The most developmentally truncated fishes show extensive *Hox* gene loss and miniaturized genomes

Martin Malmstrøm[1,2,*], Ralf Britz[3], Michael Matschiner[1,2], Ole K. Tørresen[1], Renny K. Hadiaty[4], Norsham Yaakob[5], Heok H. Tan[6], Kjetill S. Jakobsen[1], Walter Salzburger[1,2] & Lukas Rüber[7,8,*]

[1]Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, Norway. [2]Zoological Institute, University of Basel, Basel, Switzerland. [3]Department of Life Sciences, Natural History Museum, London, UK.[4]Museum Zoologicum Bogoriense, Research Center for Biology, Indonesian Institute of Sciences, Cibinong, Indonesia. [5]Forest Research Institute Malaysia, Kuala Lumpur, Malaysia. [6]Lee Kong Chian Natural History Museum, National University of Singapore, Singapore. [7]Naturhistorisches Museum der Burgergemeinde Bern, Bern, Switzerland. [8]Institute of Ecology and Evolution, University of Bern, Bern, Switzerland. *Corresponding authors Ma.M. and L.R.

Abstract

Hox genes play a fundamental role in regulating the embryonic development of all animals. Manipulation of these transcription factors in model organisms has unraveled key aspects of evolution, like the transition from fin to limb. However, by virtue of their fundamental role and pleiotropic effects, simultaneous knockouts of several of these genes pose significant challenges. Here, we report on evolutionary simplification in two species of the dwarf minnow genus *Paedocypris* using whole genome sequencing. The two species feature unprecedented *Hox* gene loss and genome reduction in association with their massive developmental truncation. We also show how other genes involved in the development of musculature, nervous system, and skeleton have been lost in *Paedocypris*, mirroring its highly progenetic phenotype. Further, we identify two mechanisms responsible for genome streamlining: severe intron shortening and reduced repeat content. As a naturally simplified system closely related to zebrafish, *Paedocypris* provides novel insights into vertebrate development.

Introduction

The developmental mechanisms that determine how genotypes translate into phenotypes and how selection acts on genotypes to shape morphological differences are central to our understanding of the diversity of living organisms[1,2]. Model organisms have been instrumental in our quest to decipher how genetic differences are associated with morphological and physiological disparity[3], and improved technologies for genetic modifications will aid in further elucidating genotype-phenotype interrelations[4]. One interesting example is the recent work by Nakamura *et al.* (2016), which utilized CRISPR-Cas9 to knock out three *Hox13* copies (*Aa, Ab,* and *D*) in zebrafish (*Danio rerio*) to investigate the role of these genes in the transition from fins to limbs. Although experiments like these are now feasible, the pleiotropic effects of genes involved in fundamental developmental processes, such as those of the large *Hox* gene family, present challenges and limitations to the phenotypic variation that can easily be induced in model organisms through genetic engineering. Examining naturally occurring extreme phenotypes in close relatives of model organisms thus provides a novel source of phenotypic and genotypic variation that is becoming increasingly important in improving our understanding of the molecular basis of evolutionary changes[5].

Discovered around 200 years ago[6], the zebrafish has been used as a molecular model organism since the 1980s and is currently one of the most important model systems for studying vertebrate development, genome evolution, toxicology, physiology, behavior, and disease[7,8]. Comparative efforts have so far focused on closely related *Danio* species[9,10], but recent studies have revealed that the diversity of zebrafish mutants is surpassed by the range of phenotypic variation among several of its related species in the wild[3,11,12]. Some of these other members of Cyprinidae (e.g *Paedocypris*, *Sundadanio*, and *Danionella*) are characterized by developmental truncation[13-16] and morphological novelties[16-18] and thus offer additional, underappreciated potential for comparative studies and promise fundamental and transformative advances in our understanding of the molecular underpinnings of evolutionary change leading to novelty and adaptation at a genetic and phenotypic level[3].

The Series Otophysi, which includes Cyprinidae, represents one of the

earliest diverging teleostean lineages, and the phylogeny of this lineage remains controversial[13,19-23]. One particularly difficult taxon to place confidently is the recently discovered miniaturized dwarf minnow genus *Paedocypris*[13,15,21,22,24] found in the highly acidic blackwater of endangered peat swamp forests in Southeast Asia. This genus of tiny vertebrates includes the world's smallest fish, maturing at ~8 mm[25]. *Paedocypris* exhibits an extreme case of organism-wide progenesis or developmental truncation, resulting in an anatomical adult condition closely resembling that of a 7.5 mm zebrafish larva, with over 40 bones not developed[15]. To investigate the genomic signatures of developmental truncation, we sequenced and compared the genomes of two representatives of the genus *Paedocypris*; *P. carbunculus* and *P. micromegethes*. The genome signatures of these two species allow the distinction of genus-specific from species-specific genomic changes and thus enable identification of features associated with the extreme phenotype of *Paedocypris*. By comparing their genomes with those of other teleosts, including the closely related zebrafish, we identify *Paedocypris*-specific genomic signatures of developmental truncation in the form of loss of various key developmental genes, mirroring their progenetic phenotype. We further demonstrate how the *Paedocypris* genome size has been reduced through shorter genes due to significantly shorter introns compared to zebrafish, while exon lengths and gene numbers have remained relatively unchanged. We find that the *Paedocypris* genomes are comparable in size and gene structure (i.e short introns and compact genomes) to those of the two pufferfish model organisms; *Takifugu rubripes* and *Dichotomyctere nigroviridis*, which have the smallest vertebrate genomes known[26]. Additionally, we show that the accumulation of transposable elements (TEs), especially DNA transposons, is very low, following the diversification of the genus *Paedocypris*, and propose a potential mechanism for enhanced transposon-silencing activity through duplication of the PIWI-like (*PIWIl1*) gene in this lineage. Highly progenetic fish species like *Paedocypris* will be important resources for future studies on vertebrate development, presenting a novel opportunity to investigate phenotype–genotype relationships in early vertebrate development and the genetic mechanisms of developmental truncation.

Results

Loss of *Hox*- and other developmental genes

Hox genes encode transcription factors essential in body patterning along the anterior-posterior body axis during early development of all animals[27]. Although a large number of genes is involved in developmental processes, *Hox* genes are especially interesting as they remain organized in conserved clusters and are ordered along the chromosomes according to where and when they are activated[28].

As all known *Paedocypris* species have a progenetic phenotype, we investigated whether this condition is reflected in their *Hox* gene repertoire. We compared 70 *Hox* gene transcripts, encoded by the 49 *Hox* genes in zebrafish, to detect syntenic *Hox* cluster regions in the two *Paedocypris* genomes (Methods Supplementary Note and Supplementary Table 1). For *Hox* clusters *Aa*, *Ab*, and *Bb*, we recover all genes on one to three contiguous sequences in synteny with *Danio rerio* (Figure 1a). Two scaffolds in both *Paedocypris* species cover the *HoxBa* cluster; however, functional copies of *hoxB10a* and *hoxB7a* cannot be identified in this region. In total, 10 of the zebrafish *Hox* genes appear to be absent in *Paedocypris*, and their predicted positions are illustrated as red lines in Figure 1a. While the remaining *Hox* clusters are recovered as more fragmented, most of the genetic regions are covered by contiguous sequences containing intact copies of other *Hox* genes in at least one of the two *Paedocypris* species. None of the 10 *Hox* genes could be detected in other parts of the genomes by sequence similarity searches, strengthening the hypothesis that these genes are indeed lost in this lineage. Although extensive *Hox* gene loss has been reported for other cyprinids, these are secondary losses, following genome duplication events[29]. Figure 1b shows the positions of both the missing and the present *Hox* genes in *Paedocypris* compared to currently available otophysan genomes, illustrating the most plausible reconstruction of *Hox*-cluster evolution in this fish lineage.

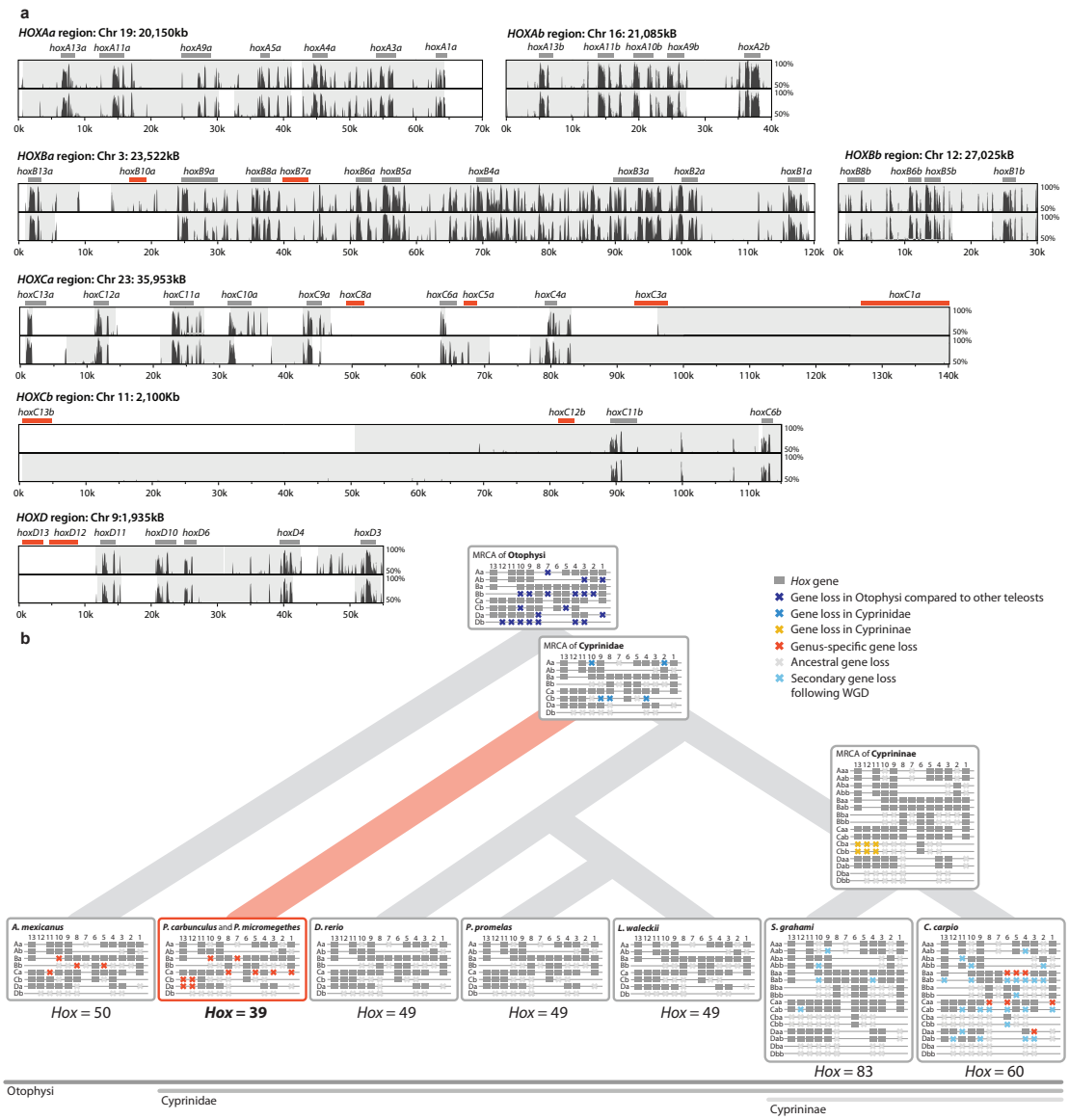


Figure 1 | Sequence similarity plots and reconstructed evolutionary history of teleost *Hox* genes. a) Pairwise sequence similarity between zebrafish (*D. rerio*) and *P. carbunculus* (top) and *P. micromegethes* (bottom) for syntenic *Hox* regions. Sequence similarity from 50–100% is shown with green bars representing the location of complete copies of recovered *Hox* genes, while red bars represent the expected location of the unrecovered genes. Grey shading represent contiguous sequences in either of the *Paedocypris* species with sufficient similarity to be mapped to the zebrafish chromosomes. b) Cladogram of selected otophysan species including hypothetical ancestors and ancestral losses (topology according to Stout *et al.* 2017). Genus-specific *Hox* gene losses are shown in red. Only intact gene copies are shown. *Hox* cluster information from other species, and the ancestral states, are based on Henkel *et al.* (2012)[29], Pascual-Anaya *et al.* (2013)[30], Yang *et al.* (2016)[31], and Ensembl[32].

Given its progenetic phenotype and the fact that ~20% of the zebrafish *Hox* gene repertoire is apparently absent in *Paedocypris*, we also investigated the presence

of other key genes involved in various developmental pathways. We first determined the overlap in gene space between the two *Paedocypris* species, the zebrafish, and the two tetraodontid models *Di. nigroviridis* and *T. rubripes,* as these pufferfishes have similarly small genomes as the *Paedocypris* species. Figure 2a shows the number of shared orthogroups for all species, identified using OrthoFinder[33], and highlights the number of orthogroups without orthologs in *Paedocypris* (numbers in bold). We used the set of genes from these 1,160 orthogroups to identify a comprehensive list of 1,581 genes that had gene ontologies associated with different system development pathways and pattern specific processes in *D. rerio* (Methods and Supplementary Table 2). These 1,581 genes were then used as queries to screen the genomic sequences and annotated protein sets of *Paedocypris*. Fourteen of these genes, primarily involved in skeletal- (*ucmab*, *disc1* and *lbh*), muscle- (*ccdc78* and *marcksa*), and nervous system (*nkx2.9*, *pho*, *cntf*, *thfrsf1b*, *tnfrsf19* and *lepa*) development, were not found in either of the *Paedocypris* species (Figure 2b). Adjacent flanking genes in zebrafish could, however, be identified in *Paedocypris* for all 14 genes (Supplementary Table 3).

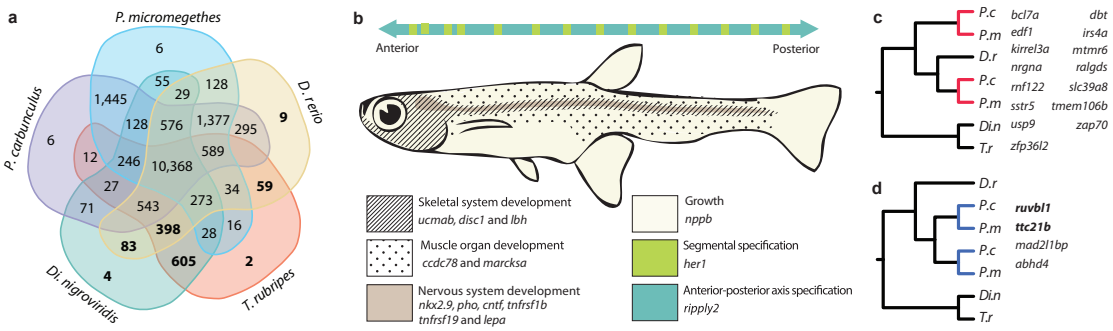


Figure 2 | Gene loss and duplication of developmental process genes. a) Five-species comparison of shared orthogroups for the identification of genes lost in *Paedocypris*. Bold numbers represent orthogroups without any *Paedocypris* orthologs. Orthogroups were identified using OrthoFinder[33] on the basis of full protein datasets for all included species. b) Lost developmental pathway genes and schematic representation of phenotypically affected body segments in *Paedocypris*. c) Gene duplicates retained in *Paedocypris*. d) Genus specific gene duplications. Genes in bold are associated with a truncated phenotype in *D. rerio*.

Because phenotypic changes can also result from gene duplication[34,35] , we identified genes with two copies in *Paedocypris*, but only a single copy in

175 zebrafish and pufferfishes. Based on the topology of gene trees generated for
176 these genes, we differentiated between genes originating from duplication
177 events predating the divergence of *Paedocypris* and *Danio*, where only one copy
178 is retained in zebrafish (Figure 2c), and *Paedocypris*-specific gene duplication
179 events (Figure 2d). Interestingly, although only four genes could be identified as
180 recent duplicates in *Paedocypris*, two of these (*ttc21b* and *ruvbl1*) are associated
181 with deformed phenotypes in zebrafish, including shortening of the anterior-
182 posterior axis and decreased head size[32].
183
184 **Genome miniaturization through intron shortening**
185 In parallel to their miniaturized body size, the two *Paedocypris* species show a
186 surprising evolutionary trajectory in terms of genome miniaturization.
187 Compared to the genome sizes of zebrafish (~1,5 Gb)[7,32] and other cyprinid
188 fishes (0.81–3.5 Gb)[36], we find that the two *Paedocypris* species have
189 substantially smaller genomes (0.45–0.52 Gb), yet similar numbers of genes
190 (Table 1). Comparative analyses of vertebrate genomes have shown that genome
191 size reduction is typically characterized by shorter introns and reduced repeat
192 content[37,38]. However, substantial loss of protein-coding genes[38,39], large
193 segmental deletions facilitated by fission of macrochromosomes[38], and a reduced
194 rate of large insertions have also been demonstrated to play a role in reducing or
195 constricting genome size[37,40].
196 As the *Paedocypris* genome sizes are comparable to those of tetraodontid
197 pufferfishes, we compared the gene repertoire of *Di. nigroviridis* and *T. rubripes*
198 to that of *Paedocypris* and *D. rerio* to determine what genomic features are
199 shared among these lineages. In order to obtain a consistent gene set for
200 comparative analysis of gene-, exon-, and intron lengths, we categorized the full
201 gene set of all five species into orthologous groups using the software
202 OrthoFinder[33]. We identified 10,368 orthogroups containing genes from all five
203 species (Figure 2a), which included 16,142 genes in zebrafish, 15,287 in *P.*
204 *carbunculus*, 14,181 in *P. micromegethes*, 14,529 in *Di. nigroviridis*, and 14,393 in
205 *T. rubripes*. The total gene length and the proportions of exonic and intronic
206 regions for this gene set are shown in Figure 3a. Although the accumulative
207 length of this gene set is significantly smaller in both *Paedocypris* species and the

208 two tetraodontids than in the zebrafish, the proportional change is more subtle,
209 with the gene set constituting on average 29–30% of the total genome length in
210 *Paedocypris* and the tetraodontids and 40% in zebrafish. Further, even though
211 we detected minor differences in this gene set regarding the average number of
212 exons per gene in *Paedocypris* (10.04) compared to *D. rerio* (11.21), the observed
213 47% reduction in total exon length of *Paedocypris* in relation to zebrafish cannot
214 be attributed to exon loss alone, as the two tetraodontids show very similar
215 results to *Paedocypris* with even higher average exon count (11.89). However,
216 the majority of the total gene length reduction observed in *Paedocypris* and
217 tetraodontids is due to an 80–84% reduction in overall intron size (Figure 2a).
218 To rule out that the overall reduction in intron size is driven by a highly deviant
219 fraction of the *Paedocypris* and tetraodontid gene sets, we also investigated the
220 distribution of gene-, exon-, and intron-lengths of this common gene set (Fig 2b).
221 We observe that zebrafish has substantially fewer short genes compared to all
222 other species, with an average gene length 4.3–5.5 longer than that of
223 *Paedocypris* and tetraodontids. This is not unexpected, as previous studies have
224 reported a lineage-specific expansion of intron size[41], resulting in an additional
225 peak of intron lengths between 1,000–2,000 bp in zebrafish, shown in Figure 3b.
226 Based on our results, it is apparent that the reduced average gene length in
227 *Paedocypris* is driven by a substantial shift towards consistently shorter introns,
228 similar to, but not as extreme as, in the two tetraodontids. This is further
229 illustrated by the calculated median sizes depicted in Fig. 3b.

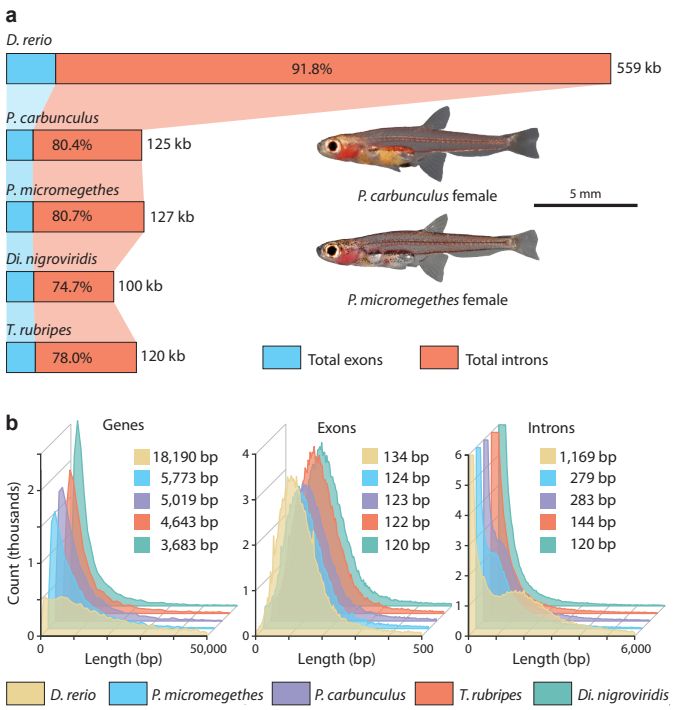


Figure 3 | Comparative analyses of repeat content and gene space in zebrafish and *Paedocypris*. a) Total gene length of 14–16,000 genes belonging to common orthogroups, and the proportional contribution of exonic and intronic regions, in zebrafish, *Paedocypris*, and tetraodontids. b) Frequency plot and median values of gene-, exon-, and intron lengths for the common gene set. In all distributions, the values for both *Paedocypris* species are significantly lower than in zebrafish, but significantly greater than those of both tetraodontids. (Wilcoxon Rank Sum test, $p < 10^{-15}$). Significant results were not detected for intron- and exon lengths between the two *Paedocypris* species; however, in terms of gene lengths, *P. micromegethes* was reported to be significantly longer.

Repeat content reduction and genome evolution

The genome size of an organism evolves through the relative rate of insertions and deletions and through the effect of natural selection that either favors or eliminates these changes. Genome shrinkage has thus been postulated to evolve through a bias in the rate of insertions relative to deletions[42], but the impact of this mechanism, and its importance in genome size evolution is still debated[43]. Importantly, even though deletions indeed appear to be more frequent than insertions, the latter tend to include significantly more base pairs, resulting in the gradual increase in genome size in eukaryotes[42]. Although several other types of mutational activity can promote genome-size expansion, self-replicating mobile elements (*i.e.* transposons) have been identified as the most prominent

contributor in this regard[44,45], and recent studies have confirmed a strong correlation between genome size and the amount of transposable elements (TEs) in teleost fishes and other vertebrates[46,47]. Because the more compact genomes have previously been shown to harbor fewer repeats[38], we investigated the degree to which this was the case in *Paedocypris*. Using a *Danio*-specific repeat library, we determined the total amount of TEs; DNA-transposons, long- and short interspersed repeats (LINEs and SINEs), and long terminal repeats (LTRs) in the genomes of *Paedocypris*, zebrafish, and four other cyprinid species (*Cyprinus carpio*, *Pimephales promelas*, *Sinocyclocheilus grahami*, and *Leuciscus waleckii*) in addition to the characiform cave tetra (*Astyanax mexicanus*) using Repeatmasker[48]. The repeat landscape graphs, illustrating the relative amount of each TE class and the Kimura distance[49] for each of these are shown in Figure 4a along with a time-calibrated phylogeny of these eight species, as inferred using BEAST[50] (see Methods). Figure 4a also shows the percentage of each repeat type of various age categories, calculated following Kapusta, *et al.* 2017)[47]. The total number of repetitive elements of each class is shown in Figure 4b, clearly illustrating that even though the proportional differences in repeat content is similar between *Paedocypris* and the other cyprinids, the total number of elements is much lower.

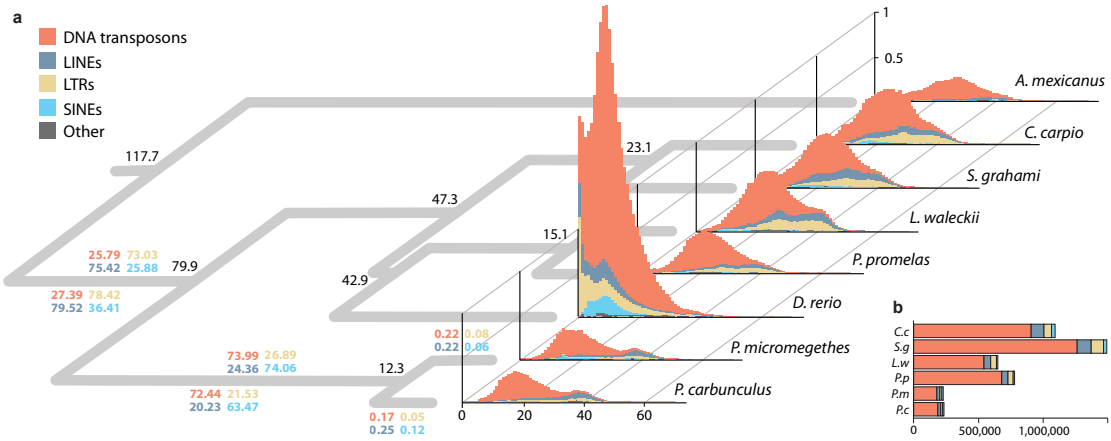
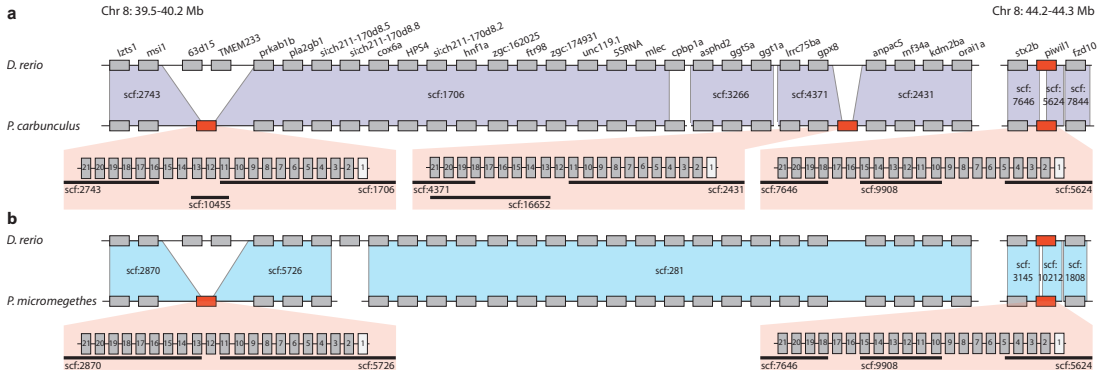


Figure 4 | Repeat landscape graphs showing the prevalence and relative age of the four main repeat classes in *Paedocypris* compared to six other teleost genome assemblies. a) Phylogeny with divergence times for the internal nodes for the eight species analyzed. Colored numbers on the branch leading to *Paedocypris* show the percentages of each of the corresponding repeat classes that originate from the time interval corresponding to this branch (117.7–79.9, 79.9–12.3, and more recent than 12.3 Ma). Repeat landscapes represent transposable elements of the four main classes as well as the unclassified ones. The x-axis

279 indicates the Kimura distance[49] as a proxy for time while the y-axis gives the relative coverage of
280 each repeat class based on the genome size. b) Total number of repetitive elements in each of the
281 cyprinid genomes (excluding zebrafish); colored by class.
282
283 Consistent with previous reports[46], we find the that the zebrafish genome is
284 dominated by repetitive elements, constituting 58.1% of its genome. In contrast,
285 only 15.3 –15.5% of the *Paedocypris* genomes consist of repetitive elements
286 (transposable elements, satellites, simple repeats, and low complexity regions), a
287 percentage that is comparable to that of the other cyprinids investigated (12.3–
288 19.01%). However, in terms of TE content, the *Paedocypris* species have
289 considerably fewer and shorter elements that comprise in total 7.24 and 7.30%
290 of their genomes, compared to the four other cyprinids in which we find on
291 average a genomic proportion of 13.81% (not including zebrafish in which these
292 constitute 53.28% of the genome)(Supplementary Table 4). Based on the timing
293 of the divergence events leading to crown *Paedocypris* and the calculated
294 substitution rate (see Methods), we investigated the proportions of "ancient" and
295 "lineage specific" TEs. Interestingly, only 0.05 – 0.25% of the *Paedocypris*
296 transposons appear to have been incorporated into their genomes after these
297 two species diverged (Figure 4).
298 As *Paedocypris* thus appears to have reversed the general trend of DNA
299 gain[43,43] towards gradual DNA loss, as indicated by their low TE content, we
300 investigated whether this could have been achieved through silencing of
301 transposon activity[40]. We thus explored the genomic content of the PIWI-like
302 genes (*PIWIl1* and *PIWIl2*), which are known for silencing transposons in
303 zebrafish and other vertebrates[51,52]. Interestingly, we identified additional copies
304 of *PIWIl1* in both *Paedocypris* genomes, located in a conserved region ~4 Mb
305 upstream of *PIWIl1* on zebrafish chromosome 8, as shown in Figure 5. One of
306 these gene duplications is shared between the two *Paedocypris* species and
307 therefore appears to have evolved prior to their divergence, while an additional,
308 third copy is found exclusively in *P. carbunculus* (Fig. 5a). These lineage-specific
309 duplications suggest that the gradual genome miniaturization has been, at least
310 in part, attained through increased transposon silencing, leading to a bias of DNA
311 loss over DNA gain. This needs to be tested further, utilizing genomic sequences

312 from other closely related cyprinids like *Danionella* and *Sundadanio*. Probably
313 resulting from the lineage-specific whole genome duplication, a single duplicate
314 of the *PIWIl1* gene was also found in *C. carpio,* but no extra copies could be
315 detected in the other cyprinids; *P. promelas*, *L. waleckii*, or *S. grahami*. No
316 additional copies of *PIWIl2* were detected in either *Paedocypris* or the other
317 cyprinids.



318

319 **Figure 5 | Synteny of the *PIWIl1* loci and surrounding regions in the zebrafish and**
320 ***Paedocypris* genomes.** a) Synteny between zebrafish (top) and *P. carbunculus*, showing the
321 original *PIWIl1* copy (right red rectangle) and the two additional copies (left and middle red
322 rectangles) detected in this genome assembly. b) Same as a) for *P. micromegethes*, showing the
323 single, presumably ancestral, duplicated copy of *PIWIl1* in this genome. Transparent red boxes
324 depict the exon-intron structure of all *PIWIl1* genes in *Paedocypris* with black lines representing
325 the scaffolds covering each region. Not all exons could be recovered for all gene copies.

326

327 **Chromosome fusion**

328 Another interesting feature of the miniaturized genome of *Paedocypris* is the
329 reduced number of chromosomes compared to that of zebrafish. While most
330 Cypriniformes have ≥ 24 chromosomes[53] in the haploid metaphase and zebrafish
331 has 25, *P. carbunculus* only features 15 chromosomes[54]. Since we did not find any
332 indication of large-scale chromosome loss, as would be indicated by substantial
333 gene loss, this discrepancy implies that genome miniaturization in *Paedocypris*
334 spp. has been accompanied by chromosomal fusion since they diverged from
335 their last common ancestor with zebrafish. In order to identify regions on
336 different chromosomes in the zebrafish genome that are now tightly linked in the
337 *Paedocypris* genome, we looked for syntenic regions in *Paedocypris* that contain
338 at least six genes from two different chromosomes in zebrafish. Figure 6 shows
339 the three scaffolds identified in each of the *Paedocypris* species fulfilling these

criteria. In both species, nine genes from chromosomes 9 and 21 in zebrafish are co-localized on a single scaffold. For chromosome 11 in zebrafish, we find good evidence in different regions for a fusion with chromosome 1 in *P. carbunculus* but also chromosome 2 in *P. micromegethes*, suggesting that either all three chromosomes may have fused or that species-specific fusion events may have occurred. Although the remaining two putative fusion events can only be detected in one of the species, there are no contradicting results between the two *Paedocypris* species, indicating that fusion of the chromosome pairs 6 and 19 and 22 and 25 is likely to have occurred in the last common ancestor of both species.

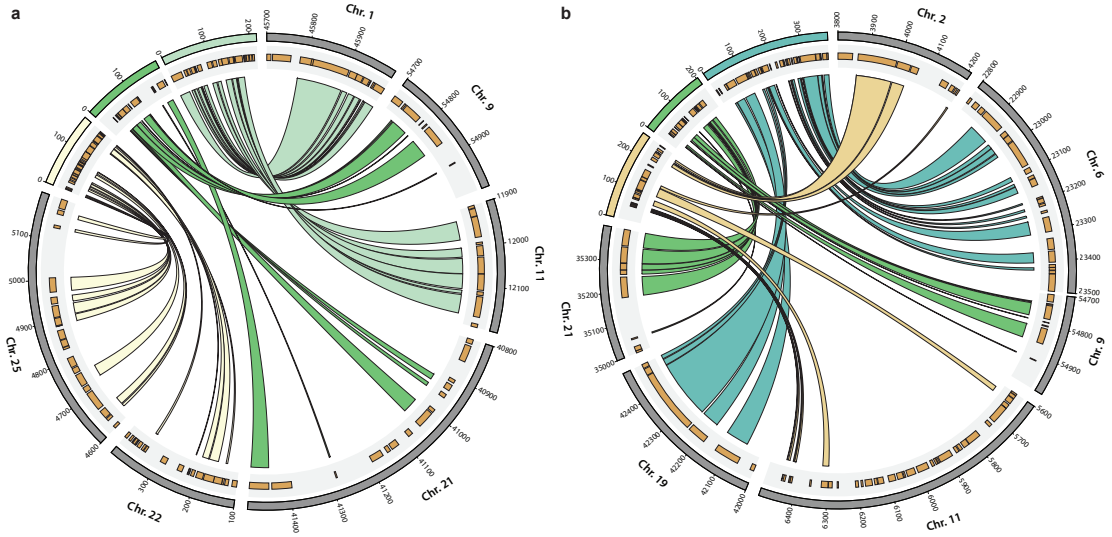


Figure 6 | Syntenic regions between zebrafish chromosomes and *Paedocypris* scaffolds. Syntenic regions between zebrafish (*D. rerio*) and *P. carbunculus* (a) or *P. micromegethes* (b). Links between *Paedocypris* scaffolds and *D. rerio* chromosomes indicate orthologous genes. Only the relevant parts of *D. rerio* chromosomes are shown and the locations of all genes in these regions are illustrated in the gray shaded areas. Numbers on scaffolds and chromosomes are given as kbp.

Discussion

Miniature body size has evolved several times within Cyprinidae[13]. However, little is known about the molecular underpinnings of these miniaturization events, whether these are similar in several instances of miniaturization, and to what extent the parallel evolution of genomic and phenotypic traits is coupled. Given the identification of several classical signatures of genome reduction, including reduced TE content and smaller introns, we conclude that the small

365 *Paedocypris* genomes represent a derived state, having evolved from an ancestor
366 with a substantially larger genome and a larger body size. Reduced genome size
367 has also been reported in other vertebrates[37,38], yet these reductions include
368 large segmental deletions, chromosome fissions, and massive losses of protein
369 coding genes, which we do not observe in the miniaturized genomes of
370 *Paedocypris*. Moreover, while reduced genome sizes in birds and bats can
371 probably be explained by the adaptation to flight through the need for higher
372 metabolic rates[55], the causes of genome size reduction in fishes seem in general
373 less apparent. However, the reduced genome sizes in *Paedocypris* may also have
374 evolved as a response to a selection for higher metabolic rate, mediated in this
375 case by its extreme habitat (*i.e.* low oxygen concentrations and pH < 4.0) rather
376 than rapid locomotion.
377 Small body size has arisen among miniature cyprinids in two different
378 ways, either as proportionate dwarfism, resulting in miniature but otherwise
379 identical copies of their larger ancestors, or through developmental truncation
380 or progenesis leading to larval appearance of the tiny mature fish[13,14].
381 Evolutionary novelties such as highly modified pelvic and pectoral girdles[14], an
382 anterior shift in genital pore and anus[13], and fang-like structures forming from
383 the jaw bones[16], however, appear to be restricted to developmentally truncated
384 miniatures (*e.g. Paedocypris*, *Danionella,* and *Sundadanio*). The absence of
385 novelties in proportioned dwarfs would suggest that developmental truncation
386 plays an important role in escaping evolutionary and developmental constraints
387 imposed by the species' *Bauplan* and in opening up new evolutionary avenues
388 for drastic morphological change[14,16,18].
389 This study represents the first report on genomes of highly
390 developmentally truncated fish species. It illustrates how the unique features of
391 *Paedocypris* at the phenotypic level (*i.e.* extreme miniaturization and dramatic
392 developmental truncation) are paralleled at the genomic level, in the form of
393 substantially smaller genomes via reduced intron lengths and lower repeat
394 content, and also through the loss of numerous genes related to development,
395 including highly important *Hox* genes. Although a causal link between
396 miniaturized body size and reduced genome size remains to be established, our
397 study highlights the potential to investigate such connections in a comparative

398 framework, including the sequencing of additional species representing other
399 miniature genera. Nevertheless, these naturally simplified species, having
400 escaped the developmental, and, as we report here, genetic constraints of the
401 cyprinid *Bauplan*, present a novel opportunity to investigate phenotype–
402 genotype relationships in vertebrate development. These naturally occurring
403 extreme phenotypes will greatly aid future research on model organisms and our
404 quest to understand how phenotypic diversity is generated during vertebrate
405 evolution and development.
406 Recent investigation of the Gulf pipefish genome[56] has revealed a loss of the
407 last *Hox7* paralog in this species. This has revived the disputed hypothesis that
408 the lack of ribs in both pufferfish and the pipefishes can be attributed to the fact
409 that none of these species have any remaining *Hox7* paralogs[57]. Interestingly, the
410 *Paedocypris* species show a similar loss of the last *Hox7* paralog (*hoxb7a*), yet the
411 two species investigated here do feature ribs. The ribs are, however, reduced and
412 remain poorly ossified[14], suggesting that although the *Hoxa7/Hoxb7* genes are
413 not the only essential genes for rib development, they do appear to have an
414 influence on their development, as suggested by earlier experiments in mice[58].
415 Finally, the observed loss of *Hox13* genes (*Cb* and *Da*) in the two
416 *Paedocypris* species is of special interest as these genes represent the
417 termination of the patterning system. *Hoxd13* was also one of the genes targeted
418 in the successful CRISPR-Cas9 knockout experiment in the zebrafish[59], that
419 showed that the *Hox* genes play a very similar role in patterning during
420 development of both fins bones, and thus elucidated the evolutionary transition
421 from fins to limbs. We propose that future studies of similar nature should
422 consider using *Paedocypris* species as a complementary system to the zebrafish
423 model, due to its already limited *Hox* gene repertoire, yet relatively close
424 phylogenetic relationship with zebrafish.
425
426 ## Materials and methods
427 ### Specimens used
428 *P. carbunculus* was caught at the type locality in Pangkalanbun, Kalimantan
429 Tengah, on the island of Borneo and *P. micromegethes* was caught in Sibu,
430 Sarawak, Malaysia, on the island of Borneo. Both species were caught using dip

431 nets. Immediately upon capture, specimens were killed by an overdose of
432 anesthesia (MS222) following guidelines by the American Society of
433 Ichthyologists and Herpetologists (ASIH) (http://www.asih.org/pubs/; issued
434 2013). Individuals were preserved in 96% ethanol for subsequent DNA analyses.
435
436 **DNA isolation**
437 A single whole specimen of each *Paedocypris* species, stored on 96% ethanol,
438 was used for isolation of high molecular genomic DNA with the EZNA Tissue
439 DNA Kit (Omega Bio-Tek, Norcross, Gerogia, USA), following manufacturer`s
440 instructions. Sample identifiers were LR12004 (*Paedocypris carbunculus*) and
441 LR7898 (*Paedocypris micromegethes*), supplied by Lukas Rüber.
442
443 **Sequencing library preparation**
444 Genomic DNA samples were fragmented to lengths of ~550 bp by sonication on a
445 Covaris E220 (Life Technologies, Carlsbad, California, USA) with the following
446 settings: 200 cycles for 45 seconds with Peak Incident Power of 175 W and
447 frequency sweeping mode. All sequencing libraries were constructed following
448 Illumina's TrueSeq PCR-free library preparation protocol for 550 bp fragments.
449
450 **Whole genome sequencing**
451 Based on expected genome sizes of 315-350 Mb[54], both *Paedocypris* genomes
452 were sequenced to ~90× coverage on the Illumina HiSeq 2500 platform, with
453 the Illumina 500 cycles kit (Rapid mode) with on-board clustering. Per species,
454 this produced 151-155 M paired reads of 250 bp each.
455 The Kapa Library quantification kit for Illumina (Kapa Biosciences, Wilmington,
456 Massachusetts, USA) was used to find the correct molarity (nM) before
457 sequencing.
458
459 **Genome assembly**
460 Both *Paedocypris* genomes were initially assembled using two different assembly
461 programs, the "Overlap-Layout-Consensus" based Celera Assembler[60] and the
462 "de Bruijn graph" based DISCOVAR *de novo*[61]. We then used the Metassembler
463 software[62] to merge and optimize these two assemblies, producing a

464 reconsolidated single assembly with superior quality for each of the two
465 *Paedocypris* species (Table 1, Supplementary Note, and Supplementary Table 5).
466
467 **Assembly quality assessment**
468 The quality of the different genome assemblies was assessed by comparing the
469 proportion of conserved genes detected, as a measure of gene-space
470 completeness. We used the program BUSCO v2.0 (Benchmarking Universal
471 Single-Copy Orthologs)[63], which searches for 4,584 highly conserved single-copy
472 actinopterygian genes. Results are listed in Table 1.
473 We also assessed the assembly quality of the three different assembly
474 versions with the software FRC^{bam} [64], which identifies "features" (incorrectly
475 mapped reads, incorrect insert size, coverage issues, etc.) in each of the
476 assemblies, and ranks the different assembly versions according to the number
477 of detected features. Additional information on the execution of these programs
478 is available in the Supplementary Note and the resulting graphs are presented in
479 Supplementary Figs. 1 and 2.
480
481 **Annotation**
482 Structural and functional annotation of both reconsolidated *Paedocypris*
483 genomes was performed with two iterative rounds of the MAKER2 (v 2.31.8)[65-67]
484 pipeline, following the instructions by Sujai Kumar (available at
485 https://github.com/sujaikumar/assemblage/blob/master/README-
486 annotation.md). Numbers of genes annotated are listed in Table 1. See
487 Supplementary Note for in-depth descriptions of additional software and
488 commands used.
489
490 **Table 1**
491 Assembly statistics, gene-space completeness-, and annotation metrics.

	P. carbunculus	***P. micromegethes***
Total assembly length (bp)	430,790,821	414,707,736
Estimated assembly coverage (CA / DDN)	67× / 73×	76× / 91×
Number of scaffolds	18,953	15,158
N50 scaffold length	59,252	61,901
Longest scaffold (bp)	653,275	678,207
BUSCOs found[2]	4,113	4,073
BUSCOs complete[2]	3,719	3,564

Total gene count[3]	25,567	25,453

492 [1]Out of 4,584 highly conserved Actinopterygii genes
493 [2]Annotation Edit Distance (AED) < 1.0 or containing a protein family (Pfam) domain
494 [3]Based on OrthoFinder[33] comparative analysis between zebrafish (*D. rerio*) and both *Paedocypris* species
495

496 **_Hox_ gene search**

497 In addition to the two *Paedocypris* genomes, *Hox* gene content was also
498 investigated in the genomes of *Pimephales promelas*
499 (GCA_000700825.1_FHM_SOAPdenovo_genomic.fna) and *Leuciscus waleckii*
500 (GCA_900092035.1_Amur_ide_genome_genomic.fa) based on BLAST[68] searches,
501 using 70 zebrafish *Hox* transcripts from Ensembl as queries (including some
502 truncated variants and isoforms), representing the 49 unique *Hox* genes
503 (Supplementary Table 1). The similarity stringency threshold used in these
504 searches was $1e^{-20}$. We also conducted additional searches using Exonerate
505 v2.2.0[69] for genes whose presence could not be established based on the BLAST
506 search. For those genes that could not be detected using either method,
507 additional searches were conducted using the orthologous protein sequence
508 from the cave tetra (*Astyanax mexicanus*). Contiguous sequences from both
509 *Paedocypris* genomes were extracted from the scaffolds assembled with
510 Metassembler, spanning all hit regions plus 10 kb upstream and downstream of
511 each hit. These scaffold sequences were aligned to the orthologous sequences
512 extracted from the zebrafish genome (GRCz10) using mVista[70].
513

514 **Identification of lost and expanded developmental genes**

515 In order to obtain a complete list of genes from *D. rerio* that were associated with
516 various developmental pathways that could be compromised in *Paedocypris*, we
517 started out with the orthogroups found not to contain orthologs from either
518 *Paedocypris* species, as identified using OrthoFinder[33]. We further utilized the
519 gene ontology information associated with all genes belonging to these
520 orthogroups, and identified key gene ontology terms: GO:0009948 (anterior–
521 posterior axis formation), GO:0009950 (dorsal–ventral axis formation),
522 GO:0040007 (growth), GO:0007517 (muscle organ development), GO:0007399
523 (nervous system development), GO:0001501 (skeletal system development), and
524 GO:0007379 (segment specification). This data set contained 1,581 unique genes
525 from *D. rerio*, and 64 of these genes could not be detected in the *Paedocypris*

526 genome sequences using TBLASTN[68] with a similarity cutoff of $1e^{-20}$, and their
527 presence was further examined by running Exonerate[69] (v2.2.0). Based on the
528 Exonerate results, reciprocal BLAST searches, annotation, and identification of
529 flanking genes, we could confidently determine that 14 of these genes were
530 indeed not present in either of the *Paedocypris* genomes. Using expression data
531 from Ensembl[32], we reconstructed the hypothetically affected body segments in
532 *Paedocypris*, as illustrated in Figure 2b.
533 We further investigated whether we could confidently identify
534 developmental genes that are now duplicated in *Paedocypris* but not in *D. rerio*.
535 From the full list of orthogroups, we identified 138 groups that were represented
536 by a single copy in the two tetraodontid pufferfishes and in *D. rerio*, but by two
537 apparent copies in both *Paedocypris* species. The alignments of these genes were
538 then screened for missing data, and based on manual inspection of these
539 alignments, only alignments with < 35% missing data were included in further
540 analysis. Out of the 35 genes that met this criterion, 19 could be confirmed
541 through annotation, and had a gene tree consistent with the two hypotheses
542 outlined in Figure 2c and 2d.
543
544 **Gene space evolution**
545 In order to assess the changes of gene-, exon-, and intron sizes in *Paedocypris* we
546 first identified the proteomic overlap of *Paedocypris*, *D. rerio*, *Di. nigroviridis*, and
547 *T. rubripes* by running the software OrthoFinder[33] on the complete protein sets
548 of these five species. We used the full protein sets from Ensembl[32] (v. 80): *D.*
549 *rerio* (GRCz10) , *Di. nigroviridis* (TETRAODON8), and *T. rubripes*(FUGU4).
550 However, as some of the *D. rerio* genes have more than one protein or transcript
551 in the Ensembl database, the output from BioMart (31,953 genes and 57,349
552 proteins) was filtered so that only the longest protein sequence from each gene
553 was used in the analysis, and genes without protein sequences were removed.
554 This resulted in a set of 25,460 genes with a single protein prediction. For the
555 *Paedocypris* species, the "standard" gene sets resulting from the annotation were
556 used as input. These sets were filtered to include only genes with AED
557 (Annotation Edit Distance) scores < 1 or those with a Pfam domain. By using only
558 genes belonging to the 10,368 orthogroups found to contain orthologs from all

559 these species (Fig. 2a), we obtained a comprehensive but conservative dataset as
560 the basis for these analyses. Information about each of the corresponding genes
561 in *D. rerio*, *Di. nigroviridis*, and *T. rubripes* was obtained from BioMart, and
562 included the Ensembl gene and protein ID, and the chromosome name in
563 addition to the start and stop position for each gene, transcript, and exon. Intron
564 sizes were then calculated on the basis of exon positions, using a custom script
565 ("gene_stats_from_BioMart.rb"). In some cases, the sum of exons and introns did
566 not equal the total length of a gene, which appears to be caused by inconsistency
567 in the registration of UTR regions in the Ensembl database for individual genes.
568 In these cases, to be conservative with regard to intron length estimates, the
569 gene length was shortened to correspond to the sum of the exons and the
570 corresponding introns between these.
571 Intron and exon lengths for the two *Paedocypris* species were calculated
572 in a similar manner, but on the basis of the "standard" filtered annotation file in
573 "gff" format produced as part of the annotation pipeline. Also for these species,
574 the intron lengths were determined on the basis of identified exons, with another
575 custom script ("gene_stats_from_gff.rb"). Gene-, exon-, and intron length
576 histograms were plotted with the R package ggplot2. Custom Ruby scripts are
577 available for download at https://github.com/uio-cees/Paedocypris_gene_stats
578
579 **Repeat content analysis**
580 The repeat contents of the two *Paedocypris* genomes and the model organism
581 genomes were assessed using RepeatMasker (v4.0.6)[48] with the *Danio*-specific
582 repeat library included in the program, and with the "-s" setting to increase
583 sensitivity. The following model organism genome assemblies were used in this
584 comparison: *D. rerio* (Danio_rerio.GRCz10.dna.toplevel.fa), *P. promelas*
585 (GCA_000700825.1_FHM_SOAPdenovo_genomic.fna), *L. waleckii*
586 (GCA_900092035.1_Amur_ide_genome_genomic.fa), *Sinocyclocheilus grahami*
587 (GCA_001515645.1_SAMN03320097.WGS_v1.1_genomic.fna), *Cyprinus carpio*
588 (GCA_001270105.1_ASM127010v1_genomic.fna), and *A. mexicanus*
589 (Astyanax_mexicanus.AstMex102.dna.toplevel.fa). Repeat landscape graphs (Fig.
590 4a) were plotted with the R package ggplot2 based on the ".aligned" output file
591 from RepeatMasker. The proportion of repeats originating from specific time

592 intervals was calculated using the parseRM.pl script[47] (available at
593 https://github.com/4ureliek/Parsing-RepeatMasker-Outputs), using the
594 *Paedocypris*-specific substitution rate and the "--age" setting set to "12.31,79.98",
595 according to divergence times estimated with BEAST2. All four repeat classes
596 were analyzed individually using the "--contain" setting.
597
598 **Calculation of substitution rate for *Paedocypris***
599 The *Paedocypris*-specific substitution rate was calculated on the basis of a whole
600 genome alignment of the two *Paedocypris* species. These alignments were
601 created by mapping the sequencing reads of *P. micromegethes* to the *P.*
602 *carbunculus* assembly using BWA (Burrows-Wheeler Aligner) (v. 0.7.12)[71] and
603 SAMTOOLS (v. 1.3.1)[72,73]. The number of nucleotide differences in each of the
604 18,953 alignments (one per scaffold) was identified using a custom script
605 ("find_variable_sites.rb", available for download at https://github.com/uio-
606 cees/Paedocypris_gene_stats). The substitution rate per million years
607 (0.001288) was then calculated as the sum of total differences (13,415,043)
608 divided by the total number of aligned sites (422,829,036) and two times the
609 estimated crown age of *Paedocypris* (12.31), as inferred by the BEAST v.2.4.5[50]
610 analysis.
611
612 **Phylogenetic inference of selected otophysan species**
613 To allow the estimation of divergence times between the two *Paedocypris*
614 species and six other otophysan taxa for which genomic resources were available
615 (*D. rerio*, *A. mexicanus, C. carpio*, *P. promelas*, *S. grahami*, and *L. waleckii*), we
616 followed the pipeline for phylogenetic marker selection presented in Malmstrøm
617 et al. (2016 and 2017)[74,75] with few modifications. These modifications included
618 the following changes: Marker selection began with a set of 3,238 cave fish (*A.*
619 *mexicanus*) exons, for which at least five orthologs were known among the seven
620 species *D. rerio*, *Gadus morhua, Gasterosteus aculeatus*, *Oreochromis niloticus*,
621 *Oryzias latipes*, *Poecilia formosa*, and *T. rubripes*, according to version 87 of the
622 Ensembl database. This marker set was then used to identify potential orthologs
623 from the genomes of the eight selected otophysan taxa based on TBLASTN[68]
624 searches followed by a strict filtering procedure. Compared to Malmstrøm et al.

625 (2016), we applied a lower dN/dS threshold of 0.25 to exclude markers
626 potentially affected by positive selection, and we removed all markers for which
627 no homologs could be detected in one or more of the eight otophysan genomes.
628 We also applied stricter thresholds on clock-like evolution of candidate markers,
629 so that all genes with an estimated coefficient of rate variation above 0.8 as well
630 as those with a mean mutation rate above 0.0004 per site per million years were
631 excluded. We identified 138 genes with a total alignment length of 135,286 bp
632 which were subsequently used for analysis with BEAST v.2.4.5[50]. Since the
633 topology of otophysan taxa has previously been resolved with a more
634 comprehensive phylogenetic dataset[21], we here focused on the inference of
635 divergence times only, by using the topology inferred by Stout et al. (2016) as a
636 starting tree and excluding all of BEAST2's operators on the tree topology.
637 Divergence times were estimated by calibrating the most recent common
638 ancestor of Cypriniformes and Characiformes with a lognormal distribution
639 centered at 121 Ma (standard deviation on log scale: 0.1) according to the results
640 of Malmstrøm et al. (2016). We performed two replicate BEAST2 analyses with
641 800 million MCMC iterations, of which the first 100 million were discarded as
642 burnin. Convergence was assessed based on similarity of parameter traces
643 between run replicates and effective sample sizes (ESS) greater than 200. A
644 maximum clade credibility (MCC) summary tree with node heights according to
645 mean age estimates was produced with TreeAnnotator v.2.1.2[50].
646
647 **Identification of PIWI-like genes**
648 We investigated the presence of PIWI-like genes in the genomes of the two
649 *Paedocypris* species and the other cyprinids using Exonerate[69] (v2.2.0) with the
650 longest transcripts available for the two PIWI-like homologs from zebrafish;
651 *PIWIl1* (ENSDARG00000041699) and *PIWIl2* (ENSDARG00000062601). Regions
652 containing sequences spanning more than three introns were aligned to the
653 zebrafish exons using mafft[76] as implemented in AliView[77] (v1.17.1) . Intron
654 sequences were aligned manually based on the established exon structure, using
655 the full-length scaffold sequences. Local gene synteny to zebrafish chromosome
656 8, surrounding the putative *PIWIl1* copies, was confirmed through reciprocal

657 BLAST searches using both the MAKER2 annotated proteins and proteins
658 predicted by GeneScan (online version)[78] as queries.
659
660 **Identification of chromosomal rearrangements (fusions)**
661 As *P. carbunculus* has been shown to have a haploid chromosome count of 15
662 [54], potential chromosomal fusions were investigated on the basis of disrupted
663 synteny of zebrafish chromosomes in relation to *Paedocypris*.
664 We identified putative homologous regions between the zebrafish genome
665 assembly and each of the *Paedocypris* species' genome assemblies by using
666 MCScanX[79]. In short, the predicted proteins for each *Paedocypris* species were
667 merged with predicted proteins from zebrafish into a single file, and BLASTP[68]
668 was executed with this file as both query and target, thus identifying putative
669 homologs both within each species and between. The genomics positions of the
670 proteins were extracted from the annotation files, and the BLASTP results and
671 the genomic positions were provided to MCScanX for identifying the putative
672 homologous regions.
673
674 **Acknowledgments**
675 Fieldwork in the peat swamp forests in Malaysia and Indonesia was funded by
676 the Natural Environmental Research Council (NERC; NE/F003749/1, to L.R. and
677 R.B.), National Geographic (8509-08, to L.R.), and the North of England
678 Zoological Society (to L.R.). Fieldwork in Sarawak was conducted under permits
679 issued by the Economic Planning Unit, Prime Minister's Department, Malaysia
680 UPE 40/200/19/2534) and the Forest Department Sarawak (NCCD.970.4.4[V]-
681 43) and fieldwork in Indonesia was conducted under permits issued by the
682 Indonesian Institute of Sciences (LIPI) and the Kementerian Negara Riset dan
683 Teknology (RISTEK; 3/TKPIPA/FRP/SM/III/2012). We thank E. Adamson H.
684 Budianto H. Ganatpathy, S. Lavoué, M. Lo, H. Michael, and S. Sauri for their help in
685 the field. All computational work was performed on the Abel Supercomputing
686 Cluster (Norwegian metacenter for High Performance Computing (NOTUR) and
687 the University of Oslo) operated by the Research Computing Services group at
688 USIT, the University of Oslo IT-department. Sequencing library creation and
689 high-throughput sequencing were carried out at the Norwegian Sequencing

Centre (NSC), University of Oslo, Norway. This work was funded by grants from the Naturhistorisches Museum der Burgergemeinde Bern to L.R. and the Research Council of Norway (RCN grants 199806 and 222378) to K.S.J. H.H.T acknowledges funding from the National University of Singapore (NUS, R-154-000-318-112) and Lee Kong Chian Natural History Museum. W.S. acknowledges funding from the European Research Council (ERC) and the Swiss National Science Foundation (SNF).

Competing financial interests

The authors declare no competing financial interests.

References

1. Carroll, S. P., Hendry, A. P., Reznick, D. N. & Fox, C. W. Evolution on ecological time-scales. *Func. Ecol.* **21,** 387–393 (2007).
2. Moczek, A. P., Sultan, S., Foster, S., Ledón-Rettig, C., Dworkin, I., Nijhout, H. F., Abouheif, E. & Pfennig, D. W. The role of developmental plasticity in evolutionary innovation. *Proc. R. Soc. B.* **278,** 2705–2713 (2011).
3. Braasch, I., Peterson, S. M., Desvignes, T., McCluskey, B. M., Batzel, P. & Postlethwait, J. H. A new model army: emerging fish models to study the genomics of vertebrate Evo-Devo. *J. Exp. Zool.* **324,** 316–341 (2014).
4. Shalem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR-Cas9. *Nat. Rev. Genet.* **16,** 299–311 (2015).
5. 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).
6. Hamilton, F. *An Account of the Fishes Found in the River Ganges and Its Branches*. Hurst, Robinson, and Co, Edinburgh (1822).
7. Howe, K., Clark, D. M., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., Collins, J. E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J. C., Koch, R., Rauch, G-J., White, S., Chow, W. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496,** 498–503 (2013).
8. Kettleborough, R. N. W. Busch-Nentwich, E. M., Harvey, S. A., Dooley, C. M., de Bruijn, E., van Eeden, F., Sealy, I., White, R. J., Herd, C., Nijman, I. J., Fényes, F., Mehroke, S., Scahill, C., Gibbons, R.,n Wali, N., Carruthers, S., Hall, A., Yen, J., Cuppen, E. & Stemple, D. L. A systematic genome-wide analysis of zebrafish protein-coding gene function. *Nature* **496,** 494–497 (2013).
9. McMenamin, S. K., Bain, E. J., McCann, A. E., Patterson, L. B., Eom, D. S., Zachary P. Waller, Z. P., Hamill, J. C., Kuhlman, J. A., Eisen, J. S. & Parichy, D. M. Thyroid hormone-dependent adult pigment cell lineage and pattern in zebrafish. *Science* **345,** 1358–1361 (2014).
10. McCluskey, B. M. & Postlethwait, J. H. Phylogeny of zebrafish, a "model

733 species," within *Danio,* a 'model genus'. *Mol. Biol. Evol.* **32,** 635–652
734 (2015).
735 11. Patterson, L. B., Bain, E. J. & Parichy, D. M. Pigment cell interactions and
736 differential xanthophore recruitment underlying zebrafish stripe
737 reiteration and *Danio* pattern evolution. *Nat. Commun.* **5,** 5299 (2014).
738 12. Parichy, D. M. The natural history of model organisms: advancing biology
739 through a deeper understanding of zebrafish ecology and evolution. *eLife*
740 **4,** e05635 (2015).
741 13. Rüber, L., Kottelat, M., Tan, H. H., Ng, P. K. L. & Britz, R. Evolution of
742 miniaturization and the phylogenetic position of *Paedocypris*, comprising
743 the world's smallest vertebrate. *BMC Evol. Biol.* **7,** 38 (2007).
744 14. Britz, R. & Conway, K. W. Osteology of *Paedocypris*, a miniature and highly
745 developmentally truncated fish (Teleostei: Ostariophysi: Cyprinidae). *J.*
746 *Morphol.* **270,** 389–412 (2009).
747 15. Britz, R., Conway, K. W. & Rüber, L. Miniatures, morphology and molecules:
748 *Paedocypris* and its phylogenetic position (Teleostei, Cypriniformes). *Zool.*
749 *J. Linnean Soc.* **172,** 556–615 (2014).
750 16. Britz, R. & Conway, K. W. *Danionella dracula*, an escape from the
751 cypriniform Bauplan via developmental truncation? *J. Morphol.* **277,** 147–
752 166 (2015).
753 17. Conway, K. W. & Britz, R. Sexual dimorphism of the Weberian apparatus
754 and pectoral girdle in *Sundadanio axelrodi* (Ostariophysi: Cyprinidae). *J.*
755 *Fish Biol.* **71,** 1562–1570 (2007).
756 18. Conway, K. W., Kubicek, K. M. & Britz, R. Morphological novelty and modest
757 developmental truncation in *Barboides*, Africa's smallest vertebrates
758 (Teleostei: Cyprinidae). *J. Morphol.* **14,** 33–18 (2017).
759 19. Fang, F., Norï n, M., Liao, T. Y., Källersjö, M. & Kullander, S. O. Molecular
760 phylogenetic interrelationships of the south Asian cyprinid genera *Danio*,
761 *Devario* and *Microrasbora* (Teleostei, Cyprinidae, Danioninae). *Zool.*
762 *Scripta* **38,** 237–256 (2009).
763 20. Tang, K. L., Agnew, M. K., Vincent Hirt, M., Sado, T., Schneider, L. M.,
764 Freyhof, J., Sulaiman, Z., Swartz, E., Vidthayanon, C., Miya, M., Saitoh, K.,
765 Simons, A. M., Wood, R. M. & Mayden, R. L. Systematics of the subfamily
766 *Danioninae* (Teleostei: Cypriniformes: Cyprinidae). *Mol. Phylogenet. Evol.*
767 **57,** 189–214 (2010).
768 21. Stout, C. C., Tan, M., Lemmon, A. R., Lemmon, E. M. & Armbruster, J. W.
769 Resolving Cypriniformes relationships using an anchored enrichment
770 approach. *BMC Evol. Biol.* **16,** 244 (2016).
771 22. Arcila, D., Ortí, G., Vari, R., Armbruster, J. W., Stiassny, M. L. J., Ko, K. D.,
772 Sabaj, M. H., Lundberg, J., Revell, L. J. & Betancur-R, R. Genome-wide
773 interrogation advances resolution of recalcitrant groups in the tree of life.
774 *Nat. Ecol. Evol.* **1,** 0020 (2017).
775 23. Hirt, M. V., Arratia, G., Chen, W. & Mayden, R. L. Effects of gene choice, base
776 composition and rate heterogeneity on inference and estimates of
777 divergence times in cypriniform fishes. *Biol. J. Linnean Soc.* (2017).
778 doi.org/10.1093/biolinnean/blw045
779 24. Mayden, R. L. & Chen, W.-J. The world's smallest vertebrate species of the
780 genus *Paedocypris*: a new family of freshwater fishes and the sister group
781 to the world's most diverse clade of freshwater fishes (Teleostei:

782 Cypriniformes). *Mol. Phylogenet. Evol.* **57,** 152–175 (2010).
783 25. Kottelat, M., Britz, R., Hui, T. H. & Witte, K.-E. *Paedocypris*, a new genus of
784 Southeast Asian cyprinid fish with a remarkable sexual dimorphism,
785 comprises the world's smallest vertebrate. *Proc. Biol. Sci.* **273,** 895–899
786 (2006).
787 26. Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli,
788 E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D.,
789 Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M.,
790 Levy, M., Boudet, N. *et al.* Genome duplication in the teleost fish *Tetraodon*
791 *nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431,**
792 946–957 (2004).
793 27. Mallo, M. & Alonso, C. R. The regulation of *Hox* gene expression during
794 animal development. *Development* **140,** 3951–3963 (2013).
795 28. Gaunt, S. J. The significance of *Hox* gene collinearity. *Int. J. Dev. Biol.* **59,**
796 159–170 (2015).
797 29. Henkel, C. V., Burgerhout, E., de Wijze, D. L., Dirks, R. P., Minegishi, Y.,
798 Jansen, H. J., Spaink, H. P., Dufour, S., Weltzien, F.-A., Tsukamoto, K. & van
799 den Thillart G. E. E. J. M. Primitive duplicate *Hox* clusters in the European
800 Eel's genome. *PLoS ONE* **7,** e32231–9 (2012).
801 30. Pascual-Anaya, J., D'Aniello, S., Kuratani, S. & Garcia-Fernàndez, J.
802 Evolution of *Hox* gene clusters in deuterostomes. *BMC Dev. Biol.* **13,** 26
803 (2013).
804 31. Yang, J., Chen, X., Bai, J., Fang, D., Qiu, Y., Jiang, W., Yuan, H., Bian, C., Lu, J.,
805 He, S., Pan, X., Zhang, Y., Wang, X., You, X., Wang, Y., Sun, Y., Mao, D., Liu, Y.,
806 Fan, G., Zhang, H. *et al.* The *Sinocyclocheilus* cavefish genome provides
807 insights into cave adaptation. *BMC Biol.* **14,** 1–13 (2016).
808 32. Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S.,
809 Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G.,
810 Gordon, L., Hourlier, T., Hunt, S. S., Janacek, S. H., Johnson, N., Juettemann,
811 T., Kähäri, A. K., Keenan, S. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,**
812 D662–669 (2015).
813 33. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole
814 genome comparisons dramatically improves orthogroup inference
815 accuracy. *Genome Biol.* **16,** 157 (2015).
816 34. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes.
817 *Genome Res.* **20,** 1313–1326 (2010).
818 35. Peterson, T. & Müller, G. B. Phenotypic novelty in EvoDevo: the distinction
819 between continuous and discontinuous variation and its importance in
820 evolutionary theory. *Evol. Biol.* **43,** 314–335 (2016).
821 36. Gregory, T. R. Animal genome size database (Available at
822 http://genomesize.com).
823 37. Neafsey, D. E. & Palumbi, S. R. Genome size evolution in pufferfish: a
824 comparative analysis of diodontid and tetraodontid pufferfish genomes.
825 *Genome Res.* **13,** 821–830 (2003).
826 38. Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., Lee, C., Storz, J. F., Antunes,A.,
827 Greenwold, M. J., Meredith, R. W., Ödeen, A., Cui, J., Zhou, Q., Xu, L., Pan, H.,
828 Wang, Z., Jin, L., Zhang, P., Hu, H., Yang, W. *et al.* Comparative genomics
829 reveals insights into avian genome evolution and adaptation. *Science* **346,**
830 1311–1320 (2014).

831 39. Hughes, A. L. & Friedman, R. Genome size reduction in the chicken has
832 involved massive loss of ancestral protein-coding genes. *Mol. Biol. Evol.* **25,**
833 2681–2688 (2008).
834 40. Vinogradov, A. E. Evolution of genome size: multilevel selection, mutation
835 bias or dynamical chaos? *Curr. Opin. Genet. Dev.* **14,** 620–626 (2004).
836 41. Moss, S. P., Joyce, D. A., Humphries, S., Tindall, K. J. & Lunt, D. H.
837 Comparative analysis of teleost genome sequences reveals an ancient
838 intron size expansion in the zebrafish lineage. *Genome Biol. Evol.* **3,** 1187–
839 1196 (2011).
840 42. Petrov, D. A. Mutational equilibrium model of genome size evolution.
841 *Theor. Pop. Biol.* **61,** 531–544 (2002).
842 43. Gregory, T. R. Insertion–deletion biases and the evolution of genome size.
843 *Gene* **324,** 15–34 (2004).
844 44. Sun, C., Shepard, D. B. & Chong, R. A. LTR retrotransposons contribute to
845 genomic gigantism in plethodontid salamanders. *Genome Biol.* **4,** 168-183
846 (2012).
847 45. Chénais, B., Caruso, A., Hiard, S. & Casse, N. The impact of transposable
848 elements on eukaryotic genomes: from genome size increase to genetic
849 adaptation to stressful environments. *Gene* **509,** 7–15 (2012).
850 46. Chalopin, D., Naville, M., Plard, F., Galiana, D. & Volff, J. N. Comparative
851 analysis of transposable elements highlights mobilome diversity and
852 evolution in vertebrates. *Genome Biol. Evol.* **7,** 567–580 (2015).
853 47. Kapusta, A., Suh, A. & Feschotte, C. Dynamics of genome size evolution in
854 birds and mammals. *Proc. Natl. Acad. Sci. USA* **114,** E1460–E1469 (2017).
855 48. Smith, A.F.A., Hubley R., & Green, P. RepeatModeler Open-1.0. (2015)
856 (Available at http://www.repeatmasker.org)
857 49. Kimura, M. A simple method for estimating evolutionary rates of base
858 substitutions through comparative studies of nucleotide sequences. *J. Mol.*
859 *Evol.* **16,** 111–120 (1980).
860 50. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard,
861 M. A., Rambaut, A. & Drummond, A. J. BEAST 2: A software platform for
862 Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10,** e1003537 (2014).
863 51. Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi-piRNA pathway
864 provides an adaptive defense in the transposon arms race. *Science* **318,**
865 761–764 (2007).
866 52. Levin, H. L. & Moran, J. V. Dynamic interactions between transposable
867 elements and their hosts. *Nat. Rev. Genet.* **12,** 615–627 (2011).
868 53. Yu, X., Zhou, T., Li, K., Li, Y. & Zhou, M. On the karyosystematics of cyprinid
869 fishes and a summary of fish chromosome studies in China. *Genetica* **72,**
870 225–235 (1987).
871 54. Liu, S., Hui, T. H., Tan, S. L. & Hong, Y. Chromosome evolution and genome
872 miniaturization in minifish. *PLoS ONE* **7,** e37305 (2012).
873 55. Zhang, Q. & Edwards, S. V. The Evolution of intron size in amniotes: a role
874 for powered flight? *Genome Biol. Evol.* **4,** 1033–1043 (2012).
875 56. Small, C. M., Bassham, S., Catchen, J., Amores, A., Fuiten, A. M., Brown, R. S.,
876 Jones, A. G. & Cresko, W. A. The genome of the Gulf pipefish enables
877 understanding of evolutionary innovations. *Genome Biol.* **17,** 258 (2016).
878 57. Amores, A., Suzuki, T., Yan, Y.-L., Pomeroy, J., Singer, A., Amemiya, C. &
879 Postlethwait, J. H. Developmental roles of pufferfish *Hox* clusters and

880 genome evolution in ray-fin fish. *Genome Res.* **14,** 1–10 (2004).
881 58. Chen, F., Greer, J. & Capecchi, M. R. Analysis of *Hoxa7/Hoxb7* mutants
882 suggests periodicity in the generation of the different sets of vertebrae.
883 *Mech. Dev.* **77,** 49–57 (1998).
884 59. Nakamura, T., Gehrke, A. R., Lemberg, J., Szymaszek, J. & Shubin, N. H.
885 Digits and fin rays share common developmental histories. *Nature* **537,**
886 225–228 (2016).
887 60. Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A.,
888 Johnson, J., Li, K., Mobarry, C. & Sutton G. Aggressive assembly of
889 pyrosequencing reads with mates. *Bioinformatics* **24,** 2818–2824 (2008).
890 61. Weisenfeld, N. I., Yin, S., Sharpe, T., Lau, B., Hegarty, R., Holmes, L., Sogoloff,
891 B., Tabbaa, D., Williams, L., Russ, C., Nusbaum, C., Lander, E. S., MacCallum I.
892 & Jaffe, D. B. Comprehensive variation discovery in single human genomes.
893 *Nat. Genet.* **46,** 1350–1355 (2014).
894 62. Wences, A. H. & Schatz, M. C. Metassembler: merging and optimizing de
895 novo genome assemblies. *Genome Biol.* **16,** 207 (2015).
896 63. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov,
897 E. M. BUSCO: assessing genome assembly and annotation completeness
898 with single-copy orthologs. *Bioinformatics* **31,** 3210–3212 (2015).
899 64. Vezzi, F., Narzisi, G. & Mishra, B. Reevaluating assembly evaluations with
900 feature response curves: GAGE and assemblathons. *PLoS ONE* **7,** e52210
901 (2012).
902 65. Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C.,
903 Alvarado, A. S. & Yandell, M. MAKER: An easy-to-use annotation pipeline
904 designed for emerging model organism genomes. *Genome Res.* **18,** 188–
905 196 (2007).
906 66. Holt, C. & Yandell, M. MAKER2: An annotation pipeline and genome-
907 database management tool for second-generation genome projects. *BMC*
908 *Bioinform.* **12,** 491 (2011).
909 67. Campbell, M. S., Law, M. Y, Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E.,
910 Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C. J., Ware, D., Shiu, S.-H.,
911 Childs, K. L., Sun, Y., Jiang, N. & Yandell, M. MAKER-P: A tool kit for the
912 rapid creation, management, and quality control of plant genome
913 annotations. *Plant Physiol.* **164,** 513–524 (2014).
914 68. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local
915 alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
916 69. Slater, G. & Birney, E. Automated generation of heuristics for biological
917 sequence comparison. *BMC Bioinform.* **6,** 31 (2005).
918 70. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA:
919 computational tools for comparative genomics. *Nucleic Acids Res.* **32,**
920 W273–W279 (2004).
921 71. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-
922 Wheeler transformation. *Bioinformatics* **25,** 1754–1760 (2009).
923 72. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,
924 Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup.
925 The sequence alignment/map format and SAMtools. *Bioinformatics* **25,**
926 2078–2079 (2009).
927 73. Li, H. A statistical framework for SNP calling, mutation discovery,
928 association mapping and population genetical parameter estimation from

929 sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).
930 74. Malmstrøm, M., Matschiner, M., Tørresen, O. K., Star, B., Snipen, L. G.,
931 Hansen, T. F., Baalsrud, H. T., Nederbragt, A. J., Hanel, R., Salzburger, W.,
932 Stenseth, N. C., Jakobsen, K. S. & Jentoft, S. Evolution of the immune system
933 influences speciation rates in teleost fishes. *Nat. Genet.* **48**, 1204–1210
934 (2016).
935 75. Malmstrøm, M., Matschiner, M., Tørresen, O. K., Jakobsen, K. S. & Jentoft, S.
936 Whole genome sequencing data and de novo draft assemblies for 66
937 teleost species. *Sci. Data* **4,** 160132–13 (2017).
938 76. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software
939 version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,**
940 772–780 (2013).
941 77. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for
942 large datasets. *Bioinformatics* **30,** 3276–3278 (2014).
943 78. Burge, C. & Karlin, S. Prediction of complete gene structures in human
944 genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
945 79. Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., Lee, T.-H., Jin, H.,
946 Marler, B., Guo, H., Kissinger, J. C. & Paterson, A. H. MCScanX: a toolkit for
947 detection and evolutionary analysis of gene synteny and collinearity.
948 *Nucleic Acids Res.* **40,** e49 (2012).
949