Marine Genomics 30 (2016) 15–26

Contents lists available at ScienceDirect

Marine Genomics

journal homepage: www.elsevier.com/locate/margen

Method paper

Targeted sequencing for high-resolution evolutionary analyses following genome duplication in salmonid fish: Proof of concept for key components of the insulin-like growth factor axis

Fiona M. Lappin, Rebecca L. Shaw, Daniel J. Macqueen *

Institute of Biological and Environmental Sciences, University of Aberdeen, Aberdeen AB24 2TZ, United Kingdom

ARTICLE INFO

Article history:
Received 19 February 2016
Received in revised form 11 June 2016
Accepted 11 June 2016
Available online 23 June 2016

Keywords:
Sequence capture
Target enrichment
Second-generation sequencing
Whole genome duplication
Salmonid fish
Insulin-like growth factor axis

ABSTRACT

High-throughput sequencing has revolutionised comparative and evolutionary genome biology. It has now become relatively commonplace to generate multiple genomes and/or transcriptomes to characterize the evolution of large taxonomic groups of interest. Nevertheless, such efforts may be unsuited to some research questions or remain beyond the scope of some research groups. Here we show that targeted high-throughput sequencing offers a viable alternative to study genome evolution across a vertebrate family of great scientific interest. Specifically, we exploited sequence capture and Illumina sequencing to characterize the evolution of key components from the insulin-like growth (IGF) signalling axis of salmonid fish at unprecedented phylogenetic resolution. The IGF axis represents a central governor of vertebrate growth and its core components were expanded by whole genome duplication in the salmonid ancestor ~95 Ma. Using RNA baits synthesised to genes encoding the complete family of IGF binding proteins (IGFBP) and an IGF hormone (IGF2), we captured, sequenced and assembled orthologous and paralogous exons from species representing all ten salmonid genera. This approach generated 299 novel sequences, most as complete or near-complete protein-coding sequences. Phylogenetic analyses confirmed congruent evolutionary histories for all nineteen recognized salmonid IGFBP family members and identified novel salmonid-specific IGF2 paralogues. Moreover, we reconstructed the evolution of duplicated IGF axis paralogues across a replete salmonid phylogeny, revealing complex historic selection regimes - both ancestral to salmonids and lineage-restricted - that frequently involved asymmetric paralogue divergence under positive and/or relaxed purifying selection. Our findings add to an emerging literature highlighting diverse applications for targeted sequencing in comparative-evolutionary genomics. We also set out a viable approach to obtain large sets of nuclear genes for any member of the salmonid family, which should enable insights into the evolutionary role of whole genome duplication before additional nuclear genome sequences become available.

1. Introduction

During the last decade, large-scale sequencing projects have become commonplace, allowing the genomes and transcriptomes of vast numbers of species to be analysed. For example, large consortium projects such as the '1000 Plant' (Matasci et al., 2014), 'Bird 10 K' (Zhang, 2015), '5000 arthropod genomes' (i5K Consortium, 2013), 'Genome 10K' (Haussler et al., 2009) and Fish-T1K (Sun et al., 2016) are aiming to characterise vast genomic diversity within eukaryotes, while providing essential data for comparative-genomic, evolutionary and phylogenetic studies (e.g. Wickett et al., 2014; Zhang et al., 2014; Jarvis et al., 2014). While such projects generate extensive high-quality sequence data at a relatively low cost, they require sizeable investment in expert person time and infrastructure necessary to achieve their bioinformatic goals (see Wetterstrand, 2015). As a cost-effective, bioinformatically less-demanding alternative, targeted capture/enrichment and sequencing of pre-selected genomic regions offers a proven approach for researchers working on both model and non-model organisms.

The concept of targeting specific areas of the genome for sequencing is well-established and has a long history. Classically, PCR is used to analyse a small number of genes in combination with the Sanger method, or more recently, with second-generation high-throughput sequencing (Tewhey et al., 2009; reviewed in Metzker, 2010). An alternative approach has been to exploit custom-designed microarrays or solution-based hybridization platforms to enrich for sequences (i.e. sequence capture) prior to second-generation sequencing (e.g. Okou et al., 2007; Gnirke et al., 2009; Turner et al., 2009).

The development of sequence capture/enrichment methods has opened up the possibility of routinely obtaining hundreds to thousands

* Corresponding author.
E-mail address: daniel.macqueen@abdn.ac.uk (D.J. Macqueen).

http://dx.doi.org/10.1016/j.margen.2016.06.003

of target sequences at both intra and inter-specific levels, which can be employed to address a range of evolutionary or ecological questions (for reviews see Grover et al., 2012; McCormack et al., 2013; Jones and Good, 2015). Such approaches have been used extensively for population genetics in humans (e.g. Ng et al., 2009; Choi et al., 2009; Calvo et al., 2012) and non-model eukaryotes (e.g. Bi et al., 2013; Hebert et al., 2013; Tennessen et al., 2013). Unmodified sequence capture using baits designed from a limited number of well-characterized species has also proven effective for broader comparisons of species at higher taxonomic levels (e.g. Nadeau et al., 2012; Hedtke et al., 2013; Neves et al., 2013; Heyduk et al., 2015). In this respect, a particularly effective approach has been to capture extremely-conserved regions within the genome (e.g. Lemmon et al., 2012; Faircloth et al., 2012; Eytan et al., 2015; Prum et al., 2015). Moreover, with modifications to the stringency of hybridization, sequence capture can be used to obtain even highly-distant homologous sequences of interest (e.g. Li et al., 2013).

As sequence capture is based on DNA hybridization, this method can be applied to study paralogous sequences arising through relatively recent gene and/or whole genome duplication (WGD) events (Grover et al., 2012, e.g. Hebert et al., 2013; Saintenac et al., 2011; Salmon and Ainouche, 2015). In this respect, sequence capture offers a feasible method to reconstruct the evolutionary history of complex gene families in large taxonomic groups sharing ancestral WGD events. Our lab is exploiting such an approach to characterize patterns of genome and gene family evolution after a salmonid-specific WGD (ssWGD) event that occurred ~95 Ma (Macqueen and Johnston, 2014; Lien et al., 2016). Crucially, the success of this approach hinges on the fact that the average divergence of paralogous regions from the ssWGD, including protein-coding gene paralogues (Berthelot et al., 2014; Lien et al., 2016; e.g. Macqueen et al., 2010, 2013), is within the proven limits of sequence capture. Here, we applied sequence capture across all extant genera of salmonid fish, allowing a detailed evolutionary characterization of key components from the insulin-like growth (IGF) factor axis - a genetic pathway that was expanded by ssWGD and hence offers an ideal model to address post-WGD evolution.

The IGF axis is conserved in all vertebrates and its core components comprise two IGF hormones (IGF1 and IGF2), a family of IGF binding proteins (IGFBPs) and a cell-membrane IGF receptor (IGF1R) (Jones and Clemmons, 1995; Wood et al., 2005; Johnston et al., 2011). The binding of IGF hormones to IGF1R triggers intracellular signalling events that govern a range of key growth phenotypes – in turn, the interaction of IGFs with IGF1R are modulated by IGFBPs, which have a high affinity for IGFs and can inhibit or facilitate the interaction of IGFs with IGF1R, regulating the extent of IGF signalling under different physiological contexts (Jones and Clemmons, 1995). The IGF axis has proven to be of great scientific interest in salmonid fish, owing to its implications in a number of key physiological contexts, including nutritional status (e.g. Bower et al., 2008; Shimizu et al., 2011), metabolism (e.g. Pierce et al., 2006), muscle development (Bower and Johnston, 2010), oocyte maturation (Kamangar et al., 2006), rapid body size evolution (Macqueen et al., 2011) and cross-talk between growth and immunity (Alzaid et al., 2016). Additionally, the IGFBP gene family is among the best-characterized of all gene families in the context of ssWGD and hence represents an excellent model system to exploit in our study. Starting from a core set of six family members, which arose during local and WGD events in the common vertebrate ancestor (Ocampo-Daza et al., 2011; Macqueen et al., 2013), the IGFBP family was expanded during a 'teleost-specific' WGD (tsWGD) (i.e. Jaillon et al., 2004), which was later followed by ssWGD – from which eight salmonid-specific paralogue pairs have been conserved in Atlantic salmon *Salmo salar* (Macqueen et al., 2013).

Our first study aim was to verify the use of sequence capture to acquire complete coding sequences of key duplicated IGF axis components across the full phylogenetic breadth of lineages within the salmonid family. The IGFBP family is well suited for this aim, owing to the retention of a large number of ssWGD paralogues with differing degrees of sequence divergence (Macqueen et al., 2013), allowing us to test the hypothesis that sequence capture can be used to characterize gene families shaped by the ssWGD, defining conserved or distinct patterns of duplicate retention in different salmonid lineages. Our second aim was to demonstrate a useful application for such data, by reconstructing fine-scale patterns of post-ssWGD evolution using phylogenetic methods, including an examination of historic selective regimes that shaped paralogous sequence variation in different salmonid lineages. Our findings highlight outstanding value for sequence capture enrichment as a tool to acquire large-scale sequence data across the entire salmonid family phylogeny, including in relation to the inherently complex, yet undoubtedly interesting aspects of this lineages evolution following ssWGD.

2. Materials and methods

2.1. Sequence capture and assembly

2.1.1. Design of capture baits

Agilent SureSelect 120mer RNA oligomer baits used for sequence capture were synthesized at 4-fold tiling to cover complete coding sequences for nineteen Atlantic salmon IGFBP genes (Macqueen et al., 2013), as well as *IGF2* (Bower et al., 2008), i.e. the 'probe' sequences (accession numbers as follows: *IGFBP-1A1*: NM_001279140, *IGFBP-1A2*: NM_001279137, *IGFBP-2A*: JX565547, *IGFBP-2B1*: NM_001123648, *IGFBP-2B2*: NM_001279160, *IGFBP-3A1*: NM_001279147, *IGFBP-3A2*: NM_001279157, *IGFBP-3B1*: NM_001279167, *IGFBP-3B2*: NM_001279170, *IGFBP-4*: JX565554, *IGFBP-5A*: JX565555, *IGFBP-5B1*: NM_001279142, *IGFBP-5B2*: JX565557, *IGFBP-6A1*: NM_001279155, *IGFBP-6A2*: NM_001279145, *IGFBP-6B1*: JX565560, *IGFBP-6B2*: NM_001279150 and *IGF2*: NM_001146402).

2.1.2. Sequence capture and Illumina sequencing

Genomic DNA was extracted from sixteen species (see Table 1) using a QIAGEN DNeasy kit. The studied species included fifteen salmonids, covering all known genera, along with a member of Esociformes (Northern pike, *Esox lucius*) – a sister lineage to salmonids that did not undergo ssWGD (Rondeau et al., 2014; see Fig. 1A). Purity, integrity and concentration of the initial gDNA was assessed, respectively, using a Nanodrop system (Thermo Scientific) by agarose gel electrophoresis

Table 1
Details of species used for targeted sequence capture.

Species	Sampling date and location
Esox lucius	2011, Lake Coulter Reservoir, Stirling, UK
Brachymystax lenok	2005, Kuanda River, Lena basin, Siberia[a]
Thymallus baicalensis	2003, Selenga Bay, Baikal, Siberia[a]
Thymallus grubii	2007, Bureya River, Russia[a]
Coregonus renke	Unknown sample date, Millstättersee, Austria[a]
Coregonus lavaretus	2011, Carron Valley Reservoir, Stirling, UK
Oncorhynchus nerka	2011, National Research Institute of Fisheries Science[b]
Oncorhynchus kisutch	2011, Center for Aquaculture and Environmental Research[c]
Oncorhynchus tshawytscha	2011, Center for Aquaculture and Environmental Research[c]
Prosopium coulteri	Unknown, Little Bitterroot Lake, Montana, USA[a]
Parahucho perryi	2006, Koppi River, Russia[a]
Hucho taimen	2005, Muna River, Lena Basin, Russia[a]
Hucho hucho	2010, Mur River, Graz, Austria[a]
Salvelinus alpinus	2008, Scotland. Loch Ericht, UK
Salmo trutta	2009, College Mill trout farm, Almondbank, Perthshire, UK
Stenodus leucichthys	1999, Yukon River, Alaska, USA[a]

[a] Gift from Dr. Steven Weiss; Karl-Franzens University of Graz, Institute of Zoology, Universitätsplatz 2, A-8010 Graz Austria.
[b] Gift from Dr. Takashi Yada; National Research Institute of Fisheries Science, Fisheries Research Agency, Nikko, Tochigi 321–1661, Japan.
[c] Gift from Dr. Robert Devlin; Centre for Aquaculture and Environmental Research, Fisheries and Oceans Canada, Vancouver, Canada.

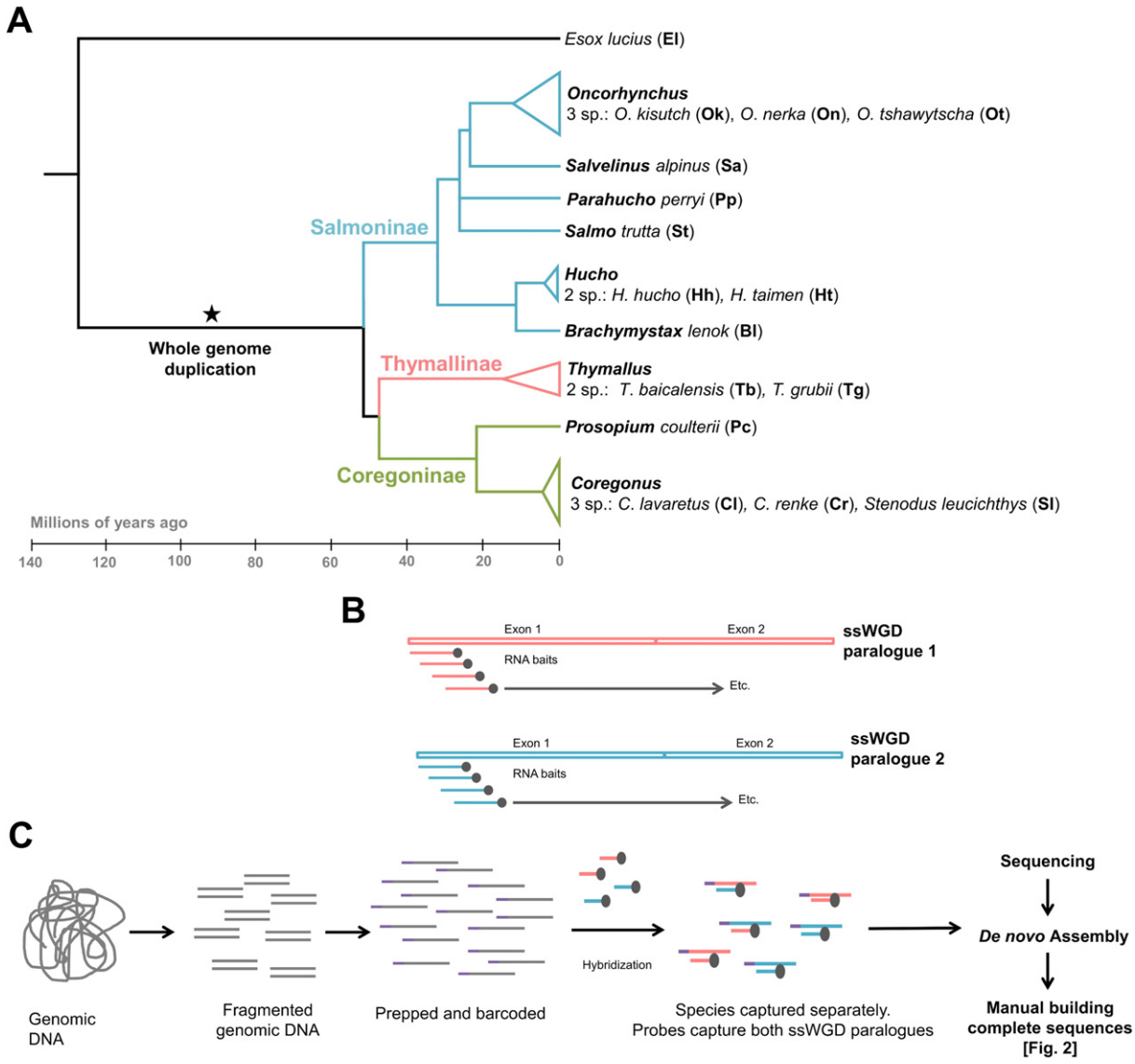

Fig. 1. Targeted sequence capture across the salmonid lineage. (A) Time-calibrated phylogenetic tree including all species used in the sequence capture study (after Macqueen and Johnston, 2014). (B) Depicts how RNA baits used for sequence capture were tiled along the coding regions of cDNAs for ssWGD paralogues for the probe sequences. (C) Depicts the entire sequence capture approach.

and using a Qubit Fluorometer with a dsDNA broad range assay kit (Life Technologies). Sixteen separate Agilent SureSelectXT libraries (3 μg of genomic DNA input, one library per species) were prepared and hybridised according to Protocol version 1.5 (SureSelectXT Target Enrichment System for Illumina Paired-End Sequencing, Nov. 2012). For amplification of pre-capture libraries, 250 ng of adapter-ligated DNA was subjected to 5 PCR cycles. 750 ng of pre-capture library was then hybridised to 2 μl of the RNA oligomer baits for 24 h at 65 °C. The captured libraries were amplified and indexed using 12 PCR cycles. Final captured libraries were quantified using the Qubit approach (described above) and their size distribution determined using an Agilent Bioanalyser (High Sensitivity DNA chip, Agilent). Captured libraries were pooled at equal concentrations and purified using Agencourt AMPure XP beads (Beckman Coulter). The size distribution and concentration of the final pool was assessed, respectively, using an Agilent Bioanalyser and by the Qubit method (described above) along with quantitative PCR (Roche LightCycler 480 II), using an Illumina KAPA Library Quantification Kit (Illumina). The captured libraries were pooled and sequenced on an Illumina HiSeq2000, using version-3 chemistry and generating 2 × 100 bp paired end reads. All steps from sequence capture to Illumina sequencing and subsequent trimming and quality control were performed at the Centre for Genome Research (University of Liverpool, UK). The sequence capture study presented here is part of a larger project (Natural Environment Research Council grant NBAF704), which used the same approach as above, but employed a much broader set of probe sequences for sequence capture (data to be published independently).

2.2. Bioinformatics

After sequence capture and Illumina sequencing, adapters were removed using Cutadapt v.1.2.1 (Martin, 2011) before further trimming was done using Sickle v. 1.21 (Joshi and Fass, 2011) with a minimum window score of 20. Reads <10 bp were excluded from further analysis. SOAPdenovo2 (Luo et al., 2012) was used to assemble the Illumina reads for each species (K = 91, M = 3) incorporating only paired end-reads. The output contigs/scaffolds were incorporated into species-specific BLAST databases (Altschul et al., 1990). A custom python script was used to query each database using BLASTn for each of the probe sequences described above and return all contigs/scaffolds in fasta format. The recovered contigs/scaffolds represented single or multiple exons homologous to the probe sequences, along with flanking genomic

regions. A manual approach was necessary to join contigs representing different exons, allowing us to build complete or larger sequences that were non-chimeric in terms of the represented ssWGD paralogues (described more in the results section and Fig. 2). Sequences were handled within BioEdit (Hall, 1999). Novel IGFBP and *IGF2* sequences generated by this approach are provided as supplementary data, along with 1633 separate SOAPdenovo2 contigs used to build the sequences (Supplementary dataset 1). We also mapped the probe sequences against the reference Atlantic salmon (*S. salar*) genome (Lien et al., 2016) using BLASTn via the SalmoBase server (http://salmobase.org/).

2.3. Phylogenetic analysis

To establish the evolutionary history of novel sequences acquired by sequence capture, seven independent phylogenetic analyses were performed, one per each of the six vertebrate IGFBP family members (IGFBP-1 through to −6), along with one for IGF2. The identity of novel captured salmonid sequences was initially assigned according to high (i.e. >90%) sequence similarity to previously characterized genes (for IGF2: Bower et al., 2008; for IGFBP subtypes: Macqueen et al., 2013). The purpose of the IGFBP trees (n = 6) was to address the relationships of salmonid paralogues within a vertebrate IGFBP family member, rather than branching arrangements between different IGFBP family members (done elsewhere: Ocampo-Daza et al., 2011; Macqueen et al., 2013). The purpose of the IGF2 tree was to test whether putative IGF2 paralogues identified by sequence capture were retained from the ssWGD. For each analysis, relevant IGFBP/IGF2 orthologues were obtained from a representative range of vertebrate lineages, including all the major lobe-finned fish groups and diverse ray-finned fish. All sequences other than novel-generated sequence capture data were obtained from Ensembl (http://www.ensembl.org/index.html) or NCBI (http://www.ncbi.nlm.nih.gov/) databases. The specific Ensembl database versions used were: *Anolis carolinensis*: AnoCar2.0; *Danio rerio*: GRCz10; *Gallus gallus*: Galgal4; *Gasterosteus aculeatus*: BROAD S1; *Homo sapiens*: GRCh38.p5; *Latimeria chalumnae*: LatCha1/GCA_000225785.1; *Lepisosteus oculatus*: LepOcu1; *Oreochromis niloticus*: Orenil1.1; *Pelodiscus sinensis*: PelSin_1.0; *Tetraodon nigroviridis*: TETRAODON 8.0/ASM18073v1 and *Xenopus tropicalis*: JGI 4.2 (GCA_000004195.1). The NCBI accession/Ensembl identifier codes of all sequences are provided within figures.

Amino acid sequences for different phylogenetic analyses were separately aligned using MAFFT v.7 (Katoh and Standley, 2013) before the GUIDANCE2 algorithm (Landan and Graur, 2008; Sela et al., 2015) was used to test for uncertainty in the alignment and identify potential regions of low confidence. For all alignments, the GUIDANCE alignment score exceeded 0.95 and >95% of aligned columns had a confidence score exceeding 0.5; a small number of columns below this confidence score were removed. Finished high-confidence amino acid alignments were 268, 295, 307, 267, 272, 203 and 215 characters in length for IGFBP-1, −2, −3, −4, −5, −6 and IGF2, respectively (Supplementary dataset 2). Each alignment was submitted to the IQ-Tree server (Nguyen et al., 2015), which identified the best-fitting of 144 tested amino acid substitution models (JTT + I + G4, JTT + G4, JTT + G4, JTT + G4, JTT + I + G4, JTT + I + G4 and JTT + I + G4 for IGFBP-1, −2, −3, −4, −5, −6 and IGF2, respectively) and performed phylogenetic analysis by maximum likelihood, employing a fast unbiased bootstrapping approach with 1000 replicates (Minh et al., 2013) to assess statistical confidence for each reconstructed node. Final phylogenetic trees were drawn using Mega v.6 (Tamura et al., 2013).

In addition, a further phylogenetic analysis was performed on nucleotide data for putative salmonid *IGFBP-6B* paralogues (complete coding sequences; rationale provided in section 3.2), covering all the data captured from the salmonid species, plus *IGFBP-6B* orthologues from *E. lucius* and *G. aculeatus*. The alignment was performed in MAFFT (Katoh and Standley, 2013) and led to an alignment of 558 nucleotide characters that was submitted to the IQ-Tree server (Nguyen et al., 2015), which identified the best-fitting of 88 tested nucleotide substitution models (K2P + G4) and performed phylogenetic analysis by

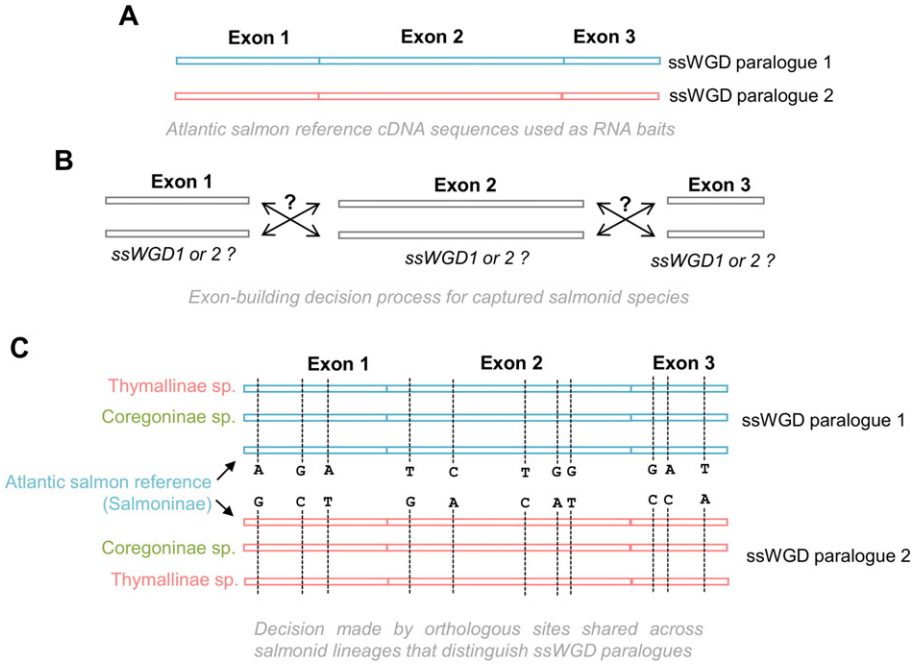

Fig. 2. Depiction of the assembly of contigs rebuilt from salmonid sequence capture data, where fragmented genomic DNA has been captured using RNA baits designed to cDNA sequences from a reference species. (A) Shows an example of two Atlantic salmon reference probe sequences, representing a pair of ssWGD paralogues with three exons. (B) Our sequence capture data would frequently recover two similar contigs, each containing exon sequences with high similarity to both reference ssWGD sequences. A manual decision is necessary to join exons into larger sequences that are truly orthologous to the original probe sequence pair (i.e. non-chimeric). (C) This decision is routinely informed with a high degree of confidence for ssWGD pairs that started diverging in the common salmonid ancestor (Macqueen and Johnston, 2014; Lien et al., 2016). In such instances, different salmonid species routinely conserve orthologous sites that distinguish two ssWGD paralogue sequences. Such paralogous differences are spread across different exons, allowing exons of different ssWGD paralogues to be correctly joined with respect to the original Atlantic salmon reference sequences.

maximum likelihood, as described above. The presented tree was rooted to the *G. aculeatus IGFBP-6B* orthologue.

2.4. Reconstruction of selective pressures

Separate codon alignments were constructed for IGFBP family members with unequivocal salmonid-specific paralogues: *IGFBP-1A*, *-1B*, *-2B*, *-3A*, *-3B*, *-5B*, *-6A* and *-6B*, along with *IGF2* (Supplementary dataset 3). The alignments were built using Pal2Nal (Suyama et al., 2006) with respect to amino acid alignments produced using MAFFT (Katoh and Standley, 2013). Each alignment included *E. lucius* as an outgroup to ssWGD and was separately uploaded to the DATAMONKEY server (Delport et al., 2010) of the maximum likelihood-based package HyPhy (Pond et al., 2005), specifying a fixed topology of species relationships (after Campbell et al., 2013; Macqueen and Johnston, 2014; see Fig. 1A). Model selection was done to determine the best-fitting of all possible general time reversible (GTR) nucleotide substitution models. All nine codon alignments were run through the BS-REL and RELAX tests (Kosakovksy Pond et al., 2011; Wertheim et al., 2014) via DATAMONKEY. The RELAX algorithm was ran twice per alignment, each time setting the 'foreground' test branch as the ancestral node to one of the two salmonid-specific paralogue pairs (in both cases, the remaining branches were set as the 'background' reference branches). Finally, we calculated d_N and d_S across each branch of the fixed salmonid phylogeny under the MG94 codon model (Muse and Gaut, 1994) crossed with the best-fitting GTR nucleotide substitution model, using parametric bootstrapping with 200 replicates to establish variation around the sampled parameters including on d_N/d_S ratios (hereafter: ω).

3. Results and discussion

3.1. Sequence capture: an efficient tool for acquiring duplicated sequences across the salmonid family

Our first study aim was to test whether sequence capture provides an effective tool to acquire complete coding sequences, including ssWGD paralogues, across different salmonid family lineages. Here, we report the success of this approach using twenty IGF axis components (*IGF2* and nineteen IGFBP family members) captured *in solution* using RNA baits designed to cover complete coding sequences from one reference species (i.e. Atlantic salmon). Specifically, we captured paired-end sequence reads from fifteen salmonid species, including all three salmonid subfamilies and their recognized genera, and one close outgroup to ssWGD (see Fig. 1). A *de-novo* approach (Luo et al., 2012) was used to assemble the captured reads, which was favoured over a reference-guided assembly owing to the evolutionary distance between the captured species. Using BLAST, we screened each assembly for the captured sequence data, which typically represented single exons of the target genes, along with flanking gDNA from introns (usually around 100 bp; note, the coding sequence of all IGFBP genes and *IGF2* is spread over four separate exons). However, less frequently we identified scaffolds spanning multiple exons when introns were short enough to be assembled.

Therefore, a manual step was necessary to join different exons and build up larger coding sequences (outlined in Fig. 2). In this respect, BLAST would routinely identify pairs of distinct, yet closely-related contigs in our data, congruent with expectations concerning the presence of ssWGD paralogues (Fig. 2A, B). In such instances, it was crucial to avoid building chimeras representing a mixture of exons from different ssWGD paralogue pairs (Fig. 2B). Crucially, the majority of salmonid-specific gene duplicates began diverging soon after the ssWGD (Lien et al., 2016) and subsequently evolved as independent genes for tens of millions of years before the extant salmonid subfamilies diverged ~45–55 Ma (Macqueen and Johnston, 2014). In such cases, paralogous sequence substitutions present in the common salmonid ancestor will have been inherited by all salmonid lineages and such sites are routinely observed as orthologous characters shared among extant salmonid species (Fig. 2C), presumably due to the action of persistent purifying selection. This property can be exploited, such that, as long as salmonid-specific paralogues have been fully verified in one reference species (here, Atlantic salmon sequences), then exons from other species can always be joined to their orthologous exons within a ssWGD paralogue pair, allowing different exons to be joined with a high degree of confidence (Fig. 2).

Using this protocol to join exons manually (after exclusion of flanking introns), a total of 299 manual nucleotide sequences were built representing the target IGF axis genes from the different salmonid species and Northern pike, of which 244 recovered >90% of the coding region (Fig. 3). The average recovery of the probe sequences across all 16 captured species was approximately 85% (SD: 5%). There was no obvious pattern suggesting that evolutionary distance among different salmonid species was an important factor affecting recovery of sequences. In other words, the recovery of data was evidently no more efficient for close relatives to the Atlantic salmon RNA bait sequences (i.e. other members of Salmoninae) in comparison to more distant salmonids (i.e. Thymallinae and Coregoninae members) (Fig. 3). In fact, a clearer observed pattern was that certain probe sequences were less likely to recover complete coding sequences than others for all species (e.g. *IGFBP-1A2* and *-3A2*). Despite this, sequence data recovered for the more distant outgroup to ssWGD, i.e. the Northern pike, which split off from the salmonid ancestor ~125 Ma (Fig. 1A) was more frequently less complete, suggesting the limits of sequence capture were being reached in some instances.

For the majority of genes characterized, BLAST searches of the Atlantic salmon genome confirmed that salmonid-specific paralogue pairs were embedded within large duplicated chromosomal regions that started diverging rapidly after the salmonid WGD and still maintain extensive collinearity (see Lien et al., 2016), specifically: Chr10 and 23 for *IGFBP-1A1* and *-1A2*; Chr14 and 03 for *IGFBP-1B1* and *1B2*; Chr25 and 21 for *IGFBP-2B1* and *-2B2*; Chr10 and 23 for *IGFBP-3A1* and *-3A2*, Chr14 and 03 for *IGFBP-3B1* and *-3B2*; Chr21 and 25 for *IGFBP-5B1* and *-5B2*; Chr22 and 12 for *IGFBP-6A*1 and *-6A2* and Chr13 and 15 for *IGFBP-6B1* and *-6B2*. The Atlantic salmon *IGF2* gene was found once on Chr10, embedded within a duplicated region that started diverging rapidly after the ssWGD (data from Lien et al., 2016). Therefore, as these genes started diverging in the salmonid ancestor, it was possible to exploit ancestral variation among the salmonid-specific duplicates to confidently rebuild complete sequences in all the tested species (i.e. correctly joining exons following the concept described in Fig. 2). Conversely, BLAST searches of the Atlantic salmon genome revealed that *IGFBP-2A*, *-4* and *-5A* duplicates are embedded in large duplicated regions that started diverging much more recently, owing to a large delay in the rediploidization process (data from Lien et al., 2016). Specifically, *IGFBP-2A*, and *-4* are located within duplicated regions on Chr16 and 17, while *IGFBP-5A* falls within regions on Chr2 and 12, which show a very high level of sequence similarity. In such cases, reference sequences from one salmonid species cannot be used to confidently rebuild exon-spanning sequences in other species, as the paralogous sites may have arisen after the lineages separated by speciation and therefore cannot be shared across all lineages. Accordingly, while sequence capture recovered *IGFBP-2A*, *-4* and *-5A* sequences (Fig. 3), in this study, we collapsed any identified variation into a single consensus sequence and excluded these genes from downstream evolutionary analyses (Section 3.3) beyond confirmatory phylogenetic reconstruction (Section 3.2).

Prior to confirmation by phylogenetic analyses (Section 3.2), our data suggested a strong maintenance in the overall structure of the IGFBP gene family across different salmonids, along with the presence of two putative paralogues of *IGF2* in Coregoninae and Thymallinae (Fig. 3). The latter case is notable, as it demonstrates that RNA baits designed to a single gene can efficiently capture multiple related ssWGD

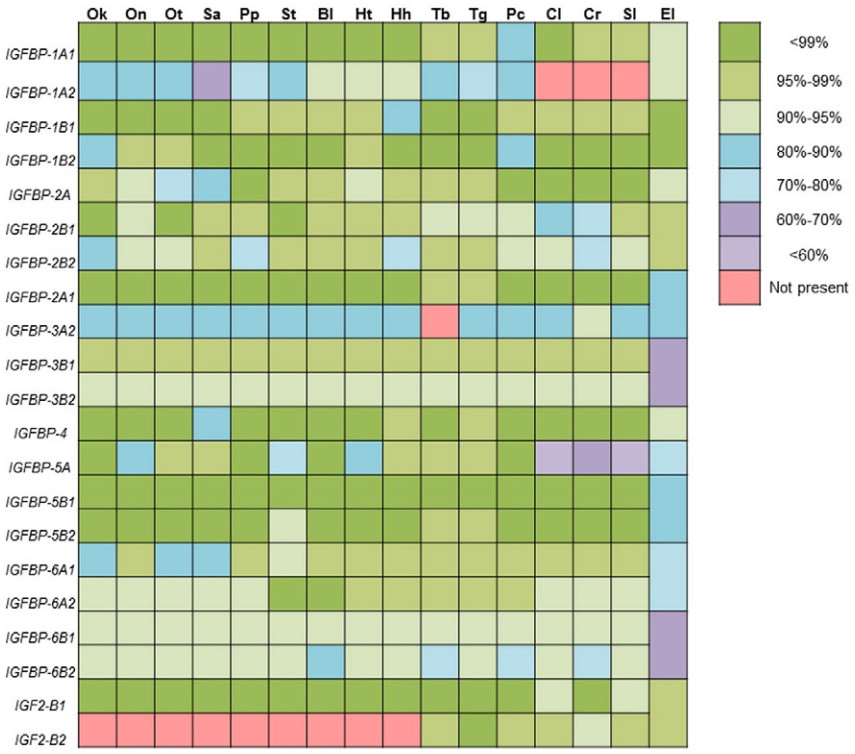

Fig. 3. Diagram depicting the proportion of the coding region recovered for the twenty IGF axis gene components targeted in our study by sequence capture.

paralogues. While it is more difficult to prove gene loss than retention, our sequence capture approach worked effectively across large evolutionary distances. Thus, with knowledge of salmonid phylogeny, it is possible to consider evidence for gene loss in an appropriate evolutionary context. Therefore, the complete absence of any *IGFBP-1A2* sequences within three separate members of *Coregonus* must be taken as strong evidence for an ancestral gene loss within the genus ancestor, especially considering that a more distant *IGFBP-1A* orthologue from the ssWGD outgroup was fully recovered using the same RNA baits (Fig. 3). On the other hand, it would be somewhat speculative to conclude a gene loss in the single additional case where a gene was totally absent in our data - *IGFBP-3A2* in *Thymallus baicalensis* – given that another member of Thymallinae retains the gene. Overall, these findings confirm that targeted sequence capture provides an efficient tool for acquiring gene coding sequences across many members of the salmonid family.

3.2. Phylogenetic validation of sequence capture approach for rebuilding salmonid-specific paralogues

To validate the success of targeted sequence capture in distinguishing ssWGD paralogues (i.e. study aim 1), we performed maximum likelihood phylogenetic analyses (done at the amino acid level) to reconstruct evolutionary histories for novel salmonid IGFBP and *IGF2* sequences obtained by sequence capture. As the captured data from Northern pike was less complete than the salmonid sequences (Section 3.1), our phylogenetic analyses were supplemented by complete sequences predicted from the genome assembly (Rondeau et al., 2014; available from NCBI). As expanded below, these analyses validate our hypothesis that sequence capture offers a useful approach to recover and characterize the evolution of gene families shaped by the ssWGD in any target salmonid lineage.

For the IGFBP family, we present one representative tree in the main paper (Fig. 4: for IGFBP-3), with all remaining trees provided as supplementary material (Fig. S1-S5), as the relevant branching patterns are mostly common to all the trees. Each IGFBP family member tree recaptured anticipated patterns of gene family evolution before the ssWGD, including the presence of all recognized teleost paralogues of IGFBP-1, -2, -3, -5 and -6 (Fig. 4, Fig. S1, S2, S4, S5) thought to be retained from the tsWGD. In particular, our new analyses, in contrast to past work considering teleost IGFBPs (Ocampo-Daza et al., 2011; Macqueen et al., 2013), included the spotted gar as a ray-finned fish outgroup to tsWGD (Braasch et al., 2016) - this species invariably branched with maximal statistical support outside two teleost IGFBP clades ('A' and 'B′ nomenclature), with all ray-finned fish being sister to a monophyletic clade of lobe-finned fish (Fig. 4, Fig. S1–S5). Moreover, most previously identified salmonid-specific paralogues retained from ssWGD (Macqueen et al., 2013) were confirmed to be ancestral salmonid characters. Specifically, for the tsWGD family members IGFBP-1A and -1B (Fig. S1), -2B (Fig. S2), -3A and -3B (Fig. 4), 5B (Fig. S4) and -6A (Fig. S5), our phylogenetic trees recovered two salmonid-specific clades, each containing the reference Atlantic salmon sequences branching within Salmoninae (as anticipated in light of species relationships, Fig. 1A). Moreover, within each salmonid clade, the three salmonid subfamilies were usually monophyletic, while relationships therein were broadly congruent with expected species relationships to the level of genera (Campbell et al., 2013; Macqueen and Johnston, 2014). Conversely, the branching arrangements separating salmonid subfamilies were more variable, which was again expected, owing to the short evolutionary time separating the most recent common salmonid ancestor to the ancestors of each subfamily (Macqueen and Johnston, 2014). The presence of the Northern pike outgroup to ssWGD (Rondeau et al., 2014), which is novel to this study, adds further weight to conclusions concerning the salmonid-specific paralogues. Specifically, the relevant pike IGFBP sequences either branched as the sister lineage to a pair of salmonid-specific paralogue clades (i.e. sister to IGFBP-1B, IGFBP-3A, IGFBP-3B, IGFBP-5B and IGFBP-6A, Fig. 4, Fig. S1, S4 and S5) or in some instances, as the sister lineage to one of the salmonid-specific

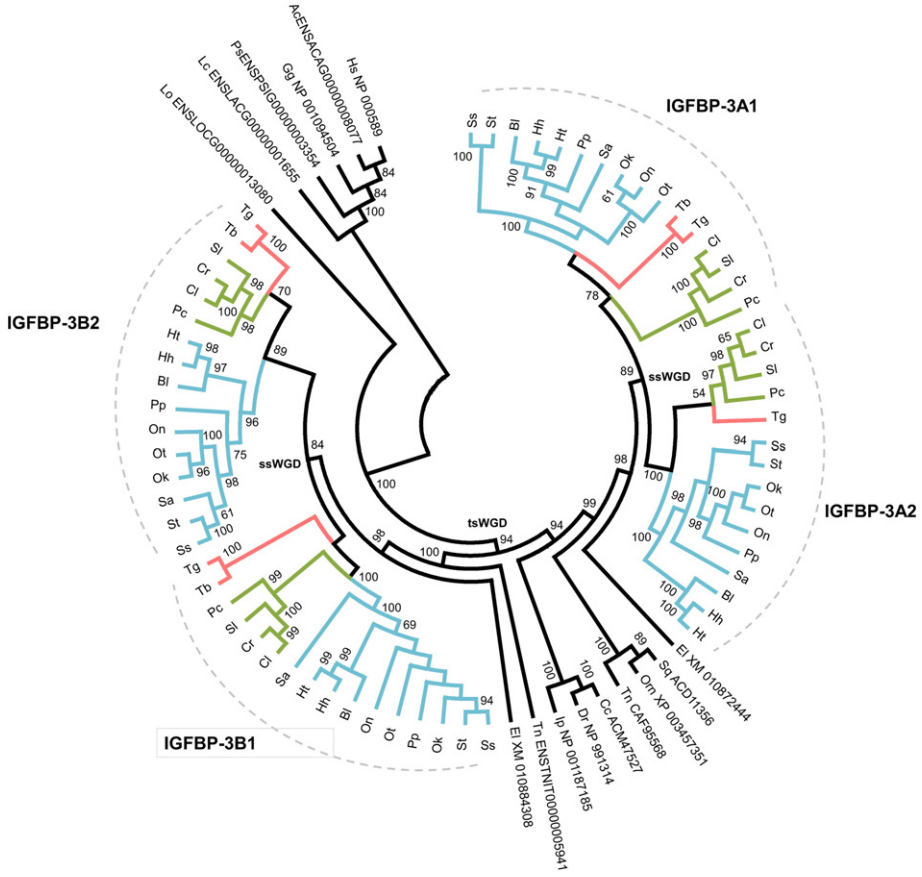

Fig. 4. Maximum likelihood topology of IGFBP-3 family member proteins including all salmonid sequences recovered during the study. The tree was built using IQ-Tree from an alignment of 307 amino acid positions under the JTT + G4 amino acid substitution model. Branch support values were gained by bootstrapping with 1000 replicates (support values >50% of sampled bootstraps is shown). The position of tsWGD and ssWGD is highlighted. Branch colours depict different salmonid subfamilies as described in Fig. 1. Abbreviations of species on tip labels are as follows: Salmoninae (blue): Ss = *Salmo salar*, St = *Salmo trutta*, Ht = *Hucho taimen*, Hh = *Hucho huho*, Bl = *Brachymystax lenok*, Pp = *Parahucho perryi*, Sa = *Salvelinus alpinus*, On = *Oncorhynchus nerka*, Ot = *Oncorhynchus tshawytscha*, Ok = *Oncorhynchus kisutch*; Thymallinae (red): Tb = *Thymallus baicalensis*, Tg = *Thymallus grubii*; Coregoninae (green): Cl = *Coregonus lavaretus*, Cr = *Coregonus renke*, Pc = *Prosopium coulteri*. Other teleosts: El = *Esox lucius*, Dr = *Danio rerio*, Ip = *Ictalurus punctatus*, Cc = *Cyprinus carpio*, Tn = *Tetraodon nigroviridis*, Sq = *Seriola quinqueradiata*, Orn = *Oreochromis niloticus*. Outgroup ray-finned fish to tsWGD: Lo: *Lepisosteus oculatus*. Lobe-finned fish (including tetrapods): Lc = *Latimeria chalumnae*, Ps = *Pelodiscus sinensis*, Ac = *Anolis carolinensis*, Hs = *Homo sapiens*, Gg = *Gallus gallus*. All species abbreviations used in the study are provided in Table S3.

paralogue clades (i.e. to IGFBP-1A2; IGFBP-2B2, Fig. S1, S2,), the former being expected and the latter representing a probable branching artefact associated with asymmetric paralogue evolution (see Section 3.4), considering that these paralogue pairs reside in verified duplicated ssWGD regions in the Atlantic salmon genome (Lien et al., 2016). The branching of putative salmonid-specific paralogues in the IGFBP-6B clade failed to recover two monophyletic clades representing each anticipated ssWGD paralogue (Fig. S5), again suggesting a branching artefact, given that these genes are present in duplicated regions retained from ssWGD in the Atlantic salmon genome (Lien et al., 2016). To confirm the branching artefact for IGFBP-6B, we performed an additional ML analysis using nucleotide characters, which should have more phylogenetic signal at the relatively modest levels of genetic distance characterizing salmonid evolution (Fig. S6). Indeed, this analysis strongly supported the presence of two ancestrally divergent salmonid IGFBP-6B paralogues retained from ssWGD that have been retained in all salmonid species (Fig. S6).

For IGF2, our phylogenetic analyses included two characterized teleost duplicates (IGF2-A and -B) thought to be retained from tsWGD, which to date have only be observed in members of the Ostariophysi, including zebrafish (Zou et al., 2009) (Fig. 5). In our tree, the spotted gar branched as the sister to one of the two teleost paralogue clades (IGF2B) with strong support (Fig. 2), which requires either that the duplication of IGF2 occurred before the tsWGD or that our analysis recovered a branching artefact. The latter is more probable, as past work identified double conserved synteny comparing the gene regions containing zebrafish *IGF2A* and *IGF2B* genes to the single-copy *IGF2* gene of human (Zou et al., 2009), which is a hallmark of a large scale duplication event, explained most parsimoniously by tsWGD. In any case, it is important to note that all salmonid IGF2 sequences formed a monophyletic group within the teleost IGF2B clade and split into two sister clades, each containing different salmonid subfamilies as monophyletic groups (Fig. 5). However, while Thymallinae and Coregoninae were present in both salmonid-specific clades, single sequences recovered in all Salmoninae members were only present in one of the clades (Fig. 5). The ssWGD outgroup species pike branched as the sister group to these two putative salmonid-specific paralogous clades (Fig. 5). As mentioned above, given that the reference Atlantic salmon *IGF2* gene sequence resides within a duplicated collinear region of the genome that started diverging in the salmonid ancestor (Lien et al., 2016), this data offers strong evidence that two *IGF2B* paralogues were retained from the ssWGD (hereafter: *IGF2-B1* and *-B2*) and while both have been retained in Thymallinae and Coregoninae, one was lost in the Salmoninae ancestor.

Overall, these analyses further verify the success of our sequence capture approach, highlighting that the manual steps required to

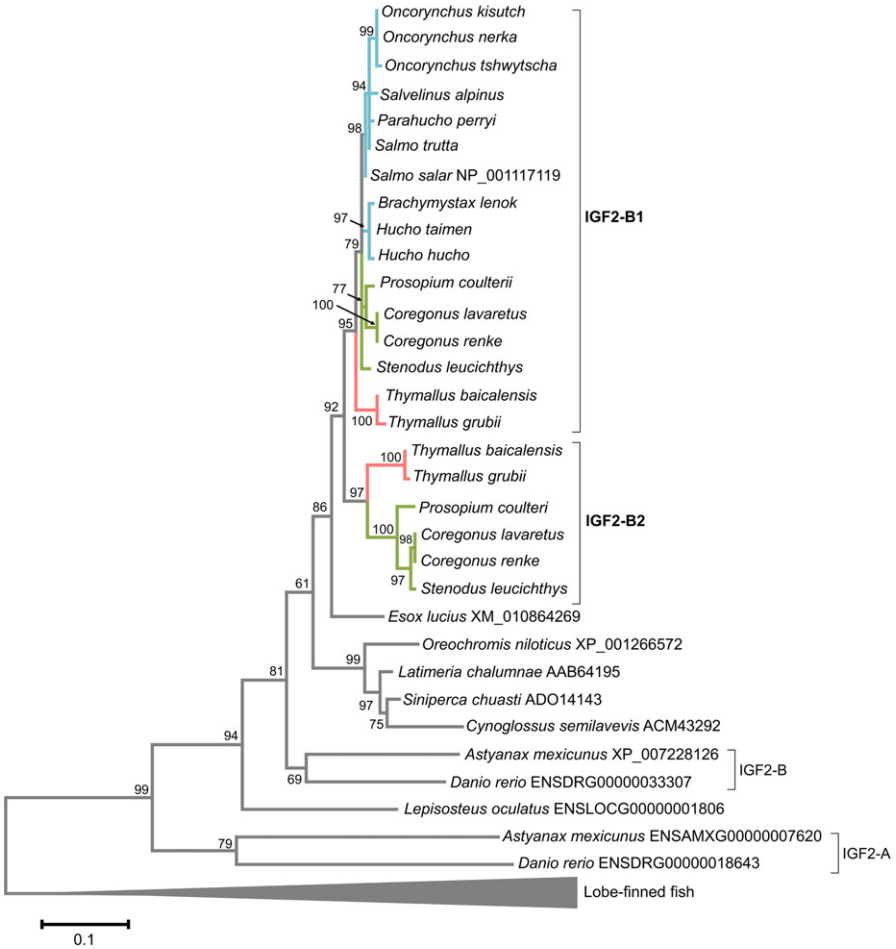

Fig. 5. Maximum likelihood tree offering strong support for novel IGF2-B paralogues retained from ssWGD. The tree was built using IQ-Tree from an alignment of 214 amino acid positions under the JTT + I + G4 amino acid substitution model. Other details are as provided in the Fig. 4 legend, except for the following additional species abbreviations: For teleosts: Sc = *Siniperca chuatsi*, Cs = *Cynoglossus semilaevis*, Am = *Astyanax mexicanus*.

rebuild coding regions of different salmonid-specific paralogues leads to phylogenetically-informative data that is congruent with expectations regarding ssWGD.

3.3. High-resolution evolutionary analysis of key components from the duplicated salmonid IGF axis

Our final study aim was to demonstrate the utility of sequence capture for comparative evolutionary analyses encompassing the salmonid lineage – hence, we explored the molecular evolution of IGFBP family members retained as two ssWGD paralogues that began diverging in the common salmonid ancestor, along with the novel salmonid *IGF2-B* paralogues (Fig. 6). To gain insights into the selective pressures underlying the sequence evolution and divergence of IGF axis ssWGD paralogues, we first estimated d_N/d_S ratios (ω) for each branch in the salmonid family tree (Fig. 6). For the ssWGD outgroup branches (Northern pike), we observed low ω estimates for all tested genes, suggesting strong purifying selection to maintain ancestral functions across the pike's evolutionary history (Fig. 6). Conversely, when taking the data as a whole, while low ω estimates were common in the salmonid phylogeny, they were frequently interspersed with higher ω values, including values exceeding one (Fig. 6). This pattern can be explained by a general background of purifying selection, with episodic periods of faster protein evolution, potentially associated with some combination of relaxed and positive selection. There is also evidence of heterogeneity comparing the historic selective regimes operating on the different salmonid IGFBP family members, including the pattern of divergence among ssWGD paralogues. For example, both ssWGD paralogues of IGFBP-5B have evidently been subjected to a greater level of purifying selection across all salmonids (i.e. mainly low estimated ω values) compared to all other tested IGFBP family members (Fig. 6). This includes the period immediately post-WGD, suggesting strong selection to preserve pre-WGD functions in both paralogues. Conversely, the other IGFBP family members had more variable ω estimates across branches, typically affecting both paralogues, which was a pattern observed to some extent for *IGF2-B* (Fig. 6). Further, IGFBP family members often had divergent ω estimates comparing the two immediate post-ssWGD paralogue branches (Fig. 6), suggesting a period of functional divergence in the common salmonid ancestor.

When interpreting these patterns, it is important to note that the ω estimates shown in Fig. 6 were often underpinned by a small number of sequence changes owing to the close relationships of many tested salmonid species and are sometimes subject to large variance in the underlying parameter estimates. In this respect, we suggest particular caution is applied when interpreting branches with ω values exceeding one, often taken as evidence of positive selection. Such high ω values, if accompanied by statistical uncertainty, cannot disentangle the relevant importance of positive selection from a relaxation in purifying selection. Moreover, it is possible for a small number of sites to be fixed by positive selection in a background of purifying selection, associated with low ω values for specific branches in a tree when assessed at the level of whole gene coding regions. Therefore, we also employed more powerful tests specifically geared towards identification of episodic positive selection or relaxation of purifying selection using branch-site models

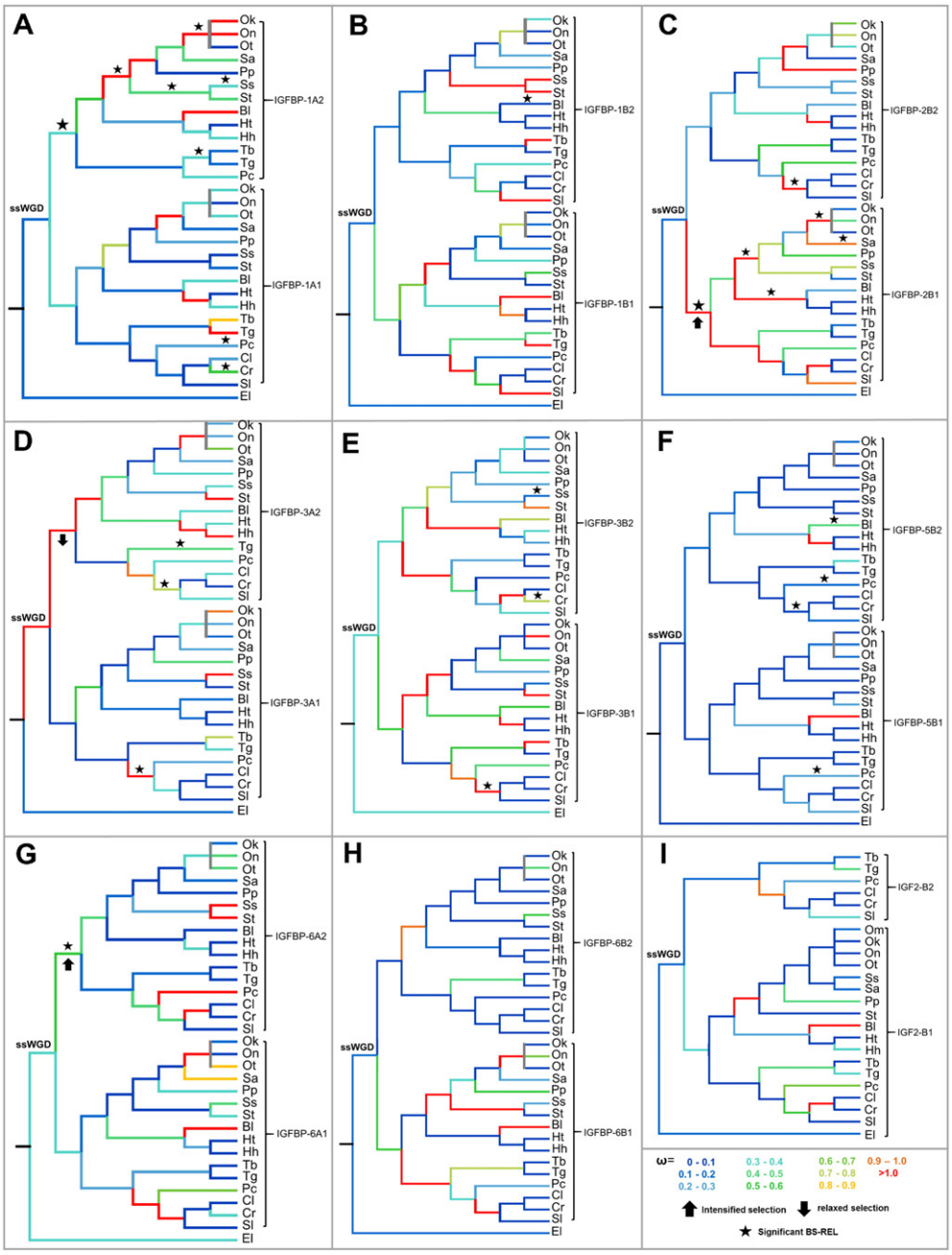

Fig. 6. Integrative analysis reconstructing the evolution of duplicated IGF axis components retained from ssWGD at unprecedented phylogenetic resolution. Maximum likelihood estimated ratios of non-synonymous (d_N) and synonymous (d_S) substitution rates (ω) are shown for all salmonid IGFBP members retaining ssWGD paralogues that started diverging in the common salmonid ancestor (**A**–**H**), along with novel identified ssWGD paralogues of *IGF2-B* (**I**). The branches are coloured to depict different ranges of ω as shown in the key (bottom right panel). Note, variation in parameter estimates contributing to ω are not shown, only the mean values derived from parametric bootstrapping (see Materials and methods); accordingly, many of the ω values are subject to large confidence intervals. In addition, significant results from the BS-REL and RELAX tests (see data in Table S1 and S2) are indicated at the relevant nodes on the trees according to symbols shown in the key on the bottom right panel. Species abbreviations are as provided in the Fig. 4 legend.

that consider variation across lineages as well as different codons in a gene. The first test, called BS-REL (Kosakovsky Pond et al., 2011), aims to detect positive selection, while the second test, called RELAX (Wertheim et al., 2014), is designed to distinguish the presence of relaxed selection from an intensification of selective pressure (data in Table S1, S2). Whereas the BS-REL test was used across the entire tree, the RELAX test requires specific nodes to be selected for testing and was done only on the immediate post-ssWGD branches. The BS-REL test provided evidence for positive selection in the ancestral salmonid branches leading to *IGFBP-1A2*, *-2B1* and *-6A2* (Fig. 6A, C, G; Table S1), with the latter two instances being coincident with evidence for an intensification in selective pressure according to the RELAX test (Table S2). These data are thus consistent with a scenario where adaptive functions became fixed in one ssWGD paralogue for three separate pairs of salmonid-specific IGFBP proteins in the ancestor to extant salmonids.

It is also interesting to note evidence for positive selection on numerous branches within different salmonid lineages for both *IGFBP-1A2* and to a lesser extent *IGFBP-1A1* (Fig. 6A). Included among these

for *IGFBP-1A2* was the branch leading into all members of the ancestrally-anadromous (i.e. having within-life capacity to migrate between fresh and seawater, including the ability to undergo smoltification) members of Salmoninae (see Alexandrou et al., 2013) and several branches therein, as well as the ancestor to the Thymallinae (Fig. 6A). In Atlantic salmon, the *IGFBP-1A2* gene was previously observed to be lowly expressed compared to its ssWGD paralogue *IGFBP-1A1*, which retains expression similar to other teleost *IGFBP-1A* genes (Macqueen et al., 2013). Thus, such repeated bouts of potential positive selection may reflect ongoing adaptive changes linked to the evolution of novel protein functions in *IGFBP-1A* that potentially accompany evolutionary divergence in gene expression. For *IGFBP-2B1*, we also observed a number of salmonid branches with evidence for positive selection, including, in addition to the ancestors to all salmonids, separate branches leading into the ancestrally-anadromous Salmoninae members as well as their sister group (*Hucho-Brachymystax*) that never evolved anadromy (Alexandrou et al., 2013) (Fig. 6C). *IGFBP-2B1* is more highly expressed than its ssWGD paralogue in Atlantic salmon and restricted to liver (Macqueen et al., 2013), which is the predominant tissue that releases IGFs and IGFBPs into the circulation to have endocrine effects on IGF signalling. Conversely, its ssWGD paralogue *IGFBP-2B2* was inferred to have undergone positive selection on the branch leading into members of Coregoninae, which independently evolved capacity for anadromy (Alexandrou et al., 2013). While the functional relevance of such evolutionary patterns remain unexplored in the context of anadromy, the role of the IGF axis in controlling seawater tolerance and salmonid smoltification is well-established (Björnsson et al., 2011), pointing towards interesting lines of future work in a comparative evolutionary context.

For *IGFBP-3A2*, a high ω value was associated with evidence for a significant relaxation in selection (Fig. 6D; Table S2). Notably, this was the only detected instance supporting a relaxation in selection in the immediate post-ssWGD paralogue branches. For *IGFBP-5B2*, despite the overall strong background of purifying selection on both ssWGD paralogues, there was nevertheless evidence of positive selection within some salmonid lineages, particularly from within the Coregoninae subfamily (Fig. 6F). *IGFBP-6B* was the only IGFBP family member without any evidence of positive selection according to the BS-REL test (Fig. 6H), along with *IGF2-B* (Fig. 6I).

Overall, these data suggest that the evolution of duplicated salmonid IGF-axis components has involved bouts of protein-level level divergence associated with shifts in selective pressures on both ssWGD paralogues in a pair. While potentially important changes among ssWGD paralogues occurred ancestrally, ongoing evolutionary divergence within salmonid lineages seem to have been equally important. While this study aimed largely to demonstrate a useful application for phylogenetically-diverse salmonid data recovered by sequence capture, it will be crucial in the future to expand on such data to decipher the actual functional relevance of such evolutionary patterns across salmonid lineages, not just for duplicated IGF axis components, but across the different functional pathways contributing to interesting aspects of salmonid physiology and development.

4. Conclusions and perspectives

We have demonstrated that sequence capture offers a routine method for generating large-scale protein-coding sequence data across salmonids, making it possible to address many genome-scale questions related to the ssWGD using phylogenomic approaches, for example, the timing of divergence of ssWGD paralogues in different lineages and subsequent patterns of paralogue retention and loss. Moreover, such data, when analysed in a framework that is cognizant of the complexities of salmonid genome evolution (Macqueen and Johnston, 2014; Lien et al., 2016), can provide a necessarily large substrate of truly orthologous nuclear sequences to resolve evolutionary relationships within salmonids to the finest possible scale of phylogenetic resolution. A conclusive salmonid phylogeny is yet to be achieved owing to several elusive inter- and intra-generic relationships, with the most compelling insights to date having been drawn from the mitochondrial genome (e.g. Crête-Lafrenière et al., 2012; Campbell et al., 2013) owing to a lack of truly-orthologous nuclear gene sequences spanning an appropriate phylogenetic breadth of taxa (Macqueen and Johnston, 2014).

Such studies are ongoing in our group and complement larger efforts to resolve whole genome sequences in model species (Berthelot et al., 2014; Lien et al., 2016). Undoubtedly, such whole genome-scale sequencing projects are providing unprecedented and remarkable insights into the evolution of salmonid genomes and will always represent the bedrock from which understanding of these fascinating fish is drawn. Nonetheless, generating a high-quality salmonid genome (which are large and highly repetitive), remains a non-trial task, meaning new genomes are unlikely to be finalized for all the major salmonid lineages for some time. In the meantime, sequence capture offers a useful option to bridge the gap between the genome-wide insights made possible by sequencing a few key salmonid genomes and the rich comparative insights gained by broadening the number of phylogenetic lineages sampled.

Acknowledgements

This study was funded by a Natural Environment Research Council grant (NERC, project code: NBAF704). FML is funded by a NERC Doctoral Training Grant (Project Reference: NE/L50175X/1). RLS was an undergraduate student at the University of Aberdeen and benefitted from financial support from the School of Biological Sciences. DJM is indebted to Dr. Steven Weiss (University of Graz, Austria), Dr. Takashi Yada (National Research Institute of Fisheries Science, Japan), Dr. Robert Devlin (Fisheries and Oceans Canada, Canada), Prof. Samuel Martin (University of Aberdeen, UK), Mr. Neil Lincoln (Environment Agency, UK) and Prof. Colin Adams/Mr. Stuart Wilson (University of Glasgow, UK) for providing salmonid material or assisting with its sampling. We are grateful to staff at the Centre for Genomics Research (University of Liverpool, UK) (i.e. NERC Biomolecular Analysis Facility – Liverpool; NBAF-Liverpool) for performing sequence capture/Illumina sequencing and providing us with details on associated methods that were incorporated into the manuscript. Finally, we are grateful to the organizers of the Society of Experimental Biology Satellite meeting '*Genome-powered perspectives in integrative physiology and evolutionary biology*' (held in Prague, July 2015) for inviting us to contribute to this special edition of Marine Genomics and hosting a really stimulating meeting.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.margen.2016.06.003.

References

Alexandrou, M.A., Swartz, B.A., Matzke, N.J., Oakley, T.H., 2013. Genome duplication and multiple evolutionary origins of complex migratory behavior in Salmonidae. Mol. Phylogenet. Evol. 69, 514–523.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Alzaid, A., Castro, R., Wang, T., Secombes, C.J., Boudinot, P., Macqueen, D.J., Martin, S.A., 2016. Cross-talk between growth and immunity: coupling of the insulin-like growth factor axis to conserved cytokine pathways in rainbow trout. Endocrinology http://dx.doi.org/10.1210/en.2015-2024.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B. … Guiguen, Y. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat. Commun. 5, 3657.

Bi, K., Linderoth, T., Vanderpool, D., Good, J.M., Nielsen, R., Moritz, C., 2013. Unlocking the vault: next-generation museum population genomics. Mol. Ecol. 22, 6018–6032.

Björnsson, B.T., Stefansson, S.O., McCormick, S.D., 2011. Environmental endocrinology of salmon smoltification. Gen. Comp. Endocrinol. 170, 290–298.

Bower, N.I., Li, X., Taylor, R., Johnston, I.A., 2008. Switching to fast growth: the insulin-like growth factor (IGF) system in skeletal muscle of Atlantic salmon. J. Exp. Biol. 211, 3859–3870.

Bower, N.I., Johnston, I.A., 2010. Transcriptional regulation of the IGF signaling pathway by amino acids and insulin-like growth factors during myogenesis in Atlantic salmon. PLoS ONE 5, e11100.

Braasch, I., Gehrke, A.R., Smith, J.J., Kawasaki, K., Manousaki, T., Pasquier, J., ... Postlethwait, J.H., 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat. Genet. 48, 427–437.

Calvo, S.E., Compton, A.G., Hershman, S.G., Lim, S.C., Lieber, D.S., Tucker, E.J., ... Christodoulou, J., 2012. Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. Sci. Transl. Med. 4, 118ra10.

Campbell, M.A., López, J.A., Sado, T., Miya, M., 2013. Pike and salmon as sister taxa: detailed intraclade resolution and divergence time estimation of Esociformes + Salmoniformes based on whole mitochondrial genome sequences. Gene 530, 57–65.

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., ... Nelson-Williams, C., 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc. Natl. Acad. Sci. U. S. A. 106, 19096–19101.

Crête-Lafrenière, A., Weir, L.K., Bernatchez, L., 2012. Framing the Salmonidae family phylogenetic portrait: a more complete picture from increased taxon sampling. PLoS One 7, e46662.

Delport, W., Poon, A.F., Frost, S.D., Pond, S.L.K., 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26, 2455–2457.

Eytan, R.I., Evans, B.R., Dornburg, A., Lemmon, A.R., Lemmon, E.M., Wainwright, P.C., Near, T.J., 2015. Are 100 enough? Inferring acanthomorph teleost phylogeny using anchored hybrid enrichment. BMC Evol. Biol. 15, 113.

Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61, 717–726.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W. … Nusbaum, C., 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27, 182–189.

Grover, C.E., Salmon, A., Wendel, J.F., 2012. Targeted sequence capture as a powerful tool for evolutionary analysis. Am. J. Bot. 99, 312–319.

Hall, T., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41, 95–98.

Haussler, D., O'Brien, S.J., Ryder, O.A., Barker, F.K., Clamp, M., Crawford, A.J. … Turner, S. 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J. Hered. 100, 659–674.

Hebert, F.O., Renaut, S., Bernatchez, L., 2013. Targeted sequence capture and resequencing implies a predominant role of regulatory regions in the divergence of a sympatric lake whitefish species pair (*Coregonus clupeaformis*). Mol. Ecol. 22, 4896–4914.

Hedtke, S.M., Morgan, M.J., Cannatella, D.C., Hillis, D.M., 2013. Targeted enrichment: maximizing orthologous gene comparisons across deep evolutionary time. PLoS ONE 8, e67908.

Heyduk, K., Trapnell, D.W., Barrett, C.F., Leebens-Mack, J., 2015. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. Biol. J. Linn. Soc. 117, 106–120.

i5K Consortium, 2013. The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J. Hered. 104, 595–600.

Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., Stange-Thomann, N., Mauceli, E., ... Roest Crollius, H., 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431, 946–957.

Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., ... Zhang, G., 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346, 1320–1331.

Johnston, I.A., Bower, N.I., Macqueen, D.J., 2011. Growth and the regulation of myotomal muscle mass in teleost fish. J. Exp. Biol. 214, 1617–1628.

Jones, J.I., Clemmons, D.R., 1995. Insulin-like growth factors and their binding proteins: biological actions. Endocr. Rev. 16, 3–34.

Jones, M.R., Good, J.M., 2015. Targeted capture in evolutionary and ecological genomics. Mol. Ecol. 25, 185–202.

Joshi, N., Fass, J. 2011. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). [Software]. (Available at) https://github.com/najoshi/sickle.

Kamangar, B.B., Gabillard, J.C., Bobe, J., 2006. Insulin-like growth factor-binding protein (IGFBP)-1,-2,-3,-4,-5, and-6 and IGFBP-related protein 1 during rainbow trout postvitellogenesis and oocyte maturation: molecular characterization, expression profiles, and hormonal regulation. Endocrinology 147, 2399–2410.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

Kosakovsky Pond, S.L., Murrell, B., Fourment, M., Frost, S.D., Delport, W., Scheffler, K., 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol. Biol. Evol. 28, 3033–3043.

Landan, G., Graur, D., 2008. Local reliability measures from sets of co-optimal multiple sequence alignments. Pac. Symp. Biocomput. 13, 15–24.

Lemmon, A.R., Emme, S.A., Lemmon, E.M., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61, 727–744.

Li, C., Hofreiter, M., Straube, N., Corrigan, S., Naylor, G.J., 2013. Capturing protein-coding genes across highly divergent species. Biotechniques. 54, 321–326.

Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Kent, M.P., Nome, T., ... Davidson, W.S., 2016. The Atlantic salmon genome provides insights into rediploidization. Nature http://dx.doi.org/10.1038/nature17164.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 1, 18.

McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C., Brumfield, R.T., 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol. 66, 526–538.

Macqueen, D.J., Kristjánsson, B.K., Johnston, I.A., 2010. Salmonid genomes have a remarkably expanded akirin family, coexpressed with genes from conserved pathways governing skeletal muscle growth and catabolism. Physiol. Genomics 42, 134–148.

Macqueen, D.J., Kristjánsson, B.K., Paxton, C.G., Vieira, V.L., Johnston, I.A., 2011. The parallel evolution of dwarfism in Arctic charr is accompanied by adaptive divergence in mTOR-pathway gene expression. Mol. Ecol. 20, 3167–3184.

Macqueen, D.J., Garcia de la serrana, D., Johnston, I.A., 2013. Evolution of ancient functions in the vertebrate insulin-like growth factor system uncovered by study of duplicated salmonid fish genomes. Mol. Biol. Evol. 30, 1060–1076.

Macqueen, D.J., Johnston, I.A., 2014. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc. Biol. Sci. 281, 20132881.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10.

Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., ... Wong, G.K.-S., 2014. Data access for the 1000 plants (1KP) project. Gigascience 3, 17.

Metzker, M.L., 2010. Sequencing technologies - the next generation. Nat. Rev. Genet. 11, 31–46.

Nadeau, N.J., Whibley, A., Jones, R.T., Davey, J.W., Dasmahapatra, K.K., Baxter, S.W., ... Jiggins, C.D., 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. Philos. Trans. R. Soc. B 367, 343–353.

Minh, B.Q., Nguyen, M.A.T., von Haeseler, A., 2013. Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. 30, 1188–1195.

Muse, S.V., Gaut, B.S., 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11, 715–724.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., ... Shendure, J., 2009. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461, 272–276.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274.

Ocampo-Daza, D., Sundström, G., Bergqvist, C.A., Duan, C., Larhammar, D., 2011. Evolution of the insulin-like growth factor binding protein (IGFBP) family. Endocrinology 152, 2278–2289.

Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J., Zwick, M.E., 2007. Microarray-based genomic selection for high-throughput resequencing. Nat. Methods 4, 907–909.

Pierce, A.L., Shimizu, M., Felli, L., Swanson, P., Dickhoff, W.W., 2006. Metabolic hormones regulate insulin-like growth factor binding protein-1 mRNA levels in primary cultured salmon hepatocytes; lack of inhibition by insulin. J. Endocrinol. 191, 379–386.

Pond, S.L.K., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676–679.

Prum, R.O., Berv, J.S., Dornburg, A., Field, D.J., Townsend, J.P., Lemmon, E.M., Lemmon, A.R., 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526, 569–573.

Rondeau, E.B., Minkley, D.R., Leong, J.S., Messmer, A.M., Jantzen, J.R., von Schalburg, K.R., ... Koop, B.F., 2014. The genome and linkage map of the northern pike (*Esox lucius*): conserved synteny revealed between the salmonid sister group and the Neoteleostei. PLoS One 9, e102089.

Saintenac, C., Jiang, D., Akhunov, E.D., 2011. Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. Genome Biol. 12, R88.

Salmon, A., Ainouche, M., 2015. Next-generation sequencing and the challenge of deciphering evolution of recent and highly polyploid genomes. In: Hörandl, E., Appelhanas, M.S. (Eds.), Next Generation Sequencing in Plant Systematics. International Association for Plant Taxonomy, 2015.

Sela, I., Ashkenazy, H., Katoh, K., Pupko, T., 2015. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res. 43, W7–W14.

Shimizu, M., Kishimoto, K., Yamaguchi, T., Nakano, Y., Hara, A., Dickhoff, W.W., 2011. Circulating salmon 28-and 22-kDa insulin-like growth factor binding proteins (IGFBPs) are co-orthologs of IGFBP-1. Gen. Comp. Endocrinol. 174, 97–106.

Sun, Y., Huang, Y., Xiaofeng, L., Baldwin, C.C., Zhou, Z., Yan, Z., ... Shi, Q., 2016. Fish-T1K (Transcriptomes of 1000 fishes) project: large-scale transcriptome data for fish evolution studies. GigaScience. 5, 18.

Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34, W609–W612.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729.

Tennessen, J.A., Govindarajulu, R., Liston, A., Ashman, T.-L., 2013. Targeted sequence capture provides insight into genome structure and genetics of male sterility in a gynodioecious diploid strawberry, *Fragaria vesca* ssp. *bracteata* (Rosaceae). G3 (Bethesda) 3, 1341–1351.

Tewhey, R., Warner, J.B., Nakano, M., Libby, B., Medkova, M., David, P.H., ... Frazer, K.A., 2009. Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat. Biotechnol. 27, 1025–1031.

Turner, E.H., Lee, C., Ng, S.B., Nickerson, D.A., Shendure, J., 2009. Massively parallel exon capture and library-free resequencing across 16 genomes. Nat. Methods 6, 315–316.

Wertheim, J.O., Murrell, B., Smith, M.D., Kosakovsky Pond, S.L., Scheffler, K., 2014. RELAX: detecting relaxed selection in a phylogenetic framework. Mol. Biol. Evol. 32, 820–832.

Wetterstrand, K.A., 2015. DNA Sequencing Costs: data from the NHGRI Genome Sequencing Program (GSP) (Available at www.genome.gov/sequencingcosts. Last accessed 18/02/2016.)

Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., ... Leebens-Mack, J., 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. 111, E4859–E4868.
Wood, A.W., Duan, C., Bern, H.A., 2005. Insulin-like growth factor signaling in fish. Int. Rev. Cytol. 243, 215–285.
Zhang, G., 2015. Genomics: bird sequencing project takes off. Nature 522.
Zhang, G., Li, C., Li, Q., Li, B., Larkin, D.M., Lee, C., ... Wang, J., 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346, 1311–1320.
Zou, S., Kamei, H., Modi, Z., Duan, C., 2009. Zebrafish IGF genes: gene duplication, conservation and divergence, and novel roles in midline and notochord development. PLoS One 4, e7026.