frontiers in Genetics

DATA REPORT
published: 05 September 2019
doi: 10.3389/fgene.2019.00776

Draft Genome and Complete *Hox*-Cluster Characterization of the Sterlet (*Acipenser ruthenus*)

Peilin Cheng[1,2†], Yu Huang[1,3,4†], Hao Du[1], Chuangju Li[1], Yunyun Lv[3,4], Rui Ruan[1], Huan Ye[1], Chao Bian[3], Xinxin You[3], Junmin Xu[3,5], Xufang Liang[2], Qiong Shi[3,4] and Qiwei Wei[1]**

[1] Key Laboratory of Freshwater Biodiversity Conservation, Ministry of Agriculture of China, Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Wuhan, China, [2] College of Fisheries, Chinese Perch Research Center, Huazhong Agricultural University, Wuhan, China, [3] Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, Academy of Marine Sciences, BGI Marine, Shenzhen, China, [4] BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, China, [5] School of Veterinary Medicine, Rakuno Gakuen University, Ebetsu, Japan

Background: Sturgeons (Chondrostei: Acipenseridae) are a group of "living fossil" fishes at a basal position among Actinopteri. They have raised great public interest due to their special evolutionary position, species conservation challenges, as well as their highly-prized eggs (caviar). The sterlet, *Acipenser ruthenus*, is a relatively small-sized member of sturgeons and has been widely distributing in both Europe and Asia. In this study, we performed whole genome sequencing, *de novo* assembly and gene annotation of the tarlet to construct its draft genome.

Findings: We finally obtained a 1.83-Gb genome assembly (BUSCO completeness of 81.6%) from a total of 316.8-Gb raw reads generated by an Illumina Hiseq 2500 platform. The scaffold N50 and contig N50 values reached 191.06 and 18.88 kb, respectively. The sterlet genome was predicted to be comprised of 42.84% repeated sequences and to contain 22,184 protein-coding genes, of which 21,112 (95.17%) have been functionally annotated with at least one hit in public databases. A genetic phylogeny demonstrated that the sterlet is situated in the basal position among ray-finned fishes and 4dTv analysis estimated that a recent whole genome duplication occurred 21.3 million years ago. Moreover, seven *Hox* clusters carrying 68 *Hox* genes were characterized in the sterlet. Phylogeny of *Hox*A clusters in the sterlet and American paddlefish divided these sturgeons into two groups, confirming the independence of each lineage-specific genome duplication in Acipenseridae and Polyodontidae.

Conclusions: This draft genome makes up for the lack of genomic and molecular data of the sterlet and its *Hox* clusters. It also provides a genetic basis for further investigation of lineage-specific genome duplication and the early evolution of ray-finned fishes.

Keywords: sterlet, sturgeon, genome, *hox*, lineage-specific whole genome duplication

OPEN ACCESS

Edited by:
Lior David,
Hebrew University of Jerusalem, Israel

Reviewed by:
Ron Dirks,
ZF-screens BV, Netherlands
Jie Mei,
Huazhong Agricultural University, China
László Orbán,
University of Pannonia, Hungary

***Correspondence:**
Qiong Shi
shiqiong@genomics.cn
Qiwei Wei
weiqw@yfi.ac.cn

†These author contributed equally to this work

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *12 October 2018*
Accepted: *23 July 2019*
Published: *05 September 2019*

Citation:
Cheng P, Huang Y, Du H, Li C, Lv Y, Ruan R, Ye H, Bian C, You X, Xu J, Liang X, Shi Q and Wei Q (2019) Draft Genome and Complete Hox-Cluster Characterization of the Sterlet (Acipenser ruthenus). Front. Genet. 10:776. doi: 10.3389/fgene.2019.00776

INTRODUCTION

Sturgeons (Acipenseridae, Acipenseriformes) have long been considered as an interesting group of fishes due to their commercial value and conservational challenges (Wei et al., 2011). They have also drawn noteworthy attention due to occupying a basal position on the phylogenetic tree of ray-finned fishes. It is estimated that the origin of sturgeons dates back to approximately 350 million years ago (Mya), which is even earlier than the origins of Holostei (bowfin and gars) and Teleostei (teleosts) (Hughes et al., 2018). Therefore, sturgeons did not the teleost-specific genome duplication (TGD) event that happened around 320 Mya (Jaillon et al., 2004). However, there are clear evidences based on molecular markers, chromosome numbers and inferred ploidy levels that they have experienced their own lineage-specific polyploidizations with one or more rounds of genome duplication (GD; Crow et al., 2012), resulting in complex genome structures and the widest range of chromosome numbers among all vertebrates (Havelka et al., 2016). However, little is known about Acipenseridae-specific GD and its consequences due to a lack of sturgeon genome sequences.

This special whole genome duplication (WGD) event has also provided new genetic material to generate phenotypic diversity among sturgeons. However, sturgeons have quite limited species diversity with exceedingly fast overall rates of body size evolution, serving as an interesting exception to the phenotypic 'evolvability' hypothesis (Rabosky et al., 2013). As one of the earliest evolved fish groups among ray-finned fishes, sturgeons still retain many shark-like features such as a cartilaginous skeleton and heterocercal tail, and the extant species look conspicuously similar to their fossil counterparts, suggesting that there has been of body-shape evolution (Rabosky et al., 2013). Therefore, sturgeons represent an ideal evolutionary group to investigate the complicated relationship between phenotypes and the polyploidy genomes caused by WGD. Meanwhile, *Hox* genes, encoding a distinct class of transcription factors associated with axial patterning and appendages development, have been often among the first list for examination to understand their roles in evolution of vertebrate body plans and novelty (Amemiya et al., 2010; Crow et al., 2012).

The sterlet (*Acipenser ruthenus*, Linnaeus, 1758) is a famous representative of sturgeon species, well-known for its relatively small body size and wide distribution in comparison to other sturgeons. Composed of 120 chromosomes, the sterlet genome has both diploid and tetraploid chromosome segments (Romanenko et al., 2015); however, various chromosomes are unequally involved in the multiple interchromosomal rearrangements after the GD event (Andreyushkova et al., 2017). In this study, we performed whole genome sequencing of the sterlet and generated a draft genome assembly of a sturgeon for the first time. We also constructed a fossil-calibrated phylogenetic tree, estimated the occurrence time of the sturgeon-specific GD (although it is unclear how many members in this family have experienced such an independent lineage-specific GD, considering that this is the first sturgeon with public genome sequences) and retrieved the complete *Hox* clusters to preliminarily reveal the early evolutionary history of ray-finned fishes.

Value of the Data

- This is the first genome report of a sturgeon. The sterlet genome was determined to be in size with a scaffold N50 of 191.06 kb. Our draft assembly contains 784 Mb (42.84% of the genome) of repeats and 22,184 protein-coding genes.
- The time-calibrated phylogenetic tree showed a most basal position of sterlet in Actinopterygii (ray-finned fishes) and dated the origin of the sterlet back to 358 Mya, which is extremely close to the Late Devonian Extinction that happened approximately 358,9 Mya.
- 4dTv analysis showed that the sturgeon-specific GD event happened about 21.3 Mya, close to the estimated occurrence time (42 Mya) of paddlefish-specific GD event, regardless of the independence of these two WGD events.
- Seven *Hox* clusters including 68 *Hox* genes were identified in the sterlet genome. Phylogeny of *Hox*A clusters of the sterlet and American paddlefish divided these sturgeons into two groups, suggesting that the WGD events happened independently in these two sturgeon species.

MATERIALS AND METHODS

Sample Processing

The sequenced sterlet (an immature juvenile, about 2.5 years old, 56.8 cm in length, weighing 0.8 kg) was artificially cultured at Taihu Station, Yangtze River Fisheries Research Institute, Chinese Academy of Fisheries Sciences, China. First, we obtained 10 mL of blood from the caudal vertebral vessels (without sacrificing the fish), but the sample was only sufficient for transcriptome sequencing. Subsequently, we had to anesthetize and sacrifice the fish to collected 30 g of skeletal muscle in order to obtain enough DNA for genome sequencing. All vouchers were deposited in China National GeneBank with accession numbers of WH20161125002-MU (muscle) and -BL (blood). All experiments were carried out in accordance with the guidelines of the Animal Ethics Committee of Yangtze River Fisheries Research Institute of Chinese Academy of Fishery Sciences (No. YFI-01).

Genome Sequencing and Assembly

We applied whole-genome shotgun sequencing to generate short paired-end reads (125 or 150 bp) by constructing a series of short-insert (270, 500, and 800 bp) or long-insert (2, 5, 10, and 20 kb) libraries (**Supplementary Figure 1**) and sequencing on a Hiseq 2500 platform (Illumina, San Diego, CA, USA). Raw reads were subsequently pre-processed by SOAPfilter software (Luo et al., 2012) to trim five bases at the 5' end of all reads and to discard the low-quality reads (quality value <20) and those reads with many nonsequenced bases (N > 10). Subsequently, the 17-mer depth frequency distribution method was employed to estimate the genome size of the sterlet using data from short-insert libraries according to the following formula: genome size = total number of k-mers/peak value of k-mer frequency distribution (Li et al., 2010). Clean reads from all the seven libraries were assembled into contigs and scaffolds using SOAPdenovo v2.04 (Luo et al.,

2012) with optimized parameters (pregraph -K 41 -d 1; contig –M 3; scaff -F; others as the default). Finally, gaps in the scaffolds were successively filled by using Kgf and GapCloser (Luo et al., 2012) with clean reads from short-insert libraries. Completeness of the final genome assembly and the entire gene set was assessed by BUSCO (Simão et al., 2015).

Repeat-Sequence Prediction and Gene Annotation

A *de novo* repeat library for the sterlet was constructed by a combination of RepeatModeler v1.05 (RepeatModeler, RRID: SCR_015027) and LTR_FINDER v1.0.6 (Xu and Wang, 2007). Known and *de novo* transposable elements (TEs) in the assembled genome were identified by RepeatMasker v4.0.6 (RepeatMasker, RRID : SCR_012954) using both the RepBase v21.01 (Jurka et al., 2005) and the *de novo* repeat library. RepeatProteinMask v3.3.0 (Chen, 2004) was then used to identify the TE relevant proteins. Meanwhile, tandem repeats were predicted by using Tandem Repeats Finder (TRF) v4.07b (Benson, 1999), and Tandem Repeats Analysis Program (Sobreira et al., 2006) was used to select candidate microsatellite markers from the TRF output.

Gene models in the sterlet genome were predicted by an integrated strategy of three methods. For homology annotation, we downloaded published protein sequences of ten representative vertebrates including zebrafish (*Danio rerio*), spotted gar (*Lepisosteus oculatus*), elephant shark (*Callorhinchus milii*), sea lamprey (*Petromyzon marinus*), medaka (*Oryzias latipes*), Nile tilapia (*Oreochromis niloticus*), three-spined stickleback (*Gasterosteus aculeatus*), Atlantic cod (*Gadus morhua*), fugu (*Takifugu rubripes*) and spotted green pufferfish (*Tetraodon nigroviridis*), and aligned them against the assembly of the sterlet genome using BLAST (Altschul et al., 1990) with tblastn mode and an e-value of 1e-5. SOLAR (Yu et al., 2006) was subsequently employed to select the best hit for each alignment. For *ab initio* prediction, the sterlet genome assembly was masked according to the previously identified repeated sequences and was then scanned using AUGUSTUS v3.2.3 (Stanke et al., 2006) and GENSCAN v1.0 (Burge and Karlin, 1997) to predict gene structures. For transcriptome-based annotation, we sequenced a blood transcriptome on a Hiseq X10 platform (Illumina), mapped the reads to the genome scaffolds using TopHat v2.0.13 (Trapnell et al., 2009) and assembled them into transcripts using Cufflinks v2.2.1 (Trapnell et al., 2010). Finally, all predicted genes from these three methods were merged and filtered by GLEAN v1.1 (Elsik et al., 2007) to create a consensus gene set.

Gene functional annotation of the sterlet genome was firstly performed by aligning all the protein sequences produced by GLEAN against public databases including Swiss-Prot, TeEMBL (Boeckmann et al., 2003) and KEGG (Kanehisa et al., 2016) using BLASTP v2.3.0+ (Altschul et al., 1990) with an e-value of 1e-5. Subsequently, motifs and domains were annotated using InterProScan (Hunter et al., 2008) by searching PANTHER (Thomas et al., 2003), Pfam (Finn et al., 2013), PRINTS (Attwood, 2002), ProDom (Bru et al., 2005) and SMART (Letunic et al., 2004) databases. Finally, InterProScan (Hunter et al., 2008) was applied to assign Gene Ontology (GO) terms and conduct a GO enrichment analysis (Ashburner et al., 2000).

Fossil-Calibrated Phylogenetic Analysis

To perform a phylogenetic analysis of the sterlet, we obtained the predicted coding sequences (CDS) from the sterlet and 14 other vertebrates, including Asian arowana (*Scleropages formosus*), coelacanth (*Latimeria menadoensis*), common carp (*Cyprinus carpio*) and Atlantic salmon (*Salmo salar*) as well as the ten species used for homology gene annotation, and used the sea lamprey as the outgroup. BLAST with blastp mode and an e-value of 1e-5 were used to build the super similarity matrix, followed by OrthoMCL (Li et al., 2003) to distinguish gene families. One-to-one orthologues were identified by Markov Chain Clustering (MCL) and were aligned by MUSCLE v3.7 (Edgar, 2004). The first nucleotide of each codon was chosen to construct a Maximum-likelihood (ML) tree using PhyML v3.0 (Guindon et al., 2010) with gamma distribution across aligned sites and HKY85 substitution model. Branch supports were evaluated by approximate likelihood ratio test (aLRT). Meanwhile, we also conducted Bayesian inference (BI) independently using MrBayes v3.2.2 (Ronquist et al., 2012) to confirm the topology deduced from ML. Totally, we performed 100,000 generations and sampled every 100 generations. The initial 20% of the runs were regarded as unreliable samples and were discarded. The rest of the samples were used to estimate the branch supports. The divergence time of the sterlet from other vertebrates was estimated by Bayesian method using MCMCtree in PAML v4.9 (Yang, 2007) with two fossil calibrations, which are *Latimeria* (Sarcopterygii, 408.0 ~ 427.9 Mya) and *Danio* (Teleostei, 151.2 ~ 252.7 Mya; Hughes et al., 2018).

4dTv Analysis to Determine the Sturgeon-Specific Genome Duplication

We performed 4-fold degenerative third-codon transversion (4dTv) analysis to test the sturgeon-specific GD by comparing the sterlet genome to Asian arowana genome. Protein sequences from the two genomes were firstly aligned using all-to-all BLAST with blastp mode and an e-value of 1e-5. Subsequently, syntenic regions between sterlet-sterlet, arowana-arowana and sterlet-arowana were identified by MCscan v0.8 (Wang et al., 2012) with default parameters. Homologous protein sequences from these syntenic regions were retrieved and converted to CDS for alignment by MUSCLE (Edgar, 2004). Lastly, 4dTv values were calculated and corrected with the HKY model in PAML package (Yang, 2007).

Hox-Cluster Identification and Phylogenetic Analysis

Reference protein sequences of complete *Hox*A cluster and partial *Hox*D cluster of American paddlefish (*Polyodon spathula*) (Crow et al., 2012) were downloaded from National Center of Biotechnology Information (NCBI). Sequences of four complete *Hox* clusters of the Indonesian coelacanth (Amemiya et al., 2010) and spotted gar (Braasch et al., 2015) were downloaded from Ensembl. The protein sequences were firstly aligned to the sterlet genome assembly by BLAST (Altschul et al., 1990) with tblastn mode and the hit sequences were further analyzed by Exonerate software (Slater

and Birney, 2005) to extract exons. *Hox* gene order and synteny were finally determined by aligning back to the genome assembly and the best hits were selected by SOLAR (Yu et al., 2006). The *Hox*A clusters from the sterlet and paddlefish, as well as *HoxA9* genes from ten vertebrates were separately aligned with MEGA v7.0.26 (Kumar et al., 2016) followed by construction of a ML phylogenetic tree.

RESULTS AND DISCUSSION

Summary of the Genome Sequencing and Assembly

We generated 316.8 Gb of pair-end raw reads (**Supplementary Table 1**) to assemble the draft genome of the sterlet. After filtering low-quality sequences, the data size of the remaining clean reads was about 248.4 Gb (**Supplementary Table 1**). The haploid genome size of the tarlet was estimated (**Supplementary Figure 2**) by a k-mer analysis (Li et al., 2010). Using all the clean reads, we produced a final genome assembly of 1.83Gb, which is quite close to the previously reported 1.87 Gb by flow cytometry (Birstein et al., 1993). The achieved draft assembly had a contig N50 of 18.88 kb and a scaffold N50 of 191.06 kb (**Table 1**).

Accordingly, the genome sequencing depth for the tarlet reached 132-fold based on the final 1.83-Gb assembly, and as much as 87.19% of the bases had an over 20-fold sequencing depth (**Supplementary Figure 3**). The total completeness of the assembly was estimated to be 81.6% by evaluation with BUSCO, including 51.9% complete and single-copy BUSCOs and another 29.7% duplicated BUSCOs. A total of 4,584 genes were searched and 302 (6.6%) of them were fragmental BUSCOs (**Supplementary Table 2**). Along with the homogeneous GC distribution of the scaffolds (**Supplementary Figure 4**), we concluded that our draft assembly of the tarlet genome was qualified for further analyses.

A Relatively High Content of Repetitive Elements

We performed repeat annotation, and a total of 784-Mb (42.84%) repeated sequences, including 726-Mb (39.68%) transposable elements (Tes) and 79 Mb (4.34%) tandem repeats, were identified in the tarlet genome assembly (**Supplementary Table 3**). These data are consistent with the dominant sub-peak ideally located at 2-fold the position of the main k-mer peak (**Supplementary Figure 2**). This repeat content was higher than those of the majority of the published fish genomes that usually contain no more than 40% repeats (Yuan et al., 2018). Interestingly, more class I (28.95%) than class II (14.93%) Tes were found in the tarlet genome (**Supplementary Table 4**), which resembled a cartilaginous species pattern (Yuan et al., 2018). In addition, as a potamodromous species dwelling mainly in freshwater, the sterlet had a relatively high DNA/TcMar-Tc1 proportion (16.58% for 130 Mb) but a relatively low microsatellites proportion (2.10% for 16 Mb) (**Supplementary Table 5**), a pattern preferred by freshwater species (**Supplementary Figure 5;** Yuan et al., 2018).

Furthermore, we identified 318 copies of *Tana1*, a new putative active *Tc1*-like transposable element (Pujolar et al., 2013) but not referred in the repeat annotation library (Romanenko et al., 2015). Our results showed that 299 of the predicted *Tana1* copies contain full-length transposases. Interestingly, the majority of these *Tana1* copies did not have internal stop codon(s) as determined in the a previous study (Pujolar et al., 2013), suggesting that this element is more likely to be active. The 299 complete *Tana1* genes were from 250 different scaffolds, with an average of 1.19 genes in each scaffold. Sequences and gene locations of the identified *Tana1* are publicly available in figshare with an accession ID of doi: 10.6084/m9.figshare.8289881.

We then calculated the number of repeats that were co-localized with the protein coding genes after gene annotation to estimate their potential functions. Our results showed that a total of 34,987 repeats (14.23 Mb in length, accounting for 1.82% of all repeats) were co-localized with 10,460 protein coding genes, among which LINE/CR1, DNA/TcMar-Tc1 and LINE/L2 were the most abundant types (**Supplementary Data Sheet 2**). The GO enrichment analysis revealed that these repeats were enriched into 52 terms. Cellular process, binding, single-organism process, metabolic process and biological regulation were the top five enriched ones (**Supplementary Figure 6**), indicating that these repeats may participate in such biological processes.

However, the distribution and location of these repeats and annotated genes on chromosomes are still awaiting identification with assistance of on-going PacBio sequencing. It seems that repetitive DNA sequences have a tendency to cluster in specific regions, such as in pericentromeric, centromeric and telomeric regions (Biltueva et al., 2017). The potential roles of repetitive sequences in chromosomal rearrangements will also be much

TABLE 1 | Statistics of assembled contigs and scaffolds.

Parameter	Contig		Scaffold	
	Size (bp)	Number	Size (bp)	Number
N90	450	257,242	1,325	38,164
N80	2,365	84,215	32,254	9,354
N70	7,919	48,548	62,161	5,330
N60	13,208	32,851	109,208	3,086
N50	18,882	22,595	191,062	1,801
Longest (bp)	223,430		5,122,172	
Total Size (bp)	1,622,894,949		1,831,554,666	
Total Number (>100 bp)	1,255,020		985,522	
Total Number (>2,000 bp)	91,019		27,173	

clearer, once a chromosome-level genome assembly is available for the sterlet.

Statistics of Gene Annotation and Phylogenetic Analysis

After masking the abundant repeats in scaffolds, we annotated 22,184 protein-coding genes with an average gene length of 21 kb using a combined strategy of *ab initio*, homology-based and transcriptome-based annotation. This predicted gene number of the sterlet genome seems to be lower than estimation, possibly due to missing data and many gaps in the draft assembly. In addition, the repetitive sequences and complex polyploidy (Romanenko et al., 2015) make it more difficult to produce a fine assembly and to predict a complete gene set. Our BUSCO analysis of the gene set showed that complete and fragmented BUSCOs accounted for 73.2% of the searched genes, and 26.8% were missing BUSCOs (**Supplementary Table 2**); we therefore inferr that the total gene number of the sterlet could reach 28,136 (with the addition of the missing BUSCOs), which is more than that of a diploid fish but less than a tetraploid species when taking the partial tetraploidy into consideration. Statistics of the gene list are provided in **Supplementary Table 6**. Length distributions of the predicted genes, CDS, exons and introns were comparable to those of spotted gar, elephant shark and many other fishes (**Supplementary Figure 7**). Of all these genes, a total of 21,112 genes (95.17%) were functionally annotated in at least one public database (find more details in **Supplementary Table 7**).

Afterwards, the predicted CDS sequences along with whole-genome CDS from other 14 examined vertebrates were clustered into gene families to determine 198 single-copy consensus orthologues from these genomes (**Supplementary Table 8**; **Supplementary Figure 8**), which were selected out for generation of the phylogenetic topology by ML (**Supplementary Figure 9**) or BI (**Supplementary Figure 10**). The two methods produced a complete coincidence of phylogenetic topology with high branch support values, suggesting that the hypothesis was well supported (**Figure 1A**). Our tree confirms the results of others (Hughes et al., 2018; Peng et al., 2007), that the sterlet is located at a base position of Actinopterygii, which serves as a sister group to all ray-finned fishes. Therefore, this phylogeny of the sterlet using numerous single-copy genes confirms its very basal position as reported by other studies. Fossil calibrations date the origin of the sterlet back to 358 Mya (**Figure 1A**), with a 95% confidence interval of 316~394 Mya (**Supplementary Figure 11**). These data are consistent with our previous comprehensive phylogeny analysis (Hughes et al., 2018), and most interestingly, this date is extremely close to the Late Devonian Extinction that happened around 358.9 Mya (McGhee et al., 1984).

Identification of an Independent WGD Event that Occurred Recently in the Sterlet

Sturgeons didn't experience the TGD event (Ravi and Venkatesh, 2018), but there are clear evidences that there was a sturgeon-specific GD event (Havelka et al., 2016). In order to identify this lineage-specific GD in the sterlet, we performed a 4dTv analysis along with Asian arowana (Bian et al., 2016), which had experienced the TGD event around 320 Mya (Jaillon et al., 2004). Our analysis displayed distinct peaks in each group of sterlet-sterlet (sturgeon-specific GD), arowana-arowana (TGD) and sterlet-arowana (speciation event), and the synonymous transversions rates (Ks values) were estimated to be 0.03 and 0.45 in the sterlet and Asia arowana, respectively (**Figure 1B**). Hence, the sturgeon-specific GD was estimated to have occurred about 21.3 Mya ([320 Mya/0.45]*0.03) d, long after the evolutionary splitting between the sturgeon and paddlefish (184 Mya; Peng et al., 2007). Hence, it that sturgeons (Acipenseridae) and paddlefish (Polyodontidae) experienced polyploidization events independently.

Characterization of the Complete *Hox* Clusters

To provide additional insights into polyploidy of the genome at the gene level after the sturgeon-specific GD event, we investigated *Hox* gene clusters in the sterlet genome. We identified seven *Hox* clusters including 68 *Hox* genes (60 intact and 8 partial/pseudo genes) in the draft assembly (**Figure 1C**, **Supplementary Data Sheet 3**). The *Hox* data seemed to be a consequence of the sturgeon-specific GD, since only four *Hox* clusters were identified in sea lamprey (43 genes), elephant shark (47 genes) and spotted gar (43 genes; Venkatesh et al., 2014). Interestingly, the possible absence of a whole *Hox*C cluster in the sterlet is similar to that in some diploid teleost such as fugu, medaka and stickleback (Pascual-Anaya et al., 2013). Furthermore, our *Hox*A based genealogy showed that, contrary to the *Hox* pattern in teleost after TGD (**Supplementary Figure 12**), *Hox*A clusters from the sterlet and paddlefish formed two separate groups (**Supplementary Figure 12**), which indicates that *Hox* genes duplicated independently after the divergence of the two families. It confirmed the independence of lineage-specific GDs in the sterlet and paddlefish, which is consistent with our above-mentioned prediction by 4dTv.

However, whether this WGD is sturgeon-specific or shared by all members of the Acipenseridae family is awaiting answers from genome sequencing of more sturgeon species. Furthermore, the present research on a complete gene-chromosome pattern of the sterlet genome is still preliminary, but this work and a previous report of sequencing 15 chromosome-specific libraries (Andreyushkova et al., 2017) provide some novel insights. We attempted to map our assembly to the spotted gar chromosomes, but the results were difficult to interpret, possibly due to the non-full-length assembly of our current draft genome, the great complexity of the sterlet chromosomes, and high sequence divergences between the two fish species. Therefore, based on our current knowledge on the sterlet genome (Romanenko et al., 2015; Andreyushkova et al., 2017), a chromosome-level assembly needs to be generated, with assistance of long-read sequencing and chromatin conformation capture technology for a better understanding of the complicated structure and evolutionary pattern of the sterlet genome.

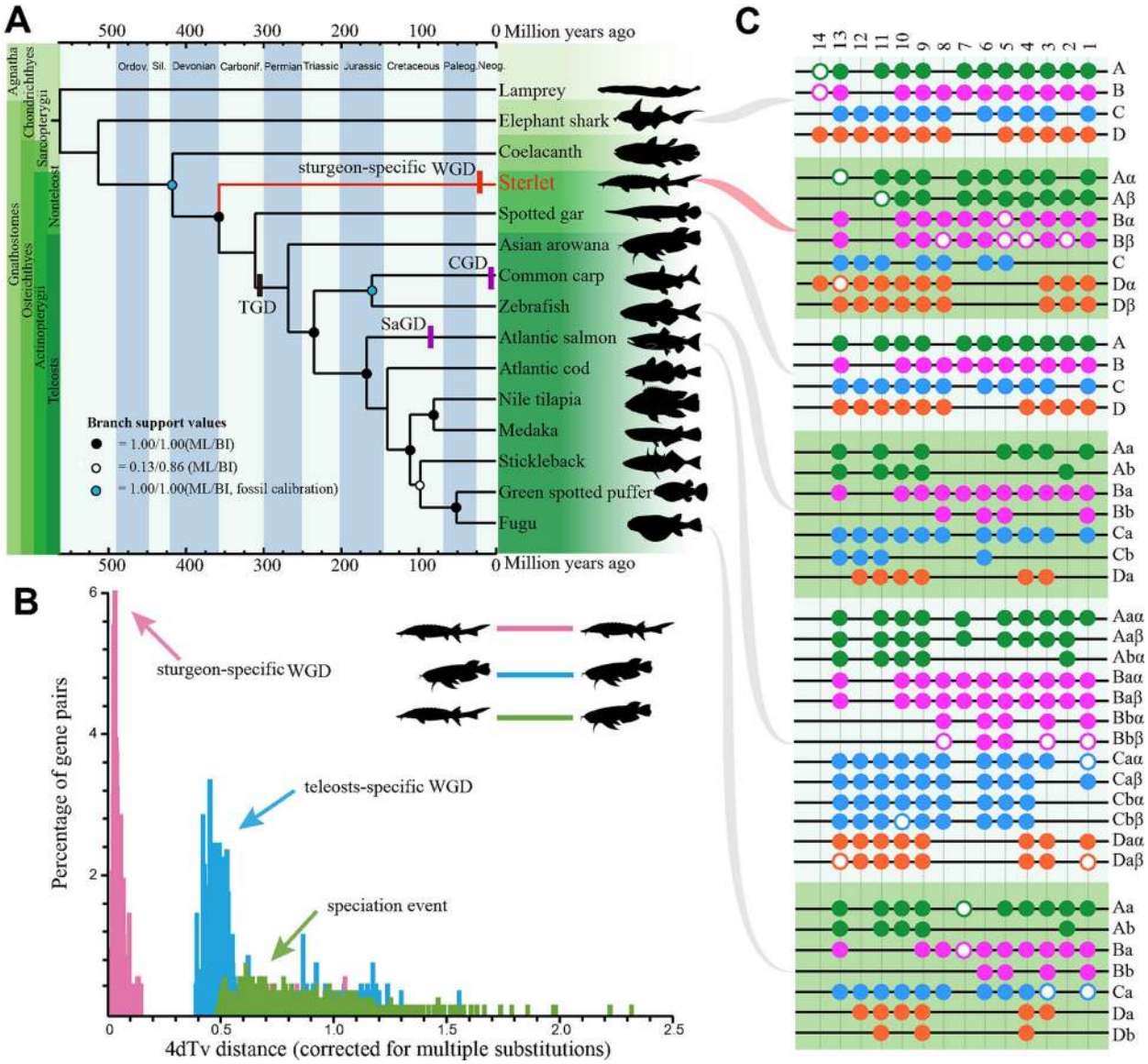

FIGURE 1 | The sterlet takes the most basal position at the phylogeny of ray-finned fish and evolved seven *Hox* clusters after the lineage-specific whole genome duplication. **(A)** The fossil-calibrated phylogenetic tree of 15 examined vertebrates including the sterlet. The phylogenetic topology was deduced by both the ML and BI methods. TGD, teleost-specific GD; CGD, carp GD; SaGD, salmonid GD. **(B)** A 4dTv comparison between Asian arowana and the sterlet. **(C)** Presence of *Hox* clusters in elephant shark (Venkatesh et al., 2014), sterlet (this study), spotted gar (Braasch et al., 2015), zebrafish (Bian et al., 2016), Atlantic salmon (Lien et al., 2016) and fugu (Bian et al., 2016). Each black line refers to a *Hox* cluster. Solid circles represent complete *Hox*A (green), *Hox*B (pink), *Hox*C (blue) and *Hox*D (orange) genes, while hollow circles stand for pseudo or partial genes. Paralogs generated by TGD were labeled with a and b, whereas paralogs produced by lineage-specific GD were named by α and β.

DATA AVAILABILITY

The datasets generated for this study can be found in the NCBI with accession PRJNA491785, SRR8371834 ~ SRR837184.

ETHICS STATEMENT

All experiments in the present study were carried out in accordance with the guidelines of the Animal Ethics Committee of Yangtze River Fisheries Research Institute of Chinese Academy of Fishery Sciences (No. YFI-01).

AUTHOR CONTRIBUTIONS

QW, HD, CL, JX, and QS, conceived and designed the project. YH, PC, YL, and CB, analyzed the data. CL, RR, HY, and XY collected and processed the samples. PC, YH, and QS wrote the

manuscript. QS, XL and QW revised the manuscript. All authors have read and approved the final manuscript and declared no competing interests.

FUNDING

The study was supported by the the National Natural Science Foundation of China (grant number NSFC 31772854), China Postdoctoral Science Foundation (grant number 2017M622560), the National Program on Key Basic Research Project (973 Program, 2015CB15072), Hubei Postdoctoral Innovation Post Project (No. 2017C08), Shenzhen Special Program for Development of Emerging Strategic Industries (No. JSGG20170412153411369) and Office of Fisheries Supervision and Management for the Yangtze River Basin, MARA, PRC (No. 171821301354051046).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00776/full#supplementary-material.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Bio.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Amemiya, C. T., Alföldi, J., Lee, A. P., Fan, S., Philippe, H., MacCallum, I., et al. (2000). The African coelacanth genome provides insights into tetrapod evolution. *Nature* 496, 311–316. doi: 10.1038/nature12027

Amemiya, C. T., Powers, T. P., Prohaska, S. J., Grimwood, J., Schmutz, J., Dickson, M., et al. (2010). Complete HOX cluster characterization of the coelacanth provides further evidence for slow evolution of its genome. *Proc. Natl. Acad. Sci. U. S. A.* 107, 3622–3627. doi: 10.1073/pnas.0914312107

Andreyushkova, D., Makunin, A., Beklemisheva, V., Romanenko, S., Druzhkova, A., Biltueva, L., et al. (2017). Next generation sequencing of chromosome-specific libraries sheds light on genome evolution in paleotetraploid sterlet (*Acipenser ruthenus*). *Genes* 8, 318. doi: 10.3390/genes8110318

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Attwood, T. K. (2002). The PRINTS database: a resource for identification of protein families. *Brief. Bioinform.* 3, 252–263. doi: 10.1093/bib/3.3.252

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Bian, C., Hu, Y., Ravi, V., Kuznetsova, I. S., Shen, X., Mu, X., et al. (2016). The Asian arowana (*Scleropages formosus*) genome provides new insights into the evolution of an early lineage of teleosts. *Sci. Rep.* 6, 24501. doi: 10.1038/srep24501

Biltueva, L. S., Prokopov, D. Y., Makunin, A. I., Komissarov, A. S., Kudryavtseva, A. V., Lemskaya, N. A., et al. (2017). Genomic organization and physical mapping of tandemly arranged repetitive DNAs in sterlet (*Acipenser ruthenus*). *Cytogenet. Genome Res.* 152, 148–157. doi: 10.1159/000479472

Birstein, V. J., Poletaev, A. I., and Goncharov, B. F. (1993). DNA content in Eurasian sturgeon species determined by flow cytometry. *Cytometry* 14, 377–383. doi: 10.1002/cyto.990140406

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31, 365–370. doi: 10.1093/nar/gkg095

Braasch, I., Gehrke, A. R., Smith, J. J., Kawasaki, K., Manousaki, T., Pasquier, J., et al. (2015). The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat. Genet.* 47, 427–437. doi: 10.1038/ng.3526

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S., and Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.* 33, D212–D215. doi: 10.1093/nar/gki034

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Sci.* 268, 78–94. doi: 10.1006/jmbi.1997.0951

Chen, N. (2004). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* 4, 4.10. doi: 10.1002/0471250953.bi0410s25

Crow, K. D., Smith, C. D., Cheng, J. F., Wagner, G. P., and Amemiya, C. T. (2012). 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. doi: 10.1093/gbe/evs067

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Elsik, C. G., Mackey, A. J., Reese, J. T., Milshina, N. V., Roos, D. S., and Weinstock, G. M. (2007). Creating a honey bee consensus gene set. *Genome Biol.* 8, R13. doi: 10.1186/gb-2007-8-1-r13

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2013). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Havelka, M., Bytyutskyy, D., Symonová, R., Ráb, P., and Flajšhans, M. (2016). The second highest chromosome count among vertebrates is observed in cultured sturgeon and is associated with genome plasticity. *Genet. Sel. Evol.* 48, 12. doi: 10.1186/s12711-016-0194-0

Hughes, L. C., Ortí, G., Huang, Y., Sun, Y., Baldwin, C. C., Thompson, A. W., et al. (2018). Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proc. Natl. Acad. Sci. U. S. A.* 115, 6249–6254. doi: 10.1073/pnas.1719358115

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2008). InterPro: the integrative protein signature database. *Nucleic Acids Res.* 37, D211–D215. doi: 10.1093/nar/gkn785

Jaillon, O., Aury, J. M., Brunet, F., Petit, J. L., Stange-Thomann, N., Mauceli, E., et al. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431, 946–957. doi: 10.1038/nature03025

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. *Genome Res.* 110, 462–467. doi: 10.1159/000084979

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Linnaeus, C. (1758). *Systema naturae per regna tria naturae*. 10th ed. (Stockholm: Laurentius Salvius). doi: 10.5962/bhl.title.542.

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., et al. (2004). SMART 4.0: towards genomic data integration. *Nucleic Acids Res.* 32, D142–D144. doi: 10.1093/nar/gkh088

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., et al. (2010). The sequence and *de novo* assembly of the giant panda genome. *Nature* 463, 311–317. doi: 10.1038/nature08696

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., et al. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature* 533, 200–205. doi: 10.1038/nature17164

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* 1, 18. doi: 10.1186/2047-217X-1-18

McGhee, G. R., Gilmore, J. S., Orth, C. J., and Olsen, E. (1984). No geochemical evidence for an asteroidal impact at Late Devonian mass extinction horizon. *Nature* 308, 629. doi: 10.1038/308629a0

Pascual-Anaya, J., D'Aniello, S., Kuratani, S., and Garcia-Fernàndez, J. (2013). Evolution of Hox gene clusters in deuterostomes. *BMC Dev. Bio.* 13, 26. doi: 10.1186/1471-213X-13-26

Peng, Z., Ludwig, A., Wang, D., Diogo, R., Wei, Q., and He, S. (2007). Age and biogeography of major clades in sturgeons and paddlefishes (Pisces: Acipenseriformes). *Mol. Phylogenet. Evol.* 42, 854–862. doi: 10.1016/j.ympev.2006.09.008

Pujolar, J. M., Astolfi, L., Boscari, E., Vidotto, M., Barbisan, F., Bruson, A., et al. (2013). Tana1, a new putatively active Tc1-like transposable element in the genome of sturgeons. *Mol. Phylogenet. Evol.* 66, 223–232. doi: 10.1016/j.ympev.2012.09.025

Rabosky, D. L., Santini, F., Eastman, J., Smith, S. A., Sidlauskas, B., Chang, J., et al. (2013). Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nat. Commun.* 4, 1958. doi: 10.1038/ncomms2958

Ravi, V., and Venkatesh, B. (2018). The divergent genomes of teleosts. *Annu. Rev. Anim. Biosci.* 6, 47–68. doi: 10.1146/annurev-animal-030117-014821

Romanenko, S. A., Biltueva, L. S., Serdyukova, N. A., Kulemzina, A. I., Beklemisheva, V. R., Gladkikh, O. L., et al. (2015). Segmental paleotetraploidy revealed in sterlet (*Acipenser ruthenus*) genome by chromosome painting. *Mol. Cytogenet.* 8, 90. doi: 10.1186/s13039-015-0194-8

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Slater, G. S. C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* 6, 31. doi: 10.1186/1471-2105-6-31

Sobreira, T. J. P., Durham, A. M., and Gruber, A. (2006). TRAP: automated classification, quantification and annotation of tandemly repeated sequences. *Bioinformatics* 22, 361–362. doi: 10.1093/bioinformatics/bti809

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439. doi: 10.1093/nar/gkl200

Thomas, P. D., Kejariwal, A., Campbell, M. J., Mi, H., Diemer, K., Guo, N., et al. (2003). PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res.* 31, 334–341. doi: 10.1093/nar/gkg115

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Venkatesh, B., Lee, A. P., Ravi, V., Maurya, A. K., Lian, M. M., Swann, J. B., et al. (2014). Elephant shark genome provides unique insights into gnathostome evolution. *Nature* 505, 174–179. doi: 10.1038/nature12826

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49. doi: 10.1093/nar/gkr1293

Wei, Q. W., Zou, Y., Li, P., and Li, L. (2011). Sturgeon aquaculture in China: progress, strategies and prospects assessed on the basis of nation-wide surveys (2007–2009). *J. Appl. Ichthyol.* 27, 162–168. doi: 10.1111/j.1439-0426.2011.01669.x

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yu, X. J., Zheng, H. K., Wang, J., Wang, W., and Su, B. (2006). Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. *Genomics* 88, 745–751. doi: 10.1016/j.ygeno.2006.05.008

Yuan, Z., Liu, S., Zhou, T., Tian, C., Bao, L., Dunham, R., et al. (2018). Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. *BMC Genom.* 19, 141. doi: 10.1186/s12864-018-4516-1

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer JM declared a shared affiliation, with no collaboration, with several of the authors, PC, XL, to the handling editor at the time of review.

Copyright © 2019 Cheng, Huang, Du, Li, Lv, Ruan, Ye, Bian, You, Xu, Liang, Shi and Wei. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.