Methodological assessment of 2b-RAD genotyping technique for population structure inferences in yellowfin tuna (*Thunnus albacares*)

Carlo Pecoraro [a,b,*,1], Massimiliano Babbucci [c,1], Adriana Villamor [a], Rafaella Franch [c], Chiara Papetti [k], Bruno Leroy [d], Sofia Ortega-Garcia [e], Jeff Muir [f], Jay Rooker [g], Freddy Arocha [h], Hilario Murua [i], Iker Zudaire [j,b], Emmanuel Chassot [b], Nathalie Bodin [b], Fausto Tinti [a], Luca Bargelloni [c], Alessia Cariani [a]

[a] *Dept. Biological, Geological and Environmental Sciences (BIGEA), University of Bologna, Via Selmi 3, 40126 Bologna, Italy*
[b] *Institut de Recherche pour le Développement (IRD), UMR MARBEC (IRD/Ifremer/UM2/CNRS) SFA, Fishing Port, BP570 Victoria, Seychelles*
[c] *Comparative Biomedicine and Food Science, University of Padova, viale dell'Università 16, 35020 Legnaro, PD, Italy*
[d] *Secretariat of the Pacific Community, Oceanic Fishery Programme, BP D5, 98848 Noumea, New Caledonia*
[e] *Departamento de Pesquerias, Instituto Politécnico Nacional-CICIMAR, Avenida IPN s/n, La Paz, BCS, Mexico*
[f] *Pelagic Fisheries Research Program, University of Hawaii, Marine Science Building 312, Honolulu, HI 96822, USA*
[g] *Department of Marine Biology, Texas A&M University, 1001 Texas Clipper Road, Galveston, TX 77553, USA*
[h] *Instituto Oceanográfico de Venezuela, Universidad de Oriente, Avda. Universidad Cerro Colorado, Cumana 6101, Venezuela*
[i] *Marine Research Division, AZTI, Herrera Kaia-Portualdea z/g, Pasaia, 20110, Gipuzkoa, Spain*
[j] *Ikerbasque Fundazioa, Maria Diaz de Haro, 3-6°, Bilbao, 48013, Bizkaia, Spain*
[k] *Section of Integrative Ecophysiology, Alfred-Wegener-Institute for Polar and Marine Research, Am Handelshafen 12, Bremerhaven 27570, Germany*

ARTICLE INFO

Keywords:
Marine fish
Population genomics
RAD sequencing
SNP
Tropical tuna
Tuna fishery

ABSTRACT

Global population genetic structure of yellowfin tuna (*Thunnus albacares*) is still poorly understood despite its relevance for the tuna fishery industry. Low levels of genetic differentiation among oceans speak in favour of the existence of a single panmictic population worldwide of this highly migratory fish. However, recent studies indicated genetic structuring at a much smaller geographic scales than previously considered, pointing out that YFT population genetic structure has not been properly assessed so far. In this study, we demonstrated for the first time, the utility of 2b RAD genotyping technique for investigating population genetic diversity and differentiation in high gene flow species. Running de novo pipeline in *Stacks*, a total of 6772 high quality genome wide SNPs were identified across Atlantic, Indian and Pacific population samples representing all major distribution areas. Preliminary analyses showed shallow but significant population structure among oceans ($F_{ST} = 0.0273$; P value < 0.01). Discriminant Analysis of Principal Components endorsed the presence of genetically discrete yellowfin tuna populations among three oceanic pools. Although such evidence needs to be corroborated by increasing sample size, these results showed the efficiency of this genotyping technique in assessing genetic divergence in a marine fish with high dispersal potential.

1. Introduction

Yellowfin tuna (*Thunnus albacares*, YFT) has relevant biological and economic importance at the global scale, being an apex predator in oceanic ecosystem and representing the second largest tuna fishery worldwide (FIGIS, 2010 2015). Currently, YFT is managed in four distinct stocks under the jurisdiction of four independent Regional Fisheries Management Organizations (RFMOs). Although a proper fish stock management needs accurate knowledge on the stock structure and its genetic variation with respect to environmental and ecological conditions (Papetti et al., 2013), YFT genetic population structure has not been resolved yet. Different studies provided discordant patterns of YFT global scale genetic differentiation (Ward et al., 1997; Ely et al., 2005; Appleyard et al., 2001), together with a genetic structuring detected at the regional level (Dammannagoda et al., 2008; Kunal et al., 2013; Li et al., 2015). This discordance was likely due to the YFT life history traits (e.g. high fecundity, large population sizes), which make detecting patterns of genetic differentiation among population samples very difficult (Ely et al., 2005; Juan Jordá et al., 2013). Moreover,

* Corresponding author at: Dept. Biological, Geological and Environmental Sciences (BIGEA), University of Bologna, Via Selmi 3, 40126 Bologna, Italy.
E-mail addresses: carlo.pecoraro2@unibo.it (C. Pecoraro), massimiliano.babbucci@unipd.it (M. Babbucci), adriana.villamor@unibo.it (A. Villamor), rafaella.franch@unipd.it (R. Franch), chiara.papetti@awi.de (C. Papetti), brunol@spc.int (B. Leroy), sortega@ipn.mx (S. Ortega-Garcia), jmuir@hawaii.edu (J. Muir), rookerj@tamug.edu (J. Rooker), farocha@udo.edu.ve (F. Arocha), hmurua@azti.es (H. Murua), emmanuel.chassot@ird.fr (E. Chassot), nathalie.bodin@ird.fr (N. Bodin), fausto.tinti@unibo.it (F. Tinti), luca.bargelloni@unipd.it (L. Bargelloni), alessia.cariani@unibo.it (A. Cariani).
[1] These two authors contributed equally to this work.

population genetic studies reporting significant differences relied upon a relatively small number of molecular markers, hence, covering only a very limited portion of the genome (Appleyard et al., 2001; Díaz Jaimes and Uribe Alcocer, 2006). Failing to detect population structure, due to limited genetic resolution of classical markers, can potentially be misleading for management purposes, driving to local overfishing and severe stock decline (Ying et al., 2011).

According to the uncertainty about both population structure and size of YFT stocks, there is an evident need for developing alternative approaches based on genomics, that allow screening a larger number of markers across the entire genome, including neutral and non neutral loci. This might enable detecting YFT population structure, quantifying the extent of spatial demographic changes and discover imprints of local adaptation, which represent priority focus for implementing any effective management plan.

The rapid advent of next generation sequencing (NGS) based genotyping methods has significantly improved our ability to analyse thousands of Single Nucleotide Polymorphism (SNP) markers across the entire genome, increasing the precision in detecting small genetic differentiation among geographical populations (Waples et al., 2008; Allendorf et al., 2010; Davey et al., 2011; Narum et al., 2013; Andrews and Luikart, 2014). Although SNPs are characterized by a low diversity due to the only four possible allelic states, this limitation is largely outweighed by their abundance, being as frequent as one SNP every few hundred base pairs (Morin et al., 2004, 2009). Moreover, SNPs are becoming the marker of choice for many applications in popu lation ecology, evolution and conservation genetics, having a high potential for genotyping efficiency, data quality and low scoring error rates, genome wide coverage and analytical simplicity (Milano et al., 2014).

Here, for the first time, we applied the 2b RAD Genotyping By Sequencing (GBS) technique (Wang et al., 2012) for testing its potential for investigating population genetic structure in a non model, large pe lagic and highly migratory fish species. This novel genomic tool is based on sequencing reduced representation libraries produced by type IIB re striction endonucleases, which cleave genomic DNA upstream and downstream of their target site, generating tags of uniform length that are ideally suited for sequencing on existing NGS platforms (Wang et al., 2012). This method permits parallel and multiplexed sample se quencing of tag libraries for the rapid discovery of thousands of SNPs across the entire individual's genome, with a very cost effective proce dure resulting in high genome coverage. The 2b RAD method allows screening in parallel almost every restriction site in the genome, where as other GBS methods can only target a subset of total restriction sites to counterbalance loss of PCR amplification and sequencing efficiency due to large size of restriction fragments. This technique also allows fine tuning the marker density by means of selective adapters in order to se quence fewer loci with higher coverage, for applications such as popula tion genetics (Puritz et al., 2014; Andrews and Luikart, 2014). Given these attributes, the 2b RAD method has the potential to discriminate the existence of genetic differentiation with a high statistical power, generating genome wide data for genetic structure analysis at different spatial scales for YFT populations.

In this study, we: i) first examine the utility of Technical Replicates (TRs) for optimizing genotyping procedure, comparing the results obtained running the *denovo_map.pl* and the *ref_map.pl* programs in *Stacks* (Catchen et al., 2011, 2013); and ii) finally assess the applicability of 2b RAD for future investigations in this highly migratory species.

2. Results and discussion

A similar number of reads was obtained among TRs, before and after quality filtering (Table 1), which underlines the reliability of this tech nique in genotyping individuals.

Among the different *Stacks* settings considered, the *m* value was the parameter that most affected the genotyping results, in particular the number of detected SNPs. Sensitivity tests performed on the TRs showed a decrease in the number of SNPs, from 5753 to 4490, when increasing *m* from 5 to 15 (Fig. 1 and Supplementary Material 1 for values with associated Standard Error). The percentage of error rate varied approximately from 1% to 5%, with a decreasing trend when in creasing the *m* values (Fig. 1 and Supplementary Material 1). The per centage of heterozygous SNPs remained constant with increasing *m* values (Fig. 1 and Supplementary Material 1).

An increase in true heterozygous SNPs calls was observed using the *bounded SNP calling model* compared to the *default SNP model* and re ducing the upper bound values, in agreement with the results obtained by Mastretta Yanes et al. (2015). In fact, reducing the upper bound on the maximum likelihood of ε decreases the possibility of calling a ho mozygote instead of a true heterozygous genotype (Catchen et al., 2013). The proper genotype calling was further checked for a sub sample of the total reads obtained, in the *Stacks* web interface, verifying the sequences alignment and monitoring the genotyping inference when the results were exported. This procedure was repeated each time when changing the different model's *upper bound* values.

By relaxing the number of mismatches within each locus (*n*) and among loci (*M*), an increase in the number of SNPs and error rate was observed (Supplementary Material 2).

Mapping 2b RAD reads against the genome of *Thunnus orientalis* allowed a high percentage of successfully mapped sequences (86.59%). The outputs obtained on the mapped data from TRs with the *ref_map.pl* program, confirmed the trends observed with the *denovo_map.pl* pro gram (Fig. 1). However, the absolute number of SNPs was lower than that obtained with the *denovo_map.pl* program, likely due to the incom pleteness of the reference genome used (the only *Thunnus* sp. genome available to date, Nakamura et al., 2013) and the phylogenetic distance between YFT and *T. orientalis*.

Aligning reads to the reference genome, before calling a locus, can filter out erroneous stacks generated by contaminants (e.g. bacteria) possibly present in very small amount in the starting gDNA sample. Moreover the error rate also showed a less evident decreasing pattern when increasing *m*, confirming however a low error rate in the genotyping call (<5%). On the contrary, the percentage of heterozygous SNPs identified using *T. orientalis* genome as reference, showed a slight increase from 35.6% to 40.7%, when higher values of *m* were used (Fig. 1).

Table 1
Details on the technical replicates: acronym (Sample ID), Oceanic origin, genomic DNA concentration in ng/μL, library concentration in nm/μL, number of raw reads obtained, retained reads after quality filtering, and their corresponding percentage.

Sample ID	Oceanic origin	gDNA ng/μL	Library nm/μL	No. raw reads	No. filtered reads	% retained reads
34_2_Y_2R1	Atlantic Ocean	333.80	185.38	2,276,239	1,772,927	78%
34_2_Y_2R2	Atlantic Ocean	333.80	207.83	2,672,917	1,914,181	72%
34_2_Y_2R3	Atlantic Ocean	333.80	197.03	2,309,039	1,805,181	78%
77_2_Y_15R1	Pacific Ocean	218.02	238.81	2,212,559	1,788,988	81%
77_2_Y_15R2	Pacific Ocean	218.02	240.53	2,292,834	1,850,871	81%
77_2_Y_15R3	Pacific Ocean	218.02	208.91	2,085,658	1,802,820	86%
51_1_Y_7R1	Indian Ocean	177.54	166.05	2,330,292	1,767,345	76%
51_1_Y_7R2	Indian Ocean	177.54	170.81	2,300,342	1,824,635	79%

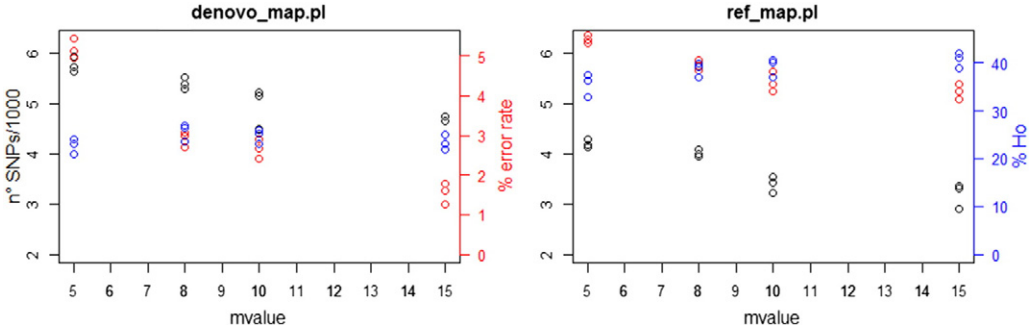

Fig. 1. Comparison between *denovo_map.pl* (left panel) and *ref_map.pl* (right panel) performance in terms of (a) number of SNPs (black dots) (b) error rate (red dots) and (c) percentage of heterozygous loci (blue dots), using different m values. Each dot represents the average value among the three individual's TRs. The three y axes (No. SNPs/1000, % error rate, % Ho) are shared between the two plots.

Based on the number of SNPs identified with the two different approaches (with or without using the reference genome), the low error rates and the consistent percentage of heterozygous SNPs obtained among TRs, the *denovo_map.pl* program was run applying the following parameter settings $m = 8$, $M = 3$, $n = 2$ and a bounded *SNP calling model* with an upper bound of 0.1 (all remaining *Stacks* settings as default), to obtain the final dataset (see Supplementary Material 3 for details about each individual). The AMOVA results (Supplementary Material 4) revealed that pooling samples into the three major oceanic regions produced the highest percentage of variation explained by groups subdivision (2.73% P value < 0.001) and at the same time very low and not significant differences were observed among populations within groups (0.97% P value > 0.01). Pooling YFT individuals in these three groups corresponding to the three oceans (Table 2), allowed to increase the sample size and to obtain more robust and reliable inferences of population structure.

The optimization process combining different r and p parameters' values of the *Stacks* population module, led to $r = 0.7$ and $p = 6$ as best middle way between the number of SNPs and the percentage of missing values obtained (Table 3).

This set of parameters produced a panel of 6772 SNPs. Pairwise F_{st} distances, calculated with this dataset, were highly significant, suggest ing genetic differences occurring among oceanic groups (Table 4).

The DAPC confirmed the genetic differentiation among oceanic basins. The graph of the Bayesian Information Criterion (BIC) values for increasing values of the number of clusters (k) showed that k = 3 corresponded to the lowest associated BIC value. In the data transformation step for PCA analysis, 30 principal components (PCs) were retained, accounting for approximately the 92% of the total genetic variability. The eigenvalues of the DAPC indicated that the first two components explained most of the variation. The resulting scatterplot (Fig. 2) showed three genetic clusters corresponding to Atlantic, Indian and Pa cific YFT groups. Moreover, the cross validation of DAPC performed on our dataset, indicated that the number of PCs associated to the highest mean success and lowest mean squared error corresponded to 35 PCs. This result supported our choice to retain 30 PCs during the dimen sion reduction step of the DAPC.

These results are in agreement with previously observed signatures of genetic heterogeneity among oceans found by Ward et al. (1997) by means of significant allele frequency differences at the locus GPI A (PGI F). Although, this scenario necessarily needs to be confirmed by in creasing the sample size, it validated 2b RAD genotyping technique as a powerful tool to assess YFT genetic structure and diversity at the global scale.

3. Conclusions

This methodological study confirmed that TRs are useful for optimiz ing genotyping procedure and that they are crucial to reduce the amount of statistical error introduced in allele frequency estimation due to PCR artefacts. We unambiguously mapped the TRs' tags against the reference genome of *T. orientalis* with a high percentage of success (86.59%), in spite of the small size of fragments (Puritz et al., 2014), and the evolutionary distance between these two species. The method ological approach showed that the lack of a reference genome, although undesirable, does not evidently compromise the reproducibility and ac curacy of the data obtained, underlying the consistence of the technique in genotyping individuals. We preliminarily demonstrated that 2b RAD is a promising tool to screen a large set of genomic loci in a marine high gene flow species, underlying the inter oceanic population genetic differentiation. Certainly, an increased sample size is needed to address estimates of genetic differentiation among YFT population samples also at a smaller local geographic scale.

4. Materials and methods

4.1. Sampling design, libraries preparation and sequencing

A total of 100 juvenile YFT (35 55 cm of fork length, FL) from Atlantic, Indian and Pacific geographic population samples (Table 5) were analysed, covering the entire species distribution (Fig. 3).

Genomic DNA (gDNA) was extracted from approximately 20 mg of tissue (skeletal muscle or finclip) using the commercial kit Invisorb® Spin Tissue Mini Kit (Invitek, STRATEC Biomedical, Germany) following the manufacturers' recommendations. Since high quality gDNA is re quired in the 2b RAD genotyping technique, its concentration and puri ty, in terms of ratios of absorbance at 260/230 nm and at 260/280 nm, were quantified by both a NanoDrop ND 1000 spectrophotometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and a Qubit 2.0 Fluorometer (Invitrogen, Thermo Fisher Scientific, Waltham,

Table 2
Summary statistics of the three *Thunnus albacares* oceanic groups. The table reports: the sampling origin (location) the sample size (N° individuals), the mean number (millions) of raw reads with the associated standard error (SE), the corresponding mean number (millions) of filtered reads (with SE), the percentage of reads retained, the mean value of unique tags, poly-morphic SNPs and number of SNPs found.

Location	No. individuals	Raw reads (mln)	Filtered reads (mln)	% of reads retained	Unique tags	Polymorphic SNPs	SNPs found
Atlantic Ocean	40	3.80 (±0.29)	3.09 (±0.37)	80%	30,776	3264	5693
Indian Ocean	20	3.98 (±0.46)	2.91 (±0.15)	79%	31,430	3695	6516
Pacific Ocean	40	3.30 (±0.26)	2.78 (±0.24)	84%	31,573	3363	5906

Table 3
Number of SNPs and percentage of missing value (NA %) obtained for the entire dataset according to the *r* and *p* parameters' values of the *Stacks* population programme.

		p					
		4		6		9	
		SNPs	NA %	SNPs	NA %	SNPs	NA %
r	0.4	8158	14.3	7871	11.49	7560	9.33
	0.7	7049	9.33	6772	6.3	6430	4.69
	0.9	5981	8.62	5673	5.44	5187	2.58

Massachusetts, USA). This procedure ensured to work with high quality samples and comparable DNA concentration.

The 2b RAD libraries were constructed for each individual following the protocol from Wang et al. (2012) with minor modifications (see below). To assess the robustness of the method and subsequent data analyses, three libraries were replicated (Technical Replicates, TRs) for two individuals (34_2_Y_2 and 77_2_Y_15) and two for an additional third specimen (51_1_Y_7). gDNA (300 ng) was digested with 2 U of the enzyme *Csp*CI (New England Biolabs, NEB, Ipswich, Massachusetts, USA) for 1 h at 37 °C. The digested DNA was ligated in a 25 μL total vol - ume reaction consisting of 0.4 μM for each of the two library specific adaptors, 0.2 mM ATP (New England Biolabs, NEB, Ipswich, Massachu- setts, USA) and 1 U T4 DNA ligase (SibEnzyme Ltd., Academ town, Siberia). To reduce marker density, one adaptor with fully degenerate 3′ overhangs NN and one with reduced 3′degeneracy NG were chosen. Sample specific barcodes were designed with Barcode Generator (http://comailab.genomecenter.ucdavis.edu/index.php/Barcode_ generator) and introduced by PCR with platform specific barcode bearing primers. 2b RAD tags were amplified by PCR in two separate 25 μL reactions, in order to minimize PCR amplification bias (Mastretta Yanes et al., 2015). Each amplification consisted of 6.25 μL of ligated DNA, 0.5 μM each primer (P4 and P6 BC, Eurofins Genomics S.r.l, Italy), 0.2 μM each primer (P5 and P7, Eurofins Genomics), 0.3 mM dNTP (New England Biolabs, NEB, Ipswich, Massachusetts, USA), 1X Phusion HF buffer and 1 U TaqPhusion high fidelity DNA poly merase (NEB). Cycling conditions were: 98 °C for 4 min; 98 °C for 5 s, 60 °C for 20 s, 72 °C for 5 s for 14 cycles, 72 °C for 5 min. The reduced num ber of amplification cycles (n = 14) is crucial to produce a negligible amount of PCR amplification errors, comparing to those needed to reach the plateau phase.

PCR products were purified with the SPRIselect purification kit (Beckman Coulter, Pasadena, California, USA), to exclude any high molecular weight DNA remaining after the enzyme digestion and any incorrect constructs that may emerge during PCR amplification. The concentration of purified individual libraries was quantified using Qubit®ds DNA BR Assay Kit (Invitrogen ThermoFisher Scientific, MA, USA) and Mx3000P qPCR instrument, and the quality checked on an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, California, USA). Individual libraries were pooled into equimolar amounts and resulting pools' quality was re verified on Agilent 2100 Bioanalyzer. Pooled libraries were sequenced on an Illumina HiSeq2500 platform with a 50 bp single read module at the Genomix4Life S.r.l. facilities (Baronissi, Salerno, Italy), which also performed data demultiplexing.

Table 4
Pairwise F_{st} values calculated among geographic pools (from Atlantic, Indian and Pacific Ocean) of yellowfin tuna are reported (below diagonal) with their associated P-values (below diagonal). Significant values after Bonferroni standard correction are in bold (nominal significant threshold $\alpha = 0.01$).

	Atlantic	Indian	Pacific
Atlantic	*	<**0.01**	<**0.01**
Indian	0.04736	*	<**0.01**
Pacific	0.02932	0.01714	*

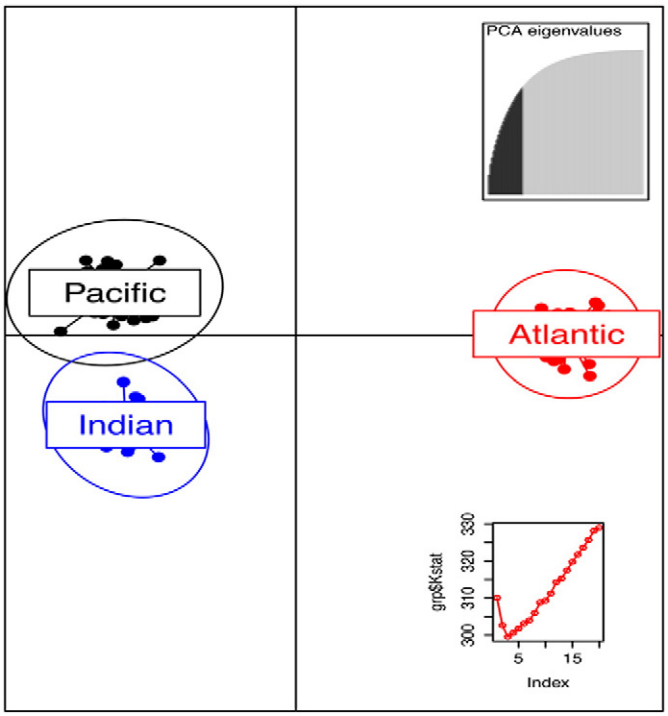

Fig. 2. Scatterplot of the DAPC results identifying three genetic clusters of *Thunnus albacares*.

4.2. Technical replicates analysis and optimization of genotyping procedure

Demultiplexed reads were returned by the sequencing facility in Fastq format and their quality was checked by FastQC (www. bioinformatics.babraham.ac.uk/projects/fastqc/). After this, a custom made Perl script was run for quality filtering and adaptors trimming of the reads, obtaining sequences of 34 bp (Fastq files available at SRA Bioproject: SRP067271). Filtered reads were analysed with the software *Stacks* v. 1.32 (Catchen et al., 2011, 2013), which allows genotype infer ence through the identification of SNP loci without a reference genome (*denovo_map.pl* program) or aligning reads against a reference genome (*ref_map.pl* program). Different settings were tested on the TRs dataset to fine tune the de novo *Stacks* pipeline parameters and to assess the consistency of results, in terms of total number of identified SNPs; the error rate calculated counting discordant genotypes between TRs and, among the concordant data, the percentage of heterozygous SNPs. Following the *Stacks* author guidelines, multiple combinations were con sidered for: a) the minimum number of identical reads necessary to call an allele (*m* value set to: 5, 8, 10, 15); b) mismatches between reads within a locus (*M* value set to: 2, 3, 4, 6); and c) mismatches among loci when comparing across individuals (*n* value set to: 0, 2, 3, 4, 6). Only one parameter was varied at a time while keeping the

Table 5
The table summarizes the sampling location, sample code and number of individual per each geographic population sample.

Sampling location	Sample code	Number of individuals
WC Atlantic Ocean	31_1	10
EC Atlantic Ocean	34_1	10
EC Atlantic Ocean	34_2	10
WC Atlantic Ocean	41_1	10
WC Indian Ocean	51_1	10
WC Indian Ocean	51_2	10
WC Pacific Ocean	71_1	10
WC Pacific Ocean	71_2	10
EC Pacific Ocean	77_1	10
EC Pacific Ocean	77_2	10

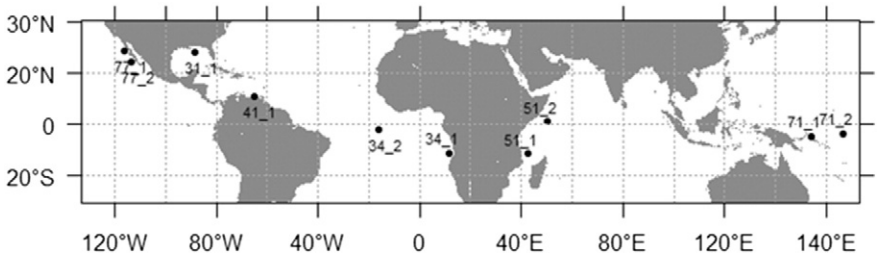

Fig. 3. Location of *Thunnus albacares* geographical population samples analysed in this study. Sample codes are given as in Table 5.

others fixed. The default values for *min_het_seqs* and *max_het_seqs* were used. In addition, we compared the default and the bounded SNP calling models (−*bound_high* value set to: 0.05, 0.1, 0.15, 0.2, 0. 5) to evaluate the percentage of heterozygous genotypes correctly assessed, in order to make the genotype calling between them as much concordant as pos sible. In the bounded SNP calling model, *Stacks* employs a multinomial based likelihood model to identify SNPs and to estimate the maximum likelihood value of the sequencing error rate ε at each nucleotide posi tion, in order to proper call each possible genotype (for details see Catchen et al., 2011, 2013).

The reads were also mapped against the genome of *T. orientalis* (GenBank accession numbers BADN01000001 BADN01133062; Nakamura et al., 2013) using CLC Genomics Workbench v. 5.1 (CLC Bio) programme. The following parameters settings were applied *length fraction* = 1.0 and *similarity fraction* = 0.9 (all remaining parameters as default), retaining only uniquely mapped reads. Mapping results were exported in SAM format and were used as input files for *refmap_map.pl* in *Stacks*. To further evaluate the robustness of the approach a similar testing was performed on mapped data, using the same settings as for the *denovo_map.pl* for *m*, *n* and −*bound_high* of the bounded SNP calling model (Fig. 2).

4.3. Preliminary analysis of YFT population structure

Once identified the *Stacks* parameter set which minimized differences among TRs, the *denovo_map.pl* program was run on the entire YFT dataset (Table 2). Using the programme *populations* available from *Stacks* software, different combinations of *p* (4, 6, 9) and *r* (0.4, 0.7, 0.9) parameters were tested, in order to investigate changes in the number of SNPs obtained, and in the percentage of missing values among samples. Following these tests, we selected from the resulting catalogue of loci only those containing one bi allelic SNP (−*F snps_l* = 1 *snps_u* = 2), and those values of *p* and *r* rendering the highest number of SNPs with the lowest percentage of missing data.

In order to increase the sample size and to improve robustness of the genetic analyses, several grouping of the geographic samples were test ed, especially due to their ocean basin distance, performing an analysis of molecular variance with the software Arlequin 3.5.1.2 (Excoffier and Lischer, 2010) with 10,000 permutations and P ≤ 0.01 significance level.

Based on the SNPs dataset and AMOVA results obtained, F_{ST} estimates for pairwise comparison among pooled samples, were calculated with the software Arlequin 3.5.1.2 using the same settings as above.

A preliminary assessment of YFT genetic structure was performed using the Discriminant Analysis of Principal Components (DAPC, Jombart et al., 2010) implemented in the R package Adegenet (Jombart et al., 2008, R version 3.1.2, R Development Core Team, 2014; http://www.r project.org). The function *find.clusters* was used to identify the optimal number of clusters (k) that maximizes the variation between groups (Jombart et al., 2010). The cross validation test was also carried out in order to validate the number of principal components (PCs) retained in the first transformation step of DAPC analysis, because a wrong choice of the number of PCs might negatively impact the DAPC results and produce unstable output due to over parameterization.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.margen.2015.12.002.

Acknowledgements

This study is part of the PhD research programme of Carlo Pecoraro performed at the University of Bologna and the Institut de Recherche pour le Développement (IRD) with the financial support of the Bolton Group and the Associazione Nazionale Conservieri Ittici e delle Tonnare. We are also grateful for the great contribution in the sampling activities received from the EMOTION project (ANR JSV7 007 01), the Internation al Seafood Sustainability Foundation, the Seychelles Fishing Authority, the Centre de Recherches Océanologiques in Abidjan and the IRD Observatoire Thonier.

References

Allendorf, F.W., Hohenlohe, P.A., Gordon, L., 2010. Genomics and the future of conservation genetics. Nat. Rev. Genet. 11, 697–709. http://dx.doi.org/10.1038/nrg2844. Andrews, K.R., Luikart, G., 2014. Recent novel approaches for population genomics data analysis. Mol. Ecol. 23, 1661–1667. http://dx.doi.org/10.1111/mec.12686.

Appleyard, S., Grewe, P., Innes, B., Ward, R., 2001. Population structure of yellowfin tuna (*Thunnus albacares*) in the Western Pacific Ocean, inferred from microsatellite loci. Mar. Biol. 139, 383–393. http://dx.doi.org/10.1007/s002270100578.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., Postlethwait, J.H., 2011. Stacks: building and genotyping loci de novo from short-read sequences. G3 Genes Genomes Genet. 1, 171–182. http://dx.doi.org/10.1534/g3.111.000240.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., Cresko, W.A., 2013. Stacks: an analysis tool set for population genomics. Mol. Ecol. 22, 3124–3140. http://dx.doi.org/10.1111/mec.

Dammannagoda, S.T., Hurwood, D.A., Mather, P.B., 2008. Evidence for fine geographical scale heterogeneity in gene frequencies in yellowfin tuna (*Thunnus albacares*) from the North Indian Ocean around Sri Lanka. Fish. Res. 90, 147–157. http://dx.doi.org/10.1016/j.fishres.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., Blaxter, M.L., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12, 499–510. http://dx.doi.org/10.1038/nrg3012.

Díaz-Jaimes, P., Uribe-Alcocer, M., 2006. Spatial differentiation in the eastern Pacific yellowfin tuna revealed by microsatellite variation. Fish. Sci. 72, 590–596. http://dx.doi.org/10.1111/j.1444-2906.2006.01188.x.

Ely, B., Viñas, J., Alvarado Bremer, J.R., Black, D., Lucas, L., Covello, K., Labrie, A.V., Thelen, E., 2005. Consequences of the historical demography on the global population structure of two highly migratory cosmopolitan marine fishes: the yellowfin tuna (*Thunnus albacares*) and the skipjack tuna (*Katsuwonus pelamis*). BMC Evol. Biol. 5, 19. http://dx.doi.org/10.1186/1471-2148-5-19.

Excoffier, L., Lischer, H.E., 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. 10, 564–567.

FAO, 2010-2015. Fisheries Global Information System (FAO-FIGIS) — Web site. Fisheries Global Information System (FIGIS). FI Institutional Websites. FAO Fisheries and Aquaculture Department [online] (Rome. Updated. http://www.fao.org/fishery/figis/en).

Jombart, T., Devillard, S., Dufour, A.B., Pontier, D., 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity 101, 92–103. http://dx.doi.org/10.1038/hdy.2008.34.

Jombart, T., Devillard, S., Balloux, F., 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11, 94. http://dx.doi.org/10.1186/1471-2156-11-94.

Juan-Jordá, M.J., Mosqueira, I., Freire, J., Dulvy, N.K., 2013. Life in 3-d: life history strategies in tunas, mackerels and bonitos. Rev. Fish Biol. Fish. 23, 135–155. http://dx.doi.org/10.1007/s11160-012-9284-4.

Kunal, S.P., Kumar, G., Menezes, M.R., Meena, R.M., 2013. Mitochondrial DNA analysis reveals three stocks of yellowfin tuna *Thunnus albacares* (Bonnaterre, 1788) in Indian Waters. Conserv. Genet. 14, 205–213. http://dx.doi.org/10.1007/s10592-013-0445-3.

Li, W., Xinjun, C., Qianghua, X., Jiangfeng, Z., Xiaojie, D., Liuxiong, X., 2015. Genetic population structure of *Thunnus albacares* in the Central Pacific Ocean based on mtDNA

COI gene sequences. Biochem. Genet. 53, 8–22. http://dx.doi.org/10.1007/s10528-015-9666-0.

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T.H., Piñero, D., Emerson, B.C., 2015. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. Mol. Ecol. Resour. 15, 28–41. http://dx.doi.org/10.1111/1755-0998.12291.

Milano, I., Babbucci, M., Cariani, A., Atanassova, M., Bekkevold, D., Carvalho, G.R., Espiñeira, M., Fiorentino, F., Garofalo, G., Geffen, A.J., Hansen Jakob, H., Helyar, S.J., Nielsen, E.E., Ogden, R., Patarnello, T., Stagioni, M., FishPopTrace Consortium, Tinti, F., Bargelloni, L., 2014. Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (*Merluccius merluccius*). Mol. Ecol. 23, 118–135. http://dx.doi.org/10. 1111/mec.12568.

Morin, P.A., Luikart, G., Wayne, R.K., the SNP workshop Group, 2004. SNPs in ecology, evolution and conservation. Trends Ecol. Evol. 19, 208–216. http://dx.doi.org/10.1016/j.tree.2004.01.009.

Morin, P.A., Martien, K.K., Taylor, L.B., 2009. Assessing statistical power of SNPs for population structure and conservation studies. Mol. Ecol. Resour. 9, 66–73. http://dx.doi. org/10.1111/j.1755-0998.2008.02392.x.

Nakamura, Y., Mori, K., Saitoh, K., et al., 2013. Evolutionary changes of multiple visual pigment genes in the complete genome of Pacific bluefin tuna. Proc. Natl. Acad. Sci. U. S. A. 110, 11061–11066. http://dx.doi.org/10.1073/pnas.1302051110.

Narum, S.R., Buerkle, C.A., Davey, J.W., Miller, M.R., Hohenlohe, P.A., 2013. Genotyping-by-Sequencing in ecological and conservation genomics. Mol. Ecol. 22, 2841–2847. http://dx.doi.org/10.1111/mec.12350.

Papetti, C., Di Franco, A., Zane, L., Guidetti, P., De Simone, V., Spizzotin, M., Zorica, B., Kec,V.C., Mazzoldi, C., 2013. Single population and common natal origin for Adriatic *Scomber scombrus* stocks: evidence from an integrated approach. ICES J. Mar. Sci. J. Cons. 70, 387–398. http://dx.doi.org/10.1093/icesjms/fss201.

Puritz, J.B., Matz, M.V., Toonen, R.J., Weber, J.N., Bolnick, D.I., Bird, C.E., 2014. Demystifying the RAD fad. Mol. Ecol. 23, 5937–5942. http://dx.doi.org/10.1111/mec.12965.

Wang, S., Meyer, E., McKay, J.K., Matz, M.V., 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. Nat. Methods 9, 808–810. http://dx.doi.org/10.1038/nmeth.2023.

Waples, R.S., Punt, A.E., Cope, J.M., 2008. Integrating genetic data into management of marine resources: how can we do it better? Fish Fish. 9, 423–449. http://dx.doi.org/10. 1111/j.1467-2979.2008.00303.x.

Ward, R.D., Elliot, N.G., Innes, B.H., Smolenski, A.J., Grewe, P.M., 1997. Global population structure of yellowfin tuna, *Thunnus albacares*, inferred from allozyme and mitochondrial DNA variation. Fish. Bull. 95, 566–575. http://dx.doi.org/10.1007/bf00347499.

Ying, Y., Chen, Y., Lin, L., Gao, T., Quinn, T., 2011. Risks of ignoring fish population spatial structure in fisheries management. Can. J. Fish. Aquat. Sci. 68, 2101–2120. http://dx.doi.org/10.1139/F2011-116.