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.