BMC Genomics

RESEARCH Open Access

Facilitating population genomics of non-model organisms through optimized experimental design for reduced representation sequencing

Henrik Christiansen[1*], Franz M. Heindler[1], Bart Hellemans[1], Quentin Jossart[2], Francesca Pasotti[3], Henri Robert[4], Marie Verheye[4], Bruno Danis[5], Marc Kochzius[2], Frederik Leliaert[3,6], Camille Moreau[5,7], Tasnim Patel[4], Anton P. Van de Putte[1,4,5], Ann Vanreusel[3], Filip A. M. Volckaert[1] and Isa Schön[4]

Abstract

Background: Genome-wide data are invaluable to characterize differentiation and adaptation of natural populations. Reduced representation sequencing (RRS) subsamples a genome repeatedly across many individuals. However, RRS requires careful optimization and fine-tuning to deliver high marker density while being cost-efficient. The number of genomic fragments created through restriction enzyme digestion and the sequencing library setup must match to achieve sufficient sequencing coverage per locus. Here, we present a workflow based on published information and computational and experimental procedures to investigate and streamline the applicability of RRS.

Results: In an iterative process genome size estimates, restriction enzymes and size selection windows were tested and scaled in six classes of Antarctic animals (Ostracoda, Malacostraca, Bivalvia, Asteroidea, Actinopterygii, Aves). Achieving high marker density would be expensive in amphipods, the malacostracan target taxon, due to the large genome size. We propose alternative approaches such as mitogenome or target capture sequencing for this group. Pilot libraries were sequenced for all other target taxa. Ostracods, bivalves, sea stars, and fish showed overall good coverage and marker numbers for downstream population genomic analyses. In contrast, the bird test library produced low coverage and few polymorphic loci, likely due to degraded DNA.

Conclusions: Prior testing and optimization are important to identify which groups are amenable for RRS and where alternative methods may currently offer better cost-benefit ratios. The steps outlined here are easy to follow for other non-model taxa with little genomic resources, thus stimulating efficient resource use for the many pressing research questions in molecular ecology.

Keywords: Biodiversity, Genome scan, Genotyping by sequencing, In silico digestion, RADseq, Southern Ocean

* Correspondence: henrik.christiansen@kuleuven.be
[1]Laboratory of Biodiversity and Evolutionary Genomics, KU Leuven, Leuven, Belgium
Full list of author information is available at the end of the article

Background

Evolutionary and ecological population genetic studies are important to understand how the diversity of life on earth is distributed, has evolved and may respond to future environmental changes [1]. A grand challenge has been to document this biodiversity and understand its role in maintaining ecosystem functionality, particularly in the ocean [2] and even more so in frontier areas such as the deep-sea and polar regions [3]. Molecular data collection has benefitted from a revolution in sequencing technologies such that genomics, where billions of nucleotides are screened simultaneously, is now an integral part of the biological toolbox [4–6]. Genome-wide data open new avenues of ecological and evolutionary research, especially to study local adaptation [7, 8]. Given ever-increasing rates of anthropogenic disturbance, it is crucial to assess spatio-temporal genomic diversity, adaptation patterns and resilience of non-model organisms [9, 10].

Similar to previous methodology shifts in population genetics (e.g. from Amplified Fragment Length Polymorphisms [AFLP] to microsatellites), the transition to novel methods requires detailed understanding of the new technology, its potential as well as its pitfalls, and careful experimental planning. While some study systems are moving towards population-specific shallow re-sequencing of whole genomes (e.g. important commercial fish species) [11, 12], many species of interest with less extensive genomic resources rely on reduced representation sequencing (RRS) techniques to subsample the genome. Among the most popular RRS techniques are Restriction site-Associated DNA sequencing (RADseq) [13] and Genotyping by Sequencing (GBS) [14]. A plethora of similar methods with unique names have been developed [5, 15–18]. Here, we follow the reasoning of Campbell et al. [18] and use the term RRS [19] to refer to all of these methods, which are attractive because they make more frugal use of sequencing volume than whole genome sequencing.

In RRS, one or several restriction endonuclease enzymes are used to first fragment the target genome into smaller portions, thus reducing sequencing costs. Millions of reads from high-throughput sequencing platforms are then aligned against either a reference genome or, alternatively, a de novo reference catalog of loci [20, 21]. Subsequently, genetic variants, most commonly single nucleotide polymorphisms (SNPs) are determined. In addition, approaches have been developed to use RRS data to create microhaplotypes [22, 23], or identify microsatellites [24] or copy number variants (CNV). The latter relies on summary statistics of the frequency of heterozygotes and the read ratio, which should differ between singleton and duplicated SNPs [25, 26]. RRS has provided many important insights across a wide range of taxa from different ecosystems, e.g. with respect to population structure and demography, as well as hybridization, landscape or seascape genomics, QTL mapping, phylogeography, and shallow phylogenies (e.g. [5, 27–32]). Limitations and problems of RRS include the potential for allele dropout, PCR duplicates, genotyping error, as well as insufficient coverage and/or low marker density (the number of genetic markers that are sequenced in relation to genome size) [5, 33, 34]. Unnecessary costs, inability to answer the research, or, in the worst case, incorrect conclusions may be the consequence. Good experimental design, however, can help avoid or mitigate some of these issues.

Effective and cost-efficient RRS experiments must be well designed. First, one should establish whether the species of interest truly represents one species or if cryptic species are present. This can be problematic in non-model taxa and has potentially large downstream implications for RRS such as high divergence but few shared loci [35, 36]. A useful complement is therefore DNA barcoding to screen for cryptic species [37, 38]. Alternatively, RRS can be specifically employed for species delimitation purposes [30, 39, 40], but this should be a deliberate choice before designing the RRS setup. For such a scenario it would be especially important to sequence many fragments thereby increasing the likelihood of capturing genetic markers that are conserved across, yet discriminatory between species. In general, the research question fundamentally determines whether the application of RRS is appropriate. For example, providing evidence for significant, evolutionary neutral genetic population structure may be easier and less expensive with a good number (> 10) of multi-allelic microsatellites [41]. However, RRS may be better suited to identify loci that are putatively affected by spatially variable selection and therefore involved in local adaptation. To this end, the density of markers that can be realized for a given species, which depends on genome size and complexity, as well as research budget, should be considered.

With low marker density one may run the risk of accepting unreasonably high rates of false positives (outliers that are not based on biological reality) in genome scans leading to biased or erroneous inferences [42, 43]. Consequently, there is debate about the usefulness of RRS (or RADseq in particular), especially for inferring local adaptation patterns [34, 44]. The genomic characteristics of a target species, most importantly its genome size and the level of linkage disequilibrium (LD), are crucial to design a RRS experiment. With little genomic information, a priori calculations may be inaccurate. Therefore, it is vital to assess, optimize and critically ponder the advantages and limitations of RRS for a given research project to avoid the creation of sequence data that are unsuitable to answer the study question and/or inefficient use of resources. A most critical point is to properly strike a balance between sequencing depth (coverage) and number of fragments, which is roughly proportional to the number of genetic markers. The estimated number of fragments generated from a genome determines the marker density (as the number of fragments translates approximately into the number of SNPs), while avoiding unnecessary "over"-sequencing of the genomic fragments, i.e. loci or RADtags, to save sequencing costs. Both excessive (> 100×) and uneven or too low (< 10×) coverage is detrimental for accurate locus reconstruction and SNP calling, particularly in de novo approaches [45]. Hence, RRS experimental procedures may benefit from thorough optimization. In this context, we used the framework of a large research project ("Refugia and Ecosystem Tolerance in the Southern Ocean") to optimize RRS for a diverse set of taxa in parallel. The Southern Ocean hosts a unique marine fauna with high levels of endemism [46, 47], but is increasingly subject to external pressures, such as warming, pollution and living resource exploitation [48–51]. Population genomic approaches are needed to understand the genetic structure and connectivity of Antarctic fauna, so that appropriate management and conservation actions can be developed (e.g. [52–54]).

In this molecular pilot experiment, we seek to investigate and optimize the applicability of RRS to a range of Antarctic non-model taxa across the animal kingdom. The target organisms are ecologically important, abundant, and widely distributed in the Southern Ocean and cover a variety of habitats – from benthos to pelagic birds. Specifically, we aim to develop economic and robust experimental setups for RRS population genomic studies in an ostracod group, two amphipod species, two bivalve species, two sea star species, two fish species, and two bird subspecies (Table 1). The outlined approach should be readily adoptable for other taxa of interest. We lay out a clear and concise protocol to follow a priori for any RRS experiment on non-model species that will help researchers to evaluate the costs, benefits, and risks of such projects.

We specifically aim to (i) collate information about the genomic properties of the target taxa; (ii) assess in silico which restriction enzymes are likely to yield the desired number of fragments; (iii) test selected restriction enzyme digestions in the laboratory; (iv) optimize restriction enzyme choice, size selection window and the number of individuals to be pooled per sequencing library (based on the previous results); and (iv) sequence and analyze test RRS libraries of promising experimental setups. These extensive pilot analyses – including literature research, computational analyses, and laboratory work – are designed to comprehensively evaluate all information for each target species or species complex. In the workflow of optimizing the setup for each target taxon, we strive to use the same restriction enzymes (or combinations) for several taxa whenever possible to reduce the costs for specifically designed barcodes and adaptors. Results shall ultimately facilitate informed decisions about whether and how RRS for each taxon could be conducted. We critically discuss these considerations and suggest alternative approaches in two cases.

Results

The optimization process of RRS experimental setups for non-model species is iterative and includes many deliberate choices that must be made based on the best available knowledge. Relatively constant variables, i.e. the number and quality of samples, the research budget and the main research question, should be considered during the entire process and flexible variables, such as restriction enzymes, size selection window and the number of

Table 1 Target taxa for a molecular pilot experiment to test and optimize the experimental setup for reduced representation sequencing (RRS)

Class	Family	Target Species	Authority
Ostracoda	Macrocyprididae	*Macroscapha opaca-tensa* species complex	Brandão et al., 2010 [55]
Malacostraca	Lysianassidae	*Charcotia obesa*	Chevreux, 1906
Malacostraca	Eusiridae	*Eusirus pontomedon*	Verheye & D'Udekem D'Acoz, 2020 [56]
Bivalvia	Laternulidae	*Laternula elliptica*	King, 1832
Bivalvia	Sareptidae	*Aequiyoldia eightsii*	Jay, 1839
Asteroidea	Astropectinidae	*Bathybiaster loripes*	Koehler, 1906
Asteroidea	Astropectinidae	*Psilaster charcoti*	Koehler, 1906
Actinopterygii	Nototheniidae	*Trematomus bernacchii*	Boulenger, 1902
Actinopterygii	Nototheniidae	*Trematomus loennbergii*	Regan, 1913
Aves	Procellariidae	*Pagodroma nivea nivea*	Forster, 1777
Aves	Procellariidae	*Pagodroma nivea confusa*	Clancey, Brooke & Sinclair, 1981

individuals to be pooled, should be adjusted to reach the desired outcome.

Genome characteristics

First, published genomic resources of our target taxa were collected. Available information is highly variable across target taxa, with typically more genomic resources available for vertebrate groups (Table 2). Genome size among ostracods varies considerably, with Macrocyprididae estimated at approximately 166 Mb (or 0.17 C) [68]. One published ostracod genome (*Cyprideis torosa*) with a genome size comparable to our target species was available [57]. Amphipods show very large variability in genome size [72] with extreme cases that dramatically exceed the size estimates of all other target taxa studied here (up to 63,198 Mb or 64.62 C, Table 2). Two amphipod reference genomes were available (*Hyalella azteca*, *Parhyale hawaiensis*) [58, 59]. In addition to these reference genomes we simulated a large (10,000 Mb) and a very large (30,000 Mb) genome for amphipods, according to genome size estimates from species of the same family (Eusiridae: 7.16 C, Lysianassidae: 27 C) [73]. For bivalves and sea stars more reference genomes were available, but not from species closely related to the target species. In both cases, we selected three reference genomes of varying size (Table 2). The Antarctic fish target species have genome size estimates available as well as a reference genome from a species from the same family (*Notothenia coriiceps*) [66]. In birds, no genome size estimates for our target species were available, but bird genome size appears to be relatively constrained between approximately 1 and 2 Gb and a reference genome from the same family has been published (*Fulmarus glacialis*) [67]. We decided to aim at 50,000 fragments as initial targets for our optimizations in all taxa, except for fish and sea stars (Table 2). In the latter

Table 2 Genomic information useful for reduced representation sequencing (RRS) optimization in target species from six organism classes

Class	Target Species	Target Fragment Number	Genome Size Estimates (C)	Genome from Related Species, Genome Size (Mb), Accession Nr., and Reference	Simulated Genomes
Ostracoda	Macrocyprididae	50,000	0.17 ± 0.003[a]	*Cyprideis torosa*, 286 Mb, GCA_905338395.1 [57]	100 Mb, 43.9% GC 500 Mb, 43.9% GC
Malacostraca	*Charcotia obesa* and *Eusirus pontomedon*	50,000	unknown (Amphipoda: 0.68–64.62[a])	*Hyalella azteca*, 551 Mb, GCA_000764305.2 [58] *Parhyale hawaiensis*, 4003 Mb, GCA_001587735.2 [59]	10,000 Mb, 38.5% GC 30,000 Mb, 40.8% GC
Bivalvia	*Laternula elliptica* and *Aequiyoldia eightsii*	50,000	unknown (0.65–5.40[a])	*Crassostrea gigas*, 558 Mb, GCA_000297895.2 [60] *Pinctada imbricata*, 991 Mb, GCA_002216045.1 [61] *Bathymodiolus platifrons*, 1658 Mb, GCA_002080005.1 [62]	1000 Mb, 35.3% GC 5000 Mb, 34.2% GC
Asteroidea	*Bathybiaster loripes* and *Psilaster charcoti*	20,000	unknown (Asteroidea: 0.54–0.96[a])	*Acanthaster planci*, 383 Mb, GCA_001949145.1 [63] *Patiria miniata*, 811 Mb, GCA_000285935.1 [64] *Patiriella regularis*, 949 Mb, GCA_900067625.1 [65]	1000 Mb, 41.3% GC 2000 Mb, 40.4% GC
Actinopterygii	*Trematomus bernacchii* and *T. loennbergii*	20,000	*T. bernacchii*: 1.12 ± 0.019[b]; 1.19[c]; 1.82[d]; *T. loennbergii*: 1.34[b]	*Notothenia coriiceps*, 637 Mb, GCA_000735185.1 [66]	1000 Mb, 40.8% GC 1800 Mb, 40.8% GC
Aves	*Pagodroma nivea nivea* and *P. nivea confusa*	50,000	unknown (0.91–2.16[a])	*Fulmarus glacialis*, 1141 Mb, GCA_000690835.1 [67]	1500 Mb, 41.2% GC 2000 Mb, 41.2% GC

For each class approximate targets for the number of fragments were defined and known genome size estimates from flow cytometry are listed. In species with unknown genome size, the range of published estimates from species from the same class is listed. Available genomes from related species and two simulated genomes per class were used for in silico digestions. The simulated genomes were simulated using the SimRAD R package based on two realistically large genome sizes with a GC content as known from related species

[a] published estimates from various species of the same class (or where indicated order), as listed on genomesize.com on 9th January 2019; Ostracoda: Macrocyprididae: Jeffery et al., 2017 [68]

[b] Auvinet et al., 2018 [69]

[c] Hardie and Hebert, 2003 [70]

[d] Morescalchi et al., 1996 [71]

Table 3 Restriction enzymes and combinations used for reduced representation sequencing (RRS) optimization

Restriction Enzyme (Combination)	Recognition Site	Approximate Fragment Number[a]	Special Features	Reference
SbfI	5′--CCTGCA|GG--3′	6000		e.g. [39, 74–76]
EcoRI	5′--G|AATTC--3′	323,000	Methylation sensitive	[13, 77]
SphI	5′--GCATG|C--3′	143,000		
PstI	5′--CTGCA|G--3′	145,000		[78, 79]
ApeKI	5′--G|CWGC--3′	940,000	Methylation sensitive, degenerate site	e.g. [80–83]
MspI	5′--C|CGG--3′	1,590,000		
MseI	5′--T|TAA--3′	8,100,000		[84]
SbfI_SphI		11,000		e.g. [33, 85–87]
SbfI_MspI		11,000		[88, 89]
PstI_MspI		265,000		e.g. [17, 90–92]
EcoRI_SphI		244,000		[93]
EcoRI_MspI		536,000		e.g. [94–97]

Recognition site, the approximate expected fragment number in a 1000 Mb genome, any special enzyme characteristic and empirical studies that recently used this enzyme (combination) are listed.

[a] For a 1000 Mb genome with 40% GC content and no size selection and rounded to the nearest thousand. Note that the double digest estimates are for ddRAD protocols where fragments with two different restriction sites but irrespective of orientation are retained. In the two enzyme GBS protocol this number would be halved as only fragments with the first restriction site first and the second restriction site second (and not vice versa) are retained during library construction. For more details see Peterson et al. (2012) [15] and Poland et al. (2012) [17]

target taxa, we aimed at 20,000 fragments initially, because we had more samples available and thus were interested in covering more individual samples from a wider geographic range at the expense of marker density. Note that these targets are highly study specific and depend on the budget, number of samples to be sequenced and, most importantly, exact research question of a given RRS project.

In silico digestions

We estimated how many RRS fragments twelve restriction enzymes and enzyme combinations (listed in Table 3) would produce. These estimates were conducted using various reference genomes and simulated genomes. We estimated the fragment number in total as well as in various size selection windows (see below and Additional File 2). As expected, the fragment number is influenced primarily by the type of enzyme and the genome size. The tested combinations showed promising results with fragment numbers close to our defined targets in all species, but there was not one single enzyme or enzyme combination that produced promising results in all species. We aimed at using as few different enzyme setups across species as possible and in addition opted for enzymes or enzyme combinations that were previously used in our laboratory as much as possible. Using the same setup for several RRS experiments reduces costs as the same adaptor sets can be reused multiple times. Therefore, we kept five initial setups that yielded promising fragment numbers: *EcoRI*, *PstI*, *ApeKI*, *MspI* and a double digest with *EcoRI* and *MspI*.

Empirical digestions

Based on preliminary in silico results, we tested the genome digestion by several enzymes and enzyme combinations in the laboratory. High quality bird DNA was not available, preventing empirical digestion tests for this group. Ostracod DNA was whole genome amplified and this proved problematic for the Bioanalyzer instrument, because the results indicated overloading even after multiple dilutions. In total, 75 empirical digestions were conducted, several of which produced unusable results even after repeating the experiment. The sensitivity of the Bioanalyzer to small irregularities especially in the size range below 500 bp made it impossible to infer sensible patterns in many cases (Additional File 3). We therefore must caution that Bioanalyzer results only sometimes provide useful additional information that increase the confidence in estimates obtained in silico. Nevertheless, from the successful runs it appeared that the empirical results are usually more similar to in silico digestions with genomes from related species than of simulated genomes (Fig. 1 and Additional File 3). For example, in fishes *ApeKI* was estimated to produce significantly more small than large fragments using the *N. coriiceps* reference genome, which was at least roughly confirmed through empirical digestion (Fig. 1). Here, using *EcoRI* provides few fragments overall, which

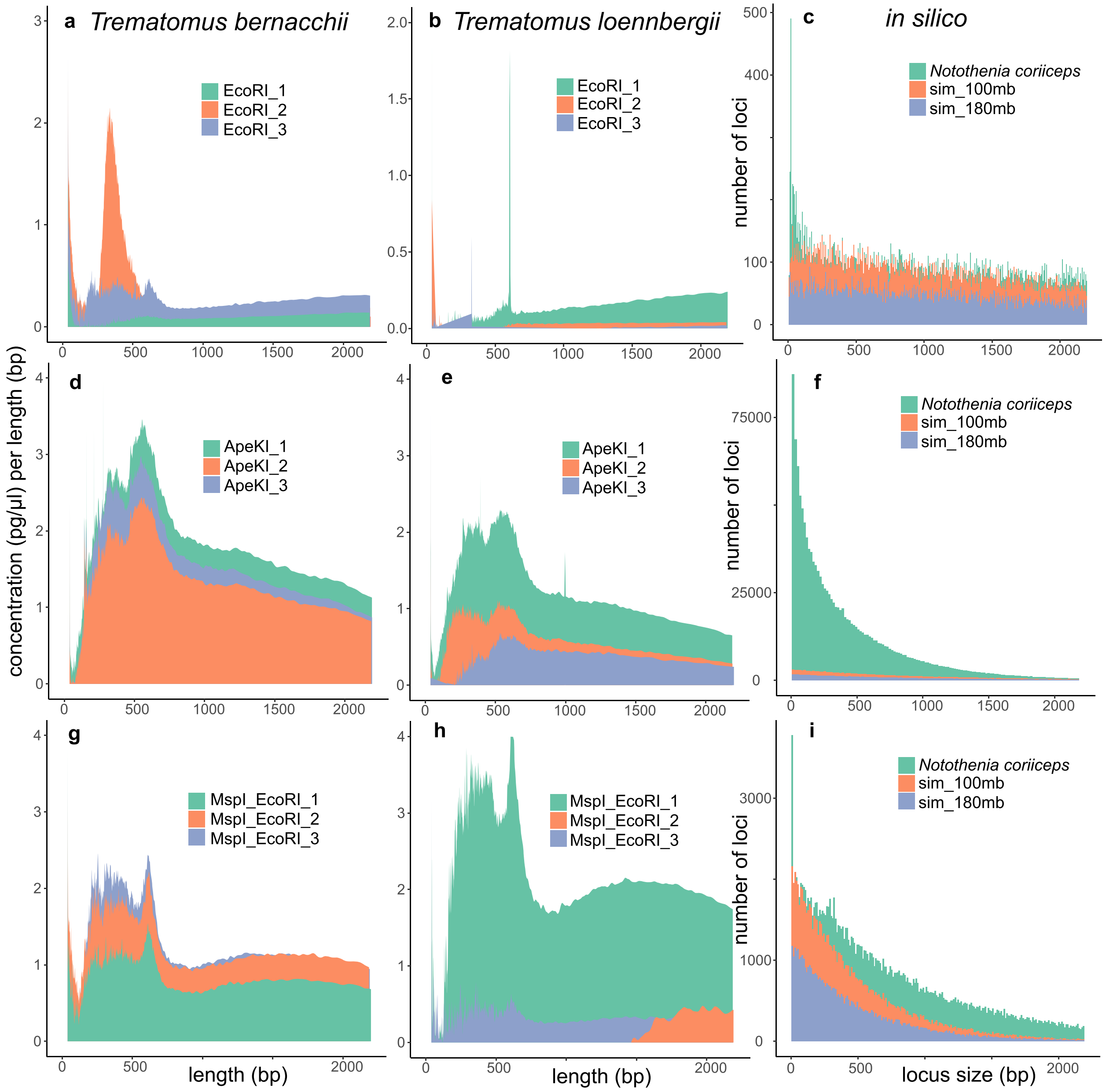

Fig. 1 Comparison of empirical and in silico restriction enzyme digestions. Empirical Bioanalyzer results with digested DNA are shown as concentration over fragment size (**a**, **b**, **d**, **e**, **g**, **h**) and estimated loci numbers over locus size from in silico digestions (**c**, **f**, **i**). The tests were conducted with restrictions enzymes *EcoRI* (**a**, **b**, **c**), *ApeKI* (**d**, **e**, **f**) and a double digest with *EcoRI* and *MspI* (**g**, **h**, **i**). Results for the fish species *Trematomus bernacchii* (**a**, **d**, **g**) and *T. loennbergii* (**b**, **e**, **h**) are shown next to in silico estimates using a related reference genome of *Notothenia coriiceps* and two simulated genomes of 100 and 180 Mb size (note that this was the absolute size used for in silico computations, but resulting estimates were extrapolated to 1000 and 1800 Mb). Results for other target taxa are shown in Additional File 3

proved difficult to accurately depict using the Bioanalyzer. In contrast, the tested double digest provided a consistent picture in four out of six replicates for the two fish species (Fig. 1). Here, we also noted a pronounced spike at around 650 bp and the size window was therefore deliberately kept lower (see below and Tables 4 and 5).

RRS setup

With all information gathered thus far, we proceeded to optimize the RRS experimental setup for each of the target taxa. We planned the same setup for species from the same class, when the genomic differences between those species were unknown (in Bivalvia and Asteroidea), or when they were related and therefore likely to

Table 4 Reduced representation sequencing (RRS) setups for seven individually optimized protocols

Class	Target Species	Restriction Enzyme (Combination)	Size Window (bp)	Assumed Genome Size (Mb)	Coverage[a]	Marker Density[a] (bp per 1 SNP)
Ostracoda	Macrocyprididae	*ApeKI*	200–350	250	31.9×	1533
Malacostraca	*Charcotia obesa*	*SbfI_MspI*	200–320	27,000	32.5×	168,503
	Eusirus pontomedon	*EcoRI_SphI*	200–260	7000	32.8×	44,045
Bivalvia	*Laternula elliptica* and *Aequiyoldia eightsii*	*ApeKI*	200–260	3000	30.2–39.0×	17,385 – 22,472
Asteroidea	*Bathybiaster loripes* and *Psilaster charcoti*	*ApeKI*	200–300	500	27.1–33.5×	2598 – 3212
Actinopterygii	*Trematomus bernacchii* and *T. loennbergii*	*EcoRI_MspI*	200–450	1500	27.5×	7352
Aves	*Pagodroma nivea nivea* and *P. nivea confusa*	*PstI*	200–300	1500	31.4×	9056

These setups were optimized in order to be run on a HiSeq 2500 platform (Illumina). The choice of restriction enzyme(s) and size window was optimized to obtain approximately 30× coverage (or half that value in a worst-case scenario) with the assumed genome size (conservatively estimated based on available information, see Table 2). Marker density (the number of bp per sequenced SNP) was estimated as a comparable measure to the metastudy by Lowry et al. (2017) [34]

[a] assuming 200 million reads of 125 bp length spread over 96 individuals and 0.01 SNP/bp

have similar genomic properties (Actinopterygii and Aves). In contrast, we designed two different setups in Malacostraca, because the genomes of *C. obesa* and *E. pontomedon* may have very different sizes (Tables 2 & 4). Experimental setups, i.e. restriction enzymes and size selection window, were furthermore tuned to suit a sequencing experiment with the HiSeq 2500 or 4000 platforms, respectively. The choice of the sequencing platform can be modified based on instrument availability and budget. In the following, results for use with a HiSeq 2500 platform are listed (Table 4), the same results for a HiSeq 4000 platform can be found in Additional File 4 (further calculations e.g. aimed at using a NovaSeq platform can be obtained by adjusting the R code; see: https://github.com/notothen/radpilot). The setup for optimizing results as listed here also includes the consideration that it would be cost-efficient to use the same enzyme or enzyme combinations for several species whenever possible, because adaptors can then be reused. Therefore, when several enzymes (or combinations) seemed promising according to in silico digestion, we attempted to choose setups that were also promising in other target species. For ostracods, we assumed a genome size of 250 Mb and 500 Mb as worst-case scenario. Using the *C. torosa* reference genome, a digest with *ApeKI* and size selection of 200–350 bp would yield 31.9× coverage (or half of that in the worst-case scenario). With this setup and genome size, we would achieve an estimated marker density of approximately one SNP every 1.5 kb. In amphipods, different setups per species are required. Given the highly uncertain genome size of 27,000 Mb for *C. obesa* and 7000 Mb for *E. pontomedon* (based on same family estimates) [72], double digest RADseq experiments with *SbfI* and *MspI* and *EcoRI* and *SphI*, respectively, would yield the desired coverage. Marker density in both cases is expected to be low, due to the large genome size (Table 4). Because of uncertainty with respect to genome size and an anticipated low marker density, we stopped RRS optimization in amphipods and instead explored alternatives. For both bivalve species, a genome digestion with *ApeKI* and size selection of 200–260 bp seemed promising with all three reference genomes and would yield around one SNP per 20 kb. Similarly, in sea stars we found setups with *ApeKI* and a slightly wider size selection that should yield good results, although results varied depending on the reference genome used. For the Antarctic fishes of the genus *Trematomus*, a double digest setup with *EcoRI* and *MspI* in a size window of 250–450 bp should yield desired coverage and marker density. Regarding the snow petrels, a setup with *PstI* and 200–300 bp size selection seemed appropriate, yielding one SNP every 9 kb. Overall, results indicate that with only three enzyme choices, it should be possible to achieve the desired coverage and marker density in five of our six target classes (excluding Malacostraca as discussed above) (Table 4).

RRS test libraries

Pilot libraries with optimized setups were sequenced, yielding a total of 531 million (M) reads. After demultiplexing and quality control, 422 M reads were retained. These reads were spread relatively evenly across libraries, species, and individuals (average and standard deviation across all taxa and libraries: 4.5 ± 2.1 M reads). All but five individuals received more than 1 M reads and most individuals received more than 3 M reads. We created de novo catalogs from these reads using Stacks [21, 98, 99] with varying *M* and *n* parameters [45]. Optimal parameters varied ($M = n = 3–6$) among taxa (Table 5 & Additional File 7). Results from this parameter

Table 5 Setup and results of five test libraries for reduced representation sequencing (RRS) from eight species/groups

Library Nr.	Class	Target Species	Protocol, Enzyme and Size Window (bp)	N + controls	Stacks parameter *M* and *n*	Expected Nr. of Fragments[a]	Obtained Loci[b]	Expected Coverage	Obtained Coverage[b]
1	*Ostracoda*	Macrocyprididae	GBS, *ApeKI*, 200–350	8 + 2	6	65,244	69,817 (± 63,114)	31.9×	28.2× (±5.4)
2	*Bivalvia*	*Laternula elliptica* *Aequiyoldia eightsii*	GBS, *ApeKI*, 200–260	8 + 2 10 + 2	4	53,399 – 69,027	125,305 (± 22,828) 143,551 (± 28,676)	30.2–39.0×	21.6× (±5.1) 20.0× (±2.6)
3	*Asteroidea*	*Bathybiaster loripes* *Psilaster charcoti*	GBS, *ApeKI*, 200–300	10 + 2 14 + 2	5	62,272 – 76,988	82,945 (± 43,521) 115,608 (± 30,589)	27.1–33.5×	21.0× (±6.8) 27.6× (±4.6)
4	*Actinopterygii*	*Trematomus bernacchii* *T. loennbergii*	ddRAD, *EcoRI_MspI*, 200–450	10 + 2 10 + 2	3	81,605	21,121 (± 3539) 23,609 (± 2362)	27.5×	42.3× (±13.5) 49.6× (±6.2)
5	*Aves*	*Pagodroma nivea nivea*	GBS, *PstI*, 200–300	6 + 2	3	66,258	140,972 (± 26,444)	31.4×	10.0× (±0.4)

Restriction enzyme(s) and size window was optimized for number of fragments and coverage as in Table 4, these estimates are listed here again as expected values and compared to empirical results regarding average (and standard deviation of) number of loci and average (and standard deviation of) coverage per sample based on data processing using Stacks v2.4 with optimized parameters

[a] only estimates from real (not from simulated) genomes listed

[b] as reported in the output file gstacks.log.distribs after using denovo_map.pl with $m = 3$ and $M = n$ as listed in column six

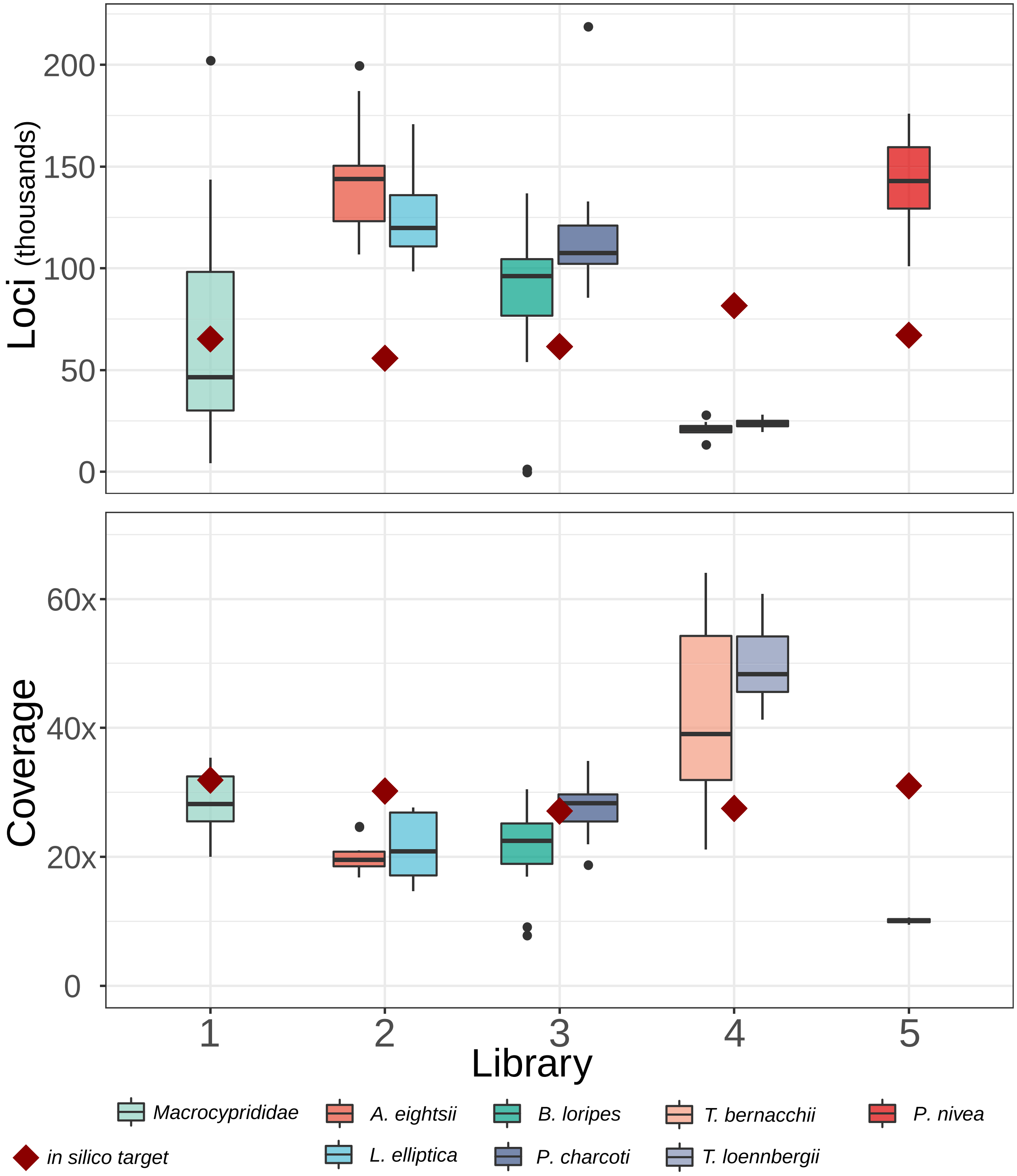

Fig. 2 The number of loci and coverage as estimated and realized in five test libraries. In silico estimates (dark red diamonds) and empirical values from reduced representation sequencing (RRS) libraries containing DNA from eight target species are shown. Boxplots show the median, quartiles, and outliers across individuals (*N* = 8–14). Libraries were prepared as listed in Table 5

optimization also revealed varying levels of diversity, e.g. sea stars showed relatively high levels of polymorphism, while the bird library produced many loci but few SNPs (Additional File 7). Comparing the unfiltered numbers of loci and coverage across individuals underlined the inverse relation of these two variables (Table 5, Fig. 2). In ostracods, our target estimates were matched best. In bivalves and sea stars, more loci than expected were

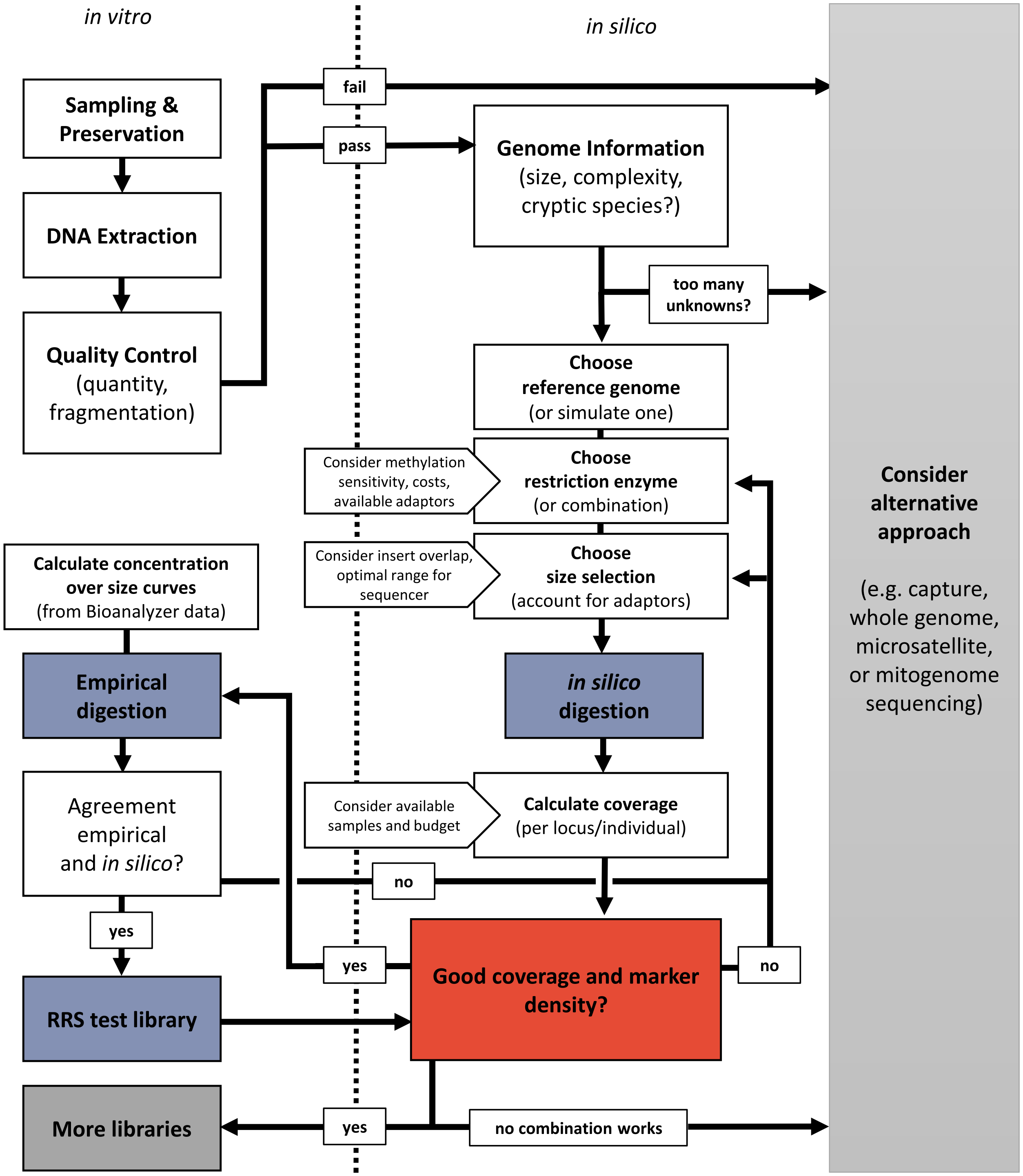
in vitro
in silico
Sampling & Preservation
DNA Extraction
Quality Control (quantity, fragmentation)
fail
pass
Genome Information (size, complexity, cryptic species?)
too many unknowns?
Choose reference genome (or simulate one)
Consider methylation sensitivity, costs, available adaptors
Choose restriction enzyme (or combination)
Consider insert overlap, optimal range for sequencer
Choose size selection (account for adaptors)
Calculate concentration over size curves (from Bioanalyzer data)
Empirical digestion
in silico digestion
Agreement empirical and in silico?
Consider available samples and budget
Calculate coverage (per locus/individual)
no
yes
yes
Good coverage and marker density?
no
RRS test library
More libraries
yes
no combination works
Consider alternative approach
(e.g. capture, whole genome, microsatellite, or mitogenome sequencing)

Fig. 3 The iterative process of reduced representation sequencing (RRS) optimization. Empirical (in vitro, left of the dotted line) and computational (in silico, right of dotted line) analyses are part of this process. Core procedures to identify suitable experimental setups are in silico and empirical enzyme digestion and sequencing of a pilot RRS library (blue boxes). The coverage and marker density that can be achieved with a given setup needs to be repeatedly checked and fine-tuned (red box). We used 30 × coverage and a density of at least one SNP per 10,000 bp as target values but emphasize that these values need to be defined depending on the specific goals of a research project

sequenced at the expense of coverage, although coverage was still reasonable. Two individuals of *B. loripes* had low coverage due to low initial numbers of reads, indicating errors during library preparation or degraded input DNA. The fish libraries contained considerably less loci than expected at high coverage, while the opposite was true for the bird library. The latter also showed very uniform low coverage at approximately 10×. Overall, these results show promise for full scale RRS libraries with sufficiently high coverage in four of five libraries.

Discussion

High-throughput sequencing methods promise new avenues of ecological and evolutionary research in non-model organisms. We provide a detailed workflow to evaluate and optimize reduced representation sequencing (RRS) techniques for any animal species of interest (Fig. 3). This approach is reproducible and ensures that researchers are well-informed about the advantages and drawbacks of RRS for their research question. Different RRS setups (i.e. various species and libraries constructed via different protocols, enzymes and size selection windows) were successfully sequenced together on one HiSeq lane. Most individuals included in this multi-library-multiplex received adequate sequencing effort, which has been problematic in other studies that pooled individuals directly after ligation [33]. From our experience (including this and previous studies in our laboratory; see e.g. [33, 83, 87]) it seems that careful, repeated quantification and standardization of DNA from every individual before and after PCR are key to achieve equivalent sequencing effort across individuals. A pilot sequencing experiment can then yield valuable insights before proceeding with sequencing at a larger scale. Here, more loci than expected were assembled in most taxa (ostracods, bivalves, sea stars) at sufficiently high per locus coverage. This highlights the value of choosing parameters conservatively, e.g. under- rather than overestimating the number of sequencing reads. The fish library yielded fewer loci than expected at higher coverage. Pooling more individuals, increasing the size window, or changing the restriction enzyme setup altogether including new optimization are future options to further optimize this project, although the current setup also yields useful data. The bird library produced coverage that is directly at the advised limit of 10× [45]. This may be partly related to low quality input DNA, which was mostly extracted from feathers. Alternative sampling and/or DNA extraction protocols and further testing are needed before sequencing full scale libraries for snow petrels. Overall, a few key properties determine the feasibility and cost of RRS in non-model organisms.

Predictability of reduced representation experiments

Planning a genome reduction through restriction enzyme digestion starts with an imperative question: how large is the target genome? Non-model species often lack information on genome size, which complicates RRS optimization [100]. If genome size appears relatively conserved across species within a taxonomic class (e.g. Asteroidea), it can be assumed that the species of interest from this class has similar genome size. Some imprecisions regarding the exact size have only limited effects on overall accuracy. Alternatively, in other groups, such as amphipods, genome size is highly variable, spanning two orders of magnitude [72, 73, 101]. In this case, using an inaccurate genome size estimate has the potential to dramatically impact the parameters one aims to optimize. In addition, very large genomes are often highly repetitive, which significantly hampers downstream bioinformatics and population genomics [74, 102, 103]. Therefore, with the current state of knowledge, we opted to exclude amphipods from our trial RRS libraries. Estimating genome size with flow cytometry or conducting a series of test libraries could be alternative ways forward.

For ostracods, bivalves, sea stars, and birds more loci were found than expected. This might indicate that genome sizes were consistently larger than expected. Another, likely explanation is that the enzymes used (*ApeKI*, *PstI*) produce more fragments than in silico digestions predicted. For example, the number of fragments resulting from four base cutters may be more difficult to predict as they sometimes produce so many fragments that effectively the entire genome would be sequenced [100]. The five-base recognition site of *ApeKI* features a degenerate base, which may have a similar effect. The methylation sensitivity of *ApeKI* may also provide more genomic markers in genic regions [104]. It is unclear, however, how general this prediction holds across metazoans. Finally, some of the excess loci recovered may be artefacts from library preparation, PCR duplicates, or incorrect locus assembly [21]. Rigorous downstream filtering and/or comparison of several, differently filtered datasets may help determine the true biological signal. Whatever the reason, the higher-than-expected number of loci still lead to sufficient coverage, except in the bird library. The latter is likely related to low quality/quantity of input DNA. Few bird samples were available, some only as feathers, which yielded very little DNA. Whole genome amplification (WGA) could be an option to increase yield for RRS as successfully applied in ostracods (this study) and insects [105].

Finally, even with reliable genome size estimates and well-tested enzymes, the empirical results may differ from in silico expectations. In *Trematomus* fishes, approximately half of the expected sites were found, despite well-known genome size [69–71]. Genomic architecture may play an important role in affecting the number of cut sites per restriction enzyme. We used the draft genome of a related species from the same family to estimate the number of fragments. The endemic Antarctic notothenioid fishes, however, are characterized by frequent chromosomal rearrangements and large numbers of transposable elements [69, 106, 107]. The genus *Trematomus* constitutes an example of a relatively recent marine adaptive radiation [108, 109]. Therefore, in this particular case, the genome of a closely related species may provide relatively poor accuracy for cut site estimations.

We have tested various enzymes and enzyme combinations that have been successfully used in RRS studies (Table 3). Yet, many previous studies achieved overall relatively little marker density, which is problematic if looking for genome-wide adaptation patterns [34]. With increasing output of sequencers, aiming at higher marker density is not an unachievable goal. Genome size, restriction enzyme characteristics and genomic complexity influence the predictability. Altogether, our results highlight the importance of conducting test libraries before embarking on larger, multi-library sequencing projects. In our case, *ApeKI* together with a narrow size window seems robust and powerful to create many genomic fragments (and thus sufficiently high marker density) across taxa with small to medium genome size. Using the same restriction enzyme for several projects drastically reduces cost as the same custom-made barcodes and adaptors can be used.

Decision making for population genomics

As we illustrated here, there are many experimental choices that may lead to inefficient or "broken" [34] RRS experiments. Given the publication bias towards successful applications [110], it is likely that a large number of unsuccessful applications of this technology to non-model species exist. It is crucially important that researchers actively engage in the decision-making process when choosing restriction enzymes, size selection windows, and the number of individuals to be pooled per sequencing lane. Furthermore, the research objectives and budget should be critically evaluated and matched. In other words, investigating genome-wide polygenic adaptation patterns in a non-model species with large, complex genome may simply not be feasible on a small budget. The number of individuals to be included is another aspect that weighs in on these considerations and latest developments in the field enable the inclusion of this parameter in in silico simulations [111]. In situations where sampling is not restricted, inferences of spatial genetic structure for example may benefit more from wider geographic sampling coverage than from higher marker density. If sampling more localities is unfeasible as may be the case in the Antarctic realm, it can be beneficial to instead invest in high density sequencing (as in several markers per linkage group). With sufficient genome coverage even advanced coalescent modeling is possible using RRS data [112].

We recommend following a few guiding principles when planning RRS for population genomics (but see also e.g. [5, 45, 100]). First, clear targets with respect to the number of individuals to be screened in a project (and/or in follow-up projects) and the marker density needed for the research objective should be defined. Determining the necessary marker density is difficult and depends on the degree of linkage disequilibrium [34]. We aimed for and achieved in three out of five cases a marker density of at least one SNP per 10,000 bp or 100 SNPs per Mbp. How valuable these marker densities are will only become apparent after full scale sequencing projects and depends also on factors that cannot be controlled through the experimental setup. However, our optimization approach yielded marker densities considerably higher (median 68 SNPs per Mbp) than in the survey of Lowry et al. (2017) [34] (median 4.08 tags per Mbp). Second, in silico estimations of how these targets can be reached and approximations of the associated costs should be obtained. The number of markers and individuals must be matched to reach a certain coverage (e.g. an average target of 30×). Subsequently, it is useful to briefly evaluate the trade-offs and benefits of RRS and other methods. If a promising combination of RRS method, enzyme, size selection, and sequencing effort is found, it is often worthwhile to conduct a pilot experiment before running the full sequencing experiment (Fig. 3). However, it is also advisable to stick to one approach afterwards and not change for example the sequencing platform, the size window or other properties of the setup that will otherwise reduce comparability between datasets. Finally, it is also important to thoroughly test and optimize the bioinformatic processing and data filtering to obtain a robust population genomic dataset [21, 113].

Alternative approaches

In some cases, RRS might not be the right choice for molecular ecological research (Fig. 3). A plethora of other genomic or genetic methods exists, which may offer more appropriate cost-benefit ratios. SNP genotyping arrays are a common and highly reproducible alternative, but usually only for species with more genomic resources (which exist for some Antarctic taxa; see e.g. [114]). Similarly, whole genome resequencing is providing the most extensive datasets which can be used for a wide range of analyses [11, 12, 115]. However, this is still too costly for many research projects, especially if information across many individuals and/or localities is needed. Another option is to focus on the expressed part of the genome and use a form of sequence capture enrichment (e.g. [103, 116, 117]) or RNAseq [118], or both [119, 120]. These approaches are versatile and can provide valuable information, even for museum samples [121, 122]. However, substantial expertise and prior investment in the development of custom methods is necessary for species that have not been investigated yet. With a limited budget and research objectives that do not depend on whole genome scans for selection, more classical molecular approaches are sometimes a good alternative. Nuclear microsatellite markers remain powerful to describe population structure and can be multiplexed and screened in large numbers. These markers can also benefit from high-throughput sequencing [123, 124]. Mitogenome sequencing and assembly using long-range PCR is another useful approach, particularly for phylogeographic applications [125, 126]. The amphipod and bird species evaluated here may currently be more amenable to such methods instead of RRS.

Conclusions

An extensive evaluation and optimization protocol allowed us to identify whether RRS is a suitable option for population genomics in a range of Antarctic animals. We have achieved promising results in some classes (ostracods, bivalves, sea stars, and fishes) that will be further developed soon. In other cases (amphipods and birds/degraded samples) alternative strategies such as mitogenome, capture sequencing or microsatellites seem more appropriate. The detailed considerations outlined here are a guideline for researchers to make informed decisions about the use of RRS or alternative methods. This is particularly important for species where genomic information remains scarce.

Methods

Specimen sampling

Samples of all target species were available from recent expeditions to the Southern Ocean (Additional File 1). For ostracods, we used existing DNA extractions of Macrocyprididae from the Southern Ocean that were already taxonomically identified and described [55, 127]. The amphipod target species were collected during RV *Polarstern* [128] expedition ANTXXIX-3 PS81. More details on *Eusirus pontomedon* (note that we initially included these specimens tentatively as *Eusirus* aff. *perdentatus*, but the taxonomy was updated during the course of this project) are provided in [56], while details of investigated *Charcotia obesa* are given in [129]. The bivalves *Laternula elliptica* and *Aequiyoldia eightsii* were sampled by scuba diving in the shallow water of Potter Cove (King George Island, western Antarctic Peninsula; by F. Pasotti) and Rothera station (Adelaide Island, West Antarctic Peninsula; courtesy of the British Antarctic Survey) in 2016. Two sea star species (*Bathybiaster loripes* and *Psilaster charcoti*) were collected during international expeditions with RRS *James Clark Ross* and RV *Polarstern* to the South Orkney Islands (JR15005 in 2016, PS77 in 2011), the Weddell Sea (PS81 in 2013), West Antarctic Peninsula (PS77 in 2011), and with RV *L'Astrolabe* to Adélie Land (REVOLTA 1 in 2010). Emerald rockcods (*Trematomus bernacchii*) were sampled in 2014 around James Ross Island with gill nets [130]. Scaly rockcods (*Trematomus loennbergii*) were sampled in the Ross Sea as bycatch of the exploratory Antarctic toothfish (*Dissostichus mawsoni*) longline fishery. Dead birds and feathers of snow petrels (*Pagodroma* spp.) were sampled during the BELARE 2017–2018 expedition in the vicinity of the Princess Elisabeth Station, and additional samples were obtained from Signy and Adelaide Islands as courtesy of the British Antarctic Survey. Samples were stored frozen, dried, or in > 90% ethanol until DNA extraction.

Genomic resources

Prior to computational analyses, genomic information was collated for all target species or, if such information was not available, from the closest related species. Published reference genomes were collected from the literature and online resources, such as GenBank and Ensembl [131]. In addition, genome size estimates were retrieved from genomesize.com [132] and other published estimates based on flow cytometry (e.g. [69]). Genome size estimates as C values were transformed to Mb for comparison (1 pg = 978 Mb) [133].

In silico genome digestion analyses

We used SimRAD to computationally digest genomic DNA at sites matching a restriction enzyme recognition site [134]. In total, seven restriction enzymes and combinations thereof were tested (Table 3). These were chosen based on what is commonly used in comparable studies and to cover a variety of enzymes ranging from very common (*MseI*, *MspI*, *ApeKI*) to medium (*EcoRI*, *SphI*, *PstI*) and rare cutters (*SbfI*). Reference genomes from related species as well as two simulated genomes per taxonomic class were used for these in silico digestions. Simulated genomes were generated randomly using SimRAD, but with GC content as in the available reference genome(s) and with two different fixed sizes per taxonomic class to cover the approximate range of genome sizes known for this class (Table 2). The total number of fragments that these enzymes (or enzyme combinations for double digest setups) produced were estimated, as well as the number of fragments in various size selection windows (between 210 and 260, 240–340, 0–100, 100–200, 200–300, 300–400, 400–500, 500–600, 600–700, 700–800, and 800–900 bp). Approximate targets for the number of fragments in each species of interest were defined (Table 2) and restriction enzyme and size selection combinations that provided fragment numbers close to our target numbers (50,000 ± 10,000 or 20,000 ± 10,000) were retained for downstream testing. After narrowing down the enzyme choice and conducting empirical digestion analyses, we ran additional in silico digestions for a final optimization of the size window and thus number of fragments for each specific case. During these fine-tuning analyses we tested as many different size selection windows as needed (in some cases > 20 additional size windows between 50 and 250 bp width) to find a suitable estimate of the number of fragments.

Empirical genome digestion analyses

Laboratory experiments were conducted with promising restriction enzymes to complement results from computational analyses. For each species, DNA from three individuals was used to test two or three restriction enzymes or enzyme combinations. Genomic DNA was extracted using either the commercial DNA extraction kits NucleoSpin Tissue (Macherey-Nagel) or DNeasy Blood & Tissue (Qiagen) and following the manufacturer's guidelines, or with a standard salting out protocol [86], or, for the bivalves, with a standard cetyl trimethylammonium bromide (CTAB) protocol. Subsequently, DNA quality and quantity were checked using the fluorescence assay Quant-iT PicoGreen dsDNA (Thermo Fisher Scientific Inc.), an Infinite M200 microplate reader (Tecan Group Ltd.) and 1% agarose gel electrophoresis. Whenever possible, only high-quality DNA extractions were used. Because of their small size, extractions from individual ostracods yielded insufficient quantities of DNA for downstream protocols, and sample numbers per locality were very low. Hence, the entire genomic DNA of ostracods was amplified using the REPLI-G kit (Qiagen) for whole genome amplification of 1 μL extracted DNA with high-fidelity polymerase Phi 20 and multiple displacement amplification following the manufacturer's protocol. For this purpose, extractions with the highest DNA concentrations from different species of Macrocyprididae, mainly of the *Macroscapha tensa-opaca* species complex, were selected [127]. For all target species, 100 ng genomic DNA of three biological replicates per species was digested with 10 units of a selected restriction enzyme at 37 °C (*EcoRI*, *MspI* and *PstI*) or 75 °C (*ApeKI*) for 2 h in a total volume of 10 μL. Reactions were purified with CleanPCR (GC Biotech) according to the manufacturer's protocol. Between 1 and 5 ng of the purified digested DNA was loaded on a High Sensitivity DNA chip (Agilent Technologies) and run on an Agilent 2100 Bioanalyzer System. Results were exported from the 2100 expert software (Agilent) as XML files and read into R v4.0.4 [135] using the bioanalyzeR package v0.5.1 [136]. Additional R packages used in this project were here v1.0.1 [137], seqinR v1.0–2 [138], the tidyverse packages [139], ggsci v2.9 [140], and gridExtra v2.3 [141] (see also more details under: https://github.com/notothen/radpilot). Because it is not possible to accurately standardize the number of fragments in an empirical digest without knowledge of the true genome size, we compared the shape of the curves of produced fragments (number of loci or DNA concentration vs. locus size or length) between in silico and empirical digests (Fig. 1 and Additional File 3).

RRS setup optimization

In order to choose a promising restriction enzyme and size selection combination, we calculated the sequencing coverage per fragment as follows:

$$coverage = \frac{\sum sequencing\ reads}{\sum individuals \sum genomic\ fragments}$$

We conservatively aimed at a coverage of approximately 30× for each fragment per individual, higher than other minimum recommendations [15, 36, 45]. Given that the accuracy of our genome size estimates is unknown, we aimed for relatively high coverage, so that in a "worst-case scenario", where the genome size is actually twice as large as we estimated (or any other factor leads to twice as many fragments as assumed), we would still reach a coverage of approximately 15×. The number of individuals per sequencing library was set to 96, corresponding to one PCR plate. Sequencing with a HiSeq 4000 platform (Illumina) should conservatively yield approximately 300 million reads per sequencing lane, while on a HiSeq 2500, we expect approximately 200 million reads. These coverage calculations were applied to fragment numbers from in silico results based on available reference genomes and extrapolated to a final, conservative estimate of genome size based on the best available knowledge (Table 4). This extrapolation is likely not biologically accurate but serves as a conservative correction factor. We then used in silico estimates again to further tweak the size window of a chosen restriction enzyme or enzyme combination in each target species to achieve the desired coverage, while considering the size range in which the two HiSeq machines work best. Finally, we estimated the number of SNPs across the genome as a measure of marker density (analogous to [34]) for a chosen enzyme and size selection setup and sequencing machine, assuming one SNP every other 100 bp. The latter estimate is based on our own experience, predominantly from fish genomes (but see also e.g. [142]). If an estimate of the naturally occurring SNP density across the genome is known for the target species or a related species, then this should be used. We provide an R function where any estimate can be used as input for marker density calculations. In general, all our calculations and plots should be reproducible with our spreadsheet tables and R scripts available at https://doi.org/10.5281/zenodo.5045574 and at https://github.com/notothen/radpilot.

RRS library preparation and sequencing

The information collected so far convinced us not to pursue RRS in amphipods (see discussion); they were therefore not included in the test libraries. In addition, not enough high molecular weight DNA samples of *P. nivea confusa* (one of the snow petrel subspecies) were available. Eventually, five RRS test libraries for eight target species were constructed using 6, 8, 10, or 14 individuals and two controls per species and sequenced on one lane of a HiSeq 2500 unit (see Table 5 in results section). With this setup, we attempted to realize the previously estimated fixed variables for our coverage calculations, i.e. an estimated 250 million reads spread over 94 individuals and between 53,399 and 81,605 fragments. We originally aimed at 96 individuals, but too many samples of low-quality DNA dropped out during sample preparation. In addition, the estimated number of fragments varied between target species, but the conservative estimates in all other aspects should allow for some flexibility here. The libraries were all prepared by the same person at the KU Leuven laboratory using custom protocols that are based on two main references: the original ddRAD protocol by Peterson et al. (2012) [15] and the original GBS protocol by Elshire et al. (2011) [14]. We adjusted these protocols slightly and provide a full-length description of the laboratory procedure in Additional Files 5 & 6. In both cases, the standardized high-quality DNA was first digested with restriction enzyme(s), followed by adaptor and barcode ligation, purification, PCR, another purification and finally quantification and pooling. The libraries were then sent to the KU Leuven Genomics Core (www.genomicscore.be), where all five libraries were individually size selected on a Pippin Prep unit (Sage Science), checked for quantity using qPCR, pooled, and paired-end sequenced on one lane of a HiSeq 2500 platform (Illumina).

Sequence analyses

Sequencing data were checked using FastQC v0.11.5 [143] and then demultiplexed and cleaned (options -c and -q) using the process_radtags module of Stacks v2.4 [98, 99]. Because some of our multiplexing barcodes for the *PstI* library were contained in longer *ApeKI* barcodes, we demultiplexed the *ApeKI* libraries first and captured reads that were discarded in the process. These reads were subsequently used for demultiplexing of the *PstI* library. All demultiplexing runs were conducted without barcode rescue to avoid cross-contamination between libraries. The Stacks pipeline was also used for each target species independently to create a de novo assembly and call genotypes. Building contigs from paired-end reads is not possible with GBS data in Stacks [21], because the orientation of the reads is ambiguous. In this case (libraries 1, 2, 3, 5), we concatenated the four output files per individual of process_radtags to run the pipeline as if it was single-end data. Our size selection windows were designed to avoid overlap between the two reads of one fragment, so this approach should work well, albeit creating shorter haplotypes. We used Stacks' default value for m, i.e., a minimum stack coverage of 3, which generally produces consistent results at typical coverage rates [36]. Choosing parameters M and n to control the formation of loci within and across individuals on the other hand is study dependent. We explored a parameter range of $n = M = [1 .. 9]$ following Rochette and Catchen [45] to strike a balance between over- and undermerging alleles and loci. To compare results from the different parameters only loci present in 80% of the samples (50% in the case of ostracods) were retained. Further detailed filtering would be required for downstream population genomic analyses.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07917-3.

Additional file 1. Samples used for reduced representation sequencing (RRS) optimization. DNA from these samples was used for empirical restriction enzyme digestions with different enzymes (single digest *EcoRI*, *PstI*, *MspI*, or double digest *EcoRI-MspI*) and for RRS pilot libraries. Some samples were extracted twice as replicates (marked as _rep in sample ID). Three samples per species (family in the case of ostracods) were used for empirical digestions. The amphipod (*C. obesa* and *E. pontomedon*) samples and one *T. loennbergii* were used for empirical digestions, but not included in any RRS library.

Additional file 2. In silico estimates of the number of fragments. Estimates were produced through in silico restriction enzyme digestions for reduced representation sequencing (RRS) optimized for approximately 30× coverage. The number of fragments depends on the restriction enzyme/combination, the size window, the assumed genome size, and the reference genome used for in silico digestion. Reference genomes of related species were used as well as simulated genomes; in this case the size and GC content used to simulate the genomes are listed. The number of fragments were extrapolated to the assumed genome size. Only two different enzyme and size selection setups per target species are listed here (for RRS setups optimized for HiSeq 2500 or HiSeq 4000 sequencing runs, respectively; the same as in Table 4, Table 5, Additional File 4); further estimates can be found in spreadsheets available at https://doi.org/10.5281/zenodo.5045574.

Additional file 3. Comparisons of empirical and in silico restriction enzyme digestions. Empirical Bioanalyzer results (left figure panels) with digested DNA are shown as concentration over fragment size and estimated loci numbers over locus size from in silico digestions (right figure panels) for all target taxa except fish (these are shown in Fig. 1).

Additional file 4. Reduced representation sequencing (RRS) setups for seven individually optimized protocols. These setups were optimized in order to be run on a HiSeq 4000 platform (Illumina). The choice of restriction enzyme(s) and size window was optimized to obtain approximately 30× coverage (or half that value in a worst-case scenario) with the assumed genome size (conservatively estimated based on available information, see Table 2). Marker density was estimated as a comparable measure to the metastudy by Lowry et al. (2017) [34].

Additional file 5. Reduced representation sequencing (RRS) laboratory protocol based on the protocol from Peterson et al. (2012) [15]. The protocol is scaled for use with 192 samples and with restriction enzymes *EcoRI* and *MspI*; the reagent volumes can be scaled down/up to suit other sample numbers; if other enzymes are used, the respective reaction conditions must be adjusted.

Additional file 6. Reduced representation sequencing (RRS) laboratory protocol based on the protocol from Elshire et al. (2011) [14]. The protocol is scaled for use with 192 samples and with restriction enzymes *PstI* or *ApeKI*; the reagent volumes can be scaled down/up to suit other sample numbers; if other enzymes are used, the respective reaction conditions must be adjusted.

Additional file 7. Results from parameter optimization for de novo assembly and genotyping. Eight parameter optimization series were conducted following Rochette & Catchen (2017) [45] to identify optimal parameters to genotype reduced representation sequencing (RRS) data with Stacks v2.4 (Rochette et al. 2019) [21]; one test series for each species/species complex. The Stacks parameter m was kept constant (m = 3), while parameters M and n were varied together from 1 to 9. Subsequently, only loci present in 80% of the samples were retained and for each M = n parameter the number of loci and polymorphic loci was plotted, as well as the proportion of these loci containing 0 to 10 or > 10 SNPs. In ostracods, the library contained DNA from a species-complex, resulting in very few shared loci across 80% of the samples. Therefore, in this case results based on loci shared by 50% of samples are shown. Optimal M = n values were decided in all cases with this information (and reported in Table 5). Note, however, that it is impossible to make absolute calls regarding the ideal value.

Acknowledgements
We are very grateful to everyone that contributed to field work and provided samples. In particular, we thank S. Brandão for ostracod DNA and H. Griffiths and K. Linse (BAS), C. Held (AWI), S. Mills and D. Macpherson (NIWA), H. Baird and G. Johnstone (AAD), and the MNHN in Paris for amphipod samples. We thank M. Clark (BAS) and the scientific staff involved for the generous collection of bivalve samples from Rothera station. We thank the captains, crew and science leaders who contributed to the collection of sea stars from RRS *James Clark Ross* (JR15005) and RV *Polarstern* (PS77, PS81). We thank the scientific team of IPEV (Institut Polaire Français Paul-Emile Victor) program 1124 REVOLTA (Radiation EVOLutives en Terre Adélie; chief scientists M. Eléaume, N. Améziane and G. Lecointre). For fishes, we thank K. Roche and P. Jurajda who provided fin clips from James Ross Island. The toothfish fishery observers and S. Parker (NIWA) are acknowledged for fish samples from the Ross Sea. We thank G.E. Maes and K. Herten from the KU Leuven Genomics Core for technical support and advice. We thank the two anonymous reviewers for their constructive comments.

Authors' contributions
HC and IS conceived the study. BH, FH, HC, QJ, FP, HR, MV and IS conducted laboratory work. HC analyzed the data and wrote the first draft of the manuscript. All authors contributed samples, intellectual input during project meetings and comments on the manuscript. All authors read and approved the final manuscript.

Funding
This is contribution no. 35 of the vERSO project, funded by the Belgian Science Policy Office (BELSPO, Contract no. BR/132/A1/vERSO) and contribution no. 8 of the RECTO project (BELSPO, Contract no. BR/154/A1/RECTO). TP thanks BELSPO for funding JPIO Mining Impact. Research was also supported by the Scientific Research Network "Eco-evolutionary dynamics in natural and anthropogenic communities" (grant W0.037.10 N), and the European Marine Biological Resource Center (EMBRC) Belgium, both funded by the Research Foundation – Flanders (FWO). The first author was supported by an individual grant from the former Flemish Agency for Innovation by Science and Technology, now managed through Flanders Innovation & Entrepreneurship (VLAIO, grant no. 141328). The funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The datasets supporting the conclusions of this article are available in the NCBI's Sequence Read Archive (SRA) repository, BioProject ID PRJNA674352, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA674352, and in the Zenodo repository, https://doi.org/10.5281/zenodo.5045574.

Declarations

Ethics approval and consent to participate
Animal tissues were sampled following internationally recognized standard methods in line with the Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR) and its Ecosystem Monitoring Program (CEMP) and were permitted under the Antarctic Marine Living Resources Act.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Laboratory of Biodiversity and Evolutionary Genomics, KU Leuven, Leuven, Belgium. [2]Marine Biology Group, Vrije Universiteit Brussel (VUB), Brussels, Belgium. [3]Marine Biology Research Group, Ghent University, Ghent, Belgium. [4]OD Nature, Royal Belgian Institute of Natural Sciences, Brussels, Belgium. [5]Marine Biology Laboratory, Université Libre de Bruxelles (ULB), Brussels, Belgium. [6]Meise Botanic Garden, Meise, Belgium. [7]Université de Bourgogne Franche-Comté (UBFC) UMR CNRS 6282 Biogéosciences, Dijon, France.

Received: 2 April 2021 Accepted: 26 July 2021
Published online: 21 August 2021

References

1. Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol. 2014;29(1):51–63.
2. Borja A. Grand challenges in marine ecosystems ecology. Front Mar Sci. 2014;1:1.
3. Brandt A, Gooday AJ, Brandão SN, Brix S, Brökeland W, Cedhagen T, et al. First insights into the biodiversity and biogeography of the Southern Ocean deep sea. Nature. 2007;447(7142):307–11.
4. Kelley JL, Brown AP, Therkildsen NO, Foote AD. The life aquatic: advances in marine vertebrate genomics. Nat Rev Genet. 2016;17(9):523–34. https://doi.org/10.1038/nrg.2016.66.
5. 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(2):81–92. https://doi.org/10.1038/nrg.2015.28.
6. Matz MV. Fantastic beasts and how to sequence them: ecological genomics for obscure model organisms. Trends Genet. 2017;34(2):121–32. https://doi.org/10.1016/j.tig.2017.11.002.
7. Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM. Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. Mol Ecol. 2013;22(11):2953–70.
8. Savolainen O, Lascoux M, Merilä J. Ecological genomics of local adaptation. Nat Rev Genet. 2013 Nov;14(11):807–20.
9. Funk WC, McKay JK, Hohenlohe PA, Allendorf FW. Harnessing genomics for delineating conservation units. Trends Ecol Evol. 2012 Sep;27(9):489–96. https://doi.org/10.1016/j.tree.2012.05.012.
10. Hoffmann A, Griffin P, Dillon S, Catullo R, Rane R, Byrne M, et al. A framework for incorporating evolutionary genomics into biodiversity

conservation and management. Clim Chang Responses. 2015;2(1):1–23. https://doi.org/10.1186/s40665-014-0009-x.

11. Fuentes-Pardo AP, Ruzzante DE. Whole-genome sequencing approaches for conservation biology: advantages, limitations, and practical recommendations. Mol Ecol. 2017;26(20):5369–406. https://doi.org/10.1111/mec.14264.
12. Therkildsen NO, Palumbi SR. Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. Mol Ecol Resour. 2017; 17(2):194–208. https://doi.org/10.1111/1755-0998.12593.
13. 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.
14. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011 Jan;6(5):e19379. https://doi.org/10.1371/journal.pone.0019379.
15. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One. 2012;7(5):e37135. https://doi.org/10.1371/journal.pone.0037135.
16. Toonen RJ, Puritz JB, Forsman ZH, Whitney JL, Fernandez-Silva I, Andrews KR, et al. ezRAD: a simplified method for genomic genotyping in non-model organisms. PeerJ. 2013;1:e203.
17. Poland JA, Brown PJ, Sorrells ME, Jannink J-L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One. 2012;7(2):e32253. https://doi.org/10.1371/journal.pone.0032253.
18. Campbell EO, Brunet BMT, Dupuis JR, Sperling FAH. Would an RRS by any other name sound as RAD? Methods Ecol Evol. 2018;9(9):1920–7. https://doi.org/10.1111/2041-210X.13038.
19. Altshuler D, Pollara VJ, Cowles CR, Lander ES. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature. 2000;407:513–6.
20. 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(7):499–510. https://doi.org/10.1038/nrg3012.
21. Rochette NC, Rivera-Colon AG, Catchen JM. STACKS 2: analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol Ecol. 2019;28(21):4737–54. https://doi.org/10.1111/mec.15253.
22. Baetscher DS, Clemento AJ, Ng TC, Anderson EC, Garza JC. Microhaplotypes provide increased power from short-read DNA sequences for relationship inference. Mol Ecol Resour. 2018;18(2):296–305. https://doi.org/10.1111/1755-0998.12737.
23. Willis S, Hollenbeck C, Puritz JB, Gold J, Portnoy D. Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. Mol Ecol Resour. 2017;17(5)955-65. https://doi.org/10.1111/1755-0998.12647.
24. Jansson E, Taggart JB, Wehner S, Dahle G, Quintela M, Mortensen S, et al. Development of SNP and microsatellite markers for goldsinny wrasse (*Ctenolabrus rupestris*) from ddRAD sequencing data. Conserv Genet Resour. 2016;8:201-6. https://doi.org/10.1007/s12686-016-0532-0.
25. McKinney GJ, Waples RK, Seeb LW, Seeb JE. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. Mol Ecol Resour. 2017;17(4):656–69. https://doi.org/10.1111/1755-0998.12613.
26. Dorant Y, Cayuela H, Wellband K, Laporte M, Rougemont Q, Mérot C, et al. Copy number variants outperform SNPs to reveal genotype–temperature association in a marine species. Mol Ecol. 2020;29:4765–82. https://doi.org/10.1111/mec.15565.
27. Fang B, Merilä J, Ribeiro F, Alexandre CM, Momigliano P. Worldwide phylogeny of three-spined sticklebacks. Mol Phylogenet Evol. 2018;127:613–25.
28. Franchini P, Fruciano C, Spreitzer ML, Jones JC, Elmer KR, Henning F, et al. Genomic architecture of ecologically divergent body shape in a pair of sympatric crater lake cichlid fishes. Mol Ecol. 2014;23:1828–45.
29. Gaither MR, Gkafas GA, De Jong M, Sarigol F, Neat F, Regnier T, et al. Genomics of habitat choice and adaptive evolution in a deep-sea fish. Nat Ecol Evol. 2018;2(4):680–7.
30. Pante E, Abdelkrim J, Viricel A, Gey D, France SC, Boisselier MC, et al. Use of RAD sequencing for delimiting species. Heredity (Edinb). 2015;11:450–9.
31. Ravinet M, Westram A, Johannesson K, Butlin R, André C, Panova M. Shared and nonshared genomic divergence in parallel ecotypes of Littorina saxatilis at a local scale. Mol Ecol. 2016;25:287–305.
32. Xuereb A, Benestan L, Normandeau É, Daigle RM, Curtis JMR, Bernatchez L, et al. Asymmetric oceanographic processes mediate connectivity and population genetic structure, as revealed by RADseq, in a highly dispersive marine invertebrate (Parastichopus californicus). Mol Ecol. 2018;27(10):2347–64. https://doi.org/10.1111/mec.14589.
33. Maroso F, Hillen JEJ, Pardo BG, Gkagkavouzis K, Coscia I, Hermida M, et al. Performance and precision of double digestion RAD (ddRAD) genotyping in large multiplexed datasets of marine fish species. Mar Genomics. 2018;39: 64–72. https://doi.org/10.1016/j.margen.2018.02.002.
34. Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, et al. Breaking RAD: an evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation. Mol Ecol Resour. 2017;17(2): 142–52. https://doi.org/10.1111/1755-0998.12635.
35. Burns M, Starrett J, Derkarabetian S, Richart CH, Cabrero A, Hedin M. Comparative performance of double-digest RAD sequencing across divergent arachnid lineages. Mol Ecol Resour. 2017;17(3):418–30. https://doi.org/10.1111/1755-0998.12575.
36. Paris JR, Stevens JR, Catchen JM. Lost in parameter space: a road map for stacks. Methods Ecol Evol. 2017;8(10):1360–73. https://doi.org/10.1111/2041-210X.12775.
37. Smith PJ, Steinke D, McMillan PJ, Stewart AL, McVeagh SM. Diaz De Astarloa JM, et al. DNA barcoding highlights a cryptic species of grenadier Macrourus in the Southern Ocean. J Fish Biol. 2011;78(1):355–65. https://doi.org/10.1111/j.1095-8649.2010.02846.x.
38. Christiansen H, Dettai A, Heindler FM, Collins MA, Duhamel G, Hautecoeur M, et al. Diversity of mesopelagic fishes in the Southern Ocean - a Phylogeographic perspective using DNA barcoding. Front Ecol Evol. 2018;6: 120. https://doi.org/10.3389/fevo.2018.00120.
39. Ogden R, Gharbi K, Mugue N, Martinsohn J, Senn H, Davey JW, et al. Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. Mol Ecol. 2013;22(11):3112–23. https://doi.org/10.1111/mec.12234.
40. Ceballos SG, Roesti M, Matschiner M, Fernández DA, Damerau M, Hanel R, et al. Phylogenomics of an extra-Antarctic notothenioid radiation reveals a previously unrecognized lineage and diffuse species boundaries. BMC Evol Biol. 2019;19(1):13. https://doi.org/10.1186/s12862-019-1345-z.
41. Langin KM, Aldridge CL, Fike JA, Cornman RS, Martin K, Wann GT, Seglund AE, Schroeder MA, Braun CE, Benson DP, Fedy BC, Young JR, Wilson S, Wolfe DH, Oyler-McCance SJ Characterizing range-wide divergence in an alpine-endemic bird: a comparison of genetic and genomic approaches. Conserv Genet. 2018; 19(0):1471–85. https://doi.org/10.1007/s10592-018-1115-2.
42. Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, et al. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. Am Nat. 2016;188(4):379–97. https://doi.org/10.1086/688018.
43. Whitlock MC, Lotterhos KE. Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F ST. Am Nat. 2015;186(S1):S24–36. https://doi.org/10.1086/682949.
44. Catchen JM, Hohenlohe PA, Bernatchez L, Funk WC, Andrews KR, Allendorf FW. Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. Mol Ecol Resour. 2017;17(3): 362–5. https://doi.org/10.1111/1755-0998.12669.
45. Rochette NC, Catchen JM. Deriving genotypes from RAD-seq short-read data using stacks. Nat Protoc. 2017;12(12):2640–59. https://doi.org/10.1038/nprot.2017.123.
46. Crame JA. Key stages in the evolution of the Antarctic marine fauna. J Biogeogr. 2018;45(5):986-94. https://doi.org/10.1111/jbi.13208.
47. Rogers AD. Evolution and biodiversity of Antarctic organisms: a molecular perspective. Philos Trans R Soc B Biol Sci. 2007;362(1488):2191–214. https://doi.org/10.1098/rstb.2006.1948.
48. Aronson RB, Thatje S, Mcclintock JB, Hughes KA. Anthropogenic impacts on marine ecosystems in Antarctica. Ann N Y Acad Sci. 2011;1223(1):82–107. https://doi.org/10.1111/j.1749-6632.2010.05926.x.
49. Griffiths HJ, Meijers AJS, Bracegirdle TJ. More losers than winners in a century of future Southern Ocean seafloor warming. Nat Clim Chang. 2017; 7(10):749–54.
50. Nicol S, Foster J, Kawaguchi S. The fishery for Antarctic krill - recent developments. Fish Fish. 2012;13(1):30–40.

51. Mangano MC, Sarà G, Corsolini S. Monitoring of persistent organic pollutants in the polar regions: knowledge gaps & gluts through evidence mapping. Chemosphere. 2017;172:37–45. https://doi.org/10.1016/j.chemosphere.2016.12.124.
52. Younger JL, Clucas GV, Kao D, Rogers AD, Gharbi K, Hart T, et al. The challenges of detecting subtle population structure and its importance for the conservation of emperor penguins. Mol Ecol. 2017;26(15):3883–97. https://doi.org/10.1111/mec.14172.
53. Clucas GV, Younger JL, Kao D, Emmerson L, Southwell C, Wienecke B, et al. Comparative population genomics reveals key barriers to dispersal in Southern Ocean penguins. Mol Ecol. 2018;27(23):4680–97. https://doi.org/10.1111/mec.14896.
54. Rintoul SR, Chown SL, DeConto RM, England MH, Fricker HA, Masson-Delmotte V, et al. Choosing the future of Antarctica. Nature. 2018;558(7709):233–41. https://doi.org/10.1038/s41586-018-0173-4.
55. Brandão SN. Macrocyprididae (Ostracoda) from the Southern Ocean: taxonomic revision, macroecological patterns, and biogeographical implications. Zool J Linnean Soc. 2010;159(3):567–672.
56. Verheye ML, D'Udekem D'AC. Integrative taxonomy of giant crested Eusirus in the Southern Ocean, including the description of a new species (Crustacea: Amphipoda: Eusiridae). Zool J Linnean Soc. 2020;zlaa141. https://doi.org/10.1093/zoolinnean/zlaa141.
57. Tran Van P, Anselmetti Y, Bast J, Dumas Z, Galtier N, Jaron KS, et al. First annotated draft genomes of non-marine ostracods (Ostracoda, Crustacea) with different reproductive modes. G3 Genes Genomes Genet. 2021;11(4):jkab043. https://doi.org/10.1093/g3journal/jkab043.
58. Poynton HC, Hasenbein S, Benoit JB, Sepulveda MS, Poelchau MF, Hughes DST, et al. The Toxicogenome of Hyalella azteca: a model for sediment ecotoxicology and evolutionary toxicology. Environ Sci Technol. 2018;52(10):6009–22. https://doi.org/10.1021/acs.est.8b00837.
59. Kao D, Lai AG, Stamataki E, Rosic S, Konstantinides N, Jarvis E, et al. The genome of the crustacean Parhyale hawaiensis, a model for animal development, regeneration, immunity and lignocellulose digestion. Elife. 2016;5:e200062.
60. Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, et al. The oyster genome reveals stress adaptation and complexity of shell formation. Nature. 2012;490(7418):49–54.
61. Du X, Fan G, Jiao Y, Zhang H, Guo X, Huang R, et al. The pearl oyster Pinctada fucata martensii genome and multi-omic analyses provide insights into biomineralization. Gigascience. 2017;6(8):1–12. https://doi.org/10.1093/gigascience/gix059.
62. Sun J, Zhang Y, Xu T, Zhang Y, Mu H, Zhang Y, et al. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. Nat Ecol Evol. 2017;1(5):1–7.
63. Hall MR, Kocot KM, Baughman KW, Fernandez-Valverde SL, Gauthier MEA, Hatleberg WL, et al. The crown-of-thorns starfish genome as a guide for biocontrol of this coral reef pest. Nature. 2017;544(7649):231–4. https://doi.org/10.1038/nature22033.
64. Cameron RA, Kudtarkar P, Gordon SM, Worley KC, Gibbs RA. Do echinoderm genomes measure up? Mar Genomics. 2015;22:1–9. https://doi.org/10.1016/j.margen.2015.02.004.
65. Long KA, Nossa CW, Sewell MA, Putnam NH, Ryan JF. Low coverage sequencing of three echinoderm genomes: the brittle star Ophionereis fasciata, the sea star Patiriella regularis, and the sea cucumber Australostichopus mollis. Gigascience. 2016;5(1):1–4.
66. Shin SC, Ahn DH, Kim SJ, Pyo CW, Lee H, Kim M-K, et al. The genome sequence of the Antarctic bullhead notothen reveals evolutionary adaptations to a cold environment. Genome Biol. 2014;15(9):468. https://doi.org/10.1186/s13059-014-0468-1.
67. Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science (80-). 2014;346(6215):1311–20.
68. Jeffery NW, Ellis EA, Oakley TH, Ryan GT. The genome sizes of ostracod crustaceans correlate with body size and evolutionary history, but not environment. J Hered. 2017;108(6):701–6. https://doi.org/10.1093/jhered/esx055.
69. Auvinet J, Graça P, Belkadi L, Petit L, Bonnivard E, Dettaï A, et al. Mobilization of retrotransposons as a cause of chromosomal diversification and rapid speciation: the case for the Antarctic teleost genus Trematomus. BMC Genomics. 2018;19(1):339.
70. Hardie DC, Hebert PD. The nucleotypic effects of cellular DNA content in cartilaginous and ray-finned fishes. Genome. 2003;46(4):683–706. https://doi.org/10.1139/g03-040.
71. Morescalchi A, Morescalchi MA, Odierna G, Stingo V, Capriglione T. Karyotype and genome size of zoarcids and Notothenioids (Teleostei, perciformes) from the Ross Sea: Cytotaxonomic implications. Polar Biol. 1996;16(8):559–64. https://doi.org/10.1007/BF02329052.
72. Ritchie H, Jamieson AJ, Piertney SB. Genome size variation in deep-sea amphipods. R Soc Open Sci. 2017;4:170862.
73. Rees DJ, Dufresne F, Glémet H, Belzile C. Amphipod genome sizes: first estimates for Arctic species reveal genomic giants. Genome. 2007;50(2):151–8. https://doi.org/10.1139/G06-155.
74. Deagle BE, Faux C, Kawaguchi S, Meyer B, Jarman SN. Antarctic krill population genomics: apparent panmixia, but genome complexity and large population size muddy the water. Mol Ecol. 2015;24(19):4943–59. https://doi.org/10.1111/mec.13370.
75. Hohenlohe PA, Day MD, Amish SJ, Miller MR, Kamps-Hughes N, Boyer MC, et al. Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. Mol Ecol. 2013;22(11):3002–13.
76. Rodríguez-Ezpeleta N, Bradbury IR, Mendibil I, Álvarez P, Cotano U, Irigoien X. Population structure of Atlantic mackerel inferred from RAD-seq-derived SNP markers: effects of sequence clustering parameters and hierarchical SNP selection. Mol Ecol Resour. 2016;16(4):991–1001. https://doi.org/10.1111/1755-0998.12518.
77. Jacobsen MW, Pujolar JM, Bernatchez L, Munch K, Jian J, Niu Y, et al. Genomic footprints of speciation in Atlantic eels (Anguilla anguilla and A. rostrata). Mol Ecol. 2014;23(19):4785–98. https://doi.org/10.1111/mec.12896.
78. Bolton PE, West AJ, Cardilini APA, Clark JA, Maute KL, Legge S, et al. Three molecular markers show no evidence of population genetic structure in the Gouldian finch (Erythrura gouldiae). PLoS One. 2016;11(12):1–19.
79. Herrera S, Shank TM. RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa. Mol Phylogenet Evol. 2016;100:70–9. https://doi.org/10.1016/j.ympev.2016.03.010.
80. Grewe F, Huang JP, Leavitt SD, Lumbsch HT. Reference-based RADseq resolves robust relationships among closely related species of lichen-forming fungi using metagenomic DNA. Sci Rep. 2017;7(1):9884. https://doi.org/10.1038/s41598-017-09906-7.
81. Pérez-Portela R, Bumford A, Coffman B, Wedelich S, Davenport M, Fogg A, et al. Genetic homogeneity of the invasive lionfish across the northwestern Atlantic and the Gulf of Mexico based on single nucleotide polymorphisms. Sci Rep. 2018;8(1):5062. https://doi.org/10.1038/s41598-018-23339-w.
82. Puncher GN, Cariani A, Maes GE, Van Houdt J, Herten K, Cannas R, et al. Spatial dynamics and mixing of bluefin tuna in the Atlantic Ocean and Mediterranean Sea revealed using next generation sequencing. Mol Ecol Resour. 2018;18(3):620–38.
83. Raeymaekers JAM, Chaturvedi A, Hablützel PI, Verdonck I, Hellemans B, Maes GE, et al. Adaptive and non-adaptive divergence in a common landscape. Nat Commun. 2017;8(1):267. https://doi.org/10.1038/s41467-017-00256-6.
84. Zhu F, Cui QQ, Hou ZC. SNP discovery and genotyping using genotyping-by-sequencing in Pekin ducks. Sci Rep. 2016;6(1):36223. https://doi.org/10.1038/srep36223.
85. Adenyo C, Ogden R, Kayang B, Onuma M, Nakajima N, Inoue-Murayama M. Genome-wide DNA markers to support genetic management for domestication and commercial production in a large rodent, the Ghanaian grasscutter (Thryonomys swinderianus). Anim Genet. 2017;48(1):113–5. https://doi.org/10.1111/age.12478.
86. Cruz VP, Vera M, Pardo BG, Taggart J, Martinez P, Oliveira C, et al. Identification and validation of single nucleotide polymorphisms as tools to detect hybridization and population structure in freshwater stingrays. Mol Ecol Resour. 2017;17(3):550–6. https://doi.org/10.1111/1755-0998.12564.
87. Hillen JEJ, Coscia I, Vandeputte M, Herten K, Hellemans B, Maroso F, et al. Estimates of genetic variability and inbreeding in experimentally selected populations of European sea bass. Aquaculture. 2017;479:742–9. https://doi.org/10.1016/j.aquaculture.2017.07.012.
88. Jacobsen MW, Christensen C, Madsen R, Nygaard R, Jónsson B, Præbel K, et al. Single nucleotide polymorphism markers for analysis of historical and contemporary samples of Arctic char (Salvelinus alpinus). Conserv Genet Resour. 2017;9:587–9.

89. Leaché AD, Grummer JA, Harris RB, Breckheimer I. Evidence for concerted movement of nuclear and mitochondrial clines in a lizard hybrid zone. Mol Ecol. 2017;26(8):2306–16. https://doi.org/10.1111/mec.14033.
90. Bernatchez S, Laporte M, Perrier C, Sirois P, Bernatchez L. Investigating genomic and phenotypic parallelism between piscivorous and planktivorous lake trout (Salvelinus namaycush) ecotypes by means of RADseq and morphometrics analyses. Mol Ecol. 2016;25(19):4773–92.
91. Henning F, Machado-Schiaffino G, Baumgarten L, Meyer A. Genetic dissection of adaptive form and function in rapidly-speciating cichlid fishes. Evolution (N Y). 2017;71(5):1297–312.
92. Recknagel H, Elmer KR, Meyer A. A hybrid genetic linkage map of two ecologically and morphologically divergent Midas cichlid fishes (Amphilophus spp.) obtained by massively parallel DNA sequencing (ddRADSeq). G3 Genes Genomes Genet. 2013;3(1):65–74.
93. Nunziata SO, Lance SL, Scott DE, Lemmon EM, Weisrock DW. Genomic data detect corresponding signatures of population size change on an ecological time scale in two salamander species. Mol Ecol. 2017;26(4):1060–74. https://doi.org/10.1111/mec.13988.
94. Escoda L, González-Esteban J, Gómez A, Castresana J. Using relatedness networks to infer contemporary dispersal: application to the endangered mammal Galemys pyrenaicus. Mol Ecol. 2017;26(13):3343–57. https://doi.org/10.1111/mec.14133.
95. Lozier JD, Jackson JM, Dillon ME, Strange JP. Population genomics of divergence among extreme and intermediate color forms in a polymorphic insect. Ecol Evol. 2016;6(4):1075–91.
96. Ng NSR, Wilton PR, Prawiradilaga DM, Tay YC, Indrawan M, Garg KM, et al. The effects of Pleistocene climate change on biotic differentiation in a montane songbird clade from Wallacea. Mol Phylogenet Evol. 2017;114:353–66.
97. Querejeta M, González-Esteban J, Gómez A, Fernández-González A, Aymerich P, Gosálbez J, et al. Genomic diversity and geographical structure of the Pyrenean desman. Conserv Genet. 2016;17(6):1333–44. https://doi.org/10.1007/s10592-016-0865-y.
98. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. Mol Ecol. 2013 Jun;22(11):3124–40. https://doi.org/10.1111/mec.12354.
99. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. Stacks: building and genotyping loci de novo from short-read sequences. G3 Genes Genomes Genet. 2011;1:171–82.
100. Herrera S, Reyes-Herrera PH, Shank TM. Predicting RAD-seq marker numbers across the eukaryotic tree of life. Genome Biol Evol. 2015;7(12):3207–25. https://doi.org/10.1093/gbe/evv210.
101. Krapp T, Lang C, Libertini A, Melzer RR. Caprella scaura Templeton, 1836 sensu lato (Amphipoda: Caprellidae) in the Mediterranean. Org Divers Evol. 2006;6(2):77–81.
102. Star B, Hansen MH, Skage M, Bradbury IR, Godiksen JA, Kjesbu OS, et al. Preferential amplification of repetitive DNA during whole genome sequencing library creation from historic samples. STAR Sci Technol Archaeol Res. 2016;2(1):36–45.
103. McCartney-Melstad E, Mount GG, Shaffer HB. Exon capture optimization in amphibians with large genomes. Mol Ecol Resour. 2016;16(5):1084–94. https://doi.org/10.1111/1755-0998.12538.
104. Pootakham W, Sonthirod C, Naktang C, Jomchai N, Sangsrakru D, Tangphatsornruang S. Effects of methylation-sensitive enzymes on the enrichment of genic SNPs and the degree of genome complexity reduction in a two-enzyme genotyping-by-sequencing (GBS) approach: a case study in oil palm (*Elaeis guineensis*). Mol Breed. 2016;36:154. https://doi.org/10.1007/s11032-016-0572-x.
105. de Medeiros BAS, Farrell BD. Whole genome amplification in double-digest RAD-seq results in adequate libraries but fewer sequenced loci. PeerJ. 2018;6:e5089. https://doi.org/10.7717/peerj.5089.
106. Ghigliotti L, Cheng CC-H, Ozouf-Costaz C, Vacchi M, Pisano E. Cytogenetic diversity of notothenioid fish from the Ross Sea: historical overview and updates. Hydrobiologia. 2015;761(1):373–96. https://doi.org/10.1007/s10750-015-2355-5.
107. Kim BM, Amores A, Kang S, Ahn DH, Kim JH, Kim IC, et al. Antarctic blackfin icefish genome reveals adaptations to extreme environments. Nat Ecol Evol. 2019;3(3):469–78. https://doi.org/10.1038/s41559-019-0812-7.
108. Near TJ, Dornburg A, Kuhn KL, Eastman JT, Pennington JN, Patarnello T, et al. Ancient climate change, antifreeze, and the evolutionary diversification of Antarctic fishes. Proc Natl Acad Sci U S A. 2012 Feb 28;109(9):3434–9. https://doi.org/10.1073/pnas.1115169109.
109. Lautrédou A-C, Hinsinger DD, Gallut C, Cheng C-HC, Berkani M, Ozouf-Costaz C, et al. Phylogenetic footprints of an Antarctic radiation: the Trematominae (Notothenioidei, Teleostei). Mol Phylogenet Evol. 2012 Oct;65(1):87–101. https://doi.org/10.1016/j.ympev.2012.05.032.
110. Sánchez-Tójar A, Nakagawa S, Sánchez-Fortún M, Martin DA, Ramani S, Girndt A, et al. Meta-analysis challenges a textbook example of status signalling and demonstrates publication bias. Elife. 2018;7:1–26.
111. Rivera-Colón AG, Rochette NC, Catchen JM. Simulation with RADinitio improves RADseq experimental design and sheds light on sources of missing data. Mol Ecol Resour. 2021;21(2):363–78. https://doi.org/10.1111/1755-0998.13163.
112. Liu S, Hansen MM. PSMC (pairwise sequentially Markovian coalescent) analysis of RAD (restriction site associated DNA) sequencing data. Mol Ecol Resour. 2017;17:631–41.
113. Cerca J, Maurstad MF, Rochette NC, Rivera-Colón AG, Rayamajhi N, Catchen JM, et al. Removing the bad apples: a simple bioinformatic method to improve loci-recovery in de novo RADseq data for non-model organisms. Methods Ecol Evol. 2021;2021(September 2020):805–17.
114. Humble E, Dasmahapatra KK, Martinez-Barrio A, Gregório I, Forcada J, Polikeit A-C, et al. RAD sequencing and a hybrid Antarctic fur seal genome assembly reveal rapidly decaying linkage disequilibrium, global population structure and evidence for inbreeding. G3 Genes Genomes Genet. 2018;8(8):2709–22.
115. Barrio AM, Lamichhaney S, Fan G, Rafati N. The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. Elife. 2016;5:e12081. https://doi.org/10.7554/eLife.12081.001.
116. Hoffberg S, Kieran T, Catchen J, Devault A, Faircloth BC, Mauricio R, et al. RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. Mol Ecol Resour. 2016;16:1264–78.
117. Puritz JB, Lotterhos KE. Expressed exome capture sequencing: a method for cost-effective exome sequencing for all organisms. Mol Ecol Resour. 2018;18(6):1209–22.
118. De Wit P, Pespeni MH, Palumbi SR. SNP genotyping and population genomics from expressed sequences - current advances and future possibilities. Mol Ecol. 2015;24(10):2310–23. https://doi.org/10.1111/mec.13165.
119. Linck EB, Hanna Z, Sellas A, Dumbacher JP. Evaluating hybridization capture with RAD probes as a tool for museum genomics with historical bird specimens. Ecol Evol. 2017;7(13):4755–67. https://doi.org/10.1002/ece3.3065.
120. Schmid S, Genevest R, Gobet E, Suchan T, Sperisen C, Tinner W, et al. HyR3, a versatile method combining exome capture and RAD sequencing to extract genomic information from ancient DNA. Methods Ecol Evol. 2017;8(10):1374–88. https://doi.org/10.1111/2041-210X.12785.
121. Li C, Corrigan S, Yang L, Straube N, Harris M, Hofreiter M, et al. DNA capture reveals transoceanic gene flow in endangered river sharks. Proc Natl Acad Sci U S A. 2015;112(43):13302–7. https://doi.org/10.1073/pnas.1508735112.
122. Bi K, Linderoth T, Vanderpool D, Good JM, Nielsen R, Moritz C. Unlocking the vault: next-generation museum population genomics. Mol Ecol. 2013;22(24):6018–32.
123. Bradbury IR, Wringe BF, Watson B, Paterson I, Horne J, Beiko R, et al. Genotyping-by-sequencing of genome-wide microsatellite loci reveals fine-scale harvest composition in a coastal Atlantic salmon fishery. Evol Appl. 2018;11(6):918–30.
124. Vartia S, Villanueva-cañas JL, Finarelli J, Farrell ED, Collins PC, Hughes GM, et al. A novel method of microsatellite using individual combinatorial barcoding. R Soc Open Sci. 2016;3(1):150565. https://doi.org/10.1098/rsos.150565.
125. Lait LA, Marshall HD, Carr SM. Phylogeographic mitogenomics of Atlantic cod *Gadus morhua* : Variation in and among trans- Northern cod , and landlocked fjord populations. Ecol Evol. 2018;8(13):6420–37.
126. Teacher AG, André C, Merilä J, Wheat CW. Whole mitochondrial genome scan for population structure and selection in the Atlantic herring. BMC Evol Biol. 2012;12(1):248. https://doi.org/10.1186/1471-2148-12-248.
127. Brandão SN, Sauer J, Schön I. Circumantarctic distribution in Southern Ocean benthos? A genetic test using the genus Macroscapha (Crustacea, Ostracoda) as a model. Mol Phylogenet Evol. 2010;55(3):1055–69. https://doi.org/10.1016/j.ympev.2010.01.014.
128. Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung Bremerhaven Germany. Polar Research and Supply Vessel POLARSTERN Operated by the Alfred-Wegener-Institute. J Large Scale Res Facil. 2017;3:A119.

129. d'Udekem d'Acoz C, Schön I, Robert H. The genus charcotia chevreux , 1906 in the southern ocean, with the description of a new species. Belgian J Zool. 2018;148:31–82.
130. Jurajda P, Roche K, Sedláček I, Všetičková L. Assemblage characteristics and diet of fish in the shallow coastal waters of James Ross island, Antarctica. Polar Biol. 2016;39:2299–309.
131. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, et al. The Ensembl genome database project. Nucleic Acids Res. 2002;30(1):38–41.
132. Gregory TR. Animal Genome Size Database [Internet]. 2021 [cited 2019 Jan 9]. Available from: http://www.genomesize.com
133. Dolezel J, Bartos J, Voglmayr H, Greilhuber J. Nuclear DNA content and genome size of trout and human. Cytom A. 2003;51(2):127–8.
134. Lepais O, Weir JT. SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. Mol Ecol Resour. 2014;14(6):1314–21.
135. R Core Team. R: a language and environment for statistical computing [internet]. Vienna, Austria: R Foundation for statistical Computing; 2021. Available from: http://www.r-project.org/
136. Foley J. bioanalyzeR: Analysis of Agilent electrophoresis data. R package version 0.5.1. 2020; Available from: https://stanford.edu/~jwfoley/bioanalyzeR.html
137. Müller K. here: A Simpler Way to Find Your Files. R package 1.0.1. 2020; Available from: https://cran.r-project.org/package=here
138. Charif D, Lobry J. SeqinR 1.0–2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman H, Vendruscolo M, editors. Structural approaches to sequence evolution: Molecules, networks, populations. New York: Springer Verlag; 2007. p. 207–32.
139. Wickham H, Averick M, Bryan J, Chang W, D'Agostino McGowan L, François R, et al. Wlecome to the tidyverse. J Open Source Softw. 2019;4(43):1686.
140. Xiao N. ggsci: scientific journal and sci-fi themed color palettes for "ggplot2". R package version 2.9. 2018; Available from: https://cran.r-project.org/package=ggsci
141. Auguie B. gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. Available from: https://cran.r-project.org/package=gridExtra
142. Gao G, Nome T, Pearse DE, Moen T, Naish KA, Thorgaard GH, et al. A new single nucleotide polymorphism database for rainbow trout generated through whole genome resequencing. Front Genet. 2018;9:147.
143. Andrews S. FastQC: a quality control tool for high throughput sequencing data [internet]. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.