Received Date : 12-Jul-2015

Revised Date : 31-Jan-2016

Accepted Date : 26-Feb-2016

Article type : Resource Article

Population structure of Atlantic Mackerel inferred from RAD-seq derived SNP markers: effects of sequence clustering parameters and hierarchical SNP selection

Naiara Rodríguez-Ezpeleta[1]*, Ian R. Bradbury[2], Iñaki Mendibil[1], Paula Álvarez[1], Unai Cotano[1], Xabier Irigoien[3]

[1]AZTI, Marine Research Division, Sukarrieta, Bizkaia, Spain.

[2] Department of Fisheries and Oceans, St. John's, Newfoundland, Canada

[3] Red Sea Research Center, King Abdullah University of Technology, Saudi Arabia

Keywords: RAD-seq, SNP, *Stacks*, read merging parameters, population structure, Atlantic mackerel

*Corresponding author:

Naiara Rodriguez-Ezpeleta, AZTI, Marine Research Division, Txatxarramend ugartea z/g, Sukarrieta, 48395, Bizkaia, Spain; nrodriguez@azti.es

Running title: RAD-seq and mackerel population structure

Abstract

Restriction-site associated DNA sequencing (RAD-seq) and related methods are revolutionizing the field of population genomics in non-model organisms as they allow generating an unprecedented number of single nucleotide polymorphisms (SNPs) even when no genomic information is available. Yet, RAD-seq data analyses rely on assumptions on nature and number of nucleotide variants present in a single locus, the choice of which may lead to an under- or overestimated number of SNPs and/or to incorrectly called genotypes. Using the Atlantic mackerel (*Scomber scombrus L.*) and a close relative, the Atlantic chub mackerel (*Scomber colias*), as case study, here we explore the sensitivity of population structure inferences to two crucial aspects in RAD-seq data analysis: the maximum number of mismatches allowed to merge reads into a locus and the relatedness of the individuals used for genotype calling and SNP selection. Our study resolves the population structure of the Atlantic mackerel, but, most importantly, provides insights into the effects of alternative RAD-seq data analysis strategies on population structure inferences that are directly applicable to other species.

Introduction

Inferring the degree of genetic exchange between populations of marine fish species is key to successfully managing exploited populations, allowing the identification of conservation units, assignment of individuals to geographic regions, and detection of product mislabeling and fraud (Dichmont *et al.* 2012; Funk *et al.* 2012; Nielsen *et al.* 2012). Yet, many exploited marine fish are characterized by little intraspecific genetic structuring even over large geographical distances (Bradbury *et al.* 2008; Ward *et al.* 1994) challenging the resolution of populations and the assignment of individuals. Increasingly, the resolution of genetic differentiation is possible with a large number of genome wide polymorphic markers (Benestan *et al.* 2015; Lamichhaney *et al.* 2012; Narum *et al.* 2013), which not only allow inference of neutral population structure, but can also provide information on local adaptation or speciation events (Allendorf *et al.* 2010). Recently, advances in sequencing technologies have allowed generating large numbers of molecular markers at an unprecedented cost and speed, even for organisms for which no genomic information was previously available. The most popular of these methods combine restriction enzyme digestion of the genome with high throughput sequencing, and they are particularly relevant for non-model organisms as they allow discovering and genotyping thousands of single nucleotide polymorphisms (SNPs) in

hundreds of individuals rapidly and at low cost regardless of size of genome and prior genomic knowledge (Baird *et al.* 2008; Davey *et al.* 2011). Consequently, restriction site associated DNA sequencing (RAD-seq) and related approaches are increasingly used to identify and genotype genome-wide markers in non-model marine species to directly inform conservation and management efforts (*e.g.* Corander *et al.* 2013; Hess *et al.* 2013; Larson *et al*. 2014; Puebla *et al.* 2014).

Several approaches have been developed to identify and genotype SNPs from RAD-seq derived sequences, such as *Stacks* (Catchen *et al.* 2013; Catchen *et al.* 2011), *pyRAD* (Eaton 2014), *rainbow* (Chong *et al.* 2012) and *RADtools* (Baxter *et al.* 2011), among others. Most of them rely on merging reads given arbitrary maximum nucleotide distances to identifying putative orthologous loci from which genotypes of each individual for each identified SNP are determined. For example, when no reference genome is available, *Stacks* relies on two main mismatch parameters: M defines the maximum number of mismatches allowed between reads within the same individual to form an individual locus, and n defines the maximum number of mismatches allowed between loci of different individuals to form a catalog locus, where the catalog includes all individuals used in the analysis. In the absence of a reference genome, selecting the optimal values of M and n is impossible, as they depend upon the degree of polymorphism of the genome (and particular regions of the genome) being analyzed, the degree of paralogy within the genome, the amount of sequencing error and the depth of the sequencing performed (Catchen *et al.* 2013; Mastretta-Yanes *et al.* 2014). Yet, allowing a too small number of mismatches can lead to an overestimation of the number of loci (over-splitting), whereas allowing a too large number of mismatches may underestimate the number of loci (under-splitting). The effect of under- or overestimating the number of loci has been assessed on the number of incorrectly merged reads (Catchen *et al.* 2013), genotyping error rate (Mastretta-Yanes *et al.* 2014) and population structure (Puebla *et al.* 2014; Ravinet *et al.* 2016). Over-splitting may reduce genetic distances among individuals or between populations if they are characterized divergent alleles and previous analysis suggests in comparison the under-splitting, over-splitting seems to have the potential for greater impact on estimates of population structure (Harvey *et al.* 2015).

The Atlantic mackerel (*Scomber scombrus L.*) is a commercially important migratory pelagic fish that forms large shoals that can reach millions of individuals (Lockwood 1988). This species has traditionally been grouped in five spawning components (two in the Northwest

and three in the Northeast Atlantic) that migrate north during the feeding season (Trenkel *et al.* 2014). Inferring the degree of mixing between spawning components of Altantic mackerel is crucial to define management units and will, moreover, help assignment of the population of origin of the mackerel fishery recently established in Iceland, where no significant amounts were fished before (Hannesson 2013). Tagging experiments have demonstrated high dispersal rates (Lockwood 1988; Uriarte & Lucio 2001) suggesting considerable levels of potential gene flow among populations; yet, no mackerel tagged in the Northeast Atlantic has been recaptured in the western Atlantic (Tenningen *et al.* 2011), and population genetic analyses based on mitochondrial markers suggest limited trans-Atlantic gene flow. Within each side of the Atlantic, the separation among spawning components is less clear (Jansen & Gislason 2013). Within the Mediterranean, Zardoya et al. (2004), using the mitochondrial control region, showed differentiation between the Eastern and Western Mediterranean populations, the latter mixing with southern Northeast Atlantic mackerel. In sum, none of these studies produced conclusive results on the population structure of Atlantic mackerel, probably because they were based on one or a few markers. Yet, existing genetic resources for this species are limited to mitochondrial DNA (Nesbo *et al.* 2000; Zardoya *et al.* 2004) and microsatellite markers (Olafsdottir *et al.* 2012), which have largely been unsuccessful in resolving spatial structure.

Here we generate a novel RAD-seq derived SNP dataset for Atlantic mackerel to provide a description of the population structure in this commercially important and spatially expanding species, and to explore the sensitivity of inferences of spatial population structuring to both, sequence clustering parameters, that is, mismatch thresholds to select orthologous loci, and hierarchical SNP selection, that is, including from less to more distant individuals to identify SNPs. For that aim, we have generated RAD sequencing data of 122 Atlantic mackerel individuals that span the three geographic areas where this species inhabits: the Northwest Atlantic, the Northeast Atlantic, and the Mediterranean Sea, and of 15 individuals of a closely related species, the Atlantic chub mackerel (*Scomber colias*). We have examined (i) how parameter choice during RAD-seq data analysis influences the number of SNPs identified (i.e. over-splitting or under-splitting) and the spatial structure observed and (ii) how the analysis is influenced by hierarchical SNP selection, ranging from adjacent populations to the inclusion of a different species. Our study resolves the global population structure of Atlantic mackerel and provides insights into the effects of alternative RAD-seq data analysis strategies on population structure inferences. Our conclusions are

directly applicable to RAD-seq based population structure analyses of other species with low intraspecific genetic differentiation. We predict that the likelihood of over-splitting impacting the resolution of population structure will increase with the level of divergence among populations (or other groups) compared. We would therefore expect to see the impact of over-splitting (i.e. increases in F_{IS} and reductions in F_{ST} or detectable population structure) at high levels of stringency and perhaps at even moderate stringency in the multiple species analysis. The inclusion of hierarchical levels of structure within the dataset considered here offers unprecedented opportunity to explore the impact of stringency on both levels of diversity, and population structure detected.

Materials and Methods

Tissue sampling

Adult *S. colias* from the Gulf of Cádiz and adult *S. scombrus* from the Gulf of Saint Lawrence (Canada), the Bay of Biscay, the Adriatic sea, the Tyrrhenian sea and the western Mediterranean sea were obtained from scientific surveys and commercial fisheries. Sampling locations and number of samples per location are shown in Fig 1. From each fish, a ~1cm^3 muscle tissue sample was excised and immediately stored in 96% molecular grade ethanol at -20ºC until DNA extraction.

DNA extraction and RAD-seq library preparation

Genomic DNA was extracted from about 20 mg of muscle tissue using the Wizard® Genomic DNA Purification kit (Promega, WI, USA) following manufacturer's instructions for "Isolating Genomic DNA from Tissue Culture Cells and Animal Tissue". Extracted DNA was suspended in Milli-Q water and concentration was determined with the Quant-iT dsDNA HS assay kit using a Qubit® 2.0 Fluorometer (Life Technologies). DNA integrity was assessed by electrophoresis, migrating about 100 ng of GelRed™-stained DNA on an agarose 1.0% gel. Restriction-site-associated DNA libraries were prepared following the methods of Etter et al. (2011). Briefly, starting DNA (ranging from 500 to 750ng, depending on integrity) was digested with the *SbfI* restriction enzyme and ligated to modified Illumina P1 adapters

containing 5bp unique barcodes. Pools of 33 individuals were sheared using the Covaris® M220 Focused-ultrasonicator™ Instrument (Life Technologies) and size selected to 300-500 bp by cutting agarose migrated DNA. After Illumina P2 adaptor ligation, each library was amplified using 14 PCR cycles. Each pool was paired-end sequenced (100 bp) on an Illumina HiSeq2000.

RAD-tag data analysis

Generated RAD-tags were analyzed using *Stacks* version 0.9999 (Catchen *et al.* 2013) – note that although newer versions of stacks have been released (http://catchenlab.life.illinois.edu/stacks/), the improvements introduced newer Stacks versions do not affect the calculations performed for this study (see Supplementary Material). Quality filtering and demultiplexing was performed with the *process_radtags* module with default parameters. Only individuals with a > 900,000 retained reads were kept. Putative orthologous tags (stacks) per individual were assembled using *ustacks* with a minimum depth of coverage required to create a stack (m) of 5 and a maximum nucleotide mismatches (M) allowed between stacks of 2 or 4. Catalogs of loci were assembled based on three nested subsets of individuals (all samples of both species, only *S. scombrus* or only Mediterranean *S. scombrus*) using *cstacks;* the number of mismatches allowed between sample tags when generating the catalog (n) was 3 or 6. Matches of individual RAD loci to the catalog were searched using *sstacks*. From each generated catalog (using all, only *S. scombrus* or only Mediterranean *S. scombrus* samples), SNPs present in RAD loci found in at least 75% of the individuals under study (all, only *S. scombrus* or only Mediterranean *S. scombrus*) were selected and exported into *PLINK* format using *populations*. Using *PLINK* version 1.07 (Purcell *et al.* 2007), SNPs with a minimum allele frequency (MAF) smaller than 0.05, a genotyping rate smaller than 0.01 and which failed the Hardy Weinberg equilibrium (HWE) test at $p < 0.05$ in at least two populations were excluded for further analyses. Each genotype dataset was exported to *Structure*, *Bayescan* and *Genepop* formats using *PGDSpider* version 2.0.5.2 (Lischer & Excoffier 2012). In total, 24 genotype subsets where created including 6 catalog/SNP selection combinations and 4 stacks parameter (M=2 or 4 and n=3 or 6) combinations.

Genetic diversity and population genetic analyses

F_{IS} per population and F_{ST} per pair of populations were calculated on each genotype dataset following the Weir & Cockerham (1984) formulation as implemented in *Genpop 4.3* (Rousset 2008). F_{ST} outliers were identified using Bayescan with default parameters and a false discovery rate of 0.05 (Foll & Gaggiotti 2008). Principal component analyses (PCA) were performed with the R package *adegenet* (Jombart & Ahmed 2011) without any a priori population definition. Differences among pairs of groups within the PCA where quantified as the average size of group 1 and group 2 divided by the size of the ellipse including individuals from groups 1 and 2. For each genotype dataset, 10 subdatasets of 5,000 randomly chosen SNPs were created and analyzed with the Bayesian clustering approach implemented in *STRUCTURE* (Pritchard *et al.* 2000). For each value of K (number of potential ancestral populations, which ranged from 1 to the number of presumed populations + 1), the genetic ancestry of each individual was estimated based on the admixture model without any prior population assignment; estimations were obtained from the 300,000 iterations that followed a burn-in period of 100,000 iterations. The 10 subdatasets obtained for each value of K were analyzed with *CLUMPP* (Jakobsson & Rosenberg 2007) to identify common modes, and results were plotted using *DISTRUCT* (Rosenberg 2004). Best K was identified according to the Evanno method (Evanno *et al.* 2005) as implemented in *StructureHarvester* (Earl & vonHoldt 2012).

Results

RAD-tag processing and SNP discovery and genotyping

The number of quality filtered RAD tags obtained per individual ranges from 909,095 to 11,804,229 with an average of 2,922,343 (Fig. 2). Nearly 98% of the reads (ranging from 92% to 99% per individual) where used for stack formation. The mean depth coverage is 50x.

Expectedly, the number of RAD loci obtained with increasing the mismatch parameter (M) decreases, as more stacks can be merged into a single locus. The average number of RAD loci per individual, which is not proportional to the number of quality filtered reads, is 56,464 (M=2) or 55,118 (M=4) in *S. scombrus* and 60,620 (M=2) or 58,723 (M=4) in *S. colias* (Fig. 2). The number of RAD loci in the catalog that occur in at least 75% of the individuals ranges

from 23,146 to 32,581 depending on the combination of i) individuals used to create the catalog, ii) individuals used to select loci and filter SNPs and iii) values used for parameters M and n (Fig. 3). Expectedly, number of RAD loci present in at least 75% of the individuals increases when increasing the number of individuals used to select them. Although decreasing the number of RAD loci per individual and in the overall catalog, increasing M from 2 to 4 increases the number of RAD loci (and consequently SNPs) present in at least 75% of the individuals, due to the fact that more common loci can be found when these are composed of more different alleles than when alleles of the same loci are spread into different loci. Increasing n from 3 to 6 reduces number of loci, both in the overall catalog and when the only the RAD tags present in at least 75% of the individuals are selected; this is due to the fact that smaller values of n make more stacks be split into different RAD loci and lager values of n make more stacks be merged into the same RAD loci.

The number of SNPs selected after the filtering steps ranges from 6,688 to 29,394 depending on the combination of i) individuals used to create the catalog, ii) individuals used to select loci and filter SNPs and iii) values used for parameters M and n (Fig. 3). Including *S. colias* for loci selection and SNP filtering produces a larger amount of SNPs even though the number of RAD loci is not so large. Interestingly, even when using the same individuals for catalog building and SNP selection, the number of SNPs obtained with each parameter combination is drastically different, being the number of SNPs obtained with M=4 almost twice the number of SNPs obtained with M=2. Whether this drastic difference in the number of SNPs has an effect on population structure inferences will be explored in the next sections.

Effect of stacks parameters and samples considered at each step on estimated F_{IS} and F_{ST}

We evaluated the combined effect of samples used to create the catalog, samples used to select the SNPs, M and n parameters on two measures reflecting slightly different processes: the inbreeding coefficient of each population reflecting deviations from HWE, F_{IS}, and the level of population differentiation calculated per pair of populations, F_{ST}. The range of F_{IS} and F_{ST} values is large and shows some degree of correlation with the use of certain combination of samples and parameters (Fig. 4, Tables S1 and S2). In general, M=4 produces higher F_{IS} than M=2 whereas the opposite effect is observed for F_{ST}. This may be due to higher values of M erroneously merging non-orthologous stacks, which increases deviations

from HWE, and reduces the potential inclusion of informative SNPs. Within the same value of M, n=3 produces higher F_{IS} than n=6, whereas no remarkable differences are observed on F_{ST} values. Both, F_{IS} and F_{ST} values are lower when all individuals (including *S. colias*) are used for SNP selection due to the fact that selecting the SNPs including *S.colias* produces a large number of low MAF SNPs within *S. scombrus*. When selecting the SNPs on the same individuals, F_{IS} values are lower when more individuals are used for catalog building. Also, when building the catalog on the same individuals, selecting SNPs on more individuals produces lower F_{IS} values likely due to larger sample sizes allowing reductions in deviation from HWE. Considering only *S. scombrus*, individuals used for catalog building or SNP selection do not affect F_{ST} values. In summary, F_{IS} is more affected than F_{ST} by the different combinations of individuals used for each step and M and n parameter combinations. Importantly, although notable differences in absolute F_{IS} and F_{ST} values are observed when using M=2 or M=4, relative F_{IS} and F_{ST} values remain constant (Tables S1 and S2), suggesting that population structure interpretation would not be affected by parameter choice. Yet, when it comes to outlier loci detection, some differences among the different SNP selection procedures are observed. M=4 and n=3 produce more outliers than M=2 and n=6 respectively, which is in line with the behavior observed for F_{IS}. Concerning individuals used for catalog building, no outliers are identified in any of the Mediterranean only datasets, whereas similar outlier SNPs subsets are obtained when using all individuals or only *S. scombrus* individuals to construct the *S. scombrus* dataset (Table S3).

Effect of stacks parameters and samples considered at each step on inferred population structure of Atlantic mackerel

Principal Component Analyses (PCA) performed on each dataset including all *S. scombrus* individuals reveal three main genetically differentiable groups: the Canadian samples, the Bay of Biscay samples and the Mediterranean samples. Within the latter, an additional distinction can be made between a group composed by the Western Mediterranean and Tyrrhenian samples and a group composed by the Adriatic samples; this separation can also be observed when only the Mediterranean *S. scombrus* individuals are considered (Fig. 5). Interestingly, the use of different subsets of individuals for catalog building and SNP selection does not alter the pattern observed as virtually the same image is obtained whatever the combination of samples is used for catalog building for either, a final analysis with all *S.*

scombrus samples or with only the Mediterranean *S. scombrus* samples. Although M=2 seems to provide an increased differentiation with respect to M=4 within the Mediterranean samples, quantitative measures of the differences among groups do not show clear trends on which parameter combination gives the highest or lowest differentiation among groups (Fig. S1). Again, regardless of the individuals used for the analysis or parameter combinations chosen, the relative differences among pairs of groups remain, suggesting that individual selection and parameter combinations should not affect interpretations of population structure inferences (Table S4).

The Bayesian clustering approach to infer the genetic ancestry of each individual is consistent whichever the combination of parameters of individuals used for catalog building or SNP selection is used (Fig. 6). Interestingly, using the same n value (between individual distance) for catalogs built from different sample subsets or different n values for the same catalog provide the same result, meaning that, although this parameter should be selected according to the expected evolutionary distance among the individuals used to build the catalog (Catchen *et al.* 2011), its effect is minor compared to the effect of the M (within individual distance) parameter. Yet, differences among the Mediterranean samples can only be appreciated with M=2 when only the Mediterranean individuals are included and are more clear with M=2 than with M=4 when all *S. scombrus* individuals are included. Thus, unlike interpretations based on F_{ST} and PCA based population structure inferences, interpretations based on structure plots can be affected by individual selection and parameter choice for SNP discovery and genotyping based in RAD-sequencing data. Indeed, performing a *Structure* analysis based on SNPs selected including *S. colias* individuals results in a non-differentiation among the *S. scombrus* individuals and the same effect is observed when SNPs to be included in the analysis are not filtered for MAF (not shown). Although these are extreme cases that imply considering a different species that diverged about 10 Mya (Miya *et al.* 2013) in the analysis or including thousands of monomorphic SNPs, they are good illustrations on how biased individual and SNP selection for analysis can affect population structure inferences based on thousands of markers.

Discussion

RAD-seq data produces a robust population structure inference of Atlantic mackerel

The use of hundreds or thousands of genome-wide polymorphic markers increases resolution in population structure inferences, allowing the detection of intraspecific genetic differentiation where single or a few marker based inferences fail. The advent of RAD-seq and related approaches for high-throughput SNP discovery and genotyping has revolutionized the particularly challenging field of demographic inferences of marine fish species (Benestan *et al.* 2015; Hemmer-Hansen *et al.* 2014; Pujolar *et al.* 2014), most of which are characterized by large population sizes and significant gene flow among populations. Our RAD-seq data based population structure inferences strongly support genetic differentiation within the highly migratory Atlantic mackerel, clearly distinguishing Northwest Atlantic, Northeast Atlantic and Mediterranean samples. The highest differentiation is observed among the Northwest and Mediterranean samples (average F_{ST}=0.039), then among the Northeast and Mediterranean samples (average F_{ST}=0.0201) and finally among the Northwest and Northeast samples (average F_{ST}=0.0157). Previous studies, based on one or two mitochondrial markers, found inconclusive or incongruent results for this species; for example, the high differentiation of Northwest Atlantic samples with respect to the Northeast Atlantic ones was supported by the mitochondrial cytochrome b gene, but not by the mitochondrial D-loop region, whereas the opposite was observed for the differentiation within the Northeast Atlantic (Nesbo *et al.* 2000). Our study contradicts previous findings based on the D-loop region that suggested gene flow between the Mediterranean and Atlantic Ocean and supports the Adriatic samples being differentiated from the Western Mediterranean samples (Zardoya *et al.* 2004), although subtly (average F_{ST}=0.0033). Once the overall picture of mackerel population structure settled, the resources produced in this study will now allow tackling the population structure of this species at a more local level; in particular, deciphering the genetic differentiation among the spawning components within each inferred subpopulation, the Northwest Atlantic, the Northeast Atlantic and the Mediterranean Sea, will be determinant for achieving a sustainable fisheries management for this species.

RAD-seq data analysis strategy affects number and nature of SNPs selected for population structure inferences

The number of obtained RAD loci per individual indicates that the *S. scombrus* and *S. colias* genomes have an estimated number of *SbfI* restriction sites of about 28,000 and 30,000 respectively, suggesting a slightly larger number of *SbfI* cut sites in the genome of *S. colias*. These numbers, as well as the number of RAD loci that occur in the majority of individuals, are similar to those found in other fish species such as sticklebacks (Hohenlohe *et al.* 2010), cichlids (Wagner *et al.* 2013) or hamlets (Puebla *et al.* 2014), which further confirms that, for fish species, this enzyme is suitable to discover a high number of polymorphic markers with high coverage by sequencing about one million reads per individual. Yet, despite the homogeneous number of RAD loci obtained between other studies and our own, and among the different parameters used for read merging within our study, the number of SNPs inferred highly varies depending on certain conditions. For example, taking only the *S. scombrus* samples into account, the number of SNPs obtained with a lower value of M is about half (average of 7,500 SNPs) of that obtained with a higher value of M (average of 14,000 SNPs). Additionally, using *S. colias* for catalog building and SNP selection results in a significantly higher number of SNPs (ranging from 14,708 to 29,394 depending on the combination of M and n parameters used), which is explained by the fact that numerous fixed positions in *S. scombrus* are considered SNPs (MAF>0.05) when an alternative base is found in *S. colias* and vice versa. Interestingly, these differences do not substantially change the main population structure conclusions derived from F_{ST}, PCA and *Structure* based inferences, being the three main Atlantic Mackerel populations (Northwest Atlantic, Northeast Atlantic and Mediterranean) clearly differentiated in all three analyses types for all combinations of parameters and individuals used for catalog building and SNP selection tested. Yet, *Structure* analyses are the most affected by the choice of M value, with inferences based on M=2 more clearly distinguishing the Adriatic samples from to the other Mediterranean samples than inferences based on M=4. This was unexpected and counter to our predictions as previous work suggests over-splitting may reduce the number of alleles and separate population specific divergent alleles reducing detectable population structure (Harvey *et al.* 2015). As already suggested by the higher values obtained for F_{IS}, it is possible that the under-splitting caused by use of M=4 produces some erroneously called genotypes and thus more SNPs deviating from HWE, and *Structure* is likely more sensitive to deviations from HWE assumption than the PCA (Pritchard *et al.* 2000). However, when both species are included in

the analysis, there is evidence that over-splitting may reduce the resolution of population structure. Analysis of the population structure within *S. scombrus* samples using the SNPs selected based on the whole dataset, *i. e.*, including *S. colias*, results in a non- differentiation among populations, probably due to the presence of nucleotide positions that would otherwise have failed the MAF or HWE filters. Additionally, SNPs identified as outliers are different depending on the parameter combination used to identify orthologous loci, which is expected given the differences in the total number of SNPs identified by each approach. This suggests that, although overall structure is robust to alternative analysis procedures because obtained from the main signal in the data, inferences based on specific markers should be validated using alternative approaches such as genotyping in additional samples of same origin.

Conclusions

Our study shows RAD-seq as a powerful approach to detect population structure in a highly migratory pelagic fish such as the Atlantic mackerel. We have applied alternative RAD data processing approaches combining different mismatch parameters for read merging for orthologous loci inference, different samples for loci catalog building and different samples for SNP filtering, and have analyzed the resulting datasets using alternative population differentiation inference methods (F_{ST}, PCA and *Structure*). Although the main conclusions are robust to the different RAD data processing approaches and inference methods, we pinpoint substantial outcome differences resultant from the use of alternative analysis strategies that may affect derived biological interpretations. Interestingly, we observe impacts of both under- and over-splitting of loci on observable population structure, the nature of which seems associated hierarchical SNP selection and the inclusion of a closely related species. As such it seems impossible to conclude that either under- or over-splitting is preferred when attempting to reduce influences on resolvable population structure and the optimal choice will be dataset dependent. The analysis procedures described here and comparisons therein are directly applicable to population genetic inferences of other species based on RAD-seq data and to the various assembly and genotyping tools using for analyzing RADseq data.

Acknowledgements

We wish to thank Iñigo Krug, Inma Martín, Naiara Serrano and Iñaki Rico (AZTI), Eneko Bachiller (IMR) and Craig T. Michel (KAUST) for technical assistance and five anonymous reviewers and the editor for useful comments on the manuscript. We are also grateful to Marianna Giannoulaki (Hellenic Centre for Marine Research, Greece), Valentina Tirelli (Istituto Nazionale Di Oceanografia E Geofisica Sperimentale, Italy), Encarnación Garcia Rodríguez (IEO, Murcia, Spain) and François Grégoire (Fisheries and Oceans Canada) for sharing samples. This project was supported by the Department of Agriculture, Fisheries and Food of the Basque Country and the General Secretary of Fisheries of the Spanish Government. This manuscript is contribution 758 from the Marine Research Division of AZTI.

Author contributions

XI, PA, UC and NRE designed the study. NRE analyzed the data and wrote the manuscript. IB contributed to the analysis of the data. IM performed laboratory experiments. All authors revised and approved the final version of the manuscript.

Data accessibility

Demultiplexed and quality filtered RAD-tags used in this study are available at the U.S. National Center of Biotechnology Information (NCBI) Sequence Read Archive (SRA) accession number SRP069121.

Supporting Information

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

Table S1. F_{IS} values per population calculated on different datasets.

Table S2. F_{ST} values per pairs of populations calculated on different datasets.

Table S3. Values of differences among pairs of populations within the PCA.

Table S4. Outliers identified for the different *S. scombrus* datasets

Fig S1. Graph of differences among pairs of populations calculated from the PCA.

Appendix I. Comparison between Stacks versions 0.99 and 1.32

References

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nat Rev Genet* **11**, 697-709.

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

Baxter SW, Davey JW, Johnston JS, *et al.* (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PloS one* **6**, e19315.

Benestan L, Gosselin T, Perrier C, *et al.* (2015) RAD-genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species; the American lobster (Homarus americanus). *Mol Ecol*.

Bradbury IR, Laurel B, Snelgrove PV, Bentzen P, Campana SE (2008) Global patterns in marine dispersal estimates: the influence of geography, taxonomic category and life history. *Proc Biol Sci* **275**, 1803-1809.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Mol Ecol* **22**, 3124-3140.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda)* **1**, 171-182.

Corander J, Majander KK, Cheng L, Merila J (2013) High degree of cryptic population differentiation in the Baltic Sea herring Clupea harengus. *Mol Ecol* **22**, 2931-2940.

Chong Z, Ruan J, Wu CI (2012) Rainbow: an integrated tool for efficient clustering and assembling RAD-seq reads. *Bioinformatics* **28**, 2732-2737.

Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* **12**, 499-510.

Dichmont CM, Ovenden JR, Berry O, Welch D, Buckworth RC (2012) *Scoping current and future genetic tools, their limitations and their applications for wild fisheries management* CSIRO, Brisbane.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**, 359-361.

Eaton DA (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* **30**, 1844-1849.

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (2011) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. *Methods Mol Biol* **772**, 157-178.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14 , 2611 - 2620. *Molecular ecology* **14**, 2611-2620.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**, 977-993.

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution* **27**, 489-496.

Hannesson R (2013) Sharing the Northeast Atlantic mackerel. *ICES Journal of Marine Science: Journal du Conseil* **70**, 259-269.

Harvey MG, Judy CD, Seeholzer GF, *et al.* (2015) Similarity thresholds used in DNA sequence assembly from short reads can reduce the comparability of population histories across species. *PeerJ* **3**, e895.

Hemmer-Hansen J, Therkildsen NO, Pujolar JM (2014) Population genomics of marine fishes: next-generation prospects and challenges. *Biol Bull* **227**, 117-132.

Hess JE, Campbell NR, Close DA, Docker MF, Narum SR (2013) Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. *Mol Ecol* **22**, 2898-2916.

Hohenlohe PA, Bassham S, Etter PD, *et al.* (2010) Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS Genet* **6**, e1000862.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806.

Jansen T, Gislason H (2013) Population structure of Atlantic mackerel (Scomber scombrus). *PloS one* **8**, e64744.

Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070-3071.

Lamichhaney S, Martinez Barrio A, Rafati N, *et al.* (2012) Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proc Natl Acad Sci U S A* **109**, 19345-19350.

Larson WA, Seeb LW, Everett MV, *et al.* (2014) Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (Oncorhynchus tshawytscha). *Evol Appl* **7**, 355-369.

Lischer HE, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298-299.

Lockwood SJ (1988) *The Mackerel - Its biology, assessment and the management of a fishery.*

Mastretta-Yanes A, Arrigo N, Alvarez N, *et al.* (2014) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Mol Ecol Resour*.

Miya M, Friedman M, Satoh TP, *et al.* (2013) Evolutionary Origin of the Scombridae (Tunas and Mackerels): Members of a Paleogene Adaptive Radiation with 14 Other Pelagic Fish Families. *PloS one* **8**, e73535.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Mol Ecol* **22**, 2841-2847.

Nesbo CL, Rueness EK, Iversen SA, Skagen DW, Jakobsen KS (2000) Phylogeography and population history of Atlantic mackerel (Scomber scombrus L.): a genealogical approach reveals genetic structuring among the eastern Atlantic stocks. *Proc Biol Sci* **267**, 281-292.

Nielsen EE, Cariani A, Aoidh EM, *et al.* (2012) Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nat Commun* **3**, 851.

Olafsdottir G, Olafsson K, Skirnisdottir S, *et al.* (2012) Isolation and characterization of thirty microsatellite loci for Atlantic mackerel (Scomber scombrus L.). *Conservation Genetics Resources*.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.
Puebla O, Bermingham E, McMillan WO (2014) Genomic atolls of differentiation in coral reef fishes (Hypoplectrus spp., Serranidae). *Mol Ecol* **23**, 5291-5303.
Pujolar JM, Jacobsen MW, Als TD, *et al.* (2014) Assessing patterns of hybridization between North Atlantic eels using diagnostic single-nucleotide polymorphisms. *Heredity* **112**, 627-637.
Purcell S, Neale B, Todd-Brown K, *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575.
Ravinet M, Westram A, Johannesson K, *et al.* (2016) Shared and nonshared genomic divergence in parallel ecotypes of Littorina saxatilis at a local scale. *Molecular ecology* **25**, 287-305.
Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* **4**, 137-138.
Rousset F (2008) Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. *Mol Ecol Resour* **8**.
Tenningen M, Slotte A, Skagen D (2011) Abundance estimation of Northeast Atlantic mackerel based on tag recapture data–A useful tool for stock assessment? . *Fish Res* **107**, 68-74.
Trenkel VM, Huse G, MacKenzie BR, *et al.* (2014) Comparative ecology of widely distributed pelagic fish species in the North Atlantic: Implications for modelling climate and fisheries impacts. *Progress in Oceanography* **129, Part B**, 219-243.
Uriarte A, Lucio P (2001) Migration of adult mackerel along the Atlantic European shelf edge from a tagging experiment in the south of the Bay of Biscay in 1994. *Fisheries Research* **50**, 129-139.
Wagner CE, Keller I, Wittwer S, *et al.* (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular ecology* **22**, 787-798.
Ward RD, Woodmark M, Skibinski D (1994) A comparison of genetic diversity levels in marine, freshwater and anadromous fishes. *Journal of Fish Biology* **44**, 213-232.
Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 358-1370.
Zardoya R, Castilho R, Grande C, *et al.* (2004) Differential population structuring of two closely related fish species, the mackerel (Scomber scombrus) and the chub mackerel (Scomber japonicus), in the Mediterranean Sea. *Molecular ecology* **13**, 1785-1798.

Figure legends

Fig 1. Map showing the locations where samples of *S. scombrus* (in red) and *S. colias* (in black) were collected. Sampling year and number of individuals (N) analyzed are shown per sampling location.

Fig 2. Boxplots depicting median, first and third quartile and standard deviation of quality filtered sequencing reads (above) and loci obtained per group (below) when allowing a maximum of 2 (in orange) or 4 (in blue) mismatches between stacks to create a loci. COL

indicates *Scomber colias* individuals. Remaining individuals are *S. scomber* from Canada (CAN), Bay of Biscay (BOB), Adriatic Sea (ADR), Tyrrhenian Sea (TYR) and Western Mediterranean Sea (WME).

Fig 3. Number of loci present in at least 75% of the individuals (above), and SNPs remaining after filtering steps (below) for each of the combinations parameters M (2 in orange; 4 in blue) and n (3 in dark color; 6 in light color) and individual subsets utilized for catalog creation (first part of the name) and SNP selection (second part of the name). Dataset notation is all: all samples; sco: only *S. scombrus* samples; med: only Mediterranean *S. scombrus* samples.

Fig 4. F_{IS} per population and F_{ST} per pairs of populations calculated on different datasets constructed using different values of M and n, as indicated by the shape of the points, and based on different individual sets to construct the catalog (first part of the name) and to select tags and SNPs (second part of the name), as indicated by the color of the points. Population and dataset notations as in Fig 1 and 2 respectively.

Fig 5. Principal Component Analysis (PCA) of allele frequencies. Each plot shows the first two principal components of the PCA obtained from datasets built using different M and n values (rows), and different individuals for catalog building and SNP selection (columns). Each dot represents one sample and is colored according to the area of origin. Ovals represent 95% inertia ellipses.

Fig 6. Graphical representation of the Bayesian clustering approach obtained from datasets built using different M and n values (rows) and different individuals for catalog building and SNP selection (columns). Each bar represents an individual and each color, its inferred membership in each of the K (2 or 3) potential ancestral populations. K=2 and K=3 are shown for being the best supported K values according to the Evanno method (Evanno *et al.* 2005).