



## The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization

Kang Du, Matthias Stöck, Susanne Kneitz, Christophe C. Klopp, Joost Woltering, Mateus Contar Adolphi, Romain Feron, Dmitry Prokopov, Alexey Makunin, Ilya Kichigin, et al.

### ► To cite this version:

Kang Du, Matthias Stöck, Susanne Kneitz, Christophe C. Klopp, Joost Woltering, et al.. The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization. *Nature Ecology & Evolution*, 2020, 4 (6), pp.841-852. 10.1038/s41559-020-1166-x . hal-02640938

**HAL Id: hal-02640938**

**<https://hal.inrae.fr/hal-02640938>**

Submitted on 6 Jan 2021


**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



OPEN

# The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization

Kang Du<sup>1,2</sup>, Matthias Stöck<sup>3</sup> , Susanne Kneitz<sup>1</sup>, Christophe Klopp<sup>4,5</sup> , Joost M. Woltering<sup>6</sup>, Mateus Contar Adolphi<sup>1</sup>, Romain Feron<sup>7</sup> , Dmitry Prokopov<sup>8</sup> , Alexey Makunin<sup>8</sup> , Ilya Kichigin<sup>8</sup>, Cornelia Schmidt<sup>1</sup>, Petra Fischer<sup>1</sup>, Heiner Kuhl<sup>3</sup>, Sven Wuertz<sup>3</sup>, Jörn Gessner<sup>3</sup>, Werner Kloas<sup>3</sup>, Cédric Cabau<sup>4,5</sup>, Carole Iampietro<sup>9</sup>, Hugues Parrinello<sup>10</sup>, Chad Tomlinson<sup>11</sup>, Laurent Journot<sup>10</sup>, John H. Postlethwait<sup>12</sup>, Ingo Braasch<sup>13</sup>, Vladimir Trifonov<sup>8</sup>, Wesley C. Warren<sup>14</sup>, Axel Meyer<sup>16</sup> , Yann Guiguen<sup>15</sup>  and Manfred Scharl<sup>15</sup>  

**Sturgeons seem to be frozen in time. The archaic characteristics of this ancient fish lineage place it in a key phylogenetic position at the base of the ~30,000 modern teleost fish species. Moreover, sturgeons are notoriously polyploid, providing unique opportunities to investigate the evolution of polyploid genomes. We assembled a high-quality chromosome-level reference genome for the sterlet, *Acipenser ruthenus*. Our analysis revealed a very low protein evolution rate that is at least as slow as in other deep branches of the vertebrate tree, such as that of the coelacanth. We uncovered a whole-genome duplication that occurred in the Jurassic, early in the evolution of the entire sturgeon lineage. Following this polyploidization, the rediploidization of the genome included the loss of whole chromosomes in a segmental deduplication process. While known adaptive processes helped conserve a high degree of structural and functional tetraploidy over more than 180 million years, the reduction of redundancy of the polyploid genome seems to have been remarkably random.**

Vertebrate genome evolution has been strongly impacted by polyploidization events<sup>1,2</sup>. Early on, vertebrate ancestors experienced two rounds (1R and 2R) of whole-genome duplications (WGDs)<sup>3</sup>. The evolutionary history of the ~30,000 species of teleost fish, which make up more than 99% of all ray-finned fishes (Actinopterygia), is defined by a third WGD (3R) that occurred in their common ancestor about 320 million years ago (Ma), but not in the basal fish (bichirs, reedfish, sturgeons, paddlefishes, bowfins and gars), the land vertebrates or their sarcopterygian forbearing relatives (coelacanths and lungfishes). Some teleost groups, such as salmonids and carps, independently underwent another round (4R) of WGD. Interestingly, among the basal fishes only the sturgeon lineage is known to be prone to polyploidization events and includes many-ploid species, some with up to 380 chromosomes.

Sturgeon genomes, however, are a missing puzzle piece for understanding vertebrate ancestry. Sturgeons are a group of ray-finned fish that diverged from the actinopterygian stem before the teleost-specific 3R duplication and after the ancient 2R event<sup>4,5</sup>. After their divergence from the other ray-finned fish, the various lineages of *Acipenseriformes* (sturgeon and paddlefish) experienced several polyploidization events<sup>6</sup>, resulting in

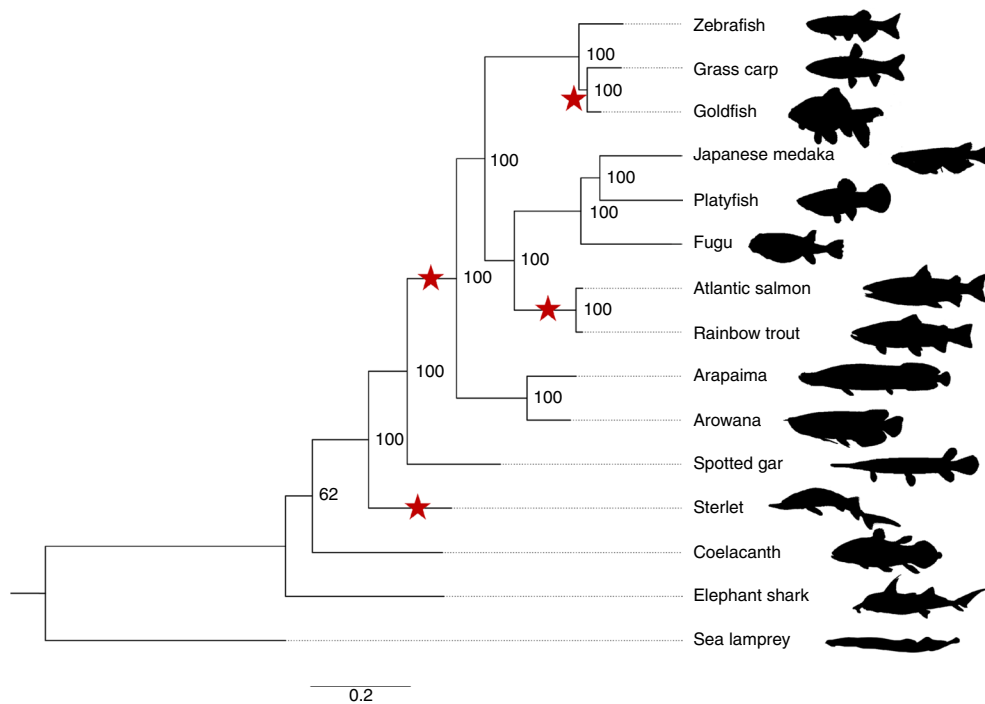
karyotypes, comprising between ~120 chromosomes in some species, and ~360 chromosomes in species that are considered dodecaploid<sup>7</sup>. The genomic basis for this parallelism between basal and derived fish lineages to acquire WGDs is not clear. While teleost lineages that experienced more recent 4R events are still recognizable apparent tetraploids, the other teleost lineages retained on average only 17% of gene duplicates from the ancient 3R ohnologues<sup>5</sup>. The evolutionary trajectories and forces driving species from polyploids to meiotic diploids are the subject of major adaptive hypotheses and their empirical evaluations<sup>8,9</sup>.

The genomic state of sturgeons is much less clear. They are often seen as ancient polyploids. On the basis of some cytogenetic and microsatellite data, others have considered sturgeons to be functional diploids<sup>10</sup> as result of an evolutionary process, where the gene content of a tetraploid species degenerates to become functionally diploid but maintains twice as many chromosomes, which form regular bivalents<sup>11</sup>. Such far-reaching redundancy reduction leads one to question their polyploidy state<sup>12</sup>.

Because sturgeons branched off early from modern fishes, their genomes may harbour traces of the ancient vertebrate ancestors<sup>13</sup>. Notably, their early embryonic development is of the classical

<sup>1</sup>Physiological Chemistry, Biocenter, University of Wuerzburg, Wuerzburg, Germany. <sup>2</sup>Developmental Biochemistry, Biocenter, University of Wuerzburg, Wuerzburg, Germany. <sup>3</sup>Leibniz-Institute of Freshwater Ecology and Inland Fisheries, IGB, Berlin, Germany. <sup>4</sup>Plate-forme Bio-informatique Genotoul, Mathématiques et Informatique Appliquées de Toulouse, INRA, Castanet-Tolosan, France. <sup>5</sup>SIGENAE, GenPhySE, Université de Toulouse, INRA, ENVT, Castanet-Tolosan, France. <sup>6</sup>Lehrstuhl für Zoologie und Evolutionsbiologie, Department of Biology, University of Konstanz, Konstanz, Germany. <sup>7</sup>Department of Ecology and Evolution, University of Lausanne, and Swiss Institute of Bioinformatics, Lausanne, Switzerland. <sup>8</sup>Institute of Molecular and Cellular Biology, Siberian Branch of the Russian Academy of Sciences, Novosibirsk State University, Novosibirsk, Russia. <sup>9</sup>INRAE, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France. <sup>10</sup>Montpellier GenomiX (MGX), c/o Institut de Génomique Fonctionnelle, Montpellier, France. <sup>11</sup>McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA. <sup>12</sup>Institute of Neuroscience, University of Oregon, Eugene, OR, USA. <sup>13</sup>Department of Integrative Biology, Michigan State University, East Lansing, MI, USA. <sup>14</sup>Bond Life Sciences Center, University of Missouri, Columbia, MO, USA. <sup>15</sup>INRA, UR1037 LPGP, Fish Physiology and Genomics, Rennes, France. <sup>16</sup>The Xiphophorus Genetic Stock Center, Department of Chemistry and Biochemistry, Texas State University, San Marcos, TX, USA. <sup>17</sup>Hagler Institute for Advanced Study and Department of Biology, Texas A&M University, College Station, TX, USA.

✉e-mail: [matthias.stoeck@igb-berlin.de](mailto:matthias.stoeck@igb-berlin.de); [phch1@biozentrum.uni-wuerzburg.de](mailto:phch1@biozentrum.uni-wuerzburg.de)



**Fig. 1 | Phylogeny of sterlet and related species.** Species tree built using RAXML on the basis of 47 one-to-one orthologues. The sea lamprey was used as the outgroup. The topology of the tree was confirmed by MrBayes (see also Supplementary Fig. 2). Red stars indicate WGDs after the 1R/2R event; numbers at branches indicate bootstrap support values based on 100 resampled data sets; the scale bar indicates the average substitutions per site; the dotted lines associate the taxon names with the branch ends.

amphibian type and very different from that of all modern fish<sup>14,15</sup>, reflecting the basal divergence of the lineage.

Sturgeons are distributed from subtropical to subarctic rivers, lakes and coastlines of Eurasia and North America<sup>16</sup>. They are long-lived and reproduce late, usually not before reaching an age of ten years. In many sturgeon species, adults migrate repeatedly from the sea into freshwater to spawn<sup>17</sup>. Sturgeons are celebrities among fishes because of their pre-ovulation female gametes, known as caviar. Habitat destruction, the lack of river connectivity, pollution<sup>16,18</sup> and the 2,000-year-old rural caviar production<sup>19</sup> culminated in ongoing devastating overexploitation that drove most sturgeon species into a threatened status (<https://www.iucnredlist.org/>). Because wild caviar can no longer be traded legally, sturgeon aquaculture has gained high economic importance, and in turn can contribute to the protection of wild populations by providing a safe market supply.

Despite their ancient lineage, peculiar biological features and economic value, sturgeon genomes have remained largely unexplored owing to their dauntingly polyploid state<sup>20</sup>. We therefore sequenced the sterlet sturgeon, *Acipenser ruthenus*, a species with only 120 chromosomes, and present here an annotated chromosome-scale genome assembly. We found that this genome represents an ancient WGD, which remained close to tetraploidy owing to the slow evolutionary rate and serves as a good representative of the ancestral actinopterygian genome. In contrast to other polyploid fish, deduplication after the sterlet WGD involves the loss of entire homeologous chromosomes (segmental rediploidization). Adaptive processes in the retention of duplicate genes are only partly responsible for determining the gene content, and they worked in parallel with stochastic events to shape the genomic landscape of the tetraploid sterlet sturgeon.

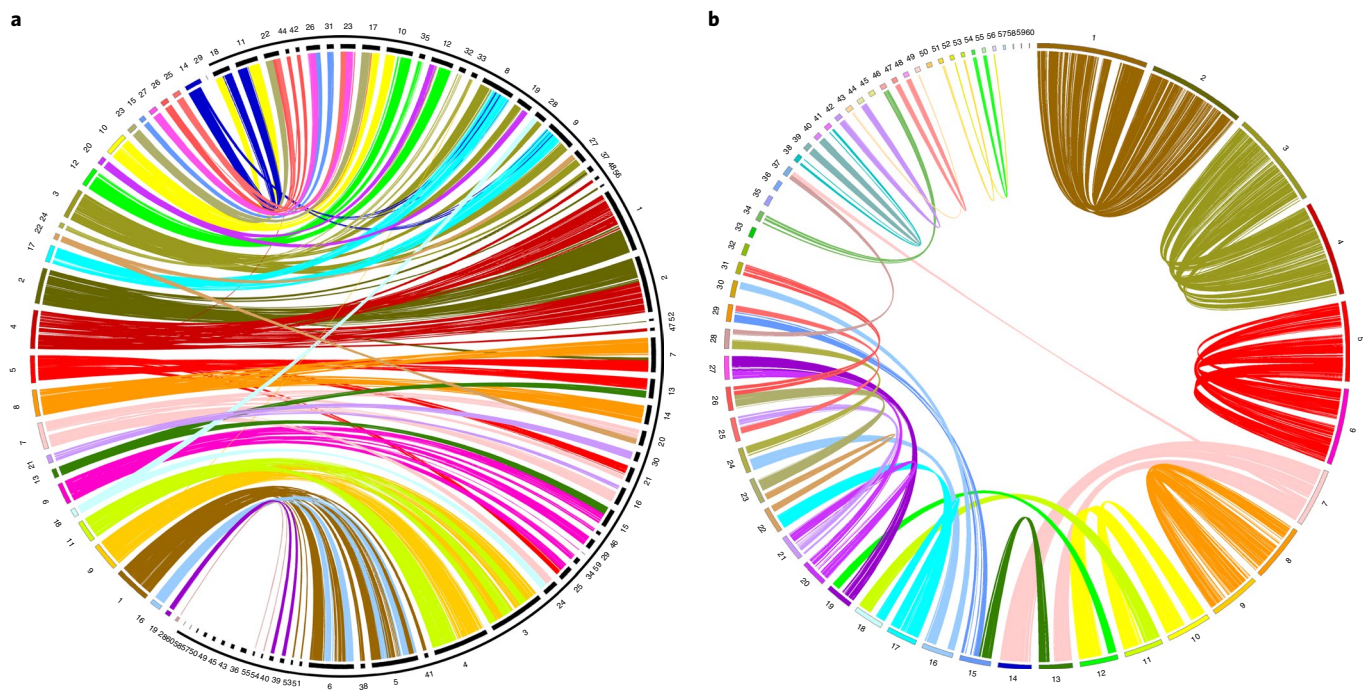
## Results

**Genome assembly and annotation.** Polyploid genomes are extremely challenging for de novo assembly because of the coexistence

of ohnologous and allelic sequences of each original locus with various degrees of sequence similarities. To generate a high-quality reference sturgeon genome, we produced 42-fold coverage of Illumina sequences, 54-fold coverage with PacBio long reads and 20-fold coverage of Hi-C sequences of the estimated 1.8-gigabase (Gb) genome of a male *A. ruthenus*<sup>21</sup>. For the assembly process, we considered possible complications owing to the simultaneous presence of polyploidy and heterozygosity (Supplementary Note 1). After reduplication and Hi-C scaffolding, we produced a 1.8-Gb assembly with a final N50 scaffold size of 42.4 megabases (Mb) (Supplementary Fig. 1, and Supplementary Tables 1 and 2). The 60 largest scaffolds correspond to 120 chromosomes of the sterlet karyotype. The chromosome number of *A. ruthenus* can vary, however, by two to four small chromosomes, indicating the occurrence of B chromosomes<sup>22</sup>. B chromosomes are enigmatic accessory elements to the regular chromosome set. They are found in some but not all individuals within a population and are considered to be either non-functional, beneficial or harmful<sup>23</sup>. Scaffold 60 consists mainly of interspersed repetitive DNA (83.9%) and contains only three corrupted gene remnants, thus probably representing a fully assembled B chromosome (Supplementary Table 2 and Supplementary Note 2).

Genome annotation combined gene evidence from homology annotation, de novo annotation and transcripts with a previously established pipeline<sup>24</sup>. We predicted 47,424 protein-coding genes. BUSCO analysis revealed that the annotation contains 2,543 (98.3%) out of 2,586 conserved and complete vertebrate genes (Supplementary Table 3).

**Ancient origin and slow evolution.** Sturgeons are one of the most deeply diverging groups of bony fishes and have been referred to as both the Leviathans and Methuselahs of freshwater fish. They appear in the fossil record between 250 and 200 Ma, near the end of the Triassic. Our phylogenomic trees place the sterlet sturgeon basal to the other ray-finned fishes (Fig. 1 and Supplementary Fig. 2),



**Fig. 2 | Homology and homeology relationships of sterlet chromosomes.** **a**, Chord diagram displaying the gene orthologies between 29 spotted gar chromosomes (left, coloured) and 60 sterlet chromosomes (right, black, bracketed by outer black partial circle) on the basis of 21,085 orthologous pairs (pairwise synteny was confirmed by the criterion of at least four orthologous genes, arranged in a row with the largest gap being fewer than 15 genes). **b**, Chord diagram depicting homeology relationships of 60 sterlet chromosomes on the basis of 9,301 ohnologue pairs (pairwise synteny was confirmed by the criterion of at least five ohnologues, arranged in a row with the largest gap being fewer than 15 genes). The chromosomes are ordered by size.

in agreement with the current tree of life<sup>25–28</sup>. Divergence time inference based on 275 one-to-one orthologues revealed that the sterlet lineage had already diverged from the actinopterygian fish 345 (295–400) Ma during the Upper Devonian or Carboniferous period (Supplementary Fig. 3), in the range of earlier estimates<sup>28</sup>.

Because extant sturgeons show remarkably little morphological change compared with fossils from the Triassic and because most of the 27 extant species differ relatively little except in body size<sup>29</sup>, Charles Darwin called them living fossils<sup>30</sup>. We therefore asked whether the morphological stasis in sturgeons is matched by a slowly evolving genome as inferred from the slower substitution rates of several mitochondrial and nuclear genes<sup>31</sup>. Calculations of pairwise distances from phylogenetic trees (Supplementary Table 4) revealed that proteins in sterlet are indeed evolving much more slowly than in teleosts, including basal species such as arowana and arapaima. The rate of protein evolution is even slower than in gar, and similar to those basal lineages such as coelacanth or elephant shark (Fig. 1, Supplementary Tables 4–6 and Supplementary Note 3).

The repeat content (40.3%) and transposable element (TE) composition (Supplementary Table 7) of the sterlet genome are comparable to those in other fish (teleosts, gar, elephant shark and coelacanth) studied so far<sup>32</sup>. Despite representing an old, slowly evolving lineage, the inferred transposon activity revealed a recent expansion of all major types of TEs (Supplementary Fig. 4a). The presence of TEs in sterlet transcriptomes, in particular of endogenous retrovirus long terminal repeat (EVR-LTR) retrotransposons and transfer-RNA short interspersed nuclear elements (tRNA-SINEs), indicates that the sterlet retains some active transposons (Supplementary Fig. 4b). The mobilome of the sterlet sturgeon thus seems to be similar to that of many modern fish genomes, including fast-evolving teleosts. This situation contrasts notably with the slow evolution of sterlet protein-coding genes, but

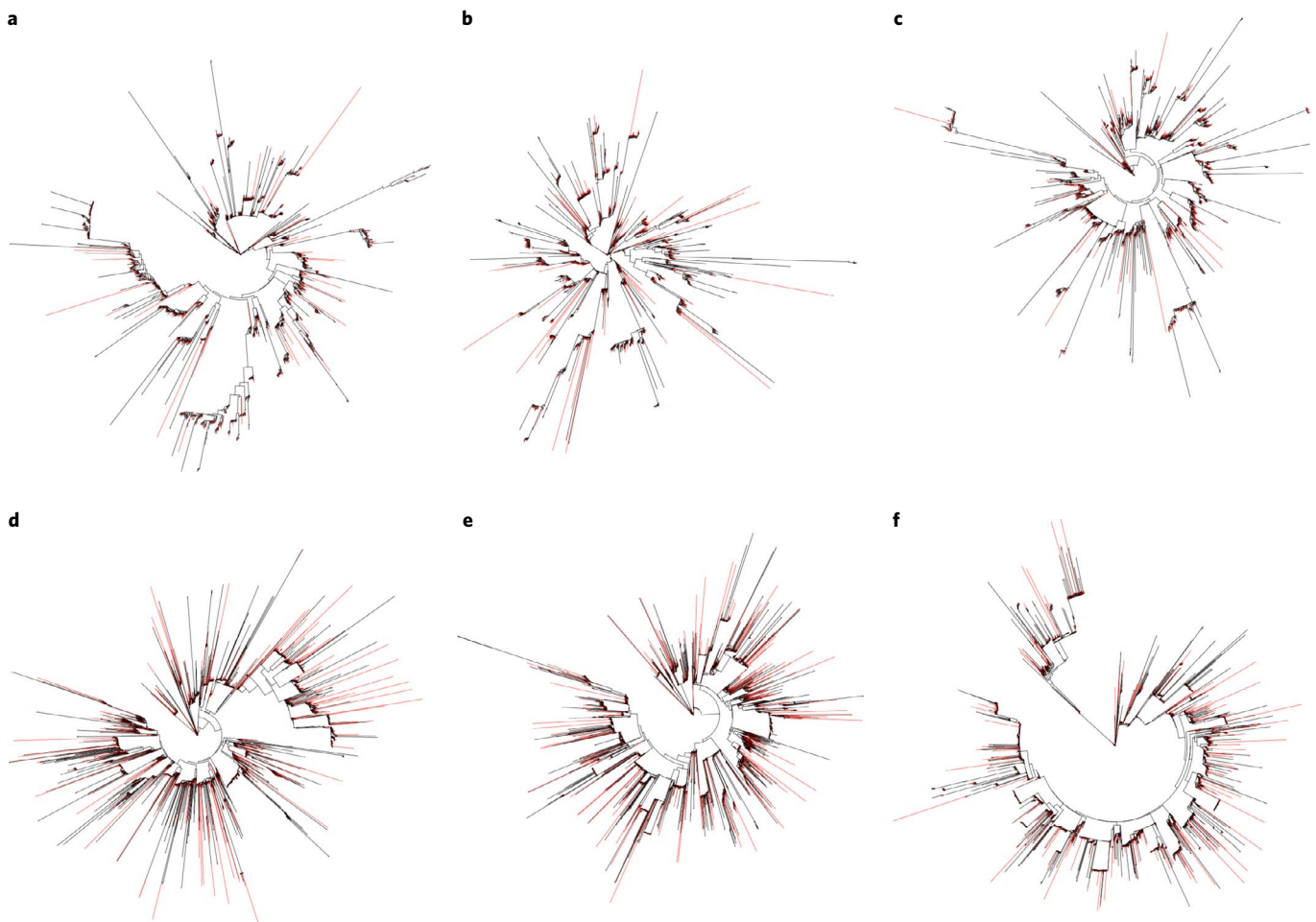
recently expanding TEs and slow protein evolution also occur in the coelacanth genome<sup>33</sup>.

**The sterlet WGD and its initial rediploidization.** Cytogenetic and microsatellite data supported the notion that polyploidy is a general feature of sturgeons. We identified 11,765 genes that have two copies in sterlet but only a single-copy orthologue in gar, coelacanth or elephant shark. We further identified in sterlet 9,914 high-fidelity ohnologue pairs with positional orthology (Supplementary Table 8). A comparison with gar revealed double conserved synteny for 8,752 genes (Supplementary Table 9). This all indicates a WGD in the sterlet lineage (Ars3R) (Supplementary Fig. 5).

To estimate the timing of the Ars3R event, we calculated the pairwise synonymous substitutions per synonymous site (dS) value among sterlet ohnologue pairs (median, 0.064) and between sterlet and one-to-one orthologues of five other sturgeon species (<http://publicsturgeon.sigenae.org/home.html>) (Supplementary Note 4). On the basis of our timing of the sterlet–Atlantic sturgeon (*A. oxyrinchus*) divergence at 166 (115–208) Ma (Supplementary Fig. 6a) and the dS value between their orthologous pairs (median, 0.059; Supplementary Fig. 6b and Supplementary Table 10), we deduced that the sterlet WGD must have happened around 180 (124–225) Ma. Thus, the Ars3R genome duplication event is older than the salmonid WGD at 80–100 Ma (refs. <sup>34,35</sup>) and the carp–goldfish 4R estimated at 14 Ma (ref. <sup>36</sup>).

The analysis of conserved synteny between sterlet and gar revealed that most gar chromosomes have two counterparts in sterlet (Fig. 2a). When sterlet ohnologue gene pairs were mapped against the genome scaffolds, they delineated 46 scaffolds, also in a pairwise fashion. This result indicates homeologous chromosome segments, as expected from a WGD event (Fig. 2b, Supplementary Figs. 7 and 8, and Supplementary Notes 5 and 6). To confirm this conclusion, we used sequence libraries, prepared from individual





**Fig. 3 | Phylogeny of DNA/PIF-Harbinger and DNA/TcMar-Tc1 repeat families on homologous chromosomes. a**, DNA/PIF-Harbinger on homologous chromosomes 1 (red) and 2 (black). **b**, DNA/PIF-Harbinger on homologous chromosomes 3 (red) and 4 (black). **c**, DNA/PIF-Harbinger on homologous chromosomes 5 (red) and 6 (black). **d**, DNA/TcMar-Tc1 on homologous chromosomes 1 (red) and 2 (black). **e**, DNA/TcMar-Tc1 on homologous chromosomes 3 (red) and 4 (black). **f**, DNA/TcMar-Tc1 on homologous chromosomes 5 (red) and 6 (black).

microdissected chromosomes or chromosome arms of the sterlet<sup>37,38</sup>. In whole-mount in situ-hybridizations, each of these probes painted two pairs of sterlet metaphase chromosomes and chromosome arms, respectively, identifying likely ohnologous pairs. Reads from each of the libraries aligned specifically to individual scaffolds, which thereby could be assigned to either of the homeologous chromosome segments (Supplementary Figs. 9 and 10, Supplementary Note 6 and Supplementary Data 1).

Remarkably, most of the large homeologous chromosomes (1–6, 8 and 9) are conserved over their full length, while the majority of the intermediate-sized chromosomes have ohnology-relationships to two other chromosomes. The alignment of chromosomes by LAST indicated that whole chromosome arms were exchanged, most probably in reciprocal translocation events (Supplementary Figs. 8 and 11, and Supplementary Note 5).

Interestingly, the remaining 11 scaffolds, corresponding to smaller chromosomes, contain exclusively singletons or only a small region with ohnologues on another chromosome, while the remainder of the chromosome only contains singletons. Those small ohnologue regions are obviously translocations from other chromosomes (Supplementary Fig. 12a). We conclude that the entire homeologue or the majority region of the counterparts of those smaller, whole-chromosome-representing scaffolds, were lost after the Ars3R (Supplementary Fig. 13). This result indicates that a relevant part of the deduplication process in sterlet occurred by the

loss of whole chromosomes or large chromosome fragments and is segmental. This mechanistic conclusion is in contrast to the continuous and genome-wide small-scale ohnologue-by-ohnologue loss in carp/goldfish and salmonids (Supplementary Fig. 12b–d). Earlier molecular cytogenetic studies of sterlet also pointed to a karyotype that is segmental rather than ubiquitously polyploid<sup>38</sup>. Such large-scale reduction of duplicates in polyploid organisms, through the loss of whole chromosomes or large chromosome segments, has so far been reported only in autotetraploid yeasts<sup>39,40</sup>, flowering plants<sup>41</sup> and endopolyploid human cancer cells<sup>42</sup>.

Polyploidy can result from duplication of the whole genome in one organism (autopolyploidy) or from the interbreeding of two divergent species with subsequent genome doubling that restores meiotic pairing and disomic inheritance (allopolyploidy). Both of these mechanisms—interspecific hybridization and autopolyploidization—have been discussed to account for the origin of the sterlet chromosome complement, on the basis of conflicting evidence<sup>12</sup>. To clarify this controversy, we used a strategy that was employed to investigate this problem in the allopolyploid African clawed frog, *Xenopus laevis*, where the fast-evolving repeats and relics of the mobilome are specific to the allopolyploid ancestors, and thus markers for the ancestral chromosomal segments of the two parental species<sup>43</sup>. A comparison of the TE landscape of sterlet paralogous chromosomes revealed that each pair has an almost identical TE content and that individual TE families are monophyletic (Fig. 3

and Supplementary Fig. 14). The sterlet genome thus shows no evidence for allopolyploidy.

Chromosomes that have retained a homeologous partner share to a large extent even their gene order (Supplementary Figs. 7 and 8). This phenomenon has also been observed in many polyploid plant species and is called positional orthology<sup>44,45</sup>. It is explained as a consequence of multivalent pairing in meiosis. Multivalent pairing would also explain tetrasomic inheritance in sterlet, noted earlier from microsatellite studies<sup>12</sup>.

The duplication of a whole genome creates a situation, where one of the two copies is in principle dispensable. The retention of duplicates is explained by several models<sup>46</sup>. They may be preserved if one copy evolves a new positively selected function and simultaneously loses the essential function retained by the other copy (neofunctionalization) or if ancestral positively selected functions partition between the two copies (subfunctionalization)<sup>9</sup>. The gene balance hypothesis posits that ohnologues persist because the loss of one copy would lead to a detrimental change in the stoichiometry of macromolecular complexes, the interactome and signalling pathways<sup>47</sup>. The majority of duplicates, however, are predicted to become non-functional or get lost (degeneration)—for example, the ohnologue retention rate from the teleost WGD in the extant teleosts is estimated to be only 15–20%<sup>48</sup>. On the basis of non-coding microsatellites, the sterlet was proposed to have undergone extensive duplicate gene degeneration and has been classified since then even as a functional diploid species<sup>10</sup>. To estimate the duplicate retention rate, we identified 9,914 high-fidelity pairs of ohnologues and 4,175 singletons (Supplementary Note 7). This dataset represents a duplicate retention rate of 70% (Supplementary Table 11), considerably higher than in all teleosts, including the 4R salmonids (Supplementary Note 8). Considering functional terms, we found that sterlet ohnologues are enriched for transcriptional regulators (genes involved in protein turnover, signal transduction, cell proliferation and development), in agreement with predictions from the gene-balance-hypothesis<sup>47</sup>. Sterlet singletons are enriched for genes with functions in DNA metabolism, intracellular transport and mitochondria. Enrichment for such categories has been observed in other polyploids, even in plants<sup>49–51</sup> (Supplementary Table 12). Like the situation reported for rainbow trout<sup>34</sup>, we found the coding sequence of singletons to be significantly shorter than that of ohnologues (12%,  $P < 2.2 \times 10^{-16}$ ; Supplementary Fig. 15). Long genes may be over-retained as ohnologues, potentially owing to more opportunities for protein domain subfunctionalization.

In our analysis of transcriptomes from 23 different sterlet organs and developmental stages, we observed the expression of one or both genes for 9,243 of the 9,914 ohnologue pairs. We found 1,139 ohnologue pairs, which showed equal expression in all samples (Supplementary Fig. 16a). We then searched for genes with differing expression patterns among samples, which would be explained by drift models of expression change or would indicate the degeneration or neofunctionalization of one duplicate, or subfunctionalization of both copies. We found 3,230 ohnologue pairs with different expression in at least two samples (Supplementary Fig. 16b and Supplementary Note 8). From just 38 of these ohnologue pairs, only one of them was expressed but never the other in all organs tested. Such a pattern is expected if regulatory elements are degenerating in the redundant copy. For 341 ohnologue pairs, the expression of duplicates was partitioned between different organs or developmental stages. This may indicate subfunctionalization of this subset of genes.

The availability of the sterlet genome now allows the revisitation of important questions concerning the forces that affect the evolutionary fate of gene duplicates. We compared the genomes of sterlet, salmon, trout, goldfish and zebrafish, using gar as the outgroup, to find genes that were commonly retained in duplicate after the various polyploidization events (Supplementary Note 9). We found only

27 such genes (Supplementary Figs. 17a and 18, and Supplementary Table 13). This finding suggests complex, independent, lineage-specific evolutionary processes of duplicate retention.

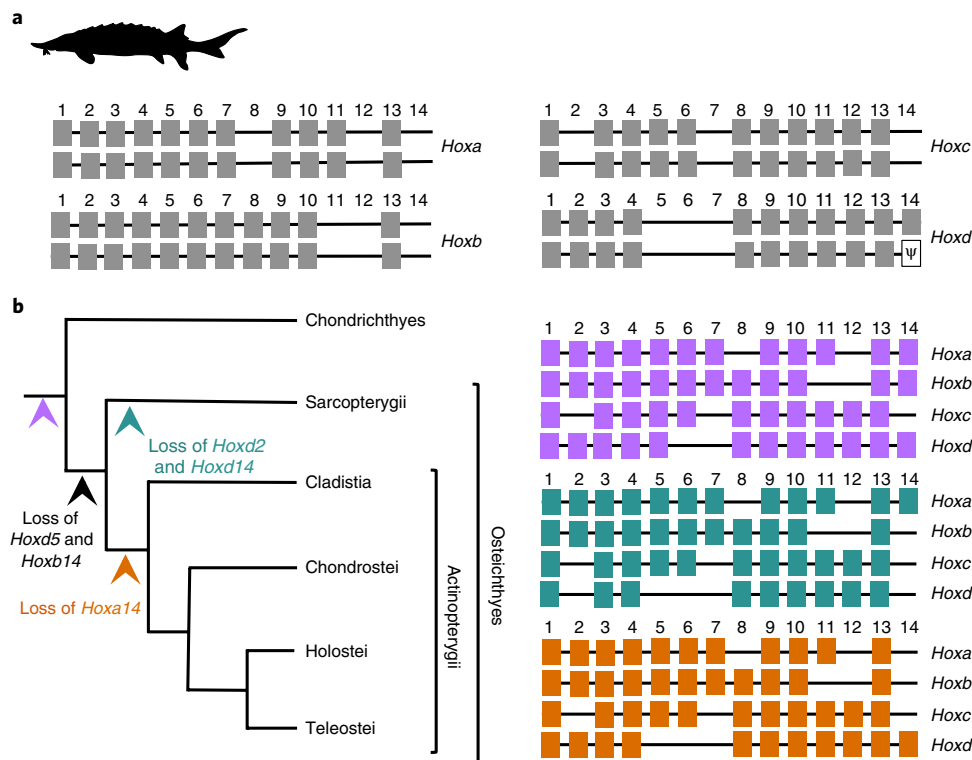
In the same set of species, we identified 191 genes that are singletons in all of them (Supplementary Note 9). Notably, 39 of these singletons are arranged in eight syntenic blocks. A similar phenomenon was seen for the commonly retained ohnologues (Supplementary Figs. 17b and 19, and Supplementary Table 14). The loss or retention of linked genes after WGDs could be explained by the functional relationships of their gene products—for example, through protein–protein interactions<sup>52</sup>. However, a search of singleton genes, embedded in syntenic blocks using the STRING<sup>53</sup> database, did not reveal such protein–protein interactions. An alternative explanation for the conservation of microsynteny is the bystander relationship<sup>54</sup>, where the regulatory region of one gene is located in neighbouring genes. Further studies are required to validate this type of physical association of genes on chromosomes over long evolutionary times rather than functional relationships of their encoded proteins.

**Genome and gene evolution. Positive selection.** Up to 210 genes (Supplementary Table 15) in sterlet are under positive selection, depending on the set of actinopterygian or vertebrate genomes, with which its full gene complement was compared (Supplementary Table 16). Positively selected genes spanned a wide spectrum of cellular and molecular functions and pathways with no particular enrichment.

When the ratios of substitution rates at non-synonymous versus synonymous sites (dN/dS values) were compared between sterlet singletons and ohnologues, we found that most retained ohnologues present higher dN/dS values than singletons (Supplementary Fig. 20), indicating relaxed purifying selection on ohnologues. This result would be expected because of ohnologue redundancy<sup>55,56</sup>. A pairwise test of dN/dS for the 9,914 ohnologue pairs revealed that 207 are under positive selection in sterlet, pointing to neofunctionalization or subfunctionalization at the protein level (Supplementary Table 17). Notably, many immune-related genes are positively selected, indicating that the sterlet host defence system may have made have especially profited from the WGD for evolutionary progress *sensu* Susumo Ohno<sup>57</sup>. A similar phenomenon was observed for duplicated immune genes in salmon<sup>58</sup>.

**Dynamics of gene family size.** We compared the rates of gene family (8,150 gene families) dynamics between phylogenetic tree branches with different WGD histories and found that gene family sizes changed much faster in branches with 4R and Ars3R than in branches with more ancient polyploidization (Supplementary Note 10). Interestingly, one of the most expanding families is the *zona pellucida* (Zp) sperm-binding proteins (ID: 4190). Zp-proteins prevent polyspermy in mammals<sup>59</sup> and provide thickness and hardness to the fish egg envelope<sup>60</sup>. A total of 116 *zp* genes were annotated in sterlet (Supplementary Table 18 and Supplementary Note 11). A similar expansion was noted in cold-adapted teleosts and explained as a protection mechanism from physical forces for the developing embryo<sup>61,62</sup>. The biological reason for the *zp* gene family expansion in sturgeon is unclear. Because sturgeons spawn on a coarse substrate often in high current velocities, a hard envelope provides protection against mechanical stress of the adhesive eggs on the spawning substrate as well as against polyspermy that would be possible through the multiple micropyles of their eggs. This biological feature might contribute to the crispness of the caviar.

**Evolution of sterlet *hox* clusters after genome tetraploidization and inference of the ancestral vertebrate *Hox* complement.** The sterlet has eight *hox* clusters containing 88 genes, reflecting the 1R/2R/3R history of its genome (Fig. 4a and Supplementary Note 12). Pseudogenization was apparent for only one *hoxd14* gene.



**Fig. 4 | Structure and evolution of *hox* clusters.** **a**, Schematic illustration of the sterlet *hox* complement. We identified 88 *hox* genes plus one pseudogenized *hoxd14* gene (indicated by psi). All *hox* clusters are retained in duplicate. **b**, Reconstruction of the ancestral actinopterygian condition and the inference of gene losses across the gnathostome phylogeny on the basis of the sterlet pretetraploidization *hox* complement in combination with that of the gar. The inferred ancestral *Hox* complements are shown in purple (likewise indicated by the purple arrowhead in the tree) for gnathostomes, in blue for Sarcopterygii and in orange for Actinopterygii.

The sterlet therefore retains the most complete 3R *hox* cluster duplicates and the highest number of 3R *hox* gene ohnologues amongst ray-finned fish. The comparison of the *hoxd* flanking gene deserts, containing long-range regulatory elements<sup>63–66</sup>, indicates high conservation of ultraconserved elements (Supplementary Fig. 21). The preservation of all duplicated *hox* clusters as well as their low divergence, including that of their regulatory regions, shows a remarkable slow evolution of these genomic loci. This stability contrasts sharply with rapidly evolving teleosts, which often show extensive remodelling of duplicated *hox* clusters<sup>4,67–73</sup>.

The *hox* gene complement in sterlets indicates an identical pretetraploidization *hox* gene arrangement and repertoire with the gar (diverging ~345 Ma). Because both species represent early-branching ray-finned fish, this similarity strengthens the scenario whereby *hoxd5* and *hoxb14* were lost in the common ancestor of bony vertebrates (Euteleostomi) and *hoxa14* in the common ancestor of actinopterygians<sup>66</sup> (Fig. 4b).

**Over-retention of glutamate receptor genes.** Glutamate receptor genes (GRGs) show particularly high ohnologue retention rates in teleosts<sup>74</sup>, which has been connected to the extraordinary cognitive abilities of many teleost species compared with other basal vertebrates. We found that 23 of 26 GRGs retained their Ars3R ohnologue, an ohnologue retention rate of 88.5% (Supplementary Fig. 22, Supplementary Table 19 and Supplementary Note 13). Compared with the genome-wide rate of 70% (9,914 ohnologs and 4,175 As3R singletons), the GRG Ars3R ohnologue retention rate is significantly higher ( $P=0.04345$ , chi-square test). GRGs have thus been convergently over-retained, following the Ars3R and teleost 3R WGD, although to a lower extent in sturgeons.

**Absence of differentiated sex chromosomes.** The relative rarity of polyploidy in animals versus plants has been ascribed to the disruption of sex determination in gonochoristic animals after genome duplication<sup>75–77</sup>. Differentiated sex chromosome pairs have largely different gene contents, to which many animals have adjusted by elaborate expression dosage compensation mechanisms. The disturbance of dosage compensation and the disruption of the chromosomal system that determines the sex ratio are thus immediate negative consequences of polyploidization<sup>78</sup>. Data from induced gynogenesis led to the common belief that all *Acipenser* species, including sterlet, have a female heterogametic (ZZ/ZW) sex chromosome system<sup>79,80</sup>. To find out if the polyploid sterlet has differentiated sex chromosomes, we searched for sex-linked sequence differences using a restriction site associated DNA (RAD) sequencing approach. A total of 176,735 markers were obtained, but none showed a bias or specificity for males or females (Supplementary Fig. 23). This result indicates that the sterlet does not have sex chromosomes with considerable sequence differentiation that would require dosage compensation and impair the occurrence of polyploidy. Our data are in agreement with the absence of differences in chromosome morphology and previous failures to isolate sex-specific molecular markers<sup>81</sup>.

## Discussion

The high-quality chromosome-level genome of the sterlet sturgeon permitted important advances in our understanding of the evolution of this lineage of ancient fish. Our results show that the sterlet lineage branched from the vertebrate tree of life about 345 Ma, shortly after the basal split between the lineage of ray-finned fish and that of lungfish, coelacanth and land vertebrates happened. While



the sterlet's slow evolutionary rate of protein-coding genes is not entirely unexpected, given the morphological stasis exhibited in the sturgeon lineage, many of the features of the sterlet's polyploid genome are much different from those of other polyploid lineages. Clearly, genomic and phenotypic evolution do not always march to the beat of the same drummer.

All sturgeons are characterized by polyploidy as a genetic hallmark and paramount feature. It has been proposed that those extant sturgeons with ~120 chromosomes (like the sterlet) represent functional diploids, which originated over 200 Ma by a WGD of a 60-chromosome diploid ancestor<sup>82</sup>. The transition between the ancestral fully tetraploid and the modern functional diploids was proposed to have been accompanied by a reduction of duplicate gene functions<sup>12</sup>. Our estimate of 180 Ma for the Ars3R provides evidence for a WGD in the ancestor of all sturgeons, and that the WGDs that led to the ~240- and ~360-chromosome species happened later, on top of the Ars3R. We found that despite the long evolutionary time that has elapsed since the sturgeon WGD, the sterlet has not returned to a diploid state by gene content or gene expression. Instead, the sterlet has retained an unexpectedly high degree of structural and functional polyploidy. This retention can be ascribed to the slow pace of molecular evolution of most fractions of the sterlet genome.

The slow evolution may also explain why the sterlet genome in several aspects represents an earlier step in the process of redundancy-reduction than the salmonid genomes, which originated from a more recent WGD. During the evolution of a polyploid genome, the initial one-to-one relationship of whole chromosomes (as still seen in the goldfish) is reduced to homeology between arms of chromosomes and then further to much smaller regions (as evident in salmonids). Sterlet seems to be in the transition towards the highly dynamic pattern of colinear duplicated blocks, but still has some fully homeologous chromosomes (Supplementary Fig. 24).

A recent wave of TE multiplication apparently swept through the sterlet genome after the Ars3R. The large-scale expansion and movement of TEs are known to increase under genomic stress<sup>83</sup>, suggesting that WGDs cause TE activation. TE expansions in the centromere induce chromosomal instability<sup>84</sup> and might have facilitated the large chromosome rearrangements of homeologue arm changes.

The timing of the Ars3R to have evolved earlier than the cypripinid and salmonid 4Rs allows comparisons of the three apparent tetraploid lineages to give insights into the processes of polyploid genome evolution. Despite its apparent evolutionary advantage as a source of genomic matter for evolution in the long term, tetraploidy seems to be an evolutionarily unstable situation. In all known instances, the initial dispensability of two sets of genes led to deduplication of the genome, with only a certain fraction of gene duplicates being retained.

The process of duplicate gene loss after the teleost, salmonid and goldfish WGDs affected the whole genome in a homogenous fashion. Unexpectedly, the sterlet genome analysis uncovered a phenomenon that creates a segmental rather than a continuous partial tetraploidy. In the sterlet, most chromosomes or chromosome arms were found to be in either a diploid or a tetraploid state. The loss of entire chromosomes can be seen as a fast stochastic process for rediploidization.

The numbers of genes that were either commonly retained or deduplicated after the WGDs in the fish lineages are substantially above random but are much lower than one would expect if strong adaptive processes determined duplicate retention or loss on the single-gene level. This conclusion, and our finding that structural features rather than protein–protein interactions are relevant for the deduplication of neighbouring genes, suggest complex processes of different lineage-specific evolutionary drivers of duplicate retention, and largely stochastic events in redundancy reduction.

In sterlet, besides the adaptive evolutionary mechanisms, neutral processes have considerably shaped its genome, most obviously manifested by the loss of whole chromosomes from homeologues pairs.

## Methods

**Experimental animals.** All fish used in this study were derived from the sterlet sturgeon population maintained at the Leibniz-Institute of Freshwater Ecology and Inland Fisheries. This stock is derived from the Danube population of *A. ruthenus*. Adult individuals were sexed by gonad morphology and gamete content. The fish were euthanized by state-of-the-art humane killing (American Veterinary Medical Association, Canadian Council of Animal Care in Science). The experiments were carried out in accordance with the European Directive 2010/63/EU and German national legislation (animal protection law, TierSchG). All experimental protocols that are part of this study were approved through an authorization (File No. ZH 114, issued 6 February 2014) of the LAGeSo, Berlin, Germany.

**Genome sequencing and assembly.** The DNA for sequencing was derived from the testis and blood of a single adult male. We generated ×42 Illumina reads (150-base-pair (bp) paired end) on a Novaseq 6000 platform with libraries produced using the TruSeqDNA PCR-Free kit. A 53.7-fold coverage of genome sequences was produced with PacBio Sequel technology. Hi-C library generation was carried out according to a protocol adapted from Foissac et al.<sup>85</sup>. A blood sample was spun down, and the cell pellet was resuspended and fixed in 1% formaldehyde. Five million cells were processed for the Hi-C library. After overnight digestion with HindIII (NEB), the DNA ends were labelled with Biotin-14-DCTP (Invitrogen), using Klenow enzyme (NEB), and then religated. Next, 1.4 µg of DNA were sheared by sonication (Covaris) to an average size of 550 bp. Biotinylated DNA fragments were pulled down using M280 Streptavidin Dynabeads (Invitrogen) and ligated to paired-end adaptors (Illumina). The Hi-C library was amplified using paired-end primers (Illumina) with 10 PCR amplification cycles. The library was sequenced using HiSeq3000 (Illumina) generating 150-bp paired-end reads at 20-fold genome coverage.

The raw Sequel BAM files were converted into subreads in fasta format using the SMRT Link software package (v.5.0.1) from Pacific Biosciences<sup>86</sup>. PacBio reads were assembled with smartdenovo (v.1.0)<sup>87</sup> with standard parameters. Contigs were polished with two rounds of racon<sup>88</sup> (v.1.3.1), using long reads aligned with minimap2 (ref. <sup>89</sup>) (v.2.7) and three rounds of pilon<sup>90</sup> (v.1.22), using 42-fold Illumina reads. The Illumina reads were aligned with bwa mem (v.0.7.12-r1039)<sup>91</sup> with standard parameters and the same file, which had been compressed, sorted and indexed with samtools view, sort and index v.1.3.1<sup>92</sup>, using standard parameters before pilon polishing. The genome size was 15% smaller than expected, and a fraction of the contigs showed twice the expected read alignment depth, indicating that chromosome parts had merged during assembly. The single- and double-copy coverage threshold was found by visual inspection of the contig coverage bimodal distribution, and the contigs were separated into two sets, corresponding to single and double coverage. A polymorphism VCF file was generated from the short read alignment file with freebayes<sup>93</sup> (v.1.1.0) under standard parameters. The VCF file shows an overall much higher variation density in double coverage contigs. PacBio long reads were used in the next steps to generate haplotypes of these variations to split the genomic locations that had been merged. Long reads were aligned to contigs, and the alignments of double coverage contigs were processed with HapCut2<sup>94</sup> (v.1.0) using the following parameters: extractHAIRS -ont 1 and HAPCUT2 -ea 1. For each contig, a haplotyped VCF file was produced. Some of these files contained more than one haplotypic segment. These contigs have been split according to the haplotypic segment information found in the VCF file, using an in-house script. The resulting haplotyped VCF files were then processed with fgbio (v.0.7.0 using standard parameters)<sup>95</sup> to generate VCF files, separated by haplotype. These VCF files and the reference were used to produce haplotypic contigs using vcf-consensus from the bcftools<sup>96</sup> package v.1.8 under standard parameters. Both contig sets, unique and split, were then merged using the Unix cat command. The Hi-C short reads were aligned to the contigs with Juicer<sup>97</sup>, and the scaffolding was performed with 3D-DNA<sup>98</sup> with parameter -r = 0. Finally, the candidate assembly was manually reviewed using the Juicebox Assembly Tools<sup>99</sup>. The contig metrics were calculated with the assemblathon\_stats.pl script.

**Repeat annotation and TE analysis.** To search for repeated elements, the sterlet genome and raw Illumina reads were used as input. The assembled genome was used in the RepeatModeler open-1.0.11 tool<sup>100</sup> with standard settings. LTR-retriever v.2.5 (ref. <sup>101</sup>) was used to search for full-length LTR elements, and the data were used as input derived from the LTRharvest<sup>102</sup> (-similar, 90; -vic, 10; -seed, 20; -seqids, yes; -minlen, 100; -maxlen, 7,000; -mintsd, 4; -maxtsd, 6; -motifms, 1) and LTR\_FINDER<sup>103</sup> (-D, 15,000; -d, 1,000; -L, 7,000; -l, 100; -p, 20; -CM, 0.9) tools. To exclude non-LTR (-linelib) and DNA transposons (-dnalib), protein sequences of these TEs from the RepeatPePs database of the RepeatMasker tool<sup>104</sup> were used. This also excluded protein sequences that were not related to TEs. The SWISS-PROT sequence library<sup>105</sup> was also used (-plantprotlib).

The sequences obtained using the previous steps were combined into a single FASTA file using CD-HIT-est<sup>106</sup> (-aS, 1; -c, 1; -r, 1; g, 1; p, 0). The resulting FASTA



file was aligned against the RepBase v.24.07 (ref.<sup>107</sup>) and FishTEDb<sup>108</sup> databases using blastn (-evalue,  $10 \times 10^{-100}$ ) and against SWISS-PROT and RepeatPeps using blastx (-evalue,  $10 \times 10^{-100}$ )<sup>109</sup> to filter incorrectly annotated sequences.

Raw reads were used in the TAREAN tool<sup>110</sup>, which is part of RepeatExplorer<sup>111</sup>. The reads were first trimmed using the fastp tool<sup>112</sup> to remove low-quality and adapter sequences (detect\_adapter\_for\_pe -g -c -l 50 -5 -3), after which RepeatExplorer was used with standard settings. We saved only satellite sequences with high confidence and added them to the library of repeated sequences. In addition, using REXdb<sup>113</sup>, a database of TE domains implemented in RepeatExplorer2, the correctness of the previous TE annotation was further verified. The content of repeated elements in the genome was estimated using RepeatMasker open-4-0-9-p2 (-s -no\_low -lib). To build the Kimura plot, the createRepeatLandscape.pl script from the RepeatMasker tool was used.

To analyse the expression of TEs, raw reads from RNA-seq were used. The reads were trimmed using fastp (-detect\_adapter\_for\_pe -g -c -5 -3) and then aligned against the FASTA file containing TE sequences obtained in the previous step using bowtie2 v.2.3.5.1 (ref.<sup>114</sup>) (-very-sensitive -dovetail). The raw read count for each superfamily was calculated. The raw counts were normalized to the total number of sequences (reads per million, the number of aligned reads for each superfamily  $\times 1000000$ /total number of reads), and then the proportion of superfamilies in the transcriptome was calculated (reads per million  $\times 100$ /total number of aligned reads). To compare the RNA-seq data with the genome proportion of the respective TE superfamily, the proportion of TEs in the genome was calculated (the number of nucleotides occupied by superfamily in the genome  $\times 100$ /total nucleotides occupied by TEs in the genome). The results were transformed to the log<sub>10</sub> values and visualized with ggplot2<sup>115</sup> and MATLAB<sup>116</sup>.

**Genome annotation.** Genome annotation was done by an in-house pipeline (Supplementary Fig. 25) improved from a previous version<sup>24</sup>. First, the pipeline assessed the assembly quality using BUSCO on the basis of the Actinopterygii odb9 database<sup>117</sup>. The parameter -long was used for the first training of AUGUSTUS v.3.2.3 (ref.<sup>118</sup>). The pipeline then identified and masked repeat elements from the assembly. Repeat elements were identified using blastx v.2.2.28+ with the protein repeat database RepeatPeps (<http://www.repeatmasker.org/>), and using RepeatMasker with two nucleotide repeat databases, one produced by RepeatModel (<http://www.repeatmasker.org/>), and the other an in-house fish repeat database combining our annotation and the one from Shao et al.<sup>108</sup>. Simple and low-complexity repeats were then softmasked, while those with known family were hardmasked. After repeat masking, the pipeline collected gene evidence from homology annotation, de novo annotation and RNA-seq annotation. For homology annotation we first pooled protein sequences from SWISS-PROT ([www.uniprot.org](http://www.uniprot.org)) and 13 Ensembl genomes (v.95, <http://www.ensembl.org>): human (*Homo sapiens*), mouse (*Mus musculus*), coelacanth (*Latimeria chalumnae*), spotted gar (*Lepisosteus oculatus*), zebrafish (*Danio rerio*), cod (*Gadus morhua*), tilapia (*Oreochromis niloticus*), medaka (*Oryzias latipes*), platyfish (*Xiphophorus maculatus*), fugu (*Takifugu rubripes*), tetraodon (*Tetraodon nigroviridis*), stickleback (*Gasterosteus aculeatus*) and sea lamprey (*Petromyzon marinus*), and reduced the redundancy using CD-HIT (<http://www.bioinformatics.org/cd-hit/>), which resulted in 544,476 proteins. These were mapped to the assembly using exonerate v.2.2.0<sup>119</sup> and Genewise2-2.0 (ref.<sup>120</sup>) respectively. Before Genewise was implemented, GenBlastA1.0.1 (ref.<sup>121</sup>) was used to roughly locate each protein on the assembly. For de novo annotation, SNAP v.2006-07-28 (<http://korflab.ucdavis.edu>) and GeneMark-ES<sup>122</sup> were independently used. For RNA-seq annotation, RNA-seq reads from juvenile male mixed organs, adult male muscle, spleen, skin, testis, female brain, liver and ovary were mapped and assembled using Tophat and cufflinks v.2.1.1 (ref.<sup>123</sup>). In parallel, HISAT2 v.2.1.0, Trinity v.2.4.0 and PASA v.2.2.0 (refs.<sup>124,125</sup>) were also used for RNA-seq read mapping and assembly. In total, 89.5% of all transcriptome reads mapped to the genome.

All gene evidence obtained from the three kinds of annotation was collected and transferred to EvidenceModeler v.1.1.1 (ref.<sup>126</sup>), where gene models confirmed by all lines of evidence were extracted as high-quality gene models. They were used for the second training of AUGUSTUS. Finally, the AUGUSTUS specially trained for sterlet took all the hints from BUSCO, repeat masking and all three annotations to predict the final set of gene models for sterlet. Some broken or artificial chimaeric gene models were found and replaced by comparing the AUGUSTUS prediction with the homology gene evidence. Low-quality gene models were removed afterwards. To assign gene symbols, their protein sequences were blasted to the SWISS-PROT database ([www.uniprot.org/e](http://www.uniprot.org/e)) (blastp v.2.2.28+ (ref.<sup>127</sup>); percentage of identical matches, >20%; e-value,  $<1 \times 10^{-5}$ ), and the symbol of the best hit was taken (<https://biobdnet-abcc.ncicrf.gov/>)<sup>128</sup>. DeepGO was used to annotate gene ontology terms for each gene<sup>129</sup>.

To annotate non-coding RNAs (ncRNAs), we adapted the method from Ensembl (<http://ensemblgenomes.org/info/data/ncrna>). tRNAs were screened using tRNAscan-SE v.2.0.3 (ref.<sup>130</sup>), and ribosomal RNAs were identified using RNAmmer<sup>131</sup>. The rest of the ncRNAs were then predicted using Infernal with Rfam v.14.1 (ref.<sup>132,133</sup>).

**Orthology assignment.** To infer gene homology among sterlet, *P. marinus* (sea lamprey), *C. milii* (elephant shark), *L. chalumnae* (coelacanth), *L. oculatus* (spotted

gar), *A. gigas* (Arapaima), *S. formosus* (arowana), *O. mykiss* (rainbow trout), *S. salar* (Atlantic salmon), *T. rubripes* (Japanese fugu), *X. maculatus* (platyfish), *O. latipes* (Japanese medaka), *C. auratus* (goldfish), *C. idellus* (grass carp) and *D. rerio* (zebrafish) (see Supplementary Table 20), we used a method that reconciles species trees for the inference of orthologues. We kept the longest protein sequence for each gene and performed an all-against-all blast using blastp v.2.2.28+ with an e-value cut-off at  $1 \times 10^{-5}$  (ref.<sup>127</sup>). Between each two protein sequences, the similarity distance was measured using H-score<sup>134</sup>, on the basis of which all protein sequences were clustered into groups (gene families) using Hcluster\_sg<sup>135</sup> with sea lamprey set as the outgroup. For each group, a gene tree was constructed using TreeBeST v.0.5 (ref.<sup>136</sup>) with the species tree guiding. Then, on the basis of the gene tree, orthology relationships among genes were determined as *n* to *m* (*n* and *m* are positive integers; there are cases where *n* = *m*) using an in-house Perl (<https://www.perl.org>) script.

**Phylogenetic analysis and divergence time estimation.** We reconstructed the phylogenomic tree for sterlet on the basis of one-to-one orthologues across 15 species. These protein sequences were first aligned using MUSCLE v.3.8.31 (ref.<sup>137</sup>); regions with bad quality were then trimmed using trimAl<sup>138</sup> with the following parameters: -gt, 0.8; -st, 0.001; -cons, 60. The resulting alignments were concatenated and transferred to RAxML v.8.2.9 (ref.<sup>139</sup>) for phylogenetic tree reconstruction. The parameter PROTGAMMAAUTO was used to select the optimal amino acid substitution model. Sea lamprey was set as the outgroup, and 100 bootstraps were performed to test for robustness.

For an additional confirmation of the phylogenomic tree, we also used MrBayes v.3.2.6 (ref.<sup>140</sup>). The Markov chain Monte Carlo algorithm was implemented in 3 runs with a total of 6 chains for 500,000 generations. Trees were sampled every 1,000 generations, and in the end the first 25% of the sampling were discarded as burn-in. After the burn-in threshold, the average standard deviation of split frequencies remained  $\leq 0.01$ .

To infer divergence time, we used MCMCTree<sup>141</sup> under a relaxed-clock model (correlated molecular clock) with approximate likelihood calculation and maximum likelihood estimation of branch lengths performed<sup>142</sup>. First, the phylogenetic tree and the coding sequences alignment were imported into baseml<sup>141</sup> to roughly estimate the substitution rate. The substitution model was determined using modelgenerator.jar<sup>143</sup>. Then mcmctree was run for the first time to estimate the gradient and Hessian. The resulting file, out.BV, was then used for the final run of MCMCTree to perform approximate likelihood calculations. The final Markov chain Monte Carlo process was run for 2,005,000 steps. The first 5,000 steps were discarded as burn-in; then 20,000 samples were collected with sampling every 100 steps. We set four fossil calibrations: *O. latipes*–*T. nigroviridis* (~96.9–150.9 Ma), *D. rerio*–*G. aculeatus* (~149.85–165.2 Ma)<sup>144</sup>, *A. gigas*–*S. formosus* (~110–156 Ma)<sup>145,146</sup> and a time for the root (<700 Ma).

**Positive selection analysis.** Protein and complementary DNA fasta files from all fish (Supplementary Table 14) were downloaded. To identify orthologous proteins, all protein sequences were compared with sterlet using inparanoid<sup>147</sup> with default settings. To match proteins and cDNA, sequences were blasted by tblastn, and only 100% hits were kept. Codon alignments for the protein–cDNA sequence pairs were constructed using pal2nal v.14 (ref.<sup>148</sup>). The resulting sequences were aligned by MUSCLE<sup>137</sup> (option: -fastaout), and poorly aligned positions and divergent regions of cDNA were eliminated by Gblocks v.0.91b (ref.<sup>149</sup>) (options: -b4, 10; -b5, n; -b3, 5; -t=c). An in-house script was used to convert the Gblocks output to paml format.

For the generation of a phylogenetic tree as input for the detection of positive selection, sequences from all homologous genes, detected by inparanoid, were concatenated after the selection of conserved blocks by Gblocks and aligned using MUSCLE. The tree was generated using Phylyp v.3.696<sup>150</sup> with *Callorhinchus milii* (comparison1–3) or *L. chalumnae* (comparison4) as the outgroup (Supplementary Table 14). For the phylogenetic analysis by maximum likelihood, we used the Environment for Tree Exploration toolkit<sup>151</sup>, which automates CodeML and SLR analyses by using preconfigured evolutionary models. For the detection of genes under positive selection in sterlet, we compared the branch-specific model bsA1 (neutral) with the model bsA (positive selection) using a likelihood ratio test (FDR  $\leq 0.05$ ). To detect sites under positive selection, naive empirical Bayes probabilities for all four classes were calculated for each site. Sites with a probability >0.95 for either site class 2a (positive selection in the marked branch and conserved in the rest) were considered. The common species tree was drawn by the interactive Tree of Life tool (iTOL, <https://itol.embl.de/>) with default settings.

**Transcriptome analysis.** Total RNA was isolated using TRIzol Reagent (Thermo Fisher Scientific) according to the supplier's recommendation, in combination with the RNeasy Mini Kit (Qiagen). To support genome annotation, the same adult female and male sterlets (from the broodstock of the Leibniz-Institute of Freshwater Ecology and Inland Fisheries) as used for the whole-genome sequencing were sampled. RNAs were obtained from six adult male (brain, testes, muscle, spleen, liver and skin) and three adult female (ovary, liver and brain) tissues. In addition, mixed RNAs (brain, heart, eyes and spleen) of one juvenile male (20 cm) were sequenced. RNA-Seq reads were used as transcriptomic

evidence for genome annotation and sex-biased expression analysis. Custom sequencing (BGI) of TruSeq libraries generated 25–30 million 100-bp paired-end reads for each sample on the Illumina HiSeq4000 platform.

For differential gene expression analysis, reads were aligned to the sterlet genome using STAR (–quantMode GeneCounts)<sup>152</sup>.

Owing to the sequence similarity between ohnologues, the mapping results were further filtered for uniquely mapped reads and reads with no mismatches to be able to obtain a reliable read assignment. To compare expression between different genes from an ohnologue pair, we used transcripts per million (TPM) values. For further analyses, genes not expressed (TPM < 5) in both ohnologues and in all included organs or ohnologue pairs without sufficient discriminating single nucleotide polymorphisms were excluded. Ohnologues were considered to be expressed at different levels if the absolute value (ohnologue1(log<sub>2</sub>TPM + 1) – ohnologue2(log<sub>2</sub>TPM + 1)) was greater than one (representing a twofold difference) in at least two different sterlet organs and developmental stages. For functional clustering, the web tool DAVID (<https://david.ncicrf.gov/>) was used, on the basis of human orthologues and all ohnologues as background.

**RAD-tag sequencing and analysis of sex-specific tags.** The genomic DNA of 31 females and 30 males was extracted from 90% ethanol-preserved fin clips using a classical phenol/chloroform protocol. The sterlet RAD-tag library was built according to standard protocols<sup>153</sup>, using SbfI as a single restriction enzyme, and sequenced on a single lane of HiSeq 2500, using the v4 SR100nt mode. The resulting read file was then demultiplexed using the process-radtags.pl script of STACKS software v.1.44 (ref. <sup>154</sup>) with default settings.

Demultiplexed reads were analysed with RADSex v.0.2.0<sup>155</sup>. RADSex sorts reads from the demultiplexed dataset into groups sharing the exact same sequence, and reads that would belong to the same polymorphic locus using standard analysis software are simply split into multiple markers. As a result, RADSex markers are non-polymorphic, thus allowing straightforward presence–absence comparison between individuals.

First, a table of depth for each RADSex marker in each individual from the dataset was generated using radsex process with default settings. The distribution of markers in males and females was then computed with radsex distrib, using a minimum depth of 10 (–min-cov 10) to consider a marker present in an individual, and a tile plot was generated from this distribution using the plot\_sex\_distribution() function from RADSex-vis (<https://github.com/RomainFeron/RADSex-vis>). The same analysis was performed with minimum depths of 1, 2 and 5, but the results were not qualitatively affected. A total of 176,735 markers were obtained that were present in at least one individual with a minimum depth of 10.

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

## Data availability

The *Acipenser ruthenus* Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession number VTUV00000000. The version described in this paper is version VTUV01000000. Genomic and transcriptomic reads are deposited in the Sequence Read Archive under accession numbers SRR110188515-10188518 and SRR110134511-11013458.

## Code availability

The in-house scripts have been deposited in Github ([https://github.com/dukecomeback/sterlet\\_Msch](https://github.com/dukecomeback/sterlet_Msch)).

Received: 19 August 2019; Accepted: 27 February 2020;

Published online: 30 March 2020

## References

- Van de Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **10**, 725–732 (2009).
- Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
- Dehal, P. & Boore, J. L. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3**, e314 (2005).
- Meyer, A. & Van de Peer, Y. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays* **27**, 937–945 (2005).
- Braasch, I. & Postlethwait, J. H. in *Polyploidy and Genome Evolution* (eds Soltis P. & Soltis, D.) 341–383 (Springer, 2012).
- Symonová, R. et al. Molecular cytogenetic differentiation of paralogs of *Hox* paralogs in duplicated and re-diploidized genome of the North American paddlefish (*Poliodon spathula*). *BMC Genet.* **18**, 19 (2017).
- Havelka, M., Hulák, M., Bailie, D., Prodöhl, P. & Flajšhans, M. Extensive genome duplications in sturgeons: new evidence from microsatellite data. *J. Appl. Ichthyol.* **29**, 704–708 (2013).
- Otto, S. P. The evolutionary consequences of polyploidy. *Cell* **131**, 452–462 (2007).
- Force, A. et al. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
- Ludwig, A., Belfiore, N. M., Pitra, C., Svirsky, V. & Jenneckens, I. Genome duplication events and functional reduction of ploidy levels in sturgeon (*Acipenser, Huso and Scaphirhynchus*). *Genetics* **158**, 1203–1215 (2001).
- Wolfe, K. H. Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* **2**, 333–341 (2001).
- Rajkov, J., Shao, Z. & Berrebi, P. Evolution of polyploidy and functional diploidization in sturgeons: microsatellite analysis in 10 sturgeon species. *J. Heredity* **105**, 521–531 (2014).
- Crow, K. D., Smith, C. D., Cheng, J.-F., Wagner, G. P. & Amemiya, C. T. An independent genome duplication inferred from *Hox* paralogs in the American paddlefish—a representative basal ray-finned fish and important comparative reference. *Genome Biol. Evol.* **4**, 937–953 (2012).
- Miller, M. J. in *Sturgeons and Paddlefish of North America. Fish & Fisheries Series Vol. 27* (eds LeBreton, G. T. O., Beamish, F. W. H. & McKinley, R. S.) 87–101 (Springer, 2004).
- Saito, T. et al. The origin and migration of primordial germ cells in sturgeons. *PLoS ONE* **9**, e86861 (2014).
- Hochleithner, M. & Gessner, J. *The Sturgeon and Paddlefishes of the World—Biology and Aquaculture* Aquatech Publication 106 (Aquatech, 2001).
- Allen, P. J., Cech, J. J. & Kültz, D. Mechanisms of seawater acclimation in a primitive, anadromous fish, the green sturgeon. *J. Comp. Physiol. B* **179**, 903–920 (2009).
- Haidvogel, G. et al. Typology of historical sources and the reconstruction of long-term historical changes of riverine fish: a case study of the Austrian Danube and northern Russian rivers. *Ecol. Freshw. Fish* **23**, 498–515 (2014).
- Saffron, I. *Caviar: The Strange History and Uncertain Future of the World's Most Coveted Delicacy* (Broadway Books, 2002).
- Cheng, P. et al. Draft genome and complete *Hox*-cluster characterization of the sterlet sturgeon (*Acipenser ruthenus*). *Front. Genet.* **10**, 776 (2019).
- Bytyutskyy, D., Srp, J. & Flajšhans, M. Use of Feulgen image analysis densitometry to study the effect of genome size on nuclear size in polyploid sturgeons. *J. Appl. Ichthyol.* **28**, 704–708 (2012).
- Fontana, F. et al. Fluorescent in situ hybridization with rDNA probes on chromosomes of *Acipenser ruthenus* and *Acipenser naccarii* (Osteichthyes Acipenseriformes). *Genome* **42**, 1008–1012 (1999).
- Valente, G. T. et al. B chromosomes: from cytogenetics to systems biology. *Chromosoma* **126**, 73–81 (2017).
- Du, K. et al. The genome of the arapaima (*Arapaima gigas*) provides insights into gigantism, fast growth and chromosomal sex determination system. *Sci. Rep.* **9**, 5293 (2019).
- Betancur-R, R. et al. Phylogenetic classification of bony fishes. *BMC Evol. Biol.* **17**, 162 (2017).
- Betancur-R, R. et al. The tree of life and a new classification of bony fishes. *PLoS Curr.* **5** <http://doi.org/dpxx> (2013).
- Near, T. J. et al. Resolution of ray-finned fish phylogeny and timing of diversification. *Proc. Natl Acad. Sci. USA* **109**, 13698–13703 (2012).
- Hughes, L. C. et al. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proc. Natl Acad. Sci. USA* **115**, 6249–6254 (2018).
- Rabosky, D. L. et al. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nat. Commun.* **4**, 1958 (2013).
- Darwin, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (Oxford Univ. Press, 1859).
- Krieger, J. & Fuerst, P. A. Evidence for a slowed rate of molecular evolution in the order Acipenseriformes. *Mol. Biol. Evol.* **19**, 891–897 (2002).
- Chalopin, D., Naville, M., Plard, F., Galiana, D. & Volff, J.-N. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol. Evol.* **7**, 567–580 (2015).
- Amemiya, C. T. et al. The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**, 311–316 (2013).
- Berthelot, C. et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat. Commun.* **5**, 3657 (2014).
- Lien, S. et al. The Atlantic salmon genome provides insights into rediploidization. *Nature* **533**, 200–205 (2016).
- Chen, Z. et al. De novo assembly of the goldfish (*Carassius auratus*) genome and the evolution of genes after whole-genome duplication. *Sci. Adv.* **5**, eaav0547 (2019).
- Andreyushkova, D. et al. Next generation sequencing of chromosome-specific libraries sheds light on genome evolution in paleotetraploid sterlet (*Acipenser ruthenus*). *Genes* **8**, 318 (2017).
- Romanenko, S. A. et al. Segmental paleotetraploidy revealed in sterlet (*Acipenser ruthenus*) genome by chromosome painting. *Mol. Cytogenet.* **8**, 90 (2015).

39. Bennett, R. J., Uhl, M. A., Miller, M. G. & Johnson, A. D. Identification and characterization of a *Candida albicans* mating pheromone. *Mol. Cell. Biol.* **23**, 8189–8201 (2003).
40. Gerstein, A. C., Chun, H.-J. E., Grant, A. & Otto, S. P. Genomic convergence toward diploidy in *Saccharomyces cerevisiae*. *PLoS Genet.* **2**, e145 (2006).
41. De Storme, N. & Mason, A. Plant speciation through chromosome instability and ploidy change: cellular mechanisms, molecular factors and evolutionary relevance. *Curr. Plant Biol.* **1**, 10–33 (2014).
42. Rajaraman, R., Rajaraman, M. M., Rajaraman, S. R. & Guernsey, D. L. Neosis—a paradigm of self-renewal in cancer. *Cell Biol. Int.* **29**, 1084–1097 (2005).
43. Session, A. M. et al. Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature* **538**, 336–343 (2016).
44. Tang, H. et al. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
45. Dewey, C. N. Positional orthology: putting genomic evolutionary relationships into context. *Brief. Bioinform.* **12**, 401–412 (2011).
46. Freeling, M. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* **60**, 433–453 (2009).
47. Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl Acad. Sci. USA* **109**, 14746–14753 (2012).
48. Hrbek, T., Seckinger, J. & Meyer, A. A phylogenetic and biogeographic perspective on the evolution of poeciliid fishes. *Mol. Phylogenet. Evol.* **43**, 986–998 (2007).
49. De Smet, R. et al. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc. Natl Acad. Sci. USA* **110**, 2898–2903 (2013).
50. Sémon, M. & Wolfe, K. H. Consequences of genome duplication. *Curr. Opin. Genet. Dev.* **17**, 505–512 (2007).
51. Conant, G. C., Birchler, J. A. & Pires, J. C. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin. Plant Biol.* **19**, 91–98 (2014).
52. Makino, T. & McLysaght, A. Positionally biased gene loss after whole genome duplication: evidence from human, yeast, and plant. *Genome Res.* **22**, 2427–2435 (2012).
53. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
54. Kikuta, H. et al. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17**, 545–555 (2007).
55. Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* **11**, 97–108 (2010).
56. Ward, R. & Durrett, R. Subfunctionalization: how often does it occur? How long does it take? *Theor. Popul. Biol.* **66**, 93–100 (2004).
57. Ohno, S. *Evolution by Gene Duplication* (Springer Science & Business Media, 2013).
58. Kjærner-Semb, E. et al. Atlantic salmon populations reveal adaptive divergence of immune related genes—a duplicated genome under selection. *BMC Genomics* **17**, 610 (2016).
59. Wassarman, P. M. Zona pellucida glycoproteins. *J. Biol. Chem.* **283**, 24285–24289 (2008).
60. Sano, K. et al. Comparison of egg envelope thickness in teleosts and its relationship to the sites of ZP protein synthesis. *J. Exp. Zool. B* **328**, 240–258 (2017).
61. Kim, B.-M. et al. Antarctic blackfin icefish genome reveals adaptations to extreme environments. *Nat. Ecol. Evol.* **3**, 469–478 (2019).
62. Cao, L. et al. Neofunctionalization of zona pellucida proteins enhances freeze-prevention in the eggs of Antarctic notothenioids. *Nat. Commun.* **7**, 12987 (2016).
63. Montavon, T. et al. A regulatory archipelago controls *Hox* genes transcription in digits. *Cell* **147**, 1132–1145 (2011).
64. Beccari, L. et al. A role for HOX13 proteins in the regulatory switch between TADs at the *HoxD* locus. *Genes Dev.* **30**, 1172–1186 (2016).
65. Woltering, J. M., Noordermeer, D., Leleu, M. & Duboule, D. Conservation and divergence of regulatory strategies at *Hox* loci and the origin of tetrapod digits. *PLoS Biol.* **12**, e1001773 (2014).
66. Braasch, I. et al. The spotted gar genome illuminates vertebrate evolution and facilitates human–teleost comparisons. *Nat. Genet.* **48**, 427–437 (2016).
67. Amores, A. et al. Zebrafish *hox* clusters and vertebrate genome evolution. *Science* **282**, 1711–1714 (1998).
68. Mungpakdee, S. et al. Differential evolution of the 13 Atlantic salmon *Hox* clusters. *Mol. Biol. Evol.* **25**, 1333–1343 (2008).
69. Martin, K. J. & Holland, P. W. Enigmatic orthology relationships between *Hox* clusters of the African butterfly fish and other teleosts following ancient whole-genome duplication. *Mol. Biol. Evol.* **31**, 2592–2611 (2014).
70. Kuraku, S. & Meyer, A. The evolution and maintenance of *Hox* gene clusters in vertebrates and the teleost-specific genome duplication. *Int. J. Dev. Biol.* **53**, 765–773 (2009).
71. Woltering, J. M. & Durston, A. J. The zebrafish *hoxdb* cluster has been reduced to a single microRNA. *Nat. Genet.* **38**, 601–602 (2006).
72. McClintock, J. M., Kheirbek, M. A. & Prince, V. E. Knockdown of duplicated zebrafish *hoxb1* genes reveals distinct roles in hindbrain patterning and a novel mechanism of duplicate gene retention. *Development* **129**, 2339–2354 (2002).
73. Takamatsu, N. et al. Duplicated Abd-B class genes in medaka *hoxAa* and *hoxAb* clusters exhibit differential expression patterns in pectoral fin buds. *Dev. Genes Evol.* **217**, 263–273 (2007).
74. Scharlt, M. et al. The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. *Nat. Genet.* **45**, 567–572 (2013).
75. Muller, H. Why polyploidy is rarer in animals than in plants. *Am. Nat.* **59**, 346–353 (1925).
76. Mable, B. 'Why polyploidy is rarer in animals than in plants': myths and mechanisms. *Biol. J. Linn. Soc.* **82**, 453–466 (2004).
77. Orr, H. A. 'Why polyploidy is rarer in animals than in plants' revisited. *Am. Nat.* **136**, 759–770 (1990).
78. Wertheim, B., Beukeboom, L. & Van de Zande, L. Polyploidy in animals: effects of gene expression on sex determination, evolution and ecology. *Cytogenet. Genome Res.* **140**, 256–269 (2013).
79. Fopp-Bayat, D., Kolman, R. & Woznicki, P. Induction of meiotic gynogenesis in sterlet (*Acipenser ruthenus*) using UV-irradiated baster sperm. *Aquaculture* **264**, 54–58 (2007).
80. Havelka, M. & Arai, K. in *Sex Control in Aquaculture* (eds Wang H.-P. et al.) 669–687 (John Wiley & Sons Ltd., 2018).
81. Keyvanshokoo, S. & Gharaei, A. A review of sex determination and searches for sex-specific markers in sturgeon. *Aquac. Res.* **41**, e1–e7 (2010).
82. Havelka, M., Kašpar, V., Hulák, M. & Flajšhans, M. Sturgeon genetics and cytogenetics: a review related to ploidy levels and interspecific hybridization. *Folia Zool.* **60**, 93–104 (2011).
83. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **8**, 272–285 (2007).
84. Klein, S. J. & O'Neill, R. J. Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Res.* **26**, 5–23 (2018).
85. Foissac, S. et al. Transcriptome and chromatin structure annotation of liver, CD4+ and CD8+ T cells from four livestock species. Preprint at *bioRxiv* <https://doi.org/10.1101/316091> (2019).
86. *SMRT Link v.5.0.1* (Pacific Biosciences of California, Inc. 2018).
87. Ruan, J. *SMARTdenovo: Ultra-fast de novo assembler using long noisy reads* (Github, accessed 10 January 2019); <https://github.com/ruanjue/smartdenovo>
88. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
89. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
90. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
91. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997> (2013).
92. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
93. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at <https://arxiv.org/abs/1207.3907> (2012).
94. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017).
95. *fgbio: Tools for working with genomic and high throughput sequencing data* (Github, 2019); <http://fulcrumgenomics.github.io/fgbio/>
96. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
97. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
98. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
99. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
100. Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-4.0* (ISB, 2013–2015); <http://www.repeatmasker.org>
101. Ou, S. & Jiang, N. LTR\_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).



102. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.* **9**, 18 (2008).
103. Xu, Z. & Wang, H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
104. Smit, A. F., Hubley, R. & Green, P. *RepeatMasker* (ISB, 2019); <http://www.repeatmasker.org/webrepeatmaskerhelp.html>
105. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
106. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
107. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
108. Shao, F., Wang, J., Xu, H. & Peng, Z. FishTEDB: a collective database of transposable elements identified in the complete genomes of fish. *Database* **2018**, 1–9 (2018).
109. Christiam, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
110. Novák, P. et al. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* **45**, e111 (2017).
111. Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792–793 (2013).
112. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
113. Neumann, P., Novák, P., Hošťáková, N. & Macas, J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob. DNA* **10**, 1 (2019).
114. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
115. Wickham, H. ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **3**, 180–185 (2011).
116. Higham, D. J. & Higham, N. J. *MATLAB Guide* Vol. 150 (Society for Industrial and Applied Mathematics, 2016).
117. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
118. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
119. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* **6**, 31 (2005).
120. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Res.* **14**, 988–995 (2004).
121. She, R., Chu, J. S.-C., Wang, K., Pei, J. & Chen, N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res.* **19**, 143–149 (2009).
122. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).
123. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
124. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
125. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
126. Haas, B. J. et al. Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
127. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
128. Mudunuri, U., Che, A., Yi, M. & Stephens, R. M. bioDBnet: the biological database network. *Bioinformatics* **25**, 555–556 (2009).
129. Kulmanov, M., Khan, M. A. & Hoehndorf, R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* **34**, 660–668 (2017).
130. Chan, P. P. & Lowe, T. M. in *Gene Prediction. Methods in Molecular Biology* Vol. 1962 (ed. Kollmar, M.) 1–14 (Humana, 2019).
131. Lagesen, K. et al. RNAmmer: consistent annotation of rRNA genes in genomic sequences. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
132. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
133. Kalvari, I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* **46**, D335–D342 (2017).
134. Cho, Y. S. et al. The tiger genome and comparative analysis with lion and snow leopard genomes. *Nat. Commun.* **4**, 3433 (2013).
135. Ruan, J. et al. TreeFam: 2008 update. *Nucleic Acids Res.* **36**, D735–D740 (2007).
136. Ponting, C. TreeBeST v0.5 (SourceForge, 2007); <http://treesoft.sourceforge.net/treebest.shtml>
137. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
138. 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).
139. Stamatakis, A. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
140. Ronquist, F. et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
141. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
142. Inoue, J., Dos Reis, M. & Yang, Z. *A step-by-step tutorial: Divergence time estimation with approximate likelihood calculation using MCMCTREE in PAML* (Citeseer, 2011).
143. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAXML web servers. *Syst. Biol.* **57**, 758–771 (2008).
144. Lin, Q. et al. The seahorse genome and the evolution of its specialized morphology. *Nature* **540**, 395–399 (2016).
145. Lundberg, J. G. & Chernoff, B. A Miocene fossil of the amazonian fish *Arapaima* (Teleostei, Arapaimidae) from the Magdalena River region of Colombia-biogeographic and evolutionary implications. *Biotropica* **24**, 2–14 (1992).
146. Kumazawa, Y. & Nishida, M. Molecular phylogeny of osteoglossoids: a new model for Gondwanian origin and plate tectonic transportation of the Asian arowana. *Mol. Biol. Evol.* **17**, 1869–1878 (2000).
147. O'Brien, K. P., Remm, M. & Sonnhammer, E. L. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33**, D476–D480 (2005).
148. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
149. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577 (2007).
150. Retief, J. D. in *Bioinformatics Methods and Protocols* (eds Misener, S. & Krawetz, S. A.) 243–258 (Humana Press, 2000).
151. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
152. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
153. Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W. & Postlethwait, J. H. Stacks: building and genotyping loci de novo from short-read sequences. *G3* **1**, 171–182 (2011).
154. Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A. & Cresko, W. A. Stacks: an analysis tool set for population genomics. *Mol. Ecol.* **22**, 3124–3140 (2013).
155. Feron, R. et al. RADSex (Github, 2018); <https://github.com/RomainFeron/RADSex>

## Acknowledgements

This work was supported by grants supplied by the Deutsche Forschungsgemeinschaft to M.Schartl (nos. DFG SCHA 408/14-1, SCHA 408/10-1 and SCHA408/16-1), J.M.W. (DFG WO-2165/2-1) and A. Meyer (DFG Me1725/24-1); by the German Federal Ministry of Food and Agriculture through the Federal Office for Agriculture and Food (grant no. 2816ERA04G) to M. Stöck, S.W., J.G. and M. Schartl; by funds from the Agence Nationale de la Recherche (grant no. ANR-13-ISR7-0005, PhyloSex project) to Y.G.; and by funds from the Russian Science Foundation to V.T., A.M., I.K. and D.P. (grant no. RSF 18-44-04007). We thank the Leibniz-IGB for supporting the sterlet genome sequencing project. The GeT core facility was supported by France Génomique National infrastructure, funded as part of the Investissement d'avenir programme, managed by the Agence Nationale pour la Recherche (contract no. ANR-10-INBS-09).

## Author contributions

M. Schartl, M. Stöck, Y.G., J.H.P. and W.C.W. conceived the study. M. Stöck, S.W., J.G. and W.K. provided the biological materials. M. Stöck, P.F. and C.S. prepared the DNA and RNA for sequencing. C.I. generated the Hi-C data. C.K., C.C. and C.T. produced the assemblies. K.D. performed the annotation. K.D., S.K., D.P., J.M.W., A. Meyer, I.B., H.K., M.C.A. and M. Schartl analysed the genome. A. Makunin, I.K. and V.T. mapped the microdissection chromosome library. R.F., Y.G., L.J., H.P. and J.H.P. did the RAD-sequencing analysis. M. Schartl, K.D., M. Stöck, J.H.P. and A. Meyer wrote the manuscript. All authors commented on the manuscript and were involved in the interpretation of the primary data.



**Competing interests**

The authors declare no competing interests.

**Additional information**

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41559-020-1166-x>.

**Correspondence and requests for materials** should be addressed to M.S. or M.S.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☒ ☐ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☒ ☐ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection	All commercial, open source and custom codes used to collect the data are described and referenced in the manuscript and are publicly available.
Data analysis	All commercial, open source and custom codes used to analyse the data are described and referenced in the manuscript and are publicly available.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The *Acipenser ruthenus* Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession VTUV000000000. The version described in this paper is version VTUV01000000. Genomic and transcriptomic reads are deposited in the sequence read archive under accession numbers SRR10188515-10188518 and SRR11013451-11013458.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	One individual male was used for whole genome sequencing, RNA-seq was done on samples from RNAs were obtained from six adult males, one juvenile male and three adult females. RAD-tags were generated from 31 females and 30 males.
Data exclusions	No data were excluded from the analysis
Replication	n/a
Randomization	n/a
Blinding	n/a

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Acipenser ruthenus, from the sterlet sturgeon population maintained at the Leibniz Institute of Freshwater Ecology and Inland fisheries (IGB), Berlin. This stock is derived from the Danube population of A. ruthenus.
Wild animals	n/a
Field-collected samples	n/a
Ethics oversight	The experiments were carried out in accordance with the European Directive 2010/63/EU and German national legislation (Animal protection law, TierSchG). All experimental protocols that are part of this study were approved through an authorization (File # ZH 114, issued 06.02.2014) of the LAGeSo, Berlin, Germany.

Note that full information on the approval of the study protocol must also be provided in the manuscript.