N., Duodu, C. P. & Skov, P. V. Growth, metabolism and respiration in Nile tilapia (Oreochromis niloticus) exposed to chronic or periodic hypoxia. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **248**, 110768 (2020).
38. Ao, J. *et al.* Genome sequencing of the perciform fish Larimichthys crocea provides insights into molecular and genetic mechanisms of stress adaptation. *PLoS Genet.* **11**, e1005118 (2015).
39. Mihaljevic, I., Popovic, M., Zaja, R. & Smital, T. Phylogenetic, syntenic, and tissue expression analysis of slc22 genes in zebrafish (Danio rerio). *BMC Genomics* **17**, 626 (2016).
40. Parry, G. The development of salinity tolerance in the salmon, *Salmo Salar* (L.) and some related species. *J. Exp. Biol.* **37**, 425–434 (1960).
41. You, X. *et al.* Mudskippers and Their Genetic Adaptations to an Amphibious Lifestyle. *Animals (Basel)* **8**, (2018).
42. Hung, T.-C. *et al.* Temperature and salinity preferences of endangered Delta Smelt (Hypomesus transpacificus, Actinopterygii, Osmeridae). *Sci. Rep.* **12**, 16558 (2022).
43. van Karnebeek, C. D. M. *et al.* NANS-mediated synthesis of sialic acid is required for brain and skeletal development. *Nat. Genet.* **48**, 777–784 (2016).
44. Mashimo, M., Kato, J. & Moss, J. Structure and function of the ARH family of ADP-ribosyl-acceptor hydrolases. *DNA Repair* **23**, 88–94 (2014).
45. Kozlov, A. P. Chapter 7 - Tumors that have Played a Role in Evolution. in *Evolution by Tumor Neofunctionalization* (ed. Kozlov, A. P.) 67–73 (Academic Press, San Diego, 2014).
46. Bastian, F. B. *et al.* The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Res.* **49**, D831–D847 (2021).
47. Diekmann, H. & Stuermer, C. A. O. Zebrafish neurolin-a and -b, orthologs of ALCAM, are involved in retinal ganglion cell differentiation and retinal axon pathfinding. *J. Comp. Neurol.* **513**, 38–50 (2009).
48. Murayama, E., Vivier, C., Schmidt, A. & Herbomel, P. Alcam-a and Pdgfr-α are essential for the development of sclerotome-derived stromal cells that support hematopoiesis. *Nat. Commun.* **14**, 1171 (2023).
49. Nelms BL, L. P. A. Chapter 9, Pax Genes. in *Transcriptional Control of Neural Crest Development.* (San Rafael (CA): Morgan & Claypool Life Sciences, 2010).
50. Yee, S. W. *et al.* Unraveling the functional role of the orphan solute carrier, SLC22A24 in the transport of steroid conjugates through metabolomic and genome-wide association studies. *PLoS Genet.* **15**, e1008208 (2019).
51. Snider, P. *et al.* Origin of cardiac fibroblasts and the role of periostin. *Circ. Res.* **105**, 934–947 (2009).
52. Tierbach, A., Groh, K. J., Schönenberger, R., Schirmer, K. & Suter, M. J.-F. Glutathione S-Transferase Protein Expression in Different Life Stages of Zebrafish (Danio rerio). *Toxicol. Sci.* **162**, 702–712 (2018).
53. Smith, J. M. Natural selection and the concept of a protein space. *Nature* **225**, 563–564 (1970).
54. Arnold, S. J., Pfrender, M. E. & Jones, A. G. The adaptive landscape as a conceptual bridge between micro- and macroevolution. *Genetica* **112-113**, 9–32 (2001).
55. Pollock, D. D. & Pollard, S. T. Parallel and convergent molecular evolution. in *Encyclopedia of Evolutionary Biology* (ed. Kliman, R. M.) 206–211 (Academic Press, 2016).
56. Orr, H. A. Adaptation and the cost of complexity. *Evolution* **54**, 13–20 (2000).
57. Paaby, A. B. & Rockman, M. V. The many faces of pleiotropy. *Trends Genet.* **29**, 66–73 (2013).

58. Zhang, J. Patterns and Evolutionary Consequences of Pleiotropy. *Annu. Rev. Ecol. Evol. Syst.* **54**, 1–19 (2023).
59. Mills, M. G., Greenwood, A. K. & Peichel, C. L. Pleiotropic effects of a single gene on skeletal development and sensory system patterning in sticklebacks. *Evodevo* **5**, 5 (2014).
60. Greenwood, A. K., Mills, M. G., Wark, A. R., Archambeault, S. L. & Peichel, C. L. Evolution of Schooling Behavior in Threespine Sticklebacks Is Shaped by the Eda Gene. *Genetics* **203**, 677–681 (2016).
61. Wang, Z., Liao, B.-Y. & Zhang, J. Genomic patterns of pleiotropy and the evolution of complexity. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 18034–18039 (2010).
62. West, G. B., Brown, J. H. & Enquist, B. J. A general model for the origin of allometric scaling laws in biology. *Science* **276**, 122–126 (1997).
63. Nigam, S. K. The SLC22 Transporter Family: A Paradigm for the Impact of Drug Transporters on Metabolic Pathways, Signaling, and Disease. *Annu. Rev. Pharmacol. Toxicol.* **58**, 663–687 (2018).
64. Grant Kinsler, Yuping Li, Gavin Sherlock, Dmitri A. Petrov. A shift from pleiotropic to modular adaptation revealed by a high-resolution two-step adaptive walk. *bioRxiv* (2024).
65. Paaby, A. B. & Rockman, M. V. Cryptic genetic variation: evolution's hidden substrate. *Nat. Rev. Genet.* **15**, 247–258 (2014).
66. Rutherford, S. L. & Lindquist, S. Hsp90 as a capacitor for morphological evolution. *Nature* **396**, 336–342 (1998).
67. Queitsch, C., Sangster, T. A. & Lindquist, S. Hsp90 as a capacitor of phenotypic variation. *Nature* **417**, 618–624 (2002).
68. Yeyati, P. L., Bancewicz, R. M., Maule, J. & van Heyningen, V. Hsp90 selectively modulates phenotype in vertebrate development. *PLoS Genet.* **3**, e43 (2007).
69. Duveau, F. & Félix, M.-A. Role of pleiotropy in the evolution of a cryptic developmental variation in Caenorhabditis elegans. *PLoS Biol.* **10**, e1001230 (2012).
70. McGuigan, K., Nishimura, N., Currey, M., Hurwit, D. & Cresko, W. A. Cryptic genetic variation and body size evolution in threespine stickleback. *Evolution* **65**, 1203–1211 (2011).
71. Rohner, N. *et al.* Cryptic variation in morphological evolution: HSP90 as a capacitor for loss of eyes in cavefish. *Science* **342**, 1372–1375 (2013).
72. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 5869–5874 (2006).
73. Hayden, E. J., Ferrada, E. & Wagner, A. Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature* **474**, 92–95 (2011).
74. Kozome, D., Sljoka, A. & Laurino, P. Remote loop evolution reveals a complex biological function for chitinase enzymes beyond the active site. *Nat. Commun.* **15**, 3227 (2024).
75. Garamszegi, L. Z. *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*. (Springer Berlin Heidelberg).
76. Uyeda, J. C., Zenil-Ferguson, R. & Pennell, M. W. Rethinking phylogenetic comparative methods. *Syst. Biol.* **67**, 1091–1109 (2018).
77. Dunn, C. W., Zapata, F., Munro, C., Siebert, S. & Hejnol, A. Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E409–E417 (2018).
78. 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).

79. Madeira, F. *et al.* Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* **50**, W276–W279 (2022).
80. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
81. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
82. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
83. Morel, B., Kozlov, A. M., Stamatakis, A. & Szöllősi, G. J. GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. *Mol. Biol. Evol.* **37**, 2763–2774 (2020).
84. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
85. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
86. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
87. DeLano, W. L. & Others. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr* **40**, 82–92 (2002).
88. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422–429 (2020).
89. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258 (2007).
90. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
91. Barua, A. & Mikheyev, A. S. An ancient, conserved gene regulatory network led to the rise of oral venom systems. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
92. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* **18**, 205–214 (2017).
93. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).
94. Bunis, D. G., Andrews, J., Fragiadakis, G. K., Burt, T. D. & Sirota, M. dittoSeq: universal user-friendly single-cell and bulk RNA sequencing visualization toolkit. *Bioinformatics* **36**, 5535–5536 (2021).