Genomic outcomes from diverse SNP panels obtained with different de novo building-loci pipelines in a varied species context: Can pipelines be influencing biological interpretations?

Adrian Casanova
Universidade de Santiago de Compostela
Francesco Maroso
Universidade de Santiago de Compostela
Andrés Blanco
Universidade de Santiago de Compostela
Miguel Hermida
Universidade de Santiago de Compostela
Nestor Rios
Universidad de la Republica Uruguay
Graciela Garcia
Universidad de la Republica Uruguay
Alice Manuzzi
Danmarks Tekniske Universitet
Lorenzo Zane
Universita degli Studi di Padova
Ana Verissimo
Universidade do Porto Centro de Investigacao em Biodiversidade e Recursos Geneticos
Jose-Luis Garcia-Marin
Universitat de Girona Facultat de Ciencies
Carmen Bouza
Universidade de Santiago de Compostela
Manuel Vera (manuel.vera@usc.es)
Universidade de Santiago de Compostela https://orcid.org/0000-0003-1584-6140
Paulino Martinez
Universidade de Santiago de Compostela

Research article

Keywords: STACKS 2, 2b-RAD v2.1 pipeline, de novo approach, Bowtie 1, reference genome approach, bivalves, fish, population genomics

Posted Date: August 7th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-50690/v1

Version of Record: A version of this preprint was published on March 2nd, 2021. See the published version at https://doi.org/10.1186/s12864-021-07465-w.

Abstract

Background

The irruption of Next-generation sequencing (NGS) and restriction site-associated DNA sequencing (RAD-seq) in the last decade has led to the identification of thousands of molecular markers and their genotyping for refined genomic screening. This approach has been especially useful for non-model organisms with limited genomic resources. Many building-loci pipelines have been developed to obtain robust single nucleotide polymorphism (SNPs) genotyping datasets using a *de novo* RAD-seq approach, i.e. without reference genomes. Here, the performances of two building-loci pipelines, STACKS 2 and Meyer's 2b-RAD v2.1 pipeline, were compared using a diverse set of aquatic species representing different genomic and/or population structure scenarios. Two bivalve species (Manila clam and common edible cockle) and three fish species (brown trout, silver catfish and small-spotted catshark) were studied. Four SNP panels were evaluated in each species to test both different building-loci pipelines and criteria for SNP selection. Furthermore, for Manila clam and brown trout, a reference genome approach was used as control.

Results

Despite different outcomes were observed between pipelines and species with the diverse SNP calling and filtering steps tested, no remarkable differences were found on genetic diversity and differentiation within species with the SNP panels obtained with a *de novo* approach. The main differences were found in brown trout between the *de novo* and reference genome approaches. Genotyped *vs* missing data mismatches were the main genotyping difference detected between the two building-loci pipelines or between the *de novo* and reference genome comparisons.

Conclusions

Building-loci pipelines seem not to have a substantial influence on population genetics inference. Anyway, we recommend being careful with certain building-loci pipeline parameters and SNP filtering steps, especially when a *de novo* approach is used. Preliminary trials with subsets of data should be performed for comparison of genetic diversity and differentiation, but always considering the specific goals of the study.

Background

Next-generation sequencing (NGS) technologies have represented a breakthrough for genomic studies [1] due to the huge reduction of sequencing cost (less than 0.02$ per Mb; [2]) and the development of a broad and versatile range of techniques for different genomic approaches [3]. By harnessing the possibilities of NGS, diverse reduced-representation genome sequencing approaches, useful to identify

and genotype thousands of markers for genomic screening, were suggested and quickly became popular [4, 5]. One of these approaches is the restriction site-associated DNA sequencing (RAD-seq), currently in a more mature phase, which includes different methods (e.g. ddRAD-seq, ezRAD, 2b-RAD) whose performances have been compared using simulations and real data [6]. RAD-seq methods require specific library preparation protocols, which exploit the ability of Restriction Enzymes (REs) to cut at specific genomic targets rendering a collection of fragments representative of a genome fraction to be compared among samples. These collections can be screened to identify and genotype a variable number of single nucleotide polymorphisms (SNPs) depending on the goals of the study for population genomics, linkage mapping or genome wide association studies, among others. The 2b-RAD method here used exploits the properties of IIB REs which produce a collection of short DNA fragments (between 33–36 bp) by cutting at both sides of the recognition site [7]. This method has the advantages of simple library preparation, short-reads to be sequenced (single-end 50 bp) and, as other methods, the number of loci can be adjusted both using REs with different recognition site frequency or by fixing nucleotides in the adaptors during library construction (i.e. selective-base ligation) [7, 8].

Genomic laboratory protocols have been set up and optimized through years by introducing modifications on RAD-seq to achieve the better results using different lab protocols (see Fig. 5 in [8]). Similarly, the bioinformatics pipelines starting from raw data, a critical issue in RAD-seq methodologies, have undergone an important refinement and diversification. Nevertheless, there is not a consensus about what is the best strategy for each scenario, despite the increasing number of studies addressed to evaluate the impact of technical and/or bioinformatics protocols [9, 10]. In a typical 2b-RAD library, hundreds of millions of reads are generated, and they need to be allocated to each multiplexed individual (dozens to hundreds in the same lane) and to each genomic position or locus in the reference genome (or RAD-tag catalogue). The rationale behind this is stacking raw reads belonging to the same locus, while discerning and separating at the same time the reads belonging to different loci. Results could be improved if a reference genome, belonging to the species itself or to other congeneric species, is available. This would enable to avoid mixing of reads pertaining to paralogous loci. In July 2020, there were reference genomes for 22 bivalve species and subspecies (19 genera) and 336 fish species (259 genera) with different assembly confidence at the NCBI database (https://www.ncbi.nlm.nih.gov/datasets/). Nevertheless, there are about 9,200 species within the 1,260 bivalve genera [11] and 35,562 recognized species within the 5,205 documented fish genera [12]. All in all, less than 0.2% of the genomes of the known eukaryotic species have been sequenced to date [13]. Although full genome sequencing assembly is becoming progressively more robust thanks to the long-read sequencing methods and assembling strategies, most of the species will have to wait for long before their genomes are assembled. Therefore, *de novo* approaches (i.e. stacking reads without a reference genome) will be the only option for many studies, although some initiatives are trying to change this perspective (e.g. Earth Biogenome Project; https://www.earthbiogenome.org/). For this reason, one of the strengths of a RAD-based method is its applicability without a reference genome [14].

There are different bioinformatics pipelines to identify a high number of SNPs and achieve confident genotypes using a RAD-seq approach. The most popular one is STACKS [15, 16] (around 3000 citations,

at Google scholar in Jul. 2020), but several other alternatives have been recently published (e.g. dDocent [17], Fast-GBS [18], TASSEL-GBS v2 [19]). Some of these pipelines are able to perform a *de novo* approach (dDocent), whereas others need a reference genome for alignment (Fast-GBS, TASSEL-GBS v2) or can address both approaches (STACKS, Meyer's 2b-RAD v2.1 pipeline [7]). Several of these alternative pipelines merge and concatenate pre-existing applications, making their design flexible and customized according to the data managed and the goals of the study, but also providing upgrading and reliable bug-fix (e.g. dDocent, Fast-GBS). Several factors should be considered for the selection of the bioinformatics pipeline to be used, among which sampling variance (e.g. distribution of reads across samples and population size), budget (e.g. total read sequencing), which determine the number of samples to be analysed and their coverage, population structure and genome architecture of the species are the most relevant. The genome of each species has its particular size, history (e.g. duplication events), polymorphism, complexity and inter-individual variability, which can hinder the identification of stacks of reads (putative RAD loci) and their variants, circumstances that should be considered when choosing the appropriate building-loci pipeline and its parameters.

Studies comparing bioinformatic pipelines and strategies already exist. Some comparisons between *de novo* and reference-based approaches are available [20, 21], and one of them tested the performance of the different strategies used to obtain accurate population genetics inferences [21]. Noticeable differences in the number of SNPs and even in some population parameters were detected among bioinformatics pipelines [21]. Other studies have evaluated the same software with different species to optimise the selection of bioinformatic parameters (STACKS 1.42, [22]; STACKS 1.44, [10]), making a common advice of doing preliminary trials to optimize the building-loci pipeline selected parameters. Published step-by-step protocols with a single species [14] also exist. A number of SNP calling comparison between STACKS 1.08 and dDocent 1.0 has been carried out using three fish species [17], while Sovic et al. [23] tested a novel pipeline (i.e. AftrRAD) *vs* STACKS and PYRAD using simulated and species datasets to assess computational efficiency and SNP calling. It is not uncommon to find large differences in the number of SNPs (e.g. of one order of magnitude) in some building-loci pipelines comparisons [17]. Recently, Wright et al. [24] compared population parameters (e.g. F_{ST}, PCoA) with SNPs obtained from three pipelines (i.e. GATK, SAMtools, STACKS) using two species with reference genome. Results showed remarkable differences in some population parameters (e.g. Hardy Weinberg Equilibrium) across bioinformatic approaches. Considering this information, a main issue that should be clarified on RAD-seq methodologies is the impact of building-loci pipelines on population genetics parameter estimations and derived conclusions using a *de novo* approach on different biological scenarios and to some extent, to be compared to a references genome approach.

In this work, two building-loci pipelines for SNP calling and genotyping: i) STACKS 2.0 (http://catchenlab.life.illinois.edu/stacks/) and ii) Meyer's 2b-RAD v2.1 pipeline (the original building-loci pipeline for 2b-RAD data) were tested using a *de novo* approach on different genomics and population genetics scenarios by using five aquatic species: (i) Manila clam (*Ruditapes philippinarum*), (ii) common edible cockle (*Cerastoderma edule*), (iii) brown trout (*Salmo trutta*), (iv) silver catfish (*Rhamdia quelen*)

and (v) small-spotted catshark (*Scyliorhinus canicula*). A range of population parameters were compared in the five species applying similar parameter settings for each pipeline (description of pipelines in Methods). The two marine bivalve species from the Order Veneroida show high polymorphism and low population structure [25, 26]; the brown trout belongs to the order Salmoniformes, which suffered a specific genome duplication event [27], and shows one of the highest population structuring among vertebrates [28]; isolated populations from different ecosystems were analysed in the silver catfish, a freshwater species from the order Siluriformes living in fluvial and costal lagoon environments [29]; finally, the small-spotted catshark (order Carcharhiniformes) is a benthic species which populations here used show low genetic differentiation [30]. To date, two of the species used in this study have a reference genome available: Manila clam (genome size: 1.123 Gb; 19 chromosomes [31]) and brown trout (2.370 Gb; 40 chromosomes [32]). Wang and Guo [33] hypothesized that bivalves with 19 chromosomes could have a tetraploid origin, due to its ability to tolerate chromosomal aneuploidies. Furthermore, gene/gene family expansions would be a rather common process in this group, likely more frequent than in other molluscs [34]. Both genomic features could pose a challenge regarding paralogous genes for stacking reads, a common problem when no reference genome is available. In addition, molluscs present the highest genetic polymorphism in animal kingdom [35], which could represent genotyping drawbacks related to the presence of null alleles. All salmonids, including brown trout, have an allotetraploid origin in process of diploidization since their origin around 90 Mya. This specific duplication should beadded to the three Whole Genome Duplications (WGDs) events in the line of teleosts from the vertebrate ancestor [36, 37], which represent a major issue regarding paralogy. This issue would not be so important in small-spotted catshark and silver catfish, with two and three older WGD events in their evolutive lines, respectively [38, 39]. The most recent, the teleost-specific 3rd WGD, dated around 300 Mya. We followed a *de novo* approach for comparison between pipelines in all species, but reference genomes in these two species were also taken as a useful reference to elucidate which build-loci pipeline provides better results with a *de novo* approach.

For all the five species, population genetics parameters (e.g. genetic diversity and population structure inference) were estimated as an essential outcome to evaluate the performance of four SNP panels after different filtering steps. From two building-loci pipelines, STACKS (STA panel onwards) and Meyer's 2b-RAD v2.1 pipeline (ALT panel onwards), and two criteria for SNP selection, common SNPs (i.e. shared between building-loci pipelines, COM panel onwards) and merged SNP (a combination of shared and exclusive SNPs from both building-loci pipelines, MER panel onwards) panels. When available, the results from reference genome approach was used to compare the results of population parameters evaluated with the *de novo* approach. In this case three additional SNP panels were obtained: the former with the reference genome approach using STACKS and the two remaining with the shared SNPs (i.e. STACKS *de novo* and 2b-RAD v2.1). Our main goal was to assess the influence of the genomic architecture and population structure on the biological conclusions obtained with the different bioinformatics pipelines, and accordingly, to propose methodological recommendations for future studies using a *de novo* approach.

Results

The number of filtered reads loaded into building-loci pipelines using the reference genome approach was lower than with the *de novo* approach. A percentage of 22.1% and 25.4% were finally used in Manila clam and in brown trout, respectively. This reduction mostly due to those reads aligning to more than one place (59.6% and 72.6%, respectively) that were filtered out (i.e. -m 1 in Bowtie 1.1.2), the remaining reads failed to align with the mismatch criteria applied (18.3% and 2.0%, respectively).

The number of initial SNPs, after the building-loci step with the *de novo* approach, ranged from 56,074 in brown trout (STA) and 125,823 in silver catfish (ALT) to 356,389 (STA) and 426,317 (ALT) in common cockle (Tables S1-S5). These figures dropped throughout the successive filtering steps (Fig. 1) up to finally being retained from 0.2% in Manila clam STA panel to 20.5% in silver catfish STA panel (Fig. 2 and Fig. 3). There was a remarkable difference at the initial number of SNPs obtained between STA and ALT pipelines in brown trout and small-spotted catshark, although the outcome after filtering was rather similar (Table 1). No comparison was made at BIAL filter (i.e. SNPs with more than two alleles excluded; Fig. 1), since triallelic SNPs are removed with STACKS by default. The proportion of missing genotypes after applying the minimum coverage filtering step was higher with ALT panel than with STA in almost all species (Fig. 4). Filtering patterns varied among species due to the different weight of each filtering step. For instance, in bivalves, where genetic polymorphism is higher, the SNP retention after the third filtering step (i.e. RAD loci with ≤ 3 SNPs per RAD-locus) was much lower than in fish species (Fig. 2 and Fig. 3), while in silver catfish and small-spotted catshark, was due to the minimum allele count (MAC). This was related to the smaller sampling size of those fish species (N = 21 and N = 28, respectively) and the higher frequency of missing data, especially in small-spotted catshark (83% after depth filter in STA and ALT panels). When comparing pipelines, no clear differences on the filtering pattern were observed through the different filtering steps (Fig. 2 and Fig. 3), except for MAC in brown trout, where more SNPs were pruned in the ALT panel. This could be related to the increment on the missing data after the minimum coverage filter, with higher frequency in ALT (59.3% vs 43.7% for ALT and STA, respectively; Fig. 4). After the third filtering step, in brown trout (N = 52) there were significantly more missing genotypes per SNP ($P < 0.01$) at ALT panel on average (28.74 ± 15.24) as compared to STA panel (21.07 ± 17.54). The two species with the lowest median coverage at this step were small-spotted catshark (median = 10x for ALT and STA panels), brown trout (median = 11x and 14x for ALT and STA panels, respectively) and Manila clam (21x for the STA panel).

Table 1

Mean (bold values) and standard deviation of population parameters for the final SNP panels using a *de novo* approach. Mean observed heterozygosity across loci and populations (H_O), mean expected heterozygosity across loci and populations (H_E), global fixation index (global F_{ST}), mean inbreeding coefficient across populations (F_{IS}), mean allelic richness across loci and populations (A_R), number of population structure units detected using STRUCTURE (STR groups) are shown. Y represents structure and N represents no structure. The complete information can be found in Supplementary Information (Tables S1-S5).

		STA	ALT	COM	MER
M. clam	H_O (± SD)	**0.120** (0.014)	**0.103** (0.005)	**0.138** (0.017)	**0.108** (0.010)
	H_E (± SD)	**0.163** (0.005)	**0.135** (0.006)	**0.170** (0.008)	**0.140** (0.000)
	Global F_{ST}	0.006	0.003	0.004	0.005
	F_{IS} (± SD)	**0.237** (0.054)	**0.251** (0.034)	**0.195** (0.058)	**0.228** (0.044)
	A_R (± SD)	**1.698** (0.010)	**1.660** (0.024)	**1.713** (0.019)	**1.668** (0.015)
	STR(Groups)	N	N	N	N
C. cockle	H_O (± SD)	**0.145** (0.010)	**0.125** (0.006)	**0.150** (0.008)	**0.133** (0.005)
	H_E (± SD)	**0.157** (0.005)	**0.140** (0.000)	0.160 (0.000)	**0.150** (0.000)
	Global F_{ST}	0.033	0.029	0.031	0.030
	F_{IS} (± SD)	**0.086** (0.028)	**0.120** (0.026)	**0.066** (0.036)	**0.114** (0.029)
	A_R (± SD)	**1.707** (0.026)	**1.683** (0.024)	**1.733** (0.024)	**1.690** (0.022)
	STR(Groups)	Y (3)	Y (3)	Y (3)	Y (3)
B. trout	H_O (± SD)	**0.243** (0.023)	**0.250** (0.035)	**0.200** (0.026)	**0.257** (0.029)
	H_E (± SD)	**0.190** (0.017)	**0.187** (0.021)	**0.170** (0.026)	**0.193** (0.023)
	Global F_{ST}	0.376	0.370	0.441	0.347
	F_{IS} (± SD)	**-0.269** (0.023)	**-0.336** (0.028)	**-0.179** (0.038)	**-0.333** (0.024)
	A_R (± SD)	**1.523** (0.041)	**1.520** (0.046)	**1.470** (0.044)	**1.533** (0.042)
	STR(Groups)	Y (2–3)	Y (2–3)	Y (2–3)	Y (2–3)
S. catfish	H_O (± SD)	**0.235** (0.049)	**0.235** (0.049)	**0.230** (0.057)	**0.235** (0.049)
	H_E (± SD)	**0.230** (0.056)	**0.230** (0.057)	**0.230** (0.057)	**0.235** (0.049)

		STA	ALT	COM	MER
	Global F_{ST}	0.451	0.453	0.464	0.450
	F_{IS} (± SD)	**-0.004** (0.032)	**-0.012** (0.036)	**-0.002** (0.024)	**-0.014** (0.038)
	A_R (± SD)	**1.690** (0.180)	**1.680** (0.170)	**1.685** (0.177)	**1.690** (0.170)
	STR(Groups)	Y (2)	Y (2)	Y (2)	Y (2)
S-s. catshark	H_O (± SD)	**0.545** (0.021)	**0.520** (0.000)	**0.460** (0.028)	**0.535** (0.007)
	H_E (± SD)	**0.355** (0.007)	**0.340** (0.000)	**0.325** (0.021)	**0.340** (0.000)
	Global F_{ST}	0.002	0.002	0.004	0.002
	F_{IS} (± SD)	**-0.528** (0.035)	**-0.541** (0.018)	**-0.406** (0.024)	**-0.544** (0.020)
	A_R (± SD)	**1.925** (0.021)	**1.905** (0.007)	**1.915** (0.021)	**1.915** (0.007)
	STR(Groups)	N	N	N	N

Table 2

Genotypic differences between shared SNPs from the different pipelines. Genotyping differences are presented as the relative frequency of total genotypes. COM panels were obtained with shared SNPs between STA and ALT panels (both *de novo* approach). RG-STA panels were obtained with shared SNPs between RG and STA (reference genome and *de novo* approach). RG-ALT panels were obtained with shared SNPs between RG and ALT (reference genome and *de novo* approach). Hom: Homozygous; MD: Missing data (i.e. missing genotype); Het: Heterozygous.

Species	SNPs Panel	Hom → Hom	Hom → MD	MD → Hom	Het → MD	MD → Het	Hom → Het	Het → Hom
Manila clam	COM	0.00062	0.01500	0.01884	0.01324	0.00569	0.00349	0.01333
	RG-STA	0.00007	0.01759	0.00217	0.00567	0.00105	0.00336	0.00369
	RG-ALT	0.00017	0.01018	0.00450	0.01376	0.00300	0.00150	0.01243
Cockle	COM	0.00004	0.00575	0.00557	0.01145	0.00273	0.00331	0.00475
Brown trout	COM	0.00002	0.01285	0.01196	0.00952	0.01161	0.00407	0.00150
	RG-STA	0	0.01231	0.00708	0.00944	0.00021	0.00020	0.00744
	RG-ALT	0	0.00179	0.00564	0.00780	0.00040	0.00009	0.00060
Silver catfish	COM	0.00019	0.00765	0.00520	0.00848	0.00382	0.00349	0.00547
Small-spotted catshark	COM	0.00098	0.04128	0.01311	0.05881	0.03522	0.01163	0.01393

The final number of SNPs ranged from 479 (STA) and 956 (ALT) in Manila clam to 21,468 (STA) and 22,481 (ALT) in silver catfish. These figures were always higher with the ALT pipeline for all species. The number of SNPs in COM panels, those from STA panel shared with ALT panel, ranged from 206 in Manila clam to 17,459 SNPs in silver catfish, while the percentage of SNPs called in STA that were found in ALT ranged from 23.9% in small-spotted catshark to 81.3% in silver catfish. The main source of variation when considering the whole COM panels genotype dataset (for instance in silver catfish $N_{\text{total COM genotypes}}$ (366,639) = N_{samples} (21) x $N_{\text{COM SNP}}$ (17,459)) was missing data, ranging from 2.6% in common cockle to 14.8% at small-spotted catshark (FigsS1-S4). The main source of missing genotype differences between pipelines was related to COM SNPs from STA pipeline that passed the minimum coverage filter (Min coverage 8x), but not were genotyped by ALT pipeline due to having a coverage lower than 3x. This situation was found in most species (causing from 46.8% of the missing genotype differences between ALT and STA in brown trout to 63.2% in small-spotted catshark), excluding Manila clam, where the main source was the removed genotypes from ALT pipeline by Min coverage filter, unlike analogous genotypes from STA pipeline that passed this filter (53.5% of the missing genotype differences). The percentage of the genotyping differences caused by a homozygous-heterozygous exchange between pipelines at the

same SNP and individual ranged from 0.5% in brown trout to 2.6% in small-spotted catshark with respect to the whole COM genotype panel (Table 2). The frequency of genotyping differences caused by different homozygotes at the same SNP and individual (e.g. AA for one pipeline and GG for the other) was negligible in almost all cases (from 0–0.098%). Finally, both *de novo* building-loci pipelines performed similarly when compared with reference genome (RG) approach (RG-STA and RG-ALT genotype comparisons; Table 2).

The population parameters evaluated (e.g. diversity levels, global F_{ST}) showed roughly similar figures among the different SNP panels in each species (Table 1). However, there was a notable exception when comparing *de novo* and reference genome approaches in brown trout, especially regarding Hardy-Weinberg tests and related population parameters (i.e. F_{IS}, Ho vs He; Table S4). Here, unlike Manila clam, the proportion of SNPs with extremely low F_{IS} (≤ -0.5) greatly differed between both approaches. The structure patterns obtained using STRUCTURE and DAPC analyses were similar between both approaches (Figs. S5-S9). The highest global F_{ST} values among populations were found in brown trout and silver catfish, as expected, and no different interpretations among panels could be extracted. Some minor discrepancies in the number of suggestive outliers among panels were detected. All suggestive outliers detected showed a positive α-value suggesting diversifying selection. The complete set of population parameters is provided in Supplementary Information (Tables S1-S5).

Discussion

In the last decade, the binomial NGS / RAD-seq has been the choice for genomic screening in many studies due to the vast number of genetic markers identified and genotyped in a single step. In this context, species with low genomic information have been targeted for population genomics and evolutionary studies broadening the opportunities for more refined approaches regarding conservation genetics and breeding programs. Nevertheless, the effect of genomic architecture, genetic diversity, and population structure into the outcomes of these techniques (number of SNPs, genotyping confidence) in the target species are essential issues to be addressed using both with simulation and real data approaches. These issues are not only important for the wet-lab protocols, but also for the bioinformatics pipelines to be used to analyse the huge amount of data produced. Technical decisions on the reduced-representation method and restriction enzymes selection to be applied when constructing libraries are critical. When a reference genome is available the number of potential loci obtained with different restriction enzymes (e.g. using ExtractSites.pl https://github.com/Eli-Meyer/2bRAD_utilities) and the uniqueness of RAD loci should be tested (e.g. using EvalFrags.pl from 2bRAD_utilities). For instance, in Manila clam was predicted that the percentage of the genome constituted by repetitive elements and combined transposable elements could exceed 70% [31]. This genomic information should be considered to make the best technical decisions. Without reference genome, different REs can be tested if the budget allows it (see Box 1 in [8]), to improve the percentage of reads to build up confident loci. In the same way, the different performances of software and bioinfomatic pipelines might depend on the species and its genomics context. These can affect not only the number of markers found, but more importantly, the

biological conclusions drawn. A repertoire of bioinformatic publications to manage the large amount of genomic data at different stages (e.g. building-loci pipelines, SNP filtering steps) has been published to serve as guidelines for researchers with limited experience in the field and advices for bioinformatic “Gordian Knots”.

The panels used in this study came from species that differ in their genomic architecture, polymorphism and population structure, and these factors could influence the results obtained also depending on the different population parameters used (e.g. global F_{ST}, allelic richness). Nevertheless, the results obtained in our study within species using the four *de novo* panels (i.e. STA, ALT, COM, MER) were roughly similar for all the population parameters evaluated. Accordingly, biological inferences would hardly change. Minor differences were found in the number of suggestive outliers detected in common cockle. In this case, the number of suggestive outliers detected could be related to the total number of SNPs of each panel. But beyond this observation, our results suggest that whatever the pipeline chosen similar results are obtained.

A practical approach to decide between pipelines with different building-loci strategies for handling Genotyping-by-sequencing (GBS) data is to assay trials with a small subset of data and check for their results using a meaningful set of population parameters, previously selected according the objectives of the study. Indeed, using the number of SNPs obtained as the main criterion [17] to decide the best building-loci pipeline to be used is not advisable, since a higher number of SNPs does not necessarily indicate a better stacking and confident RAD-seq data [10], and consequently, it might have a negative impact on the confidence of results and biological inferences. The initial number of SNPs obtained with STA and ALT pipelines across the different species tested was rather similar except for the brown trout and the small-spotted catshark. These species showed the lowest median coverage, near to the selected threshold coverage filter (8x) hence the differences on the number of putative loci from input data. While STACKS 2, with a *de novo* approach, starts with individual data demanding a number of identical reads to build a locus (see Methods), Meyer’s 2b-RAD v2.1 works with a combined subset of confident reads from all samples to build a global reference panel to which align every read. In cases with low coverage, a less demanding criterion to build loci can produce large differences in the initial number of SNPs. Nevertheless, through the filtering steps, the SNP number from building-loci pipelines converged in both species and importantly, the population parameters between SNPs panels were similar. Once chosen the pipeline, it would be recommendable to run several trials with different parameters to properly adjust them to the dataset. For instance, –M in STACKS (which defines the maximum nucleotide differences allowed between intraindividual putative loci) depends on the levels of polymorphism of the species and -m on the existing coverage [22]. In the same way as for the choice of the building-loci pipeline, it would be advisable to choose parameters taking into account the results from population outcomes, since there is not a unique pipeline suited to every situation, as already indicated [20].

After the building-loci pipeline, it is important to adjust filtering (criteria and order [9]) according to the particular scenario of each species (e.g. sequencing and genotyping errors, duplicated loci [40]). Since the filtering parameters are dataset dependent [41], the filtering criteria should be adjusted accordingly (e.g.

the stringency of MAC filtering step is sample size dependent). For instance, the number of SNPs was markedly reduced through filtering steps and the highest difference in the percentage of retained SNPs was found among species. In the study by O'Leary et al. [9] the percentage of retained SNPs ranged from 0% to 63% using the same filtering pipeline with four marine fish species. In our study, the three SNP/RAD-locus filter used to avoid inconsistent RAD-loci could not work well for highly polymorphic species or taxa (e.g. bivalves). Furthermore, the POP filter (i.e. 60% call rate per population) could be applied not so stringently since in previous studies qualitative interpretations of population parameters were maintained in most cases [21, 24], and sometimes even improved [42]. Notwithstanding, the drawback could be using larger SNP panels for similar information. If well, for some studies it is fundamental to achieve the highest density of SNPs possible (e.g. linkage disequilibrium, outlier detection and gene mining, Genome-wide association study; GWAS). The biggest difference between pipelines was found with the MAC filtering step in brown trout which could be explained by the higher average of missing genotypes per SNP (MAC is sample size dependant) and the lower coverage per RAD-locus (misclassification of heterozygotes) in the ALT pipeline. Finally, more filtering steps might be necessary, especially when working without a reference genome (e.g. F_{IS} SNP filtering step when paralogs or null alleles can be a problem to avoid misinterpretations); this is the case of HWE deviations in brown trout caused by potential paralogs whose impact can be reduced using a reference genome.

Attention should be paid to the order of the different filtering steps because this can alter the final SNP panel. When adjusting the filtering parameters, it would be advisable to consider not exclusively the number of removed SNPs at each step separately, since they could result from the interaction among filtering steps. For instance, the coverage filter determines the increase of missing data which influences the percentage of SNPs eliminated by MAC and population representation filters, according to the stringency of the coverage threshold used. Furthermore, missing data may be due to a lower coverage than the selected threshold or for not being genotyped by the building-loci software with the genotyping options selected (e.g. previously selected nucleotide frequencies range to genotype in ALT pipeline). We found that the last could be the main source of COM SNPs genotyping differences between both building-loci pipelines excluding Manila clam. This means that the ALT pipeline genotyping parameter should be improved by choosing appropriate ranges for each species. The objective of any filtering strategy is removing SNPs that are not reliable without losing informative SNPs. Different factors can influence the filtering criteria, e.g. to achieve the number of SNPs required to meet the research goals. In this sense, a panel made up with markers found by two different pipelines should ensure reliability. It was found that 67% of SNPs from STACKS panel were common with UNEAK panel, using a *de novo* approach in soybean (*Glycine max L.*) data [20]. With reference genome the overlap percentages among STACKS and other building-loci pipelines ranged from 76–96% [20]. Using a reference genome approach the percentages of shared SNPs between STACKS with SAMtools and GATK ranged from 7.3–71.4% [24]. The lowest values could be partially explained because STACKS panel recruited many more SNPs than the other building-loci pipeline. The lowest percentage of COM SNPs taking STA panel as reference in our study (i.e. 23.9% in small-spotted catshark and 43.0% in Manila clam) were in panels with less than 1000 SNPs. These low values may be explained by a strong filtering effect, on shared SNPs between pipelines. The highest

number of COM SNPs were detected when STA panels included the highest number of SNPs, around 74% in brown trout and 81% in silver catfish. Despite including a lower number of SNPs, the COM panels provided roughly similar results to the larger ones. This suggests that most informative markers are retained downstream, with the advantage of working with a reduced panel that can simplify and speed-up analyses. In the study by Díaz-Arce et al. [10] the possible effect of SNP number on F_{ST} estimation using reduced SNPs subsets was tested and similar values regarding the full panel were obtained. Moreover, estimated genotyping accuracy may be higher with SNPs shared by more than one building-loci pipeline according to Torkamaneh et al. [20]. The impact of genotypic differences between shared SNP panels was low, such as those obtained by Wright et al. [24].

Summarizing, the results obtained suggest that both building-loci pipelines are adequate and provide more confident results adjusting parameters and SNP filtering steps to the research context. Despite the differences observed in the number of SNPs among *de novo* approach panels, this seems not to affect dramatically the conclusions, at least in the biological scenarios managed in this study. When there is no reference genome a COM panel could be interesting with species with high genomic complexity. In a general way for some population parameters, to have less SNPs do not imply loss of biological information and a COM panel could increase data reliability in these cases. In the case of choosing this option differences in genotyping between pipelines should be checked, although in this study genotyping differences between pipelines were infrequent. The main source of genotyping differences in COM SNP panels was missing data and had different sources. On one hand, different genotypes can be obtained due to the different building-loci pipeline parameters to call genotypes (e.g. –alpha at STACKS) and the different alignment strategies (e.g. reads with alternative allele can be stacked into another putative loci). For missing genotype differences, it should be considered that even RAD loci showing high coverage, might have missing data if building-loci pipeline parameters involved in genotyping are not properly set up. Furthermore, small differences in the building-loci pipeline could have more influence in the number of missing genotypes when working with low coverage loci. Anyway, it would be advisable to use a few intra-library and inter-library sample-replicates to estimate genotyping errors [43] to increase the confidence in our data, especially if the RAD-seq libraries are designed with low estimated coverage per locus (e.g. around 10x), due to the impact of coverage in genotyping error rates [44].

Conclusion

The results here obtained show that selected building-loci pipeline do not have a substantial influence in the population parameters and their derived biological interpretations. The different results obtained between some *de novo* and reference genome derived panels could be solved with improved SNP filtering steps. Anyway, our results are not directly transferable to any building-loci pipelines or genomic scenarios but suggest that building-loci pipelines should be tested with a small subset of data to check on their performances. Nevertheless, we recommend paying special attention to certain building-loci pipeline parameters and SNP filtering steps when assaying with subsets of data. This should be done using population parameters consistent with the research goals to make the best decisions. Although the

recommendations previously raised are time consuming, they could improve the robustness of results and improve knowledge about the use of bioinformatics tools and datasets.

Methods

Samples studied

Four Manila clam (*R. philippinarum*) samples, three from the Adriatic Sea (Italy: Chioggia, N= 30; Porto Marghera, N= 30 [45]; and Po river mouth N= 25) and one from the Atlantic Ocean (Spain: Galicia, N= 25), were studied. Four common edible cockle (*C. edule*) samples from the European Atlantic area (Somme Bay, France, N= 30; Campelo, Spain, N= 30; Miño River mouth, Spain, N= 30; and Ría Formosa, Portugal, N= 30) were used from regions with extractive cockle activities [46]. Three samples of brown trout (*Salmo trutta*) from Duero River basin in the Iberian Peninsula (Águeda, N= 15; Omaña, N= 20; and Pisuerga, N= 17), two of them representing different mitochondrial pure native lineages (Atlantic and Duero, [47, 48]) and one from the hybrid zone (Omaña; [49]), were evaluated. Two populations of silver catfish (*R. quelen*), a Neotropical freshwater species distributed from the Northeast of Los Andes to the centre of Argentina and living in fluvial and coastal lagoon environments [50], were analysed. These samples came from Sauce Lagoon (N=10) and Uruguay River Basin (N= 11) belonging to two divergent lineages [29]. Finally, two nearby populations without genetic differentiation of small-spotted catshark (*S. canicula*) from North Sea (N= 13) and Irish Sea (N= 15) were analysed [30]. All information from samples analysed is summarized in supplementary material (Table S6).

Library preparation

DNA extraction and 2b-RAD libraries preparation using AlfI IIB RE followed the same protocol except for small-spotted catshark [30] where CspCI IIB RE was used instead. The libraries were sequenced using Illumina sequencing platforms (i.e. HiSeq 1500 for small-spotted catshark, NextSeq500 for the remaining species) following a 50 bp single-end chemistry. For details see [29, 30].

Bioinformatic Analysis

Building-loci pipelines: background

STACKS 2.0 and Meyer's 2b-RAD v2.1 were the building-loci pipelines chosen for comparing their performances using a *de novo* approach within the broad genome and population genetics species scenarios selected. Meyer's 2b-RAD v2.1 pipeline and STACKS building-loci pipelines have some similarities on their strategy; roughly, both are based on stacking reads into putative loci by sequence similarity, assuming that each locus correspond to a single place in the species genome. Nevertheless, there are many differences on how loci are built and how the user can control the existing options and genotyping strategies. STACKS, with a *de novo* approach, works firstly at the individual level demanding a number of identical reads to build a locus (i.e –m parameter in ustacks), while the 2b-RAD v2.1 pipeline works with a combined subset of samples to build a global reference panel to which align every read. For

genotyping, STACKS uses a chi square test to call a heterozygote or a homozygote (i.e. –alpha and –gt-alpha), whereas nucleotide frequencies based on allele read depth at each position and sample are used for genotyping in the 2b-RAD v2.1 pipeline . Accordingly, huge differences in the raw SNP number were obtained in preliminary analysis in our study previous analysis. Anyway, we tried to apply the highest number of common parameters in both pipelines to be consistent among comparisons.

STACKS 2.0 pipeline can be summarized as follows: (1) Raw sequence reads were demultiplexed and filtered according to different criteria such as quality, uncalled bases and read length (process_radtags); (2) reads from each individual were clustered into putative loci, and polymorphic nucleotide sites identified (ustacks); (3) putative loci were grouped across individuals and catalogues of RAD-loci, SNPs and alleles were constructed (cstacks); (4) putative loci from each individual were matched against the catalogue (sstacks); (5) the data were transposed to be oriented by locus, instead of by sample (tsv2bam); (6) all individuals were genotyped at each called SNP (gstacks); and (7) SNPs were finally subjected to population genetics filters (populations) and results written in different output files (e.g. GENEPOP[51], STRUCTURE[52] file formats).

The *de novo* approach used for the 2b-RAD v2.1 pipeline can be summarized as follows: (1) from a subset of high quality reads (Phred quality scores ≥ 30 at all positions) from all samples, a global *de novo* reference catalogue was built by clustering these reads with the BuildRef.pl script and cd-hit 4.6.8 [53,54]; (2) every read was aligned against the reference catalogue using a mapping program (in our case Bowtie 1.1.2 [55]); (3) allele frequencies were counted at each position (SAMBasecaller.pl) and genotypes determined from that information (NFGenotyper.pl); and (4) genotypes called across samples were combined into a single genotype matrix with samples as columns and loci as rows (CombineGenotypes.pl). Perl scripts belonging to the last version are publicly available (https://github.com/Eli-Meyer/2brad_utilities/) and earlier versions on request.

A reference genome-based pipeline was used for Manila clam and brown trout, the species with chromosome assembly level reference genomes. In this case, the differences in the pipelines of STACKS 2 and Meyer's 2b-RAD v2.1 are lower. For instance, STACKS 2 reduces the number of modules necessary from six to two when using reference genome and the building of putative loci is conditioned by the step using a short-read aligner. Our goal was not to compare this option between pipelines, but to take as reference the genome-based approach to be compared with the *de novo* approach as a gold standard within each pipeline.

Building-loci pipelines: analysis

After demultiplexing raw data, several filtering criteria were applied: (1) all reads were trimmed and filtered by the RE recognition site to retain only those sequences of 36 bp (or 32 bp from CII RE catshark) centred on the RE recognition site using our own Perl scripts and Trimmomatic 0.38 [56]; and (2) process_radtags (module belonging to STACKS) was used to remove reads with uncalled bases (-c option). Other parameters were species-specific (e.g. window sliding size, -w; score limit, -see Table S7A). To check raw and filtered reads quality, FastQC 0.11.7 [57] was used. STACKS input sequences were oriented in the

same orientation using our own Perl script to avoid oversplitting. To say, the set of input reads for both building-loci pipelines in each species was always the same.

At the building-loci pipeline step, the parameters considered were: (1) minimum number of identical reads to create a stack (default values were used); (2) maximum number of mismatches between RAD loci within and between individuals (-M 2/-n 2 for fish or -M 3/-n 3 for bivalves and their analogous parameters with ALT pipeline); (3) indels were discarded (i.e. –disable-gapped at different modules); (4) SNP calling model and alpha cut-off: their default values were used in the STA pipeline (STACKS 2.0), while in the ALT pipeline (Meyer's 2b-RAD v2.1) we considered a range of 0.1-0.2 to determine the genotype at each position (default values are 0.01-0.25); to say, when the frequency of the less frequent allele was lower than 0.1 the genotype was called as homozygous while frequencies higher than 0.2 were called as heterozygous; intermediate allele frequencies for the less frequent allele were called as uncertain (see Tables S7B and S7C).

When a reference genome was available, Bowtie 1.1.2 was used as short read aligner. The number of mismatches allowed between reads and the reference genome using the -v alignment mode was 2 mismatches for brown trout and 3 mismatches for Manila clam, the same as mentioned above. Only reads which aligned to a single site in the reference genome were considered (-m 1). The same parameter values were used in STACKS modules shared between reference-based genome and *de novo* approaches (i.e. gstacks and populations modules).

SNP filtering steps and creation of datasets

The raw SNP panels of the STA and ALT pipelines were filtered using the same parameters for consistency, retaining only a set of markers and alleles represented across the individuals genotyped. Filters were applied in the same order for each dataset (Fig. 1), as recommended [9]: i) SNPs with more than two alleles were excluded (BIAL filter); ii) a minimum locus coverage of eight reads was chosen (Min coverage filter); iii) RAD-loci with more than three SNPs were excluded for further analyses; iv) SNPs were retained only if the less frequent allele was represented at least three times at the whole species sample (MAC, minimum allele count, filter); v) less than 40% of missing data in each population for a SNP to be retained (POP filter); vi) SNPs were excluded when they did not adjust to Hardy-Weinberg (HW) expectations ($P < 0.01$) in more than half of the populations analysed (HW filter); and vii) only the first SNP per RAD-locus was retained when several SNPs were called in the same RAD-locus to avoid redundant information.

According to the aforementioned criteria, four SNP panels were tested and compared: (1) STA and (2) ALT SNP panels were further used to obtain (3) common (COM) and (4) merged (MER) panels. When reference genome was available three additional SNP panels were obtained (i.e. RG, RG-STA, RG-ALT). RAD-loci of these panels from both pipelines were compared to identify shared and private RAD-loci. In order to do this, cd-hit-est was used to cluster similar RAD-loci taken from the two pipeline catalogues with the same threshold of similarity (-c) used in the clustering steps of building-loci pipelines (i.e. two mismatches maximum for fish species and three mismatches maximum for bivalve species).

Furthermore, we used a -g value of 1 when clustering sequences to meet the established similarity threshold and a band alignment width (-b) of 1 to avoid previously separated sequences due to indels at specific clusters. This procedure rendered COM and MER SNP panels created using a customized Perl script (see Supplementary material). SNPs at shared RAD-loci between STA and ALT panels were selected after the sixth filtering step, since the first SNP at shared RAD-loci could be different after the full filtering pipeline (Fig. 1). MER panel was finally created by taking the SNPs from “private” ALT and STA pipelines RAD-loci (i.e. those from cd-hit-est clusters with only STA or ALT pipelines RAD-tags) plus the COM SNP panel previously obtained. Again, only the first SNP per RAD locus was retained to avoid redundant information. The GENEPOP files of all shared SNP panels were compared to quantify their genotyping differences using own Perl script (see Supplementary material, customised_scripts.zip).

Comparison of outputs and population genetics analyses.

The results of the aforementioned pipelines were compared using both quantitative (i.e. number of SNPs) and biological criteria (population genetics data). Filtered GENEPOP files were obtained using customized Perl scripts. These files were transformed for subsequent analyses using the PGDSPIDER 2.1.1.5 software [58]. Firstly, the consequences of filtering over the number of RAD-loci and SNPs were evaluated for each combination of species-pipeline; secondly, common and private RAD-loci/SNPs between the two pipelines were obtained for each species. Finally, biological interpretations from each pipeline/species were compared, including basic population genetics results (i.e. genetic diversity levels and population structure).

Observed and expected heterozygosity (Ho and He, respectively), inbreeding coefficient (F_{IS}, using 1000 bootstrap iterations to estimate their 95% confidence intervals) and allelic richness were calculated per population using R’s package *diveRsity* [59]. Global F_{ST} was calculated with Genepop R package [51]. STRUCTURE software [52], using R package ParallelStructure [60], was used to define the most likely number of population units (K) present with LOCPRIOR model with correlated allele frequencies model, testing K values from 1 to the number of sampling localities in the species dataset + 1 with 10 replicates composed by 100,000 Markov chain Monte Carlo (MCMC) replicates and a burn-in period of 10,000 steps. Structure results were parsed with Structure Harvester [61], which implements the Evanno’s method [62] to detect the most likely number of clusters according to the data. CLUMPAK [63] was used to merge runs with the same K that suggested similar patterns of structuring and to obtain cluster membership plots. As a second approach to detect population structure, a Discriminant Analysis of Principal Components (DAPC) analysis was performed based on genetic data, as implemented in R package adegenet [64,65]. The optimal number of principal components to be used was estimated with the *cross-validation* method implemented in the package and from one to three discriminant components were retained according to the amount of population structure variation they explained. Finally, outlier loci potentially under selection (OL), i.e. those showing higher or lower differentiation values (i.e. F_{ST}) across populations than the neutral background, were detected using the Bayesian approach implemented in BAYESCAN 2.1 [66] with default parameters. Loci with a Log10 posterior odds (PO) higher than 1.5 were

retained as potential outliers for later comparison among the four datasets resulting from the two pipelines.

Abbreviations

ALT: Alternative pipeline *de novo* panel; ALT pipeline:Meyer's 2b-RAD v2.1 pipeline with the selected parameters at this research; COM:Common SNP panel; DAPC:Discriminant analysis of principal components; ddRAD:Double digest restriction-site associated DNA; Gbp:Giga base pairs; GBS:Genotyping-by-sequencing; GWAS:Genome-wide association study; H_E:expected heterozygosity; H_O:observed heterozygosity; HW:Hardy-Weinberg; MAC:Minimum allele count; Mya:Million years ago; MER:Merged SNP panel; NGS:Next-generation sequencing; OL:Outlier Loci; PCoA:Principal Coordinate Analysis; PO:posterior odds; RAD-seq:Restriction site-associated DNA sequencing; REs:Restriction enzymes; SNPs:Single nucleotide polymorphisms (SNPs) ; REs:Restriction enzymes; RG:Reference genome SNP panel; RG-STA:shared SNP panel between RG and STA panels; RG-ALT:shared SNP panel between RG and ALT panels; STA:STACKS *de novo* panel; STA pipeline:STACKS 2.0 pipeline with the parameters selected at this research; WGD:Whole Genome Duplications.

Declarations

Availability of data and materials

Data are available on request to the authors.

Ethics approval and consent to participate

All samples were previously used in other published articles or in the framework of European project COCKLES (EAPA_458/2016). All specimens were collected in accordance with all bioethics standards and legislation of the different country governments where sampling was performed. The samples were shared with us for analysis.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

The work undertaken in this project was funded by Xunta de Galicia Autonomous Government (GRC2014/010), Interreg Atlantic Area (Cockles project, EAPA_458/2016) and Girona University (MPCUdG2016/060) projects. Adrián Casanova was a Xunta de Galicia fellowship (ED481A-2017/091).

Authors' contributions

MV, PM, AC, FM, AB and MH designed the study. AC, FM, AB, MH ran the different building-loci pipelines. AC and FM performed the after building-loci analysis. AB wrote Perl scripts. MV, PM, NR, GG, AM, LZ, AV, JLGM and CB provided genomic data. MV and PM supervised the study. AC wrote the initial version of the manuscript and all authors contributed to the writing and editing of the final manuscript. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge the bioinformatics support of the Centro de Supercomputación de Galicia (CESGA). We wish to thank L. Insua, M. Portela, S. Sánchez-Darriba and S. Gómez for their technical support. We thank Carlos Saavedra for Manila clam samples from Galicia.

References

1. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics. 2012;13:341. https://doi.org/10.1186/1471-2164-13-341.
2. Wetterstrand KA. DNA Sequencing Costs: Data | NHGRI. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data. Accessed 1 Jul 2020.
3. Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51. https://doi.org/10.1038/nrg.2016.49.
4. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 2008;3(10): e3376. https://doi.org/10.1371/journal.pone.0003376.
5. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 2011;12:499–510. https://doi.org/doi:10.1038/nrg3012.
6. Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. 2016;17:81–92. https://doi.org/doi:10.1038/nrg.2015.28.
7. Wang S, Meyer E, McKay JK, Matz M V. 2b-RAD: a simple and flexible method for genome-wide genotyping. Nat Methods. 2012;9:808–10. https://doi.org/10.1038/nmeth.2023.

8. Barbanti A, Torrado H, Macpherson E, Bargelloni L, Franch R, Carreras C, et al. Helping decision making for reliable and cost-effective 2b-RAD sequencing and genotyping analyses in non-model species. Mol Ecol Resour. 2020;20:795–806. https://doi.org/10.1111/1755-0998.13144.
9. O'Leary SJ, Puritz JB, Willis SC, Hollenbeck CM, Portnoy DS. These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. Mol Ecol. 2018;27:3193–3206. https://doi.org/10.1111/mec.14792.
10. Díaz-Arce N, Rodríguez-Ezpeleta N. Selecting RAD-Seq Data Analysis Parameters for Population Genetics: The More the Better? Front Genet. 2019;10:533. https://doi.org/10.3389/fgene.2019.00533.
11. Huber M. Compendium of Bivalves. A Full-color Guide to 3,300 of the World's Marine Bivalves. A Status on Bivalvia after 250 Years of Research. Hackenheim: ConchBooks; 2010.
12. Fricke R, Eschmeyer W, Fong JD. CAS - Eschmeyer's Catalog of Fishes - Species by Family. 2020.http://researcharchive.calacademy.org/research/ichthyology/catalog/SpeciesByFamily.asp. Accessed 14 Jul 2020.
13. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci. 2018;115:4325–33. https://doi.org/10.1073/pnas.1720115115.
14. Rochette NC, Catchen JM. Deriving genotypes from RAD-seq short-read data using Stacks. Nat Protoc. 2017; 12:2640–59. https://doi.org/10.1038/nprot.2017.123.
15. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: An analysis tool set for population genomics. Mol Ecol. 2013;22(11):3124–40. https://doi.org/10.1111/mec.12354.
16. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. G3. 2011;1(3):171–82. https://doi.org/10.1534/g3.111.000240.
17. Puritz JB, Hollenbeck CM, Gold JR. dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ. 2014;2:e431. https://doi.org/10.7717/peerj.431.
18. Torkamaneh D, Laroche J, Bastien M, Abed A, Belzile F. Fast-GBS: A new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. BMC Bioinformatics. 2017;18:1–7. https://doi.org/10.1186/s12859-016-1431-9.
19. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, et al. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS One. 2014;9 (2): e90346. https://doi.org/10.1371/journal.pone.0090346.
20. Torkamaneh D, Laroche J, Belzile F. Genome-wide SNP calling from genotyping by sequencing (GBS) data: A comparison of seven pipelines and two sequencing technologies. PLoS One. 2016;11(8): e0161333. https://doi.org/10.1371/journal.pone.0161333.
21. Shafer ABA, Peart CR, Tusso S, Maayan I, Brelsford A, Wheat CW, et al. Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. Methods Ecol Evol.

2017;8:907–17. https://doi.org/10.1111/2041-210X.12700.

22. Paris JR, Stevens JR, Catchen JM. Lost in parameter space: a road map for stacks. Methods Ecol Evol. 2017;8:1360–73. https://doi.org/10.1111/2041-210X.12775.
23. Sovic MG, Fries AC, Gibbs HL. AftrRAD: A pipeline for accurate and efficient de novo assembly of RADseq data. Mol Ecol Resour. 2015;15:1163–71. https://doi.org/10.1111/1755-0998.12378.
24. Wright B, Farquharson KA, McLennan EA, Belov K, Hogg CJ, Grueber CE. From reference genomes to population genomics: comparing three reference-aligned reduced-representation sequencing pipelines in two wildlife species. BMC Genomics. 2019;20:453. https://doi.org/10.1186/s12864-019-5806-y.
25. Martínez L, Freire R, Arias - Pérez A, Méndez J, Insua A. Patterns of genetic variation across the distribution range of the cockle *Cerastoderma edule* inferred from microsatellites and mitochondrial DNA. Mar Biol. 2015;162:1393–406. https://doi.org/10.1007/s00227-015-2676-y.
26. Vera M, Carlsson J, Carlsson J El, Cross T, Lynch S, Kamermans P, et al. Current genetic status, temporal stability and structure of the remnant wild European flat oyster populations: conservation and restoring implications. Mar Biol. 2016;163:239. https://doi.org/10.1007/s00227-016-3012-x.
27. Leitwein M, Guinand B, Pouzadoux J, Desmarais E, Berrebi P, Gagnaire PA. A Dense Brown Trout (*Salmo trutta*) Linkage Map Reveals Recent Chromosomal Rearrangements in the Salmo Genus and the Impact of Selection on Linked Neutral Diversity. G3. 2017;7:1365–76. https://doi.org/10.1534/g3.116.038497.
28. Ferguson A. Genetic differences among brown trout, *Salmo trutta*, stocks and their importance for the conservation and management of the species. Freshw Biol. 1989;21:35–46.
29. Ríos N, Casanova A, Hermida M, Pardo BG, Martínez P, Bouza C, et al. Population genomics in Rhamdia quelen (Heptapteridae, siluriformes) reveals deep divergence and adaptation in the neotropical region. Genes. 2020; 11: 109. https://doi.org/10.3390/genes11010109.
30. Manuzzi A, Zane L, Muñoz-Merida A, Griffiths AM, Veríssimo A. Population genomics and phylogeography of a benthic coastal shark (*Scyliorhinus canicula*) using 2b-RAD single nucleotide polymorphisms. Biol J Linn Soc. 2018;126:289–303. https://doi.org/10.1093/biolinnean/bly185.
31. Yan X, Nie H, Huo Z, Ding J, Li Z, Yan L, et al. Clam Genome Sequence Clarifies the Molecular Basis of Its Benthic Adaptation and Extraordinary Shell Color Diversity. iScience. 2019; 19:1225-37. https://doi.org/10.1016/j.isci.2019.08.049.
32. Salmo trutta assembly (NCBI). https://www.ncbi.nlm.nih.gov/assembly/GCF_901001165.1 Accessed on date 26 July 2020.
33. Wang Y, Guo X. Chromosomal rearrangement in pectinidae revealed by rRNA loci and implications for bivalve evolution. Biol Bull. 2004; 207(3):247-56. https://doi.org/10.2307/1543213.
34. Takeuchi T, Koyanagi R, Gyoja F, Kanda M, Hisata K, Fujie M, et al. Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. Zool Lett. 2016;2: 3. https://doi.org/10.1186/s40851-016-0039-2.

35. Curole JP, Hedgecock D: Bivalve Genomics: Complications, Challenges, and Future Perspectives. In Aquaculture Genome Technologies. Edited by Zhanjiang (John) Liu. Oxford: Blackwell Publishing Ltd; 2007: 525-543.
36. Pasquier J, Cabau C, Nguyen T, Jouanno E, Severac D, Braasch I, et al. Gene evolution and gene expression after whole genome duplication in fish: The PhyloFish database. BMC Genomics. 2016; 17: 368. https://doi.org/10.1186/s12864-016-2709-z.
37. Macqueen DJ, Johnston IA. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc R Soc B Biol Sci. 2014; 281:1778. https://doi.org/10.1098/rspb.2013.2881.
38. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun. 2014;5:2. https://doi.org/10.1038/ncomms4657.
39. Donoghue PCJ, Purnell MA. Genome duplication, extinction and vertebrate evolution. Trends Ecol Evol. 2005;20(6):312–9. https://doi.org/10.1016/j.tree.2005.04.008.
40. Benestan LM, Ferchaud AL, Hohenlohe PA, Garner BA, Naylor GJP, Baums IB, et al. Conservation genomics of natural and managed populations: Building a conceptual and practical framework. Mol Ecol. 2016;25:2967–77. https://doi.org/10.1111/mec.13647.
41. Hendricks S, Anderson EC, Antao T, Bernatchez L, Forester BR, Garner B, et al. Recent advances in conservation and population genomics data analysis. Evol Appl. 2018;11:1197–211. https://doi.org/10.1111/eva.12659.
42. Hodel RGJ, Chen S, Payton AC, McDaniel SF, Soltis P, Soltis DE. Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: Comparing microsatellites and RAD-Seq and investigating loci filtering. Sci Rep. 2017;7:17598. https://doi.org/10.1038/s41598-017-16810-7.
43. Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson BC. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. Mol Ecol Resour. 2015;15:28–41. https://doi.org/10.1111/1755-0998.12291.
44. Fountain ED, Pauli JN, Reid BN, Palsbøll PJ, Peery MZ. Finding the right coverage: the impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. Mol Ecol Resour. 2016;16:966–78. https://doi.org/10.1111/1755-0998.12519.
45. Milan M, Maroso F, Dalla Rovere G, Carraro L, Ferraresso S, Patarnello T, et al. Tracing seafood at high spatial resolution using NGS-generated data and machine learning: Comparing microbiome versus SNPs. Food Chem. 2019;286:413–20. https://doi.org/10.1016/j.foodchem.2019.02.037.
46. Maroso F, De Gracia CP, Iglesias D, Cao A, Díaz S, Villalba A, et al. A useful snp panel to distinguish two cockle species, cerastoderma edule and c. Glaucum, co-occurring in some european beds, and their putative hybrids. Genes (Basel). 2019;10:760. https://doi.org/10.3390/genes10100760.

47. Bouza C, Castro J, Sánchez L, Martínez P. Allozymic evidence of parapatric differentiation of brown trout (*Salmo trutta L.*) within an Atlantic river basin of the Iberian Peninsula. Mol Ecol. 2001;10:1455–69. https://doi.org/10.1046/j.1365-294X.2001.01272.x.
48. Vera M, Cortey M, Sanz N, García-Marín JL. Maintenance of an endemic lineage of brown trout (*Salmo trutta*) within the Duero river basin. J Zool Syst Evol Res. 2010;48:181–7. https://doi.org/10.1111/j.1439-0469.2009.00547.x.
49. Martínez P, Bouza C, Castro J, Hermida M, Pardo BG, Sánchez L. Analysis of a secondary contact between divergent lineages of brown trout Salmo trutta L. from Duero basin using microsatellites and mtDNA RFLPs. J Fish Biol. 2007;71:195–213. https://doi.org/10.1111/j.1095-8649.2007.01551.x.
50. Perdices A, Bermingham E, Montilla A, Doadrio I. Evolutionary history of the genus *Rhamdia* (Teleostei: Pimelodidae) in Central America. Mol Phylogenet Evol. 2002;25:172–89. https://doi.org/10.1016/S1055-7903(02)00224-5.
51. Rousset F. GENEPOP'007: A complete re-implementation of the GENEPOP software for Windows and Linux. Mol Ecol Resour. 2008;8:103–6. https://doi.org/10.1111/j.1471-8286.2007.01931.x.
52. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945-59.
53. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2. https://doi.org/10.1093/bioinformatics/bts565.
54. Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9. https://doi.org/10.1093/bioinformatics/btl158.
55. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
56. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20. https://doi.org/10.1093/bioinformatics/btu170.
57. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
58. Lischer HE, Excoffier L. PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics. 2012;28:298–9. https://doi.org/10.1093/bioinformatics/btr642.
59. Keenan K, Mcginnity P, Cross TF, Crozier WW, Prodöhl PA. DiveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. Methods Ecol Evol. 2013;4:782–8. https://doi.org/10.1111/2041-210X.12067.
60. Besnier F, Glover KA. ParallelStructure: A R Package to Distribute Parallel Runs of the Population Genetics Program STRUCTURE on Multi-Core Computers. PLoS One. 2013;8(7): e70651. https://doi.org/10.1371/journal.pone.0070651.
61. Earl DA, vonHoldt BM. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour. 2012;4:359–61.

https://doi.org/10.1007/s12686-011-9548-7.

62. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol Ecol. 2005;14:2611–20. https://doi.org/10.1111/j.1365-294X.2005.02553.x.
63. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour. 2015;15:1179–91. https://doi.org/10.1111/1755-0998.12387.
64. Jombart T. Adegenet: A R package for the multivariate analysis of genetic markers. Bioinformatics. 2008;24(11):1403–5. https://doi.org/10.1093/bioinformatics/btn129.
65. Jombart T, Ahmed I. Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 2011;27(21):3070-1. https://doi.org/10.1093/bioinformatics/btr521.
66. Foll M, Gaggiotti O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. Genetics. 2008;180:977–93. https://doi.org/10.1534/genetics.108.092221.

Figures

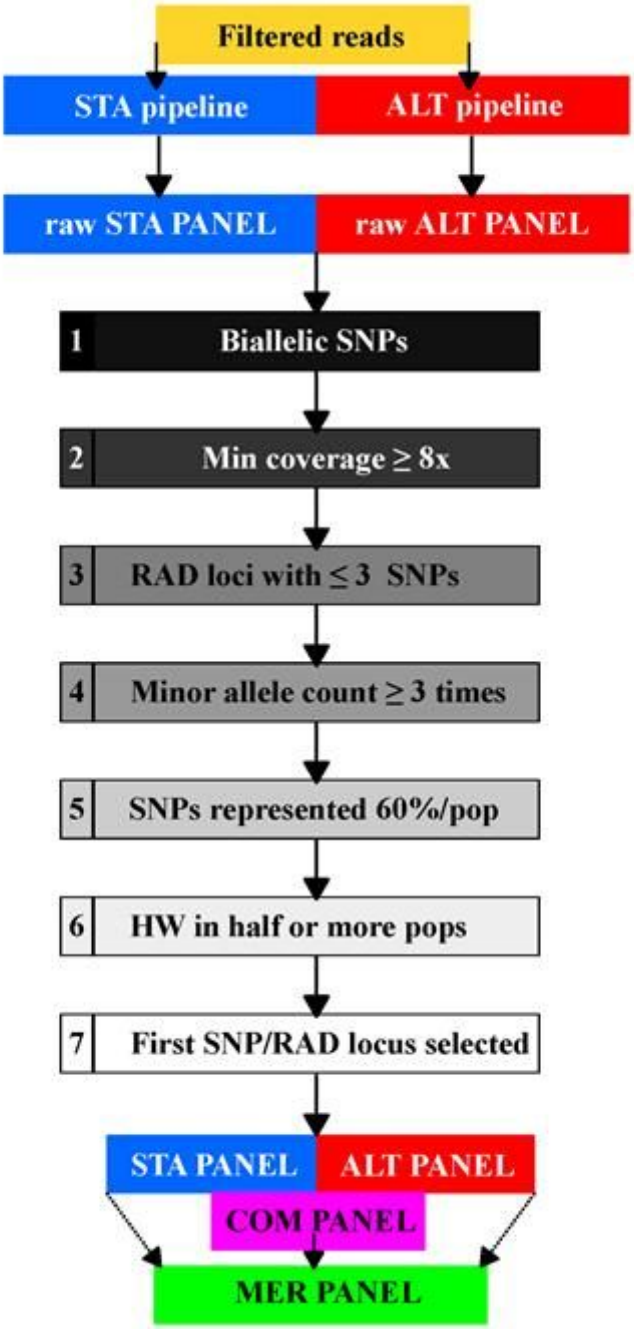

Figure 1

Scheme of filtering steps to obtain the different SNP panels: STA, ALT, COM and MER, representing STACKS, Alternative, Shared and Merged panels, respectively.

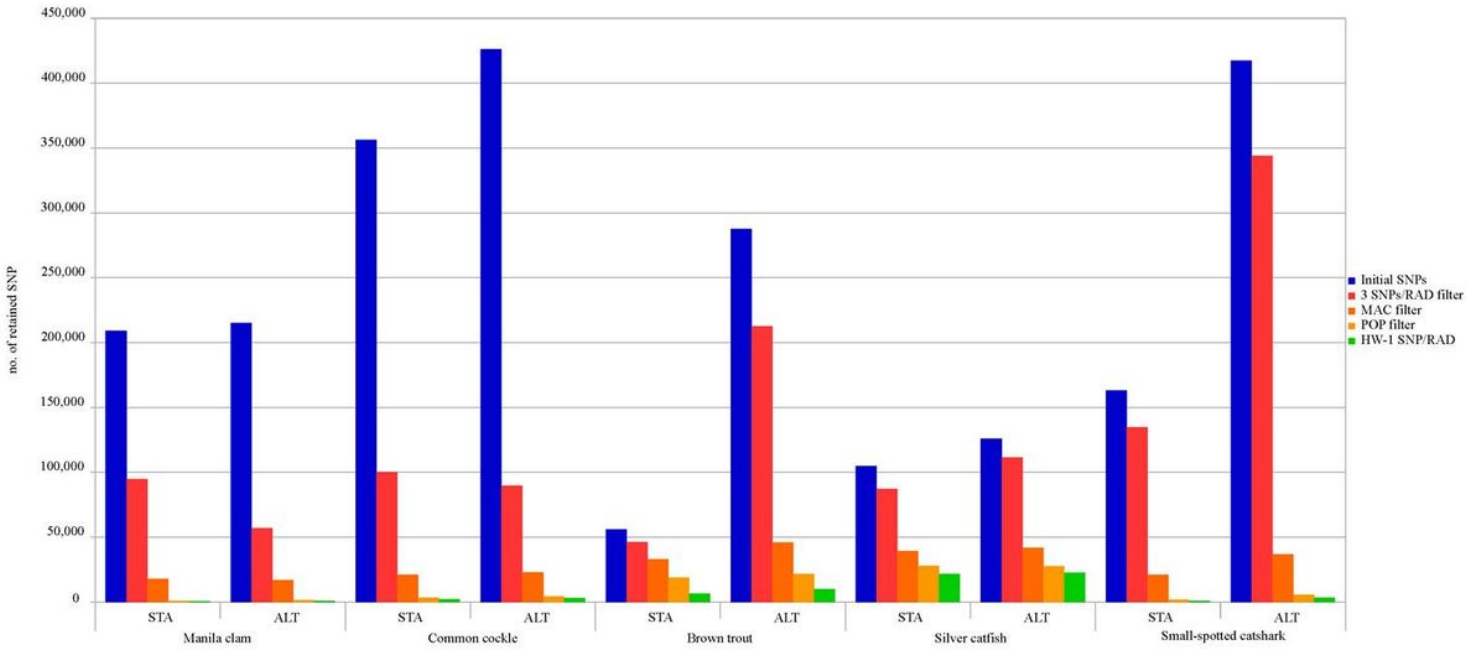

Figure 2

Number of SNPs from the initial build-loci pipelines (blue bars) to the final panels (green bars) through the different SNPs filtering steps.

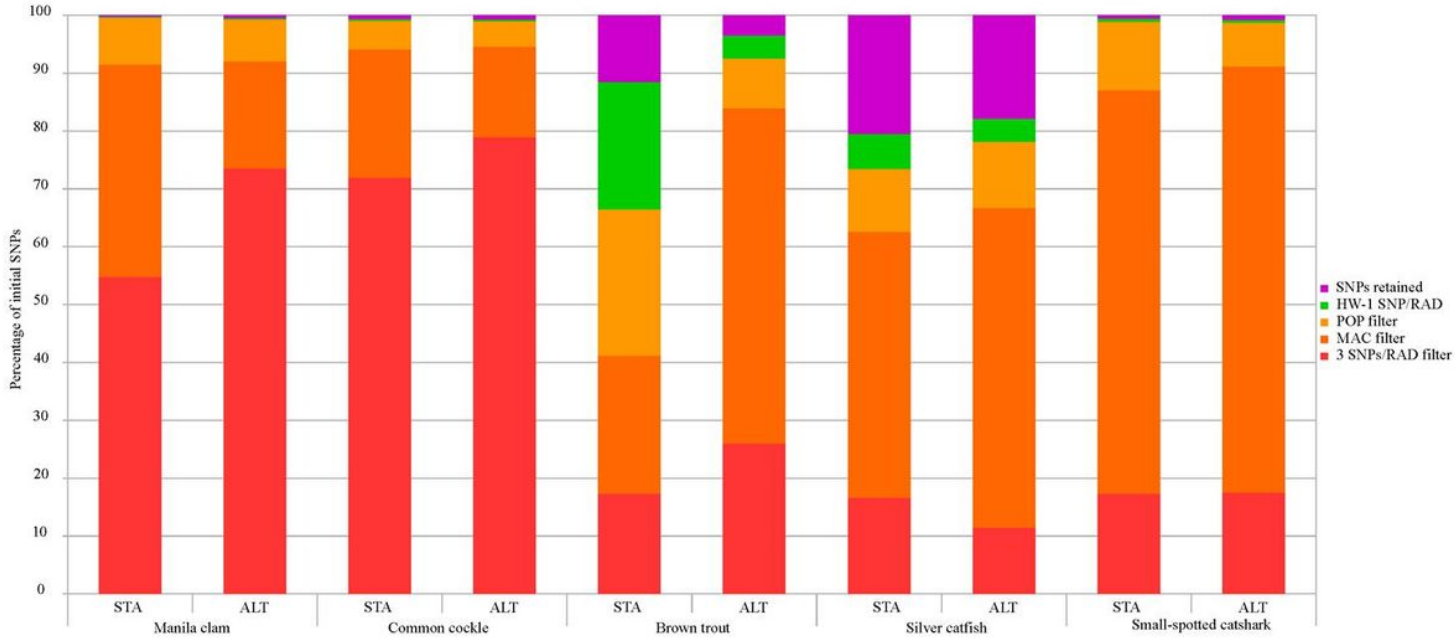

Figure 3

SNP reduction (in percentage) according to population genetics filtering steps from the initial number of SNPs to the panel finally retained (purple bar).

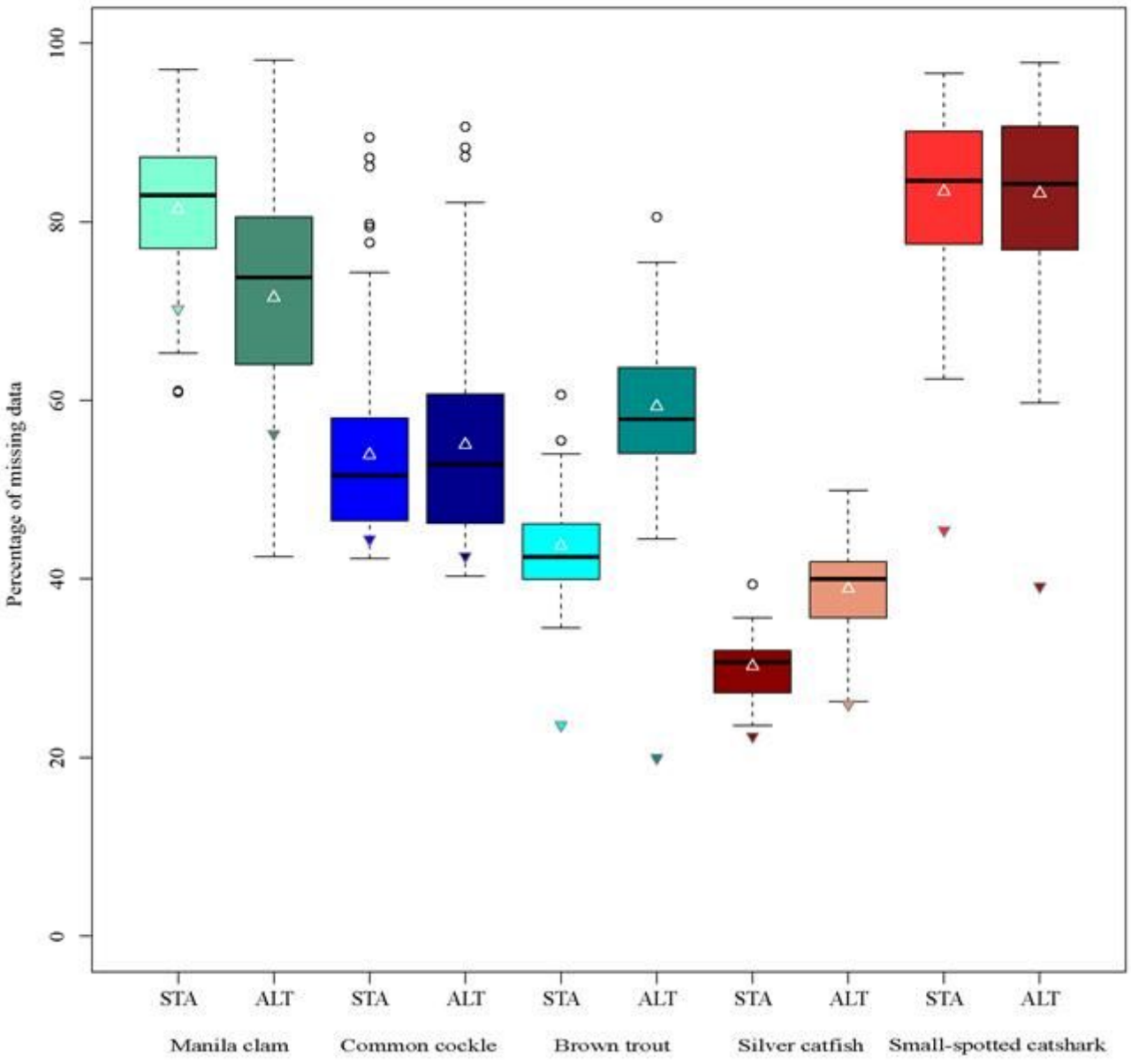

Figure 4

Percentage of missing genotypes after the Min coverage filter (8x). Boxplots were obtained with the percentage of missing genotypes through the different samples. Up and down triangles represent the percentage of missing genotypes at different SNP panels after and before 8x coverage filter, respectively.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile3customisedscriptsBMCGRev1.zip

- Additionalfile2CasanovaetalBMCGRev1.pdf
- Additionalfile1CasanovaetalBMCGRev1.pdf