The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization

Kang Du  
*University of Wuerzburg*

Chad Tomlinson  
*Washington University School of Medicine in St. Louis*

et al

Follow this and additional works at: [https://digitalcommons.wustl.edu/open_access_pubs](https://digitalcommons.wustl.edu/open_access_pubs)

Please let us know how this document benefits you.

**Recommended Citation**

[https://digitalcommons.wustl.edu/open_access_pubs/11725](https://digitalcommons.wustl.edu/open_access_pubs/11725)

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization

Kang Du¹,², Matthias Stöck³, Susanne Kneitz¹, Christophe Klopp⁴,⁵, Joost M. Woltering⁶, Mateus Contar Adolfi¹, Romain Feron⁶, Dmitriy Prokopov⁷, Alexey Makunin⁸, Ilya Kichigin⁸, Cornelia Schmidt¹, Petra Fischer¹, Heiner Kuhl³, Sven Wuertz², Jörn Gessner³, Werner Kloas³, Cédric Cabau⁴,⁵, Carole Iampietro⁹, Hugues Parrinello¹⁰, Chad Tomlinson¹¹, Laurent Journot¹⁰, John H. Postlethwait¹², Ingo Braasch¹³, Vladimir Trifonov⁸, Wesley C. Warren¹⁴, Axel Meyer⁶, Yann Guiguen¹⁵ and Manfred Schartl⁶,⁷, Dmitry Prokopov⁸

Vertebrate genome evolution has been strongly impacted by polyploidization events. Early on, vertebrate ancestors experienced two rounds (1R and 2R) of whole-genome duplications (WGDs). 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 the land vertebrates or their sarcopterygian forbearing relative about 320 million years ago (Ma), but not in other major clades (coelacanths and lungfishes). Some teleost groups, such as salmonids and carp, independently underwent another round (4R) of polyploidization events. After their divergence from the other ray-finned fish lineages, the teleost-specific 3R duplication and after the ancient 2R event, the various lineages of Acipenseriformes (sturgeon and paddlefish) experienced several polyploidization events, resulting in karyotypes, comprising between ~120 chromosomes in some species, and ~360 chromosomes in species that are considered dodecaploid. 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 as tetraploids, the other teleost lineages retained on average only 17% of gene duplicates from the ancient 3R ohologues. The evolutionary trajectories and forces driving species from polyploids to meiotic diploids are the subject of major adaptive hypotheses and their empirical evaluations.

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 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. Such far-reaching redundancy reduction leads one to question their polyploidy state.

Because sturgeons branched off early from modern fishes, their genomes may harbour traces of the ancient vertebrate ancestors. Notably, their early embryonic development is of the classical...
Sturgeons are distributed from subtropical to subarctic rivers, lakes and coastlines of Eurasia and North America. 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. Sturgeons are celebrities among fishes because of their pre-ovulation female gametes, known as caviar. Habitat destruction, the lack of river connectivity, pollution and the 2,000-year-old rural caviar production 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. 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 chromosome sets. They are found in some but not all individuals within a population and are considered to be either non-functional, beneficial or harmful. 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. 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. Scaffolds 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. 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).
in agreement with the current tree of life\textsuperscript{25–28}. 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\textsuperscript{9}.

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\textsuperscript{30}, Charles Darwin called them living fossils\textsuperscript{30}. We therefore asked whether the morphological stasis in sturgeons is matched by the slower substitution rates of several mitochondrial and nuclear genes\textsuperscript{31}. 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\textsuperscript{31}. 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\textsuperscript{32}.

**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 orthologue 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 orthologue 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 (\textit{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. \textsuperscript{34,35}) and the carp–goldfish 4R estimated at 14 Ma (ref. \textsuperscript{36}).

The analysis of conserved syntenies between sterlet and gar revealed that most gar chromosomes have two counterparts in sterlet (Fig. 2a). When sterlet orthologous 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. 2** | Homology and homeology relationships of sterlet chromosomes. \textbf{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). The chromosomes are ordered by size. **b**, Chord diagram depicting homeology relationships of 60 sterlet chromosomes on the basis of 9,301 orthologue 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.
microdissected chromosomes or chromosome arms of the sterlet\textsuperscript{37,38}. 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\textsuperscript{38}. 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\textsuperscript{39,40}, flowering plants\textsuperscript{41} and endopolyploid human cancer cells\textsuperscript{42}.

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 autopolyploidy—have been discussed to account for the origin of the sterlet chromosome complement, on the basis of conflicting evidence\textsuperscript{12}. To clarify this controversy, we used a strategy that was employed to investigate this problem in the allopolyploid African clawed frog, \textit{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\textsuperscript{43}. 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).

---

**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).
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. 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.

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. 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). 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. 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%.

On the basis of noncoding microsatellites, the sterlet was proposed to have undergone extensive duplicate gene degeneration and has been classified since then even as a functional diploid species. 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. 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 (Supplementary Table 12).

Like the situation reported for rainbow trout, 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. However, a search of singleton genes, embedded in syntenic blocks using the STRING database, did not reveal such protein–protein interactions. An alternative explanation for the conservation of microsynteny is the bystander relationship, 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. 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 use of especially profited from the WGD for evolutionary progress sensu Susumo Ohno. A similar phenomenon was observed for duplicated immune genes in salmon.

**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 pellicida* (Zp) sperm-binding proteins (ID: 4190). Zp-proteins prevent polyspermy in mammals and provide thickness and hardness to the fish egg envelope. 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. 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.
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, 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.

Over-retention of glutamate receptor genes. Glutamate receptor genes (GRGs) show particularly high ohnologue retention rates in teleosts, 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 Table 19 and Supplementary Note 13). Compared with the genome-wide rate of 70% (9,914 ohnologs and 4,175 Ars3R 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. 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 negative consequences of polyploidization. Data from induced gynogenesis led to the common belief that all Acipenser species, including sterlet, have a female heterogametic (ZZ/ZW) sex chromosome system.

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.

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 the extant sturgeons with ~120 chromosomes (like the sterlet) represent functional diplpoids, which originated over 200 Ma by a WGD of a 60-chromosome diplpoid ancestor82. The transition between the ancestral fully tetraploid and the modern functional diplpoids was proposed to have been accompanied by a reduction of duplicate gene functions83. Our estimate of 180 Ma for the Ars3R provides evidence for a WGD in the ancestor of all sturgeons, and that the deduplication of neighboring genes, suggest complex processes have considerably shaped its genome, most obviously manifested by the loss of whole chromosomes from homoeologous 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 genet 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 were approved through an authorization file (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 tests and blood of a single adult male. We generated X42 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 Poisati et al.46. 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 Illumina HiSeq4000 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 Biosciences46. PacBio reads were assembled with smartdenovo (v.1.0) with standard parameters. Contigs were polished with two rounds of racon48 (v.1.3.1), using long reads aligned with miniasm48 (v.1.22) of the assembled genome as input. The Illumina reads. The Illumina reads were aligned with bwa mem (v.0.7.12-r1039) with standard parameters and the same file, which had been compressed, sorted and indexed with samtools view, sort and index v.1.3.1, using standard parameters before polish 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 data. The VCF file with frequencies (v.1.1) under standard parameters was produced using the VCFtools11 tool. The VCF file showed 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 HapCut249 (v.1.0) using the following parameters: extracHAIRS –ont 1 and HapCut2 –ont 1 –r –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 lgho (v.0.7.0 using standard parameters)49 to generate VCF files, separated by haplotype. These VCF files and the reference were used to create haplotyped contigs using vcf-consensus from the bcf tools15 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 Juicer12, and the scaffolding was performed with 3D-DNA13 with parameter -r = 0. Finally, the candidate assembly was manually reviewed using the Juicebox Assembly Tools11. 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 tool46 with standard settings. LTR-retriever v.2.5 (ref. 103) was used to search for full-length LTR elements, and the data were used as input derived from theLTRharvest51 (similar, 90; vic, 10; -seed, 20; -seqisd, y=; -miniletrn, 100; -maxiletrn, 7,000; -minstd, 4; -maxstd, 6; -motifims, 1) and LTR_FINDER52 (d, 15,000; -d, 1,000; -l, 7,000; -l, 100; -p, 20; -CM, 0.9) tools. To exclude non-LTR (-lineib) and DNA transposons (-dnalib), protein sequences of these TEs from the RepeatPeps database of the RepeatMasker tool53 were used. This also excluded protein sequences that were not related to TEs. The SWISS-PROT sequence library105 was also used (-plantprotlib).

The sequences obtained using the previous steps were combined into a single FASTA file using CD-HIT-est106 (–s, 1; –c, 1; –r, 1; –g, 1; –p, 0). The resulting FASTA
file was aligned against the RepBase v2.4.07 (ref. [18]) and FishTEDb14 databases using blastn (-evalue, 10−09) and against SWISS-PROT and RepeatPeps using blastx (-evalue, 10−10) to filter incorrectly annotated sequences. Raw reads were assembled using the TANGANAS database Table 3 in part of RepeatExplorer18. The reads were first trimmed using the fastq tool to remove low-quality and adapter sequences (detect_adapter_for_pe -g -c 150 -5 -3) and then aligned against the FASTA file containing TE sequences obtained in the previous step using bowtie2 v2.3.5.1 (ref. [19]) (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 x 1000000/total number of reads), and then the proportion of superfamilies in the transcriptome was calculated (reads per million x 100/total number of aligned reads). To compare the RNA-seq data with the genome proportion of the respective TE superfamily, the proportion of TE sequences in the genome was calculated (the number of nucleotides occupied by superfamily in the genome x 100/total nucleotides occupied by TEs in the genome). The results were transformed to the log10 values and visualized with ggplot2 and MATLAB.

Genome annotation. Genome annotation was done by an in-house pipeline (Supplementary Fig. 25) improved from a previous version1. First, the pipeline assessed the assembly quality using BUSCO on the basis of the Actinopterygii odb9 database. The putative long was used for the first training of AUGUSTUS v3.2.3 (ref. [20]). The pipeline then identified and marked 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. Simple and low-complexity repeats were then softmasked, while those with known family names 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) and 13 Ensembl genomes (v95, http://wwwensembl.org/), human (Homo sapiens), mouse (Mus musculus), coelacanth (Lattimeria chalumnae), spotted gar (Lepisosteus oculatus), zebrafish (Danio rerio), cod (Gadus morhua), tilapia (Oreochromis niloticus), medaka (Oryzias latipes), platyfish (Xiphophorus maculatus), fugu (Takifugu rubripes), tetrodon (Tetrododon 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 v2.2.0 (ref. [21]) and GeneWise2-2.0 (ref. [21]) respectively. Before GeneWise was implemented, genBLASTA1.0.1 (ref. [22]) was used to roughly locate each protein on the assembly. For de novo annotation, SNAP v0.67-28 (http://korlin.kazusa.or.jp/gene/seq/) and GeneMark-ES22 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 v2.1.1 (ref. [22]). In parallel, HISAT2 v2.1.0, Trinity v2.4.0 and PASA v2.2.0 (ref. [23]) were also used for RNA-seq read mapping and assembly. In total, 89.3% of all transcriptome reads mapped to the genome.

All gene evidence obtained from the three kinds of annotation was collected and transferred to EVidenceModeler v1.1.1 (ref. [24]), 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 was used to construct the new gene models. To assign gene symbols, their protein sequences were blasted against the homology gene evidence. Low-quality gene models were removed afterwards. To assign gene symbols, their protein sequences were blasted using tRNAscan-SE v.2.0.3 (ref. [25]), and ribosomal RNAs were identified using 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) and RepeatModel (http://www.repeatmasker.org/) script. The resulting file, out.BV , was then used for the final run of MCMCTree to perform approximate likelihood calculations. The phylogenetic tree and the coding sequences alignment were imported into baseip2 to roughly estimate the substitution rate. The substitution model was determined using modelselection.jar. Then mcemtree 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 phylogenetic tree and coding sequences alignment were imported into baseip2 to roughly estimate the substitution rate. The substitution model was determined using modelselection.jar. Then mcemtree 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.

To analyse the expression of TEs, raw reads from RNA-seq were used. The read counts were compared for each fish tissue using Cufflinks v2.1.1 (ref. [26]). In parallel, HISAT2 v.2.1.0, Trinity v.2.4.0 and PASA v2.2.0 (ref. [27]) were used for the generation of a phylogenetic tree as input for the detection of positive selection. Sites with a probability ≤0.05 were kept. Codon alignments for the protein–cDNA sequence pairs were constructed using pal2nal v.14. The resulting sites were aligned using MUSCLE (option: -fastest), and poorly aligned positions and divergent regions of cDNA were eliminated by Gblocks v0.91b (ref. [28]) (options: -b4: 10; -b5: n; -b3: 5; -n=1). An in-house script was used to convert the Gblocks output to pam format. 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 Phylop v3.696 with a time of ~110–156 Ma145,146 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 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. The resulting sites were aligned using MUSCLE (option: -fastest), and poorly aligned positions and divergent regions of cDNA were eliminated by Gblocks v0.91b (ref. [28]) (options: -b4: 10; -b5: n; -b3: 5; -n=1). An in-house script was used to convert the Gblocks output to pam format. 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) 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 for Freshwater Ecology and Inland Fisheries) were used as the whole-genome sequencing sample. RNAs were obtained from six adult male (brain, testis, 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 between different sexes, we used transcripts per million (TPM) values. For further analyses, genes not expressed (TPM < 5) in both sexes were considered to be expressed at different levels if the absolute value (Ohnologue1(log2TPM + 1) – Ohnologue2(log2TPM + 1)) was greater than one (representing a twofold difference) in at least two different sturgeon organs and developmental stages. For functional clustering, the web tool David (https://david.ncifcrf.gov/) was used, on the basis of human orthologues and all ohnologues as background.

RAD-tag sequencing and analysis of sex-specific tags. The genome 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 protocols14, 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-radiags.pl script of STACKS software version 1.4.4 (ref. 14) with default settings. Demultiplexed reads were analysed with RADSex v0.2.0.15. 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-polyphyletic, 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 distrub, 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_depth() 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

Acipenser ruthenus Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession number VTUV00100000. The version described in this paper is version VTUV00100000. Genomic and transcriptomic reads are deposited in the Sequence Read Archive under accession numbers SRR1088515-1088518 and SRR1103451-1103458.

Code availability

The in-house scripts have been deposited in Github (https://github.com/dukecomeback/sterlet_Msch).


104. Smit, A. F., Hubley, R. & Green, P. RepeatMasker (ISB, 2019); http://www.repeatmasker.org/repetmaskerhelp.html


127. 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.Scharl (nos. DFG SCHE 408/14-1; SCHE 408/10-1 and Sche408/16-1), J.M.W. (DFG VO 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. Stock, S.W., J.G. and M. Scharl; by funds from the Agence Nationale de la Recherche (grant no. ANR-13-ISV7-0005, PhyloX Sequence 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 18-44-0007). We thank the Leibniz-IBG for supporting the sterlet genome sequencing project. The Gt 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

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.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
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
- 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 VTUV00000000. The version described in this paper is version VTUV010000000. 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

Life sciences study design

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

<table>
<thead>
<tr>
<th>Sample size</th>
<th>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.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data exclusions</td>
<td>No data were excluded from the analysis</td>
</tr>
<tr>
<td>Replication</td>
<td>n/a</td>
</tr>
<tr>
<td>Randomization</td>
<td>n/a</td>
</tr>
<tr>
<td>Blinding</td>
<td>n/a</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Materials &amp; experimental systems</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>☒ Antibodies</td>
<td>☒ ChIP-seq</td>
</tr>
<tr>
<td>☒ Eukaryotic cell lines</td>
<td>☒ Flow cytometry</td>
</tr>
<tr>
<td>☒ Palaeontology</td>
<td>☒ MRI-based neuroimaging</td>
</tr>
<tr>
<td>☒ Animals and other organisms</td>
<td></td>
</tr>
<tr>
<td>☒ Human research participants</td>
<td></td>
</tr>
<tr>
<td>☒ Clinical data</td>
<td></td>
</tr>
</tbody>
</table>

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.