MOLECULAR ECOLOGY

Molecular Ecology (2013) doi: 10.1111/mec.12234

Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing

R. OGDEN,* K. GHARBI,† N. MUGUE,‡ J. MARTINSOHN,§ H. SENN,¶ J. W. DAVEY,** M. POURKAZEMI,†† R. MCEWING,* C. ELAND,† M. VIDOTTO,‡‡ A. SERGEEV‡ and L. CONGIU‡‡

*TRACE Wildlife Forensics Network, RZSS, Edinburgh, EH12 6TS, UK, †The GenePool, School of Biological Sciences, The University of Edinburgh, Edinburgh, UK, ‡Russian Institute for Fisheries and Oceanography (VNIRO), Moscow 107140, Russia, §Joint Research Centre of the European Commission, Maritime Affairs Unit, Institute for the Protection and Security of the Citizen, Via Fermi TP 051, Ispra, VA, 21027, Italy, ¶WildGenes Laboratory, Royal Zoological Society of Scotland, Edinburgh, EH12 6TS, UK, **Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JT, UK, ††Sturgeon International Research Institute, PO Box 41635-3464, Rasht, Iran, ‡‡Department of Biology, University of Padova, Via U. Bassi 58/b 35121, Padova, Italy*

Abstract

Caviar-producing sturgeons belonging to the genus *Acipenser* are considered to be one of the most endangered species groups in the world. Continued overfishing in spite of increasing legislation, zero catch quotas and extensive aquaculture production have led to the collapse of wild stocks across Europe and Asia. The evolutionary relationships among Adriatic, Russian, Persian and Siberian sturgeons are complex because of past introgression events and remain poorly understood. Conservation management, traceability and enforcement suffer a lack of appropriate DNA markers for the genetic identification of sturgeon at the species, population and individual level. This study employed RAD sequencing to discover and characterize single nucleotide polymorphism (SNP) DNA markers for use in sturgeon conservation in these four tetraploid species over three biological levels, using a single sequencing lane. Four population meta-samples and eight individual samples from one family were barcoded separately before sequencing. Analysis of 14.4 Gb of paired-end RAD data focused on the identification of SNPs in the paired-end contig, with subsequent *in silico* and empirical validation of candidate markers. Thousands of putatively informative markers were identified including, for the first time, SNPs that show population-wide differentiation between Russian and Persian sturgeons, representing an important advance in our ability to manage these cryptic species. The results highlight the challenges of genotyping-by-sequencing in polyploid taxa, while establishing the potential genetic resources for developing a new range of caviar traceability and enforcement tools.

Keywords: *Acipenser*, caviar, DNA, fisheries, traceability, wildlife forensics

Received 29 June 2012; revision received 11 December 2012; accepted 13 December 2012

Introduction

The genus *Acipenser* contains 17 of the 25 caviar-producing fish commonly termed 'sturgeon'. Distributed at temperate latitudes throughout the northern hemisphere, sturgeons have been fished for meat for many centuries, but are now targeted primarily for the harvesting of caviar. Within Europe and Asia, exploitation of wild stocks intensified rapidly during the 20th century, and all major sturgeon fisheries are now in decline due to overfishing (Pikitch *et al.* 2005). In the Caspian Sea, historically the world's largest sturgeon fishery, the decline has continued despite increasing levels of man-

Correspondence: Rob Ogden, Fax: +44 131 3140317; E-mail: rob.ogden@tracenetwork.org

agement intervention that has resulted in outright bans on commercial fishing (Pourkazemi 2006). While disease outbreaks and hydroelectric power developments are known to have reduced Caspian Sea populations, the greatest threat to sustainability of stocks is Illegal, Unreported or Unregulated (IUU) fishing that occurred at levels surpassing legal quotas in the 1990s and has continued beyond the imposition of a fishing moratorium this century (Ruban & Khodorevskaya 2011). In other regions, such as the Adriatic Sea, within the Mediterranean, the native species, *A. naccarii,* has since May 2010 been listed by IUCN as critically endangered (IUCN 2012) and possibly extinct in the wild due to overfishing (Boscari *et al.* 2011). Similar catastrophic predictions have been made for a number of other species within the genus, which in March 2010 was identified as the most endangered group of species with 85% of sturgeons at risk of extinction (Congiu *et al.* 2011).

In parallel to this decline in wild stocks, and driven by the economic value of caviar, efforts have been made to develop commercial aquaculture programmes to produce caviar, originally in Europe and North America, but more recently also in Caspian Sea range states such as Russia and Iran (Raymakers & Hoover 2002). At the present time, it is estimated well over 50% of caviar in trade is harvested from farmed stocks (Bronzi *et al.* 2011). The capacity to produce farmed caviar could theoretically alleviate fishing pressure on wild stocks; however, as has been observed with other rare wildlife species of high commercial value, despite sturgeon farming, fishermen might still be incentivized to catch every last fish available, even in the face of increasingly stringent wildlife protection laws.

All *Acipenser* species are now listed under Appendices I or II of CITES (Convention on International Trade in Endangered Species) with the intention of controlling international trade and promoting the implementation of sustainable management policies (Raymakers 2006; CITES & UNEP 2012). Regulations within the EU, the world's largest caviar importer as well as a major producer, include a strict labelling system for all caviar products detailing the species name and country of origin, in an attempt to restrict the caviar trade to product derived from CITES quotas (permissible wild trade), approved farms or licensed repackaging companies (EC 2006). Despite such regulations, the value of caviar drives a multimillion euro black market economy in which caviar from nonsustainable sources is known to be widely traded under mislabelled packaging (European Commission 2006a,b).

To address the issue of illegal trade, it is essential that there exists a robust, legally valid method of authenticating the caviar labelling system, ideally embedded in a forensic framework. This would enable customs officials and trade authorities to test products in trade on a routine basis but also as part of intelligence-led targeted investigations. Such a testing system should allow traceability of a product to source: either the wild geographic origin or the farm where the caviar was produced. In the first instance, it is vital to be able to accurately determine the species of origin of traded caviar to assess the validity of the product label. A broad range of analytical tools have been examined for their utility in caviar identification (Rehbein *et al.* 2008), with the most promising technology for general application found to be DNA-based methods (Waldman *et al.* 2008). In fact, DNA markers have been used for caviar identification at a species level for over 15 years (DeSalle & Birstein 1996) and are still regularly employed to support enforcement action (Ludwig 2008). However, the current method of identification, DNA sequencing of the mitochondrial *cytochrome b* gene, does not provide sufficient resolution to discriminate among certain species, in particular four species commonly found in Europe and central Asia either as traded products or wild fish, namely Russian (*A. gueldenstaedtii*), Persian (*A. persicus*), Siberian (*A. baerii*) and Adriatic (*A. naccarii*) sturgeons (Ludwig 2008). Although *A. baerii* can now be distinguished from the other three species at the mitochondrial control region (Mugue *et al.* 2008), the remaining three species have so far proved indistinguishable, severely limiting Europe's ability to enforce existing regulations concerning the caviar trade. There is therefore an urgent need to discover, test and validate new molecular markers capable of species identification in this group.

Beyond species identification, it is also becoming increasingly important to be able to verify whether caviar has been produced through aquaculture or was harvested from the wild. With wild stocks continuing to decline and zero CITES quotas in place for Caspian and Azov sea sturgeons, it is likely that future trade will be completely restricted to aquaculture products. To be able to impose such constraints, any regulations must be enforceable and that would require the ability to trace back caviar to its farm of origin or at least to enable caviar to be excluded from the source described on the packaging. Nuclear genetic markers are already being used in this context in Russia, as part of a 'genetic passport' system to licence caviar-producing farms by associating caviar products with specific registered genotypes (N. Mugue & L. Congiu unpublished data). However, more genetic markers across a broader range of species are needed, if such a method is to ever be implemented commercially.

Aside from enforcing trade regulations or ensuring traceability, molecular genetic markers are also required to support conservation genetic management of the

remaining sturgeon stocks. Around the Caspian Sea, restocking programmes are releasing millions of sturgeon fry into the wild in an attempt to augment dwindling populations; however, these efforts have been largely unmanaged and are potentially leading to large skews in the genetic variability and adaptive capacity of augmented stocks, resulting in specific recommendations for genetic management published in the proceedings from the 6th International Sturgeon Symposium in 2009 (Rosenthal *et al.* 2011). There is also a lack of fundamental molecular ecological data on the remaining wild populations of sturgeon in Europe and Asia.

A number of different types of genetic markers may be applied to investigate population structure, determine geographic origin and implement traceability in fisheries (Ogden 2008). The choice of marker typically depends on the taxonomic level being examined. Traditionally, mitochondrial DNA (mtDNA) sequencing of the *cytochrome b* or *cytochrome oxidase I* genes is used for species discrimination, and the mtDNA control region for identifying highly differentiated populations and panels of microsatellite markers for resolving finer scale population structure and performing parentage assignment. More recently, single nucleotide polymorphism markers (SNPs) have become the favoured marker for population genetic analysis due to their distribution throughout the nuclear genome, their association with either neutral or adaptive variation and their relative ease of genotyping and method transfer among laboratories. These characteristics render SNPs particularly relevant to the future development of fisheries forensic techniques (Ogden 2011). The polyploid nature of sturgeon confers further advantages on the use of SNPs; genotyping microsatellite markers in polyploid species is often reduced to a dominant (presence/absence) scoring system, whereas allele copy number and codominant genotype can more easily be resolved for SNP markers.

Methods for discovering SNP markers have been hugely influenced by the increasing availability of high-throughput sequencing techniques for nonmodel species (Garvin *et al.* 2010; Davey *et al.* 2011). The potential now exists to identify very large numbers of candidate SNP markers throughout the genome, with discovery targeted towards SNPs that are informative for the particular question at hand. A common aim of these methods is to reduce the proportion of the genome that is subject to sequencing (reduced representation sequencing), to enable multiple copies of the same section of DNA to be compared and SNP markers identified. One such method, restriction site associated DNA (RAD) sequencing (Miller *et al.* 2007; Baird *et al.* 2008), is becoming increasingly popular for SNP discovery and was chosen for use in this study.

One of the advantages of the RAD approach is that it is possible to attach identifying DNA barcodes to individual samples or pools of samples during the preparation of genomic libraries, allowing them to be separated during analysis of the downstream sequence data. This ultimately allows SNP markers to be selected that show variation between individuals, populations or species (Hohenlohe *et al.* 2011) and creates the potential for simultaneous genotyping-by-sequencing (GBS) of populations. Additionally, by utilizing paired-end RAD sequencing (RAD-PE), it is possible to generate contigs of sufficient depth to discover SNPs but also, importantly, of sufficient length to enable genotyping assay design (Etter *et al.* 2011; Willing *et al.* 2011). However, in practice, the ability to accurately identify the presence of SNPs and to genotype individual samples based on RAD-PE data is complicated by the occurrence of sequencing errors inherent to high-throughput sequencing techniques. These errors, and the subsequent risk of false SNP discovery, can be mitigated to some extent by ensuring sufficient sequence read depth and by applying filters that optimize SNP calling and moderate downstream genotyping. Nevertheless, bioinformatic analysis remains a balance between conservative filtering parameters that tend to discard good-quality data and more relaxed parameters that risk error. The application of RAD sequencing to many sturgeon species carries the additional complication of polyploidy. The four species targeted in the current study belong to the group of sturgeon species with about 240 chromosomes (Ludwig *et al.* 2001). This high chromosome number is the result of two whole genome duplication (WGD) events, and for this reason, these species are considered to be evolutionary octoploid by some authors (Vasil'ev 2009; Drauch Schreier *et al.* 2011). However, between the two WGD events, a functional rediploidization took place, and these species should be better considered as functionally tetraploid (Fontana *et al.* 2008). Tetraploidy has implications for the application of filters during SNP discovery and GBS. The risk of false SNP discovery is increased due to genome duplication and subsequent alignment of near-identical duplicated regions. Hohenlohe *et al.* (2011) attempted to mitigate this effect by removing candidate loci showing excess heterozygosity and very low (subzero) inbreeding measures, calculated from a population of individually barcoded fish. Accurate GBS is made difficult by the potential for heterozygote individuals to display three different tetraploid genotypes at a bi-allelic SNP (e.g. AAAC, AACC, ACCC); while approaches for dealing with tetraploid genotype data are starting to be developed for SNP assay approaches (e.g. Serang *et al.* 2012), this area remains challenging for GBS.

Table 1 A list of all samples made available to SturSNiP by project partners. A subset of these samples (shown in parentheses) was used for single nucleotide polymorphism (SNP) discovery by RAD Sequencing. Additional samples were subsequently used to validate candidate SNPs. Samples from Russia and Iran were transferred to the UK under the appropriate CITES licences

Species	Origin	Criteria	Number
A. gueldenstaedtii	Caspian Sea	Northern extreme	115 (16)
		Southern extreme	10
A. persicus	Caspian Sea	Northern extreme	106 (16)
		Southern extreme	20
A. baerii	Siberia	River Ob	30 (16)
		River Lena	30 (16)
A. naccarii	Adriatic Sea /Hatchery	Family group:	8 (8)
		2 parents	
		6 offspring	
		Total	319

The work described here was undertaken as part of an international collaborative project entitled 'SturSNiP' (stursnip.jrc.ec.europa.eu), which aimed to address the issues identified for sturgeon conservation management at the Caspian Environment Program Regional Workshop on Sturgeon Genetics, held in Turkey in 2009 (Doukakis 2009). The broad aim was to undertake research to discover SNP markers in Russian (*A. gueldenstaedtii*), Persian (*A. persicus*), Siberian (*A. baerii*) and Adriatic (*A. naccarii*) sturgeons. Specifically, SturSNiP aimed to discover candidate SNP markers that (i) discriminate Russian (*A. gueldenstaedtii*) from Persian (*A. persicus*) sturgeons in the Caspian Sea; (ii) enable investigation of population structure in Siberian sturgeons (*A. baerii*) from the Ob and Lena rivers and (iii) display variation among individual Adriatic sturgeons (*A. naccarii*) for potential application to parentage (farm traceability) analysis.

Materials and methods

Sample collection

Authenticated reference samples (fin clips or whole juveniles) from the four target species were collated (Table 1) as either a family group or unrelated wild individuals representing a range of geographic origins (Fig. 1). *A. persicus* and *A. gueldenstaedtii* were each collected from populations at the northern and southern extremes of the Caspian Sea; all samples of north Caspian Sea (Russian water) fish were accompanied by morphological measurements and photographs (N. Mugue, unpublished data) detailing the basis for their identification as either *A. persicus* and *A. gueldenstaedtii*; samples from the South Caspian Sea fish were collected from adult sturgeons in the Sefidrud River, Iran, and identified using morphological characters (Pourkazemi 2009). Samples of the Ob' River population of Siberian sturgeons (*A. baerii*) were collected from the wild in 2002, while samples from the Lena River was obtained from wild-caught individuals, maintained at Konakovo Sturgeon hatchery near Moscow. For Adriatic sturgeons, a family group from the aquaculture breeding centre 'Azienda Agricola VIP' (Orzinuovi, BS, Italy) was used to identify markers that vary at the level of the individual with a view to generating pedigree-informative markers capable of tracing sturgeon products back to farmed origin. The family group was independently verified using a panel of eight microsatellite markers previously employed for parental allocation in the same species (Congiu *et al.* 2011). The inclusion of a family group also acted as a candidate SNP validation tool, to allow the identification of markers displaying patterns of Mendelian segregation, a characteristic of biparentally inherited markers that can be used to exclude false SNPs observed due to sequencing error or gene dupli-

Fig. 1 Geographic distribution of sampling localities across Europe and Asia, with species coded by colour: pink = Adriatic sturgeon (*Acipenser naccarii*); red = Russian sturgeon (*A. gueldenstaedtii*); green = Persian sturgeon (*A. persicus*); orange = Siberian sturgeon (*A. baerii*). See Table 1 for sample numbers per locality.

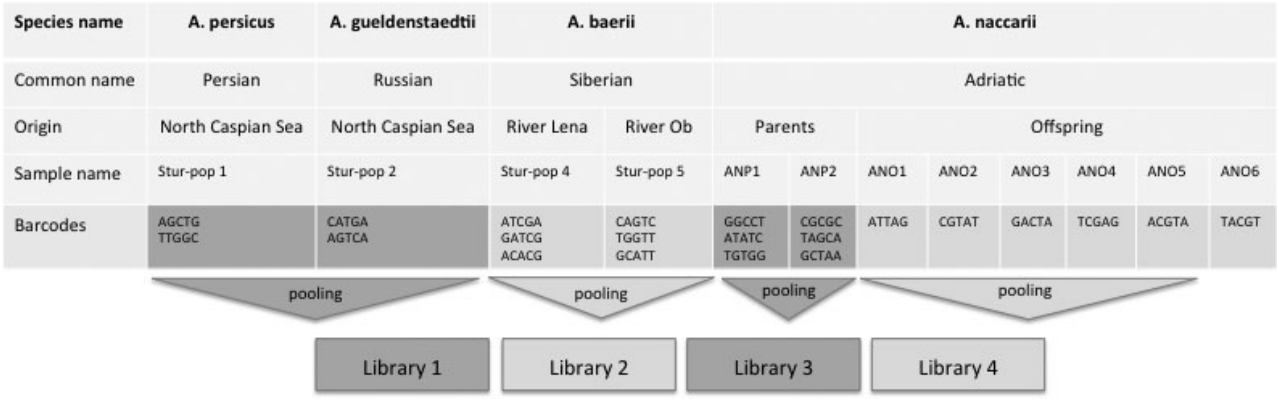

Fig. 2 Summary of the 12 individually labelled sturgeon samples produced for the SturSNiP project. The first four (Persian, Russian and Siberian (Ob and Lena)) each consisted of DNA from 16 individual fish. The Adriatic family group fish were labelled individually. The samples were then pooled into four libraries and sequenced in a single lane of an Illumina HiSeq flow cell using paired-end sequencing.

cation. Delays in permitting prevented the inclusion of the south Caspian sturgeon samples in the SNP discovery phase; these were therefore only used in candidate SNP validation and population screening.

Library preparation and sequencing

DNA was extracted (Qiagen DNeasy protocol) from fin clip or whole fingerling (family offspring) and normalized to 50 ng/uL (Table 1). For Russian (*A. gueldenstaedtii*), Persian (*A. persicus*) and Siberian (*A. baerii*) fish, the DNA from sixteen individuals was pooled in equimolar proportions prior to library preparation to create four population 'metasamples': Russian (N. Caspian), Persian (N. Caspian), Siberian (River Ob) and Siberian (River Lena). The eight family group fish (two parents and six offspring) were barcoded individually (Fig. 2). One microgram of each sample (or pool sample) as quantified by fluorometry (Qubit dsDNA BR Assay, Invitrogen) was digested with the restriction enzyme Sbf1 followed by P1 adapter ligation. Multiple barcodes within each one microgram sample were used for each population pool and in family parents to mitigate variation in barcode representation following amplification and sequencing, while family offspring samples were tagged with a single barcode, resulting in a total of four libraries representing 12 samples tagged with 22 barcodes (Fig. 2). The DNA was then sonically sheared using a Covaris S2 instrument (dust cycle 10%; intensity 5; cycles/burst 200; duration 60 s), and the size range 300–600 bp was isolated using gel excision and purification. After end-repair and A-tailing, the size-selected DNA was ligated to P2 adapters and PCR amplified to select for fragments containing the 5′ and 3′ sequences necessary for Illumina sequencing. Quality control was carried out at different stages of library preparation from quality assessment of the input samples to final QC of the sequencing library. Final RAD libraries were quantified by qPCR (Kapa Library Quantification Kit) and pooled in equimolar proportions for sequencing in a single lane of an Illumina HiSeq 2000 instrument using 100 base paired-end reads (v1 chemistry). Resulting sequence data were examined and filtered for sequence quality to remove reads that had less than 75% of bases with a quality score of 20 or more, trimmed to remove primers and collated by barcode using in-house scripts (available from the authors upon request). Raw data were submitted to the Sequence Read Archive (SRA) (see Data Accessibility).

SNP discovery

Sequence data were analysed in three separate species groups: (i) Adriatic sturgeon, (ii) Siberian sturgeon and (iii) Russian and Persian sturgeons combined. A SNP discovery pipeline was applied to each of the three groups separately, which followed that of Senn *et al.* (This Issue). Briefly, this pipeline involved cataloguing the Read 1 and cleaning the Read 1 and Read 2 data in Stacks (Catchen *et al.* 2011) using the scripts *process_radtags*, *denovo_map.pl* and a version of *sort_read_pairs.pl* modified to output FASTQ, instead of FASTA Read 2 files. FASTQ Read 2 files were then output from Stacks for tags that had at least 15 reads per sample (individual, population pool or species pool). From these Read 2s only, contigs were simultaneously assembled and bubbles called using Cortex_var v1.0.5.3 (Iqbal *et al.* 2012). During this step, cleaning was applied to the multicolour De Bruijn graphs by removing low-coverage 'supernodes' of 1× coverage. Variants were then called on the bubble files using the Cortex script *process_calls* again according to the method outlined in

Senn *et al.* (This Issue). The default estimated sequencing error rate of 0.01 was used, and as we performed a separate assembly for each tag, the genome_size parameter was set to 250 bp (the approximate footprint of the paired-end contig at a fixed tag). During variant genotyping, Cortex assumes diploidy (using parameter–experiment_typeEachColourADiploidSample); we therefore used variant calling only as a rough classifier into homozygous and heterozygous states. Stampy v1.0.14 (Lunter & Goodson 2011) was used to realign Read 2 data to the contigs generated by Cortex. These files were viewed in Integrative Genomics Viewer v1.4.2 (IGV, Robinson *et al.* 2011) along with the associated Variant Call Format (VCF) files generated by Cortex as a visual quality control of the pipeline.

In the Adriatic family group, SNPs were only selected for further validation if they were called in at least two samples, as we assume that the presence of the same SNP in multiple samples strengthens the likelihood of it being true. As only eight samples from the same family group were individually barcoded, the exclusion of loci based on heterozygote excess under Hardy–Weinberg expectations was not employed. However, fixed heterozygotes in all samples were excluded from the candidate SNP set. For Russian/Persian and Siberian sturgeon scenarios, the VCFs were parsed to gain an estimate of the possible number of polymorphic and diagnostic SNP markers in each sample.

SNP validation

To verify the existence of candidate SNPs and to evaluate the accuracy of GBS, a number of approaches were used to validate different sets of SNPs identified from the sequence data. Candidate SNP markers identified in the Adriatic sturgeon (*A. naccarii*) family were examined with respect to Mendelian inheritance in the unambiguous cases where the two parents displayed alternate homozygous genotypes, first *in silico* using the genotypes called by Cortex and then via direct Sanger sequencing of the same family group using primers designed in Read 2 flanking regions for eight of these candidate SNPs (Table S1, Supporting information).

Validation of informative markers identified for Russian/Persian sturgeon discrimination was undertaken using two empirical genotyping methods. First, a series of Kaspar probes (K-Bioscience) was designed for five candidate SNPs showing segregation between Russian (*A. gueldenstaedtii*) and Persian (*A. persicus*) sturgeons (Table S2, Supporting information). These were used to genotype additional samples of each species from the North Caspian Sea (12 Russian and 13 Persian individuals not used in SNP discovery) and samples of each species from the South Caspian Sea (six individuals per species). Second, a separate set of candidate SNPs displaying a heterozygous state in Persian sturgeon and a homozygous state in Russian sturgeon were selected. Genotyping was performed on 85 Russian and 76 Persian North Caspian samples using thirteen allele-specific primer trios (two specific forward primers, one common reverse) with the presence/absence of the allele being resolved under gel electrophoresis (Table S3, Supporting information).

Results

Sequencing and SNP discovery

The twelve separate RAD samples (four population pools and eight family members) produced 97 million 100-base read pairs of DNA sequence. Following quality control filtering, 77 million read pairs were retained for downstream bioinformatic analysis. Sequencing effort per sample varied from 1.3 to 15.3 million reads per barcode (Table 2). Variation in amount of sequence data per barcode was mitigated in pooled populations through the use of multiple barcodes per sample which served to equalize sequencing effort. In the Adriatic family, parent samples were intentionally sequenced to greater depth than their offspring; however, among-offspring variation was also pronounced (Table 2). Subsequent results are presented separately for the three predefined aims relating to Russian/Persian, Siberian and Adriatic sturgeons.

Discriminating between Russian and Persian sturgeons (A. gueldenstaedtii/A. persicus). Cataloguing of Read 2 Tags returned 233 564 loci, of which 112 714 had fifteen or more reads in at least one sample. Of the 48 731 tags with 15 or more reads in both Russian and Persian samples, 88.7% were observed in the Russian North Caspian and 59.2% in the Persian North Caspian, indicating that an imbalance in sequencing effort limited the analysis of interspecies variation (Table 2). The Stacks/Cortex pipeline returned 7346 candidate SNPs across the two samples of which 43.9% were estimated to be polymorphic in Russian, 64.5% in Persian and 9.2% estimated to be fixed between Russian and Persian sturgeon (alternate homozygotes) (Table 3).

Population structure of Siberian sturgeon (A. baerii). Cataloguing of Read 2 Tags returned 197 471 loci, of which 140 260 had fifteen or more reads in at least one sample. Of the 65 736 loci with greater than fifteen reads in both samples; 69.7 and 77.2% were observed in the rivers Lena and Ob, respectively (Table 2). The Stacks/Cortex pipeline returned 14 083 candidate SNPs, of which 49.3% were estimated to be polymorphic in the Lena, 57.9% in the Ob and 11.9% estimated to be fixed

Table 2 RAD sequence summary results for the 12 sturgeon samples. Wide variation in read number per barcode was observed (col. 3), leading to variation in the number of RAD loci observed in each sample (col. 4) across the different three identification issues (col. 5). For each of the applied issues under study, the total number of unique loci sequenced (col. 6) was much greater than the number observed in any one sample (col. 7). For the Adriatic offspring, the low representation of some samples (cols. 7 & 8) combined with selection criteria that require loci to be observed in all samples, severely limited the number of candidate SNPs retained

Sample	Replicate barcode/ Family member	No. Reads per barcode	Total reads per sample	Issue	No. unique Read 2 loci with >15 reads for at least one sample	No. Read 2 loci with >15 reads for the sample	% unique Read 2 loci observed for the sample
Persian, North Caspian	1	1 996 569	9 752 244	Species discrimination	112 714	90 156	88.7
	2	7 755 675					
Russian, North Caspian	1	2 334 886	4 697 508			60 183	59.2
	2	2 362 622					
Siberian, River Lena	1	1 299 051	9 987 936	Population discrimination	140 260	97 764	69.7
	2	1 499 761					
	3	7 189 124					
Siberian, River Ob	1	5 800 577	12 848 866			108 236	77.2
	2	4 119 565					
	3	2 928 724					
Adriatic P1	Male (3 barcodes)	9 457 673	9 457 673	Individual discrimination	126 520	94 595	74.7
Adriatic P2	Female (3 barcodes)	15 348 782	15 348 782			104 223	82.3
Adriatic F1	O1	1 836 470	1 836 470			27 857	22.0
Adriatic F1	O2	2 183 635	2 183 635			33 748	26.7
Adriatic F1	O3	2 637 201	2 637 201			41 299	32.6
Adriatic F1	O4	3 788 657	3 788 657			91 549	72.6
Adriatic F1	O5	9 887 336	9 887 336			44 957	35.5
Adriatic F1	O6	2 804 619	2 804 619			56 420	45.5

between the Lena and Ob rivers (alternate homozygotes) (Table 3).

Variation in an Adriatic sturgeon family (A. naccarii). Cataloguing of Read 2 Tags returned 222 890 loci, of which 126 520 had greater than fifteen reads in at least one sample. However, representation of loci among individual samples varied considerably with j the proportion of loci observed in each sample ranging from 22 to 75% of the total of 126 520 (Table 2). The level of locus representation among samples, combined with the criterion to only search for candidate SNPs at loci for which all eight individuals have more than 15 reads for the Read 2 data, reduced the number of candidate loci to 9346 (Table 3). After processing with the Stacks/Cortex pipeline, this resulted in 1329 polymorphic loci.

SNP validation

A total of 149 loci displayed alternate homozygous genotypes in the parental samples. However, none of the loci could be validated using patterns of Mendelian inheritance in the family group, indicating either false SNP discovery or incorrect GBS. Subsequent laboratory validation by Sanger sequencing confirmed the presence of the target SNPs at all eight loci tested, but again, alleles did not segregate as expected. This result indicated an error in the sequence genotype call for one or both parents, which was confirmed by additional sequencing of the parental samples within the family group (Fig. S1, Table S1, Supporting information). Subsequent examination of the original sequence data revealed alternate alleles present at very low frequency at three of the eight loci (Table S1, Supporting information).

Exploratory population genotyping

A total of 675 candidate SNP markers showed alternate fixed alleles between Russian and Persian samples in the sequence data (Table S4.2, Supporting information). Of those carried forward for population testing, three of the five Kaspar genotyping assays generated reproducible data with one showing clustering into the two putative species (Fig. 3); the remaining two displayed poor segregation among Russian and Caspian samples. The allele-specific primer validation study identified four loci that show clear differences in allele frequencies

Table 3 Single nucleotide polymorphism (SNP) discovery summary for the three issues targeted within the experimental design. Greater mean coverage in Siberian sturgeon corresponds to higher SNP retention throughout the pipeline; conversely, loss of candidate SNPs in Adriatic sturgeon reflects low coverage in some samples. Mean Read 2 contig length suggests sufficient flanking sequence would usually be available for downstream genotyping assay design

Category	Species discrimination (Caspian: Persian/Russian)		Population discrimination (Siberian: Lena/Ob)		Individual discrimination (Adriatic)	
Filtering steps	Count	% of total	Count	% of total	Count	% of total
Total putative RAD tag loci	233 564	100.0	197 471	100.0	222 890	100.0
$\geq$15 reads in every sample	48 731	20.9	65 736	33.3	9346	4.2
$\geq$1 variant in the RAD tag	26 151	11.2	34 598	17.5	8049	3.6
Single contig assembled	12 620	5.4	22 210	11.2	2518	1.1
Variant present is a single SNP	8734	3.7	15 546	7.9	1865	0.8
SNP present in $\geq$2 samples	7346	3.1	14 083	7.1	1329	0.6
Polymorphic candidate SNPs	7346		14 083		1329	
Mean coverage at SNP	18.2		35.4		14.24	
Mean Read 2 contig length (min–max)	222 (49–575)		214 (52–580)		232 (51–592)	
Estimated polymorphism level in each sample	64.5% (Persian)		49.3% (Lena)		n/a	
	43.9% (Russian)		57.9% (Ob)			
Estimated fixed between samples	9.2%		11.9%		n/a	

between Russian and Persian samples, although no diagnostic loci were observed (Table 4).

Discussion

From an applied perspective, the SNP markers discovered under the SturSNiP project offer the potential to develop genetic management and traceability systems in all four species at a level of resolution not previously possible. Differentiating Russian and Persian sturgeons is currently undertaken using a complex array of morphological characters, which are not present in processed fish or caviar. Moreover, the expertise required to identify fish is limited at both ends of the Caspian Sea, and an international consensus on species identifiers is urgently needed. In terms of CITES enforcement, the ability to identify traded caviar to species level is essential (Birstein *et al.* 2000), and it is hoped that the results of this work will enable the development and validation of species ID tests for trade regulation in the near future. In addition to taxonomic identification, the SNP markers can form the basis of population genetic studies that are planned to investigate the current population structure of sturgeon within the Caspian and Azov Seas and support the management of hatchery release programmes.

The Siberian sturgeon, *A. baerii*, is distributed across a very large geographic area and is extensively farmed. The availability of population-informative markers should aid research into the evolution and molecular ecology of the species, as well as provide information for the conservation of genetic diversity in wild and managed stocks. Lastly, the isolation of SNPs that show individual variability among broodstock individuals in the Adriatic sturgeon, *A. naccarii*, paves the way for the development of caviar traceability systems that can identify individual caviar-producing females and their produce. Forensic SNP DNA profiling systems based on around 50 markers in a single assay have been developed for individualization in humans (Wei *et al.* 2012). In fish, SNP-based origin assignment in wild marine species has been demonstrated using even lower numbers of markers to address specific provenance questions (Nielsen *et al.* 2012). In addition to panels of informative SNPs, these methods require the use of extensive reference population data sets from which to calculate match probability statistics. Replicating such systems to profile caviar to an individual level via the DNA recovered from interovular mucous and follicular cells surrounding unfertilized egg, would therefore require further work, but should enable caviar producers to trace individual batches of caviar through their supply chains and thus detect the illegal laundering of wild-derived caviar into the system.

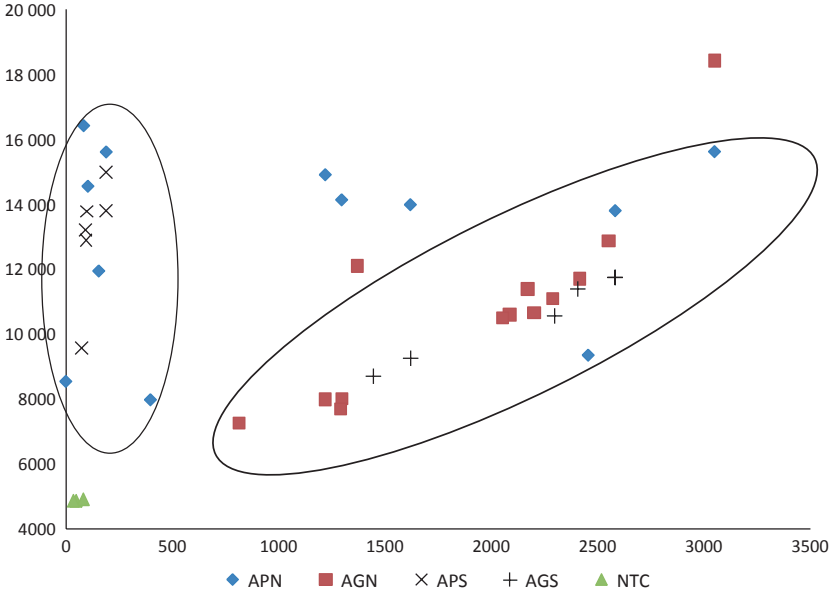

Fig. 3 Results of a Kaspar genotyping assay for single nucleotide polymorphism (SNP) #53778 (Table S2, Supporting information), a candidate marker for discriminating Russian from Persian sturgeon. APN = Acipenser persicus north Caspian; AGN = *A. gueldenstaedtii* north Caspian; APS = *A. persicus* south Caspian; AGS = *A. gueldenstaedtii* south Caspian; NTC = no template control. The plot shows two principal clusters, one containing *A. persicus* (north and south Caspian) only, and the second containing predominantly *A. gueldenstaedtii* from north and south Caspian. The presence of *A. persicus* north Caspian samples in the *A. gueldenstaedtii* cluster may be due to incomplete segregation at this marker or discrepancies between molecular and morphological differentiation.

The RAD approach enabled a relatively complex experimental design to be incorporated into a single lane of sequencing on the Illumina HiSeq2000 platform, providing sufficient data for the identification of polymorphic sites relating to three different applications at three different biological levels. However, the results of the various validation studies undertaken also demonstrate limitations of the current study for determining individual genotypes by sequencing alone in polyploid species such as sturgeon.

The total lack of correspondence between offspring and parental genotypes determined directly from the sequence data was initially surprising, as the Adriatic sturgeon family group had been included to verify Mendelian inheritance at individual loci. However, the subsequent empirical validation data identified errors in the RAD sequence genotypes of the parents, highlighting another benefit to the inclusion of family groups in SNP discovery, namely an assessment of the ability to perform GBS. While all of the candidate SNPs sequenced for Adriatic sturgeons were validated as polymorphic markers, genotyping error was also observed at all of them, suggesting that false SNP discovery due to sequence error was minimal and that in fact discrepancies were probably the result of alleles missing from the sequence data or overly conservative allele calling effectively leading to allelic dropout at most loci. The genotype caller used in this study, Cortex, as with other genotype callers, determines genotype and calculates a genotype quality score based on variables including read depth and 'minor' allele frequency. Under a simple diploid model, with no sequencing bias and individually barcoded samples, a heterozyogote genotype is expected to have a balanced number of reads for each allele, reducing the risk of genotyping error. However, in this study, these criteria were not met.

Table 4 Results of allele-specific PCR analysis to investigate the utility of candidate single nucleotide polymorphism markers for differentiating Russian from Persian sturgeon (using 85 and 76 individuals, respectively). Allele frequencies show strong differences between species. For primer and sequence details, see Table S3 (Supporting information)

Locus	Primer name (allele)	Allele frequency Persian sturgeon	Russian sturgeon
54644	Pr54644G	0.91	0.5
	Pr54644P	0.09	0.5
62734	Pr62734G	0.48	0.88
	Pr62734P	0.52	0.12
65774	Pr65774G	0.39	0.66
	Pr65774P	0.61	0.34
Hel15	Hel15 R1	0.55	0.13
	Hel15 R2	0.45	0.87

Tetraploidy leads to an expectation of heterozygotes with intermediate allele frequencies (i.e. 3:1 and 1:3 as well as 2:2), which may force genuine heterozygote genotypes to be called as homozygotes, particularly where read depth is limiting. The assumption that different alleles are amplified proportionally to their copy number in the genome is also known to be a major source of error in genotyping tetraploid sturgeons (Boscari *et al.* 2011). To complicate matters further, the use of pooled individuals under a single barcode in Siberian, Russian and Persian sturgeons will have further skewed heterozygote allele read frequencies from a 50:50 expectation, as the

data for each pooled sample are treated as a single composite individual. In addition to these issues, it is becoming evident that the production of RAD sequence data may be affected by bias in the efficiency of DNA shearing during library preparation with longer restriction fragments preferentially sheared, potentially leading to an imbalance in the representation of the alleles at a locus (Davey *et al.* This Issue). Any or all of these considerations may have affected the genotypes called from the RAD sequence data in Adriatic sturgeon. The net effect of these issues is towards conservative SNP discovery with fewer false SNPs. This may be advantageous for marker production, but the accuracy of GBS is reduced. In tetraploid species, with intermediate heterozygote genotypes, the ability to genotype at all has been severely restricted, although recent analytical approaches have been designed to address this issue for SNP genotyping assays (e.g. Serang *et al.* 2012). While solutions to individual issues may present themselves, in terms of this study, increasing sequence depth to ensure sufficient coverage of all alleles would have improved our ability to genotype accurately. This is particularly relevant to the paired-end approach used here, where sequence depth across the Read 2 contig is reduced by the staggered alignment. This was recognized from the outset, but the experimental design was driven predominantly by the objective of discovering informative SNPs for multiple downstream applications.

The apparent masking of heterozygote genotypes observed in the family group validation also limited the ability to correctly identify truly diagnostic markers, in particular for the differentiation of Russian and Persian sturgeons. Single locus validation of markers that appeared to segregate completely between these species showed that in fact these differences were frequency based. By barcoding a pool of individuals into a single population sample, there is greater potential to simultaneously discover markers and assess their variability at a population level. The experimental design employed here used 16 fish (=64 haploid genomes) for a single population barcode, which would theoretically be a reasonable population sample, assuming there was sufficient read depth at each RAD Tag to ensure that these 64 haploid genomes were represented in the sequence data. This was generally not the case, resulting in apparent cases of complete segregation reverting to allele frequency differences. Nevertheless, the discovery of molecular genetic markers that show differences between Russian and Persian sturgeons is important and a key finding of this study. Until now, these two taxa, while widely recognized as valid species (Ludwig 2008; but see Ruban & Khodorevskaya 2011), have been impossible to separate using standard phylogenetic markers, such as mtDNA gene and control region sequences (Ludwig 2008).

The work presented here represents a starting point in the development of tools for sturgeon research and management. The 14.4 billion base pairs of sturgeon sequence data produced here represent a very large genetic resource for potential use in studies of taxonomy, population structure, ecology, conservation management and product traceability. At a technical level, further sequencing of both RAD libraries and whole genomes would assist the characterization of SNP markers and, in an applied context, much remains to be done in terms of marker validation and genotyping assay development. Molecular genetic tools remain a small part of the solution to conserving one of the world's most ancient lineages of fish; however, through the use of methods such as RAD Sequencing for SNP discovery demonstrated here, it is hoped, incrementally, to improve the outlook for the survival of sturgeon in the wild.

Acknowledgements

SturSNiP has been funded by the European Commission Joint Research Centre. We thank staff of the GenePool Genomics Facility at the University of Edinburgh, especially Marian Thomson and Urmi Trivedi, for assistance with library preparation and sequencing. Part of the analyses on *A. naccarii* was performed thanks to the University of Padova grant CPDA087543/08. We are grateful to two anonymous reviewers for helping to improve the manuscript.

References

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Birstein VJ, Doukakis P, Desalle R (2000) Polyphyly of mtDNA lineages in the Russian sturgeon, Acipenser gueldenstaedtii: forensic and evolutionary implications. *Conservation Genetics*, **1**, 81–88.

Boscari E, Barbisan F, Congiu L (2011) Inheritance pattern of microsatellite loci in the polyploid Adriatic sturgeon (*Acipenser naccarii*). *Aquaculture*, **321**, 223–229.

Bronzi P, Rosenthal H, Gessner J (2011) Global sturgeon aquaculture production: an overview. *Journal of Applied Ichthyology*, **27**, 169–175.

Catchen J, Amores A, Hohenlohe P, Cresko W, Postlethwait J (2011) Stacks: building and genotyping loci de novo from short-read sequences. G3: Genes. *Genomes and Genetics*, **1**, 171–182.

CITES & UNEP (2012) Convention on international trade in endangered species of wild fauna and flora. Appenices I, II and III. Valid from 3 April 2012. http://www.cites.org/eng/app/appendices.php (accessed on 26 June 2012).

Congiu L, Pujolar JM, Forlani A *et al.* (2011) Managing polyploidy in ex situ conservation genetics: the case of the critically endangered adriatic sturgeon (*Acipenser naccarii*). *PLoS ONE*, **6**, 1–10.

Davey JW *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews. Genetics*, **12**, 499–510.

Davey JW, Cezard T, Utrilla PF, Eland C, Gharbi K, Blaxter M (This Issue) Special characters of RAD Sequencing data: implications for genotyping. *Molecular Ecology*.

DeSalle R, Birstein VJ (1996) PCR identification of black caviar. *Nature*, **381**, 197–198.

Doukakis P (2009)Report on the Caspian Environment Program Regional Workshop on Sturgeon Genetics, 16-18 June 2009, Trabzon, Turkey. Available from the World Bank, full citation pending.

Drauch Schreier A, Gille DA, Mahardja B, May B (2011) Neutral markers confirm the octoploid origin and reveal spontaneous polyploidy in white sturgeon, Acipenser transmontanus. *Journal of Applied Ichthyology* **27**(Suppl2), 24–33.

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA (2011) Local de novo assembly of RAD paired-end contigs using short sequencing reads. *PLoS ONE*, **6**, e18561.

European Commission (2006a) COMMISSION REGULATION (EC) No 865/2006 of 4 May 2006 laying down detailed rules concerning the implementation of Council Regulation (EC) No 338/97 on the protection of species of wild fauna and flora by regulating trade therein.

European Commission (2006b) New rules to combat illegal caviar trade. Commission Press Release IP/06/611, Brussels 15/05/2006.

Fontana F, Congiu L, Mudrak VA, Quattro JM, Smith TIJ, Ware K, Doroshov SI (2008) Evidence of hexaploid karyotype in shortnose sturgeon. *Genome*, **51**, 113–119.

Garvin MR, Saitoh K, Gharrett AJ (2010) Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources*, **10**, 915–934.

Hohenlohe P, Amish S, Catchen J, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources* **11**(Suppl 1), 117–122.

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, **44**, 226–232. Doi: 10.1038/ng.1028.

IUCN (2012) *The IUCN Red List of Threatened Species*. Version 2012.1. http://www.iucnredlist.org. (Accessed on 26 June 2012).

Ludwig A (2008) Identification of Acipenseriformes species in trade. *Journal of Applied Ichthyology*, **24**(S1), 2–19.

Ludwig A, Belfiore NM, Pitra C, Svirsky V, Jenneckens I (2001) Genome duplication events and functional reduction of ploidy levels in sturgeon (Acipenser, Huso and Scaphirhynchus). *Genetics*, **158**, 1203–1215.

Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, **21**, 936–939.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Mugue NS, Barminsteva AE, Rastorguev SM, Mugue VN, Barminstev VA (2008) Polymorphism of the mitochondrial DNA control region in eight sturgeon species and development of a system for DNA-based species identification. *Russian Journal of Genetics*, **44**, 793–798.

Nielsen EE, Cariani A, MacAoidh E *et al.* (2012) Gene-associated markers provide tools for tackling IUU fishing and false eco-certification. *Nature Communications*, **3**, 851.

Ogden R (2008) Fisheries forensics: the use of DNA tools for improving compliance, traceability and enforcement in the fishing industry. *Fish and Fisheries*, **9**, 462–472.

Ogden R (2011) Unlocking the potential for genomic technologies for wildlife forensics. *Molecular Ecology Resources*, **11**(S1), 109–116.

Pikitch EK, Doukakis P, Lauck L, Chakrabarty P, Erickson DL *et al.* (2005) Status, trends and management of sturgeon and paddlefish fisheries. *Fish and Fisheries*, **6**, 233–265.

Pourkazemi M (2006) Caspian Sea sturgeon conservation and fisheries: past, presnt and future. *Journal of Applied Ichthyology*, **22**(S1), 12–16.

Pourkazemi M (2009) Comprehensive study on assessment of sturgeons population genetic structure in the Caspian Sea. Iranian Fishery Research Organization. International Sturgeon Research Institute. Project final report. 315 pages.

Raymakers C (2006) CITES, the Convention on International Trade in Endangered Species of Wild Fauna and Flora: its role in the conservation of Acipenseriformes. *Journal of Applied Ichthyology*, **22**(S1), 53–65.

Raymakers BC, Hoover C (2002) Acipenseriformes: CITES implementation from Range States to consumer countries. *Journal of Applied Ichthyology*, **18**, 629–638.

Rehbein H, Molkentin J, Schubring R, Lieckfeldt D, Ludwig A (2008) Development of advanced analytical tools to determine the origin of caviar. *Journal of Applied Ichthyology*, **24** (Suppl.1), 65–70.

Robinson JT, Thorvaldsdóttir H, Winckler W *et al.* (2011) Integrative genomics viewer. *Nature Biotechnology*, **29**, 24–26.

Rosenthal H, Wei Q, Chang J, Bronzi P, Gessner J (2011) Conclusions and recommendations of the 6th International Symposium on Sturgeons (Wuhan, China, October 2009)*. *Journal of Applied Ichthyology*, **27**, 157–161.

Ruban GI, Khodorevskaya RP (2011) Caspian Sea sturgeon fishery: a historic overview. *Journal of Applied Ichthyology*, **27**, 199–208.

Senn H, Ogden R, Cezard T *et al.* (This Issue) A simple pipeline for the generation of SNP genotyping assays from RAD paired-end data without a reference genome in the Eurasian Beaver. *Molecular Ecology*.

Serang O, Mollinari M, Garcia AAF (2012) Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. *PLoS ONE*, **7**, e30906.

Vasil'ev VP (2009) Mechanisms of polyploid evolution in fish: polyploidy in sturgeons. In *Biology, Conservation, and Sustainable Development of Sturgeons* (eds Carmona R, Domezain A, García-Gallego M, Hernando JA), pp. 97–117. Springer, Berlin, Germany.

Waldman JR, Doukakis P, Wirgin I (2008) Molecular analysis as a conservation tool for monitoring the trade of North American sturgeons and paddlefish. *Journal of Applied Ichthyology*, **24**(S1), 20–28.

Wei Y-L, Li C-X, Jia J, Hu L, Liu Y (2012) Forensic identification using a multiplex assay of 47 SNPs. *Journal of Forensic Sciences*, **57**, 1448–1456.

Willing E-M, Hoffmann M, Klein JD, Weigel D, Dreyer C (2011) Paired-end RAD-seq for de novo assembly and mar-

ker design without available reference. *Bioinformatics (Oxford, England)*, **27**, 2187–2193.

R.O. was co-PI and wrote the manuscript, K.G. contributed to project design, supervised RAD-seq and edited the manuscript, N.M. & A.S. contributed to project design, provided samples and performed SNP validation, J.M. managed project for JRC and edited manuscript, M.P. provided samples and contributed to manuscript, R.M. designed and genotyped KASP probes, C.E. processed samples and prepared RAD libraries, J.D., H.S. and M.V. designed and performed bioinformatics, L.C. was co-PI, provided samples, supervised analysis and contributed to manuscript.

Data accessibility

All sequence data have been deposited in the Sequence Read Archive (SRA) and can be accessed via the following study accession link: http://www.ebi.ac.uk/ena/data/view/ERP001794. Filtered SNP loci and.vcf files are available under DRYAD entry, doi: 10.5061/dryad.39sc7.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Example of biallelic SNP selected for inheritance validation.

Table S1 Genotypes of two parental Adriatic (*A. naccarii*) sturgeon at eight loci observed following Sanger sequencing (cols. 7 & 8).

Table S2 Details of the five candidate SNP loci showing segregation between Russian (*A. gueldenstaedtii*) and Persian (*A. persicus*) sturgeon.

Table S3 Details of the SNP genotyping assays designed for thirteen of the SNPs showing differences between Russian and Persian sturgeon.

Table S4.1 Putative SNP markers discovered in Adriatic sturgeon.

Table S4.2 Putative SNP markers discovered in Russian and Persian sturgeon.

Table S4.3 Putative SNP markers discovered in Siberian sturgeon.