Fisheries Research xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Fisheries Research

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

Full length article

SNP identification and validation on genomic DNA for studying genetic diversity in *Thunnus albacares* and *Scomberomorus brasiliensis* by combining RADseq and long read high throughput sequencing

Zoila Raquel Siccha-Ramirez[a,b,*,1], Francesco Maroso[c,d,1], Belén G. Pardo[c], Carlos Fernández[c], Paulino Martínez[c,**], Claudio Oliveira[a]

[a] Departamento de Morfologia, Instituto de Biociências, UNESP, Botucatu, São Paulo, Brazil
[b] Instituto del Mar del Perú, Laboratorio Costero de Tumbes, Zorritos, Peru
[c] Departamento de Zooloxía, Xenética e Antropoloxía Física, Facultade de Veterinaria, Universidade de Santiago de Compostela, Campus de Lugo, Lugo, 27002, Spain
[d] Dipartimento di Biomedicina Comparata e Alimentazione (BCA), Università degli Studi di Padova, Italy

ARTICLE INFO

Keywords:
Scomberomorus brasiliensis
Thunnus albacares
ddRAD
Long-read high throughput sequencing
Genetic structure

ABSTRACT

A combination of a RADseq method (ddRAD) with long read high throughput sequencing (Roche 454) was tuned up in order to identify and validate a set of SNPs useful for gene diversity analysis in two important South American commercial tuna (*Thunnus albacares* and *Scomberomorus brasiliensis*). A total of 11 and 21 individuals of *T. albacares* and *S. brasiliensis*, respectively, were used for SNP identification. DNA was individually digested with two restriction enzymes (*Sbf*I and *Sph*I) and fragments between 300 and 600 bp selected. Combinatorial barcoding was used to identify individuals by including short sequences (5–7 bp) in the adaptors of each restriction site (P1 and P2). After adaptor ligation, samples were pooled and size-selected, amplified by PCR, and sequenced on a 454 GS-Junior sequencer. A total of 180,779 reads were produced with an average length and coverage of 287 bp and 26x, respectively. Sets of 60 and 79 SNPs were *in silico* selected for *T. albacares* and *S. brasiliensis*, respectively, and were tested and validated in 74 and 66 individuals, respectively, on a MassARRAY platform. A total of 36 and 47 SNPs were polymorphic and useful for population analysis. A preliminary study on two distant Brazilian populations of both species (~3000 km) with these SNPs suggested the absence of significant structure among local populations of both species. Our results demonstrate the possibility of combining ddRAD with long read high throughput sequencing for marker development in species with scarce genomic resources.

1. Introduction

Next Generation Sequencing (NGS) has revolutionized the field of genetics (Guo et al., 2016; Mardis, 2008; Metzker, 2010) allowing investigation on non-model species with unprecedented genomic coverage. An increasingly used application is the search, validation and large-scale genotyping of genetic markers using different methodologies (Helyar et al., 2012). SNPs are stable and usually bi-allelic polymorphisms (Mullikin et al., 2000), found in coding and non-coding regions (Bruneaux et al., 2013; Stölting et al., 2013), and homogeneously distributed across the genome at high densities (Du et al., 2012); thus, SNPs are the most common genetic markers for genomic screening (Vera et al., 2013). In fish, SNPs are found every ~100 bp (Pardo et al., 2008; Vera et al., 2013). These properties make SNPs ideal for comparative genomics (Bouza et al., 2012), evolutionary genomics (Du et al., 2012), fine mapping of genes associated with productive traits (Sánchez et al., 2009), genomic screening in populations for conservation and management (Albaina et al., 2013; Glaubitz et al., 2003; Kuhner et al., 2000; Vilas et al., 2015), and hybridization and impact of biological invasions (Bers et al., 2010).

Abbreviations: NGS, Next Generation Sequencing; SNP, Single Nucleotide Polymorphisms; RRL, Reduced Representation Library; RADseq, Restriction-site associated DNA sequencing; ddRAD-seq, double-digestion RADseq; RE, restriction enzyme; PCR, polymerase chain reaction; H_E, expected heterozygosity; H_O, observed heterozygosity; MAF, minimum allele frequency; H-W, Hardy-Weinberg; F_{IS}, inbreeding coefficient; F_{ST}, coefficient of relative differentiation between populations; NCBI, National Center for Biotechnology Information; BLAST, Basic Local Alignment Search Tool; WGS, whole genome shotgun; nr, non redundant; ts, transition; tv, transversion

* Corresponding author at: Departamento de Morfologia, Instituto de Biociências, UNESP, Botucatu, São Paulo, 18618-970, Brazil.
** Corresponding author.
E-mail addresses: rsicchar@gmail.com (Z.R. Siccha-Ramirez), paulino.martinez@usc.es (P. Martínez).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.fishres.2017.09.002
Received 3 March 2017; Received in revised form 30 August 2017; Accepted 4 September 2017

Reduced Representation Library (RRL) sequencing provides a high potential for efficient SNP discovery at very low cost (Sánchez et al., 2009), because sequencing is targeted to a small and adjustable genome fraction flanking the chosen RE sites (Du et al., 2012; Robledo et al., 2017). Since restriction sites are common to most individuals of a species, high throughput sequencing using DNA barcodes to label individuals, populations or species provides a set of reads at each restriction site to be explored bioinformatically. Depending on the goals of the study, SNP identification and genotyping can be jointly performed as a single step, or alternatively a subset of SNPs can be selected for further validation and genotyping using specific platforms (SNaPshot, Sequenom, Truseq, Ampliseq, *etc.*) (Davey et al., 2011; Robledo et al., 2017). This last option offers the possibility of developing a panel of adjustable SNP number to be routinely used at lower cost depending on the goals of the study. RADseq is a RRL strategy which has gained popularity in non-model organisms because it allows obtaining huge SNP genotyping data at very low cost (Baird et al., 2008; Gompert et al., 2010; Vandepitte et al., 2013). ddRAD-seq is a variant of the classical RADseq (Baird et al., 2008), which eliminates random shearing and explicitly uses size selection to recover a tunable number of regions according to the objectives of the study (Peterson et al., 2012). ddRAD not only allows high-throughput multiplexed sequencing amenable for genotyping, but also provides improved efficiency and robustness compared to classical RADseq (Zhou et al., 2014). However, the short length of reads of the commonly used platforms for RADseq hampers the subsequent development of SNP panels for genotyping in PCR-based platforms, if no reference genome is available. Consequently, combining ddRAD libraries with longer-read sequencing platforms could aid to solve this drawback making affordable the selection and validation of a subset of SNPs.

The yellowfin tuna (*Thunnus albacares,* Scombridae) is a migratory species with high commercial value found in tropical and subtropical waters all over the world (Guo et al., 2016). This species is currently overfished and appears on the Red List as near threatened species (Collette et al., 2011). A variety of studies have been done to assess the population structure of *T. albacares* using different approaches, including genetic markers (Aires-da-Silva and Maunder, 2012; Alvarado Bremer et al., 1997; Appleyard et al., 2001; Dammannagoda et al., 2008; Ely et al., 2005; Kunal et al., 2013; Ward et al., 1994; Wu et al., 2009), but its genetic structure is still controversial. The Spanish mackerel (*Scomberomorus brasiliensis,* Scombridae) (Collette et al., 1978) is a neritic tuna (Fonteles Filho, 1989) distributed from Belize to south Brazil (Carvalho-Filho, 1999) of high commercial interest in Trinidad and Tobago and Venezuela (Hodgkinson-Clarke, 1990), and in Brazil, especially in the state of Maranhão (Fonteles-Filho, 1988). Few studies have been conducted in this species (Batista and da Fabré, 2001; Maia et al., 2015), and only one aimed at evaluating its genetic structure (Gold et al., 2010). This work describes how the identification and validation of a large number of SNPs is achievable combining ddRAD and long read high throughput sequencing (*e.g.* 454 pyrosequencing) in *T. albacares* and *S. brasiliensis* for their routine genotyping on a Sequenom platform. Validated SNPs were further tested on a preliminary analysis of genetic diversity and structure in both species using two geographically separated Brazilian samples.

2. Material and methods

2.1. Biological material and DNA extraction

Samples of both species were collected in Brazil from areas separated by ~3000 km: *T. albacares* was collected in Natal (n = 50, 05°50′24.90″S, 34°59′24.34″W) and Santos (n = 24, 24°16′24.88″S, 45°49′51.79″W) and *S. brasiliensis in* Penha (n = 29, 26°47′24.14″S, 48°33′37.43″W) and Bragança (n = 37, 00°49′4.63″S, 46°30′0.40″W). A total of 11 *T. albacares and* 21 *S. brasiliensis* individuals including samples from both populations were used for *in silico* SNP discovery and subsequent validation on a Sequenom platform. Seventy-four individuals of *T. albacares* and 66 of *S. brasiliensis* were genotyped using validated SNPs for a preliminary evaluation of genetic diversity and structure of species. Representative specimens and all tissues were deposited in the fish collection of Laboratório de Biologia and Genética de Peixes of the Universidade Estadual Paulista (Botucatu, São Paulo, Brazil).

Genomic DNA was obtained from ethanol-preserved tissues. Samples were lysed in 300 µl of SSTNE extraction buffer (Blanquer, 1990) plus SDS (0.1%) and 5 µl of proteinase K (20 mg/ml) for 3 h at 55 °C. After 20 min at 70 °C, samples were treated with7.5 µl of RNAse (10 mg/ml) for 1 h at 37 °C for RNA degradation. Total DNA was purified after protein precipitation (5 M NaCl) with freezing absolute ethanol (1 ml). DNA quality (high molecular weight > 20 kb) was first evaluated on agarose gels and the DNA quantity was measured using the NanoDrop ® ND-1000 spectrophotometer (NanoDrop® Technologies Inc) and PicoGreen kit (Molecular Probes) according to the kit instructions. Finally, DNA concentration was accurately measured on a Qubit fluorometer (Life Technologies).

2.2. Library construction

A reduced fraction of the genome of the two species was sequenced using a modified ddRAD protocol (Peterson et al., 2012) in a single 454 GS Junior run (Fig. 1). Briefly, the same amount of DNA (78 ng) of each sample (11 *T. albacares* and 21 *S. brasiliensis*) was individually digested with *Sbf*I and *Sph*I restriction enzymes (RE). Subsequently adaptors for each RE were ligated to both ends of digested fragments. Adaptors included: i) complementary cohesive ends for the correspondent RE; ii) barcodes to identify individuals; and iii) primers for an intermediate PCR amplification. Barcoded samples were then pooled and run in agarose gels 1.1% for fragment selection (300–600 bp), followed by extraction using Qiagen MinElute Gel Extraction kit. After selection, the target DNA fragments were amplified by PCR to obtain the required concentration for sequencing in the 454 GS Junior platform: initial denaturation and enzyme activation at 98 °C for 30s; 14 cycles at 98 °C for 10 s (denaturation step), 65 °C for 30 s (annealing) and 72 °C for 30 s (extension); final extension at 72 °C for 5 min. The PCR products were purified using the Qiagen MinElute PCR Purification kit followed by a magnetic bead clean-up/size selection using an equal volume of Beckman Coulter AMPure XP beads. This protocol ensured that only those fragments including *SbfI* and *SphI* target sites were amplified and further sequenced.

2.3. Sequencing, assembly and SNP identification and selection

The final library was sequenced in a single shotgun run on a 454 GS Junior sequencer (Roche Diagnostics) available at the Sequencing and Functional Genomics Platform of the University of Santiago de Compostela (USC, Campus Lugo, Spain) (Fig. 1). Sequencing reads were filtered using default parameters, classified per individual according to barcodes and assembled with Newbler software (specifically designed for 454 GS series data). Alignments were then parsed with Tablet (Milne et al., 2010) in order to identify the most consistent contigs and detect SNPs in the assembled sequences. Only contigs containing a sequencing depth > 6 were retained for further analysis to reduce SNPs attributable to sequencing errors.

SNPs were selected according to the presence of enough flanking regions for primer design (± 100 bp) and the absence of other DNA polymorphism (SNPs and indels) in those regions that could interfere with primer annealing. Additionally, only those SNPs with at least three sequences of the least common allele were selected.

2.4. SNP genotyping

In silico selected SNPs were validated and genotyped using the