Received: 4 September 2020 | Accepted: 15 January 2021

DOI: 10.1111/2041-210X.13562

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Removing the bad apples: A simple bioinformatic method to improve loci-recovery in de novo RADseq data for non-model organisms

José Cerca[1,2,3] | Marius F. Maurstad[1,4] | Nicolas C. Rochette[5,6] | Angel G. Rivera-Colón[5] | Niraj Rayamajhi[5] | Julian M. Catchen[5] | Torsten H. Struck[1]

[1]Frontiers in Evolutionary Zoology, Natural History Museum, University of Oslo, Oslo, Norway; [2]Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA; [3]Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology, Trondheim, Norway; [4]Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway; [5]Department of Evolution, Ecology, and Behavior, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA and [6]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA, USA

Correspondence
José Cerca
Email: jose.cerca@gmail.com

Handling Editor: Robert Freckleton

Abstract

1. The restriction site-associated DNA (RADseq) family of protocols involves digesting DNA and sequencing the region flanking the cut site, thus providing a cost and time-efficient way for obtaining thousands of genomic markers. However, when working with non-model taxa with few genomic resources, optimization of RADseq wet-lab and bioinformatic tools may be challenging, often resulting in allele dropout—that is when a given RADseq locus is not sequenced in one or more individuals resulting in missing data. Additionally, as datasets include divergent taxa, rates of dropout will increase since restriction sites may be lost due to mutation. Mitigating the impacts of allele dropout is crucial, as missing data may lead to incorrect inferences in population genetics and phylogenetics.
2. Here, we demonstrate a simple pipeline for the optimization of RADseq datasets which involves partitioning datasets into subgroups, namely by reducing and analysing the dataset at a population or species level. By running the software Stacks at a subgroup level, we were able to reliably identify and remove individuals with high levels of missing data (bad apples) likely stemming from artefacts in library preparation, DNA quality or sequencing artefacts.
3. Removal of the bad apples generally led to an increase in loci and decrease in missing data in the final datasets.
4. The biological interpretability of the data, as measured by the number of retrieved loci and missing data, was considerably increased.

José Cerca and Marius F. Maurstad co-first authors.

Julian Catchen and Torsten H. Struck co-last authors.

KEYWORDS
ddRADseq, genetics, genome, genomics, library preparation, methods, pipeline, RADseq

1 | INTRODUCTION

The establishment of high-throughput sequencing, together with bioinformatic processing tools—the genomics revolution—has impacted biology during the last decade by resolving long-standing questions in phylogenetics (Abalde et al., 2019; Rochette et al., 2014; Struck et al., 2011), speciation and adaptation (Birkeland et al., 2020; Ravinet et al., 2017, 2018; Weber et al., 2019), and opening new venues of research such as genome-wide structural variants (Catchen et al., 2020; Faria et al., 2019). While genome-level data have become widely accessible, population-level (i.e. population genomics) and species-level (i.e. phylogenomics) inference remains challenging due to the limited number of high-quality genomes and the costs associated with sequencing and analysing large datasets.

These challenges have encouraged the development and establishment of reduced-representation sequencing (RRS), where genomic complexity is reduced by sequencing only a portion of the genome. Chief among RRS is the 'Restriction site-Associated DNA Sequencing' (RADseq; Baird et al., 2008; Davey et al., 2011), a family of techniques which involve digesting DNA using type-II restriction enzymes and sequencing the flanking regions of the cut site. Benefiting from the distribution of restriction sites over the genome, RADseq-based approaches are cost and time efficient, typically providing thousands of independent loci for population and species-level inference. For instance, Rochette et al. (2019) estimated that for the price of a single whole genome resequenced three-spined stickleback *Gasterosteus aculeatus*, >100 individuals may be sequenced at similar depth using RADseq, which would only cover ~3% of the genome.

Since RADseq-based approaches rely on the existence of cut sites along the genome, the conservation of the cut site is of critical importance for recovering shared data among different individuals (Eaton et al., 2017; Huang & Lacey Knowles, 2016; O'Leary et al., 2018). Allele dropout occurs when a given locus or allele is not sequenced in one or more individuals, and it may result from biological divergence—when a mutation modifies the cut site. Rates of allele dropout are thereby expected to be correlated with the divergence between lineages (Crotti et al., 2019; Eaton et al., 2017; O'Leary et al., 2018). However, allele dropout may also result from artefacts in the experimental design, such as sampling bias and low sequence coverage; or from problems associated with library preparation, such as issues with enzyme digestion or size selection, and human error; or from challenges in DNA extraction since, for some organismal groups, extracting DNA may still be non-trivial due to their reduced size or presence of chemical compounds which may interfere with the extraction; or from artefacts from bioinformatic analyses, such as problems associated with clustering of sequencing reads (Crotti et al., 2019; O'Leary et al., 2018). In fact, allele dropout originating from these technical artefacts can sometimes exceed dropout of biological origin under certain experimental conditions (Rivera-Colón et al., 2020). Whatever the case may be, high allele dropout translates to high rates of missing data in the dataset, which may dramatically influence allele frequency in the dataset (Arnold et al., 2013; Gautier et al., 2013; Hodel et al., 2017), or phylogenetic reconstruction (Crotti et al., 2019; Eaton et al., 2017).

Attempts to minimize the challenges posed by allele dropout include suggestions for parameter optimization and control (Paris et al., 2017; Rochette & Catchen, 2017), data-filtering and data exploration (O'Leary et al., 2018), data-cleaning thresholds (Crotti et al., 2019), and prospective and retrospective data simulation based on the reference genome available (Rivera-Colón et al., 2020). Despite these, retrieving an optimal dataset from a RADseq experiment can pose challenges. On the one hand, technical expertise involving enzyme selection, library size selection and selection of the number of PCR rounds (broadly library preparation) may be scarce for biologists working with non-model systems. On the other hand, post-sequencing approaches to filter data may lead to pruning of informative loci (Huang & Lacey Knowles, 2016; Lee et al., 2018), or to the retention of loci with particular characteristics in datasets with varying levels of species divergence (Dincă et al., 2019; Hodel et al., 2017).

Here, we suggest a simple method to mitigate allelic dropout issues in datasets where biological allele dropout is mixed with allele dropout associated with biases in library preparation, experimental design and bioinformatic processing of the data. Simply put, by processing RADseq data in subgroups, which is at the population or species level (Figure 1), users will better distinguish between biological and technical sources of allele dropout. While most RAD studies are comprised of a metapopulation, composed of several subpopulations, most analyses focus on optimizing parameters across the metapopulation as a whole. Instead, here we suggest optimizing parameters directly in each subpopulation or subspecies. Using this approach in four datasets, we identified individuals with a high degree of missing data (hereafter bad apples) and removed them from the final analysis comprising all populations (Figure 1).

2 | MATERIALS AND METHODS

2.1 | Data

To test the suggested pipeline, we used four datasets including: (a) a ddRADseq dataset comprising populations of the meiofaunal annelid *Stygocapitella zecae* (J. Cerca et al, unpubl. data; 21 samples, six populations); (b) a single-digest RADseq dataset comprising several species of *Euhadra* molluscs (Richards et al., 2017); 16 samples, four species); (c) a ddRADseq dataset comprising populations of the Antarctic sponge *Dendrilla antarctica* (Leiva et al., 2019; 62 samples,

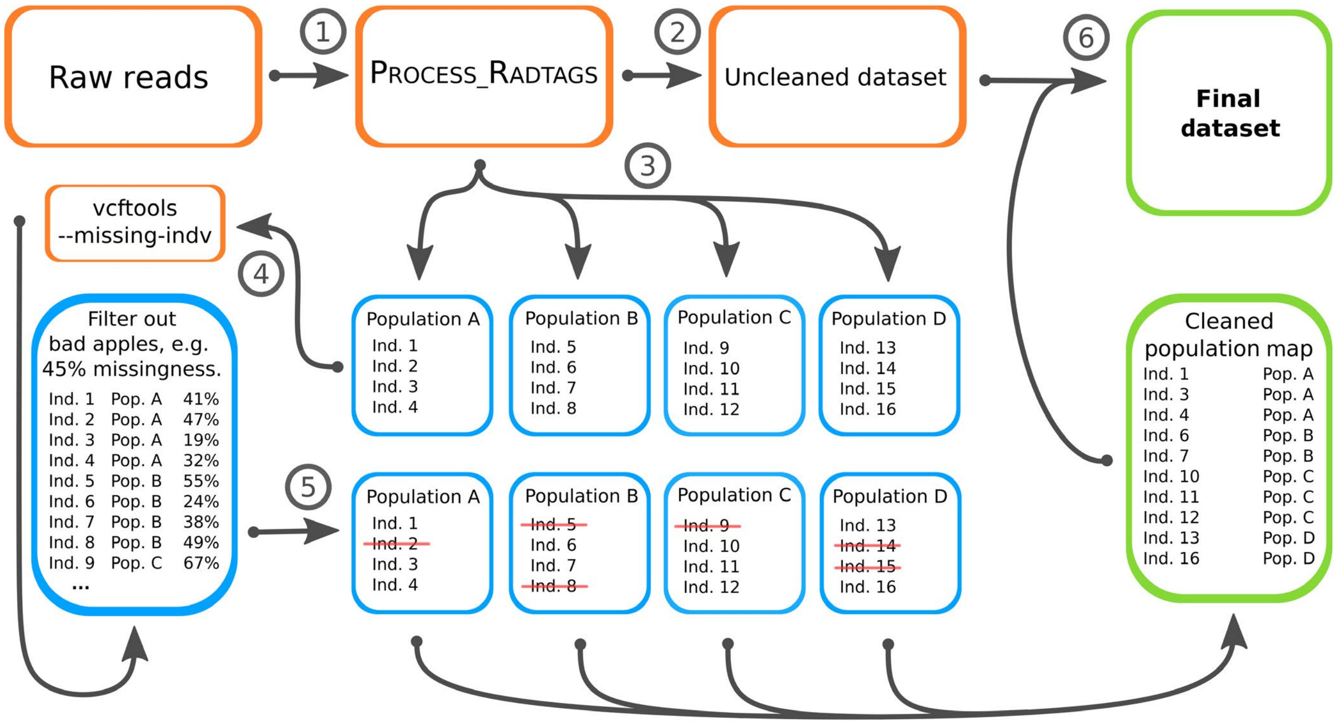

FIGURE 1 Overview of the Stacks-based bad apples pipeline. 1—raw reads are first processed with process_radtags, which removes reads without barcode and cut-site, and de-multiplexes the raw-reads; 2—a Stacks run (u/c/s/gstacks) consisting of the whole dataset; 3—identification of bad apples by running Stacks separately for each population or species (depending on the scale of the dataset); 4—determination of bad apples based on individual-level missing data, as estimated with vcftools (--missing-indv); 5—identification and removal of bad apples from the population map (the file that Stacks uses to assign individuals to populations or species); 6—using the initial Stacks run (unclean dataset) together with a cleaned population map, users retrieve the final dataset using the populations module which will filter the bad apples from the original dataset (hybrid-clean approach)

seven populations); and (d) a hyRADseq dataset of *Anthochaera phrygia* combining museum and modern samples (Crates et al., 2019; 230 samples, eight populations).

2.2 | Bioinformatic processing

A graphical overview of the bioinformatic processing is provided in Figure 1. We opted for a subgroup (i.e. population-by-population or species-by-species) analysis pipeline because analysing missing data at the level of the whole dataset would lead to the removal of whole populations—especially those which are more divergent. By running subgroup-level analyses, we explore the data at a reduced biological divergence level, thereby operating at a scale where most dropout is expected to derive from artefacts associated with molecular biology and sequencing. We processed the four datasets using the *de novo* pipeline implemented in Stacks v2.41 (Rochette et al., 2019). We began by optimizing the clustering parameters -M (number of mismatches allowed between stacks within individuals) and -n (number of mismatches allowed between stacks between individuals) following Paris et al. (2017)'s method to optimize RADseq data in Stacks. Essentially, this method involves running Stacks with different -M -n combinations and determining the number of loci obtained, thereby choosing the 'right' parameter space - i.e. avoiding over-splitting or over-merging the data. We selected -M 2 -n 2 for the final analysis of every dataset, with the exception of the *Anthochaera phrygia*, where we selected -M 3 -n 3. Using these parameters, we ran Stacks using the de novo wrapper thereby generating a dataset with the complete number of individuals available (hereafter unclean datasets; Figure 1, steps 1–2).

After obtaining the unclean datasets, we ran Stacks for each population individually for the *Stygocapitella*, *Dendrilla* and *Anthochaera* datasets (Figure 1, step 3), and for each species separately in the *Euhadra* datasets (using default parameters and applying -M 2 -n 2 for all datasets with the exception of *Anthochaera*, where -M 3 -n 3 was used; Tables S1–S4), and generated a variant call format (VCF) file for each. We obtained information on missing data for each individual using vcftools (Danecek et al., 2011) (--missing-indv option; Tables S1–S4; Figure 1, step 4). Since coverage is a common problem in genomic-level studies, we also retrieved coverage information using vcftools (--depth option; Tables S1–S4). With the whole dataset in mind, we labelled individuals as to keep or remove (bad apples), following a general strategy which included: (a) retaining a minimum of two individuals per population or species; (b) designating a threshold for missing data, based on the average missing data for each whole dataset (≥40% missing data for *Stygocapitella*, ≥30% for *Euhadra*, ≥65% for *Dendrilla*, ≥40% for *Anthochaera*). Notice that we kept some individuals with high missing data in the *Anthochaera* dataset since they are valuable historical specimens (Figure 1, step 5). We have done this to recreate the conditions of a typical ancient-DNA study. Additionally, since there was a very high range of missing data for *Dendrilla*, we also ran populations with -r 0.2 (minimum percentage of individuals in a population required to process a locus for that population) for this dataset.

After the identification of the bad apples, we generated three new datasets: clean, hybrid-clean and random. The clean dataset comprised only kept samples (that is every specimen not labelled as a bad apple) and involved rerunning the whole Stacks pipeline (starting in ustacks). The hybrid-clean dataset also included the same specimens as the clean one (all kept specimens) but involved reusing the unclean Stacks output as represented in Figure 1, step 6. Essentially, Stacks assembles loci across all samples first using u/c/s/gstacks, and then the data are filtered with the populations module. Therefore, the Stacks run behind the hybrid-clean dataset includes all individuals, and bad apples are only excluded at the filtering step. Finally, to understand the overall impact of removing specimens, we performed 10 random runs, where we removed the same number of specimens as the number of bad apples detected; however, specimens were removed haphazardly. The aim of the random dataset was to assess the effect of removing a certain number of specimens on the final dataset.

2.3 | Assessment of the results

To determine differences between the unclean, clean, hybrid-clean and random datasets at different filtering options, we generated (a) the overall number of loci (regardless of the number of SNPs per locus) provided by populations at different filtering thresholds, (b) the % of missing data also at different filtering thresholds (Table S5) and (c) explored whether there are differences between kept/removed loci. Specifically, to obtain the (a) number of loci, we ran populations with a fixed -p (minimum number of populations a locus must be present in to process a locus; -p 4 for *Stygocapitella zecae*, -p 4 for Euhadra spp., -p 2 for *Dendrilla antarctica* and -p 2 for *Anthochaera phrygia*) and with -R of 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% (minimum percentage of individuals across populations required to process a locus). We opted for a fixed -p since the number of populations was constant across datasets, but varied -R since it relies on the number of individuals, which varies between the unclean and the remaining data. To obtain (b) estimates of missing data, we obtained estimates of missing data using vcftools, as reported above for the identification of bad apples. For both the number of loci and missing data, we determined the relative difference of the evaluated datasets (clean, hybrid-clean and random) to the value of the unclean dataset. In other words, first, we obtained the difference in the number of loci and missing data between the evaluated datasets and the unclean dataset; second, we divided the obtained result with the value of the unclean dataset. The relative differences were obtained as percentages by multiplying by 100. We restricted these analyses to filtering thresholds, which generated at least 100 loci in the unclean dataset, since small denominators generate high relative values even when facing small changes. For example, in *Stygocapitella*, the filtering of -p 4 -R 1 generates only one locus in the unclean dataset and two in the hybrid-clean and clean ones. In relative terms, this is a 100% increase in loci, but, in practical terms, this does not translate to an improvement. Moreover, studies using RADseq do not typically rely on small numbers of loci, but usually consider hundreds or thousands of loci. Therefore, only filtering stages up to -R 0.6 were considered for *Stygocapitella*, -R 1.0 for *Euhadra*, and -R 0.5 for *Dendrilla* and *Anthochaera* (Table S5). Finally, (c) we explored whether removing individuals affected particular classes of loci by comparing the hybrid-clean and unclean datasets, which share the same set of assembled loci (catalogue in Stacks), since they diverge only at the filtering stage. To do so, we converted the data to a present/absence data format and plotted the number of (a) loci kept in the whole dataset, (b) all loci in the bad apples, (c) all loci in kept specimens, (d) loci kept in the bad apples and (e) removed loci in bad apples.

A concern with removing bad apples in a genomic dataset is whether users remove genetically distinct individuals that may, for various reasons, occur in nature. To explore this, we performed a principal component analysis (PCA) at a subgroup level, labelling bad apples and kept specimens. We also tested for deviations in nucleotide diversity (π) and Watterson's estimator (θ) between the uncleaned, hybrid-cleaned and cleaned datasets. The PCA was chosen since its calculation assumes average values when data are missing for a particular locus. As a result, individuals with high missing data will be represented in the middle of the PC axis. Deviation in π and θ should provide further evidence on whether we removed genetically distinct individuals. To carry out a PCA, we used the --write-random-SNP option while generating a variant call format file (vcf) using populations, so that linkage disequilibrium is removed. The vcf was then loaded to R using the package vcfR (Knaus & Grünwald, 2017), and a PCA was carried using functions included in the adegenet package (Jombart & Ahmed, 2011). For π and θ calculations, we retrieved fasta sequences using populations. Using custom perl and unix scripts, we split the fasta sequences into loci, and these, in turn, were split according to the subgroup (i.e. Locus1_populationA; Locus 1_populationB; … Locus N_population_X). From these, we selected loci with least missing data, by selecting the loci present in at least the number of kept specimens (i.e. in a population with nine individuals, including five kept specimens and four bad apples, we kept loci present in five or more specimens). These loci were then loaded in DNAsp v6 (Rozas et al., 2017), where we calculated loci-by-loci π and θ, as well as the averages for the whole population.

3 | RESULTS

3.1 | Data cleaning

There was a wide variation in terms of variant sites (SNPs) and missing data at the sample level (hereafter referred simply to missing data) when running different populations separately in all four datasets (Tables S1–S4). Nonetheless, the values for the number of variants and missing data do not correlate. For instance, in the *Stygocapitella* dataset, at one extreme the Ardtoe population has a total of 37,028 SNPs, while at the other end the Lødingen population has 2,821 SNPs (Table 1; random SNP per locus). Despite these differences, both populations have an average missingness of 53% and 55% respectively. These two, together with Henningsvær (53% missing data), are at the higher distribution of missing data (Table 1). The lowest value for missing data is 22% and found in Cutty Sark. According to the established protocol to label and remove bad apples, we removed one individual from Ardtoe with 97% missing data, one individual from Lødingen with 71% and two from Henningsvær with 60% and 94%. No individual was removed from the Cutty Sark population (Table S1). The remaining populations, Kristineberg and Musselburgh, have a number of variants within

TABLE 1 Results of the subgroup (population or species level) analyses including the number of variant sites and the average missing data. Total number of variant sites per population/species and average missingness per population/species for all individuals, all individuals below the threshold and all included individuals are given. Since there was a very high range of missing data for *Dendrilla antarctica*, we also ran populations' filtering with -r = 0.2 (minimum percentage of individuals in a population required to process a locus for that population) for this species

	Variant sites	Average missingness in %		
Population/species	**#**	**All individuals**	**All < threshold**	**All included**
Stygocapitella zecae				
Ardtoe	37,028	53	42	42
Cutty Sark	11,059	22	22	22
Henningsvær	13,893	53	28	28
Kristineberg	20,158	45	41	41
Lødingen	2,821	55	48	48
Musselburough	6,172	48	38	38
Euhadra spp.				
Euhadra aomoriensis	33,866	28	22	22
Euhadra quaesita	40,493	23	21	21
Euhadra senckenbergiana	19,126	10	10	9
Dendrilla antarctica				
Den_CIE (without -r)	5,155	74	–	–
Den_CIE (with -r 0.2)	4,389	70	41	41
Den_DEC (without -r)	4,829	68	–	–
Den_DEC (with -r 0.2)	4,083	65	44	44
Den_FIL (without -r)	3,613	77	–	–
Den_FIL (with -r 0.2)	1,581	62	52	52
Den_HM (without -r)	2,162	67	–	–
Den_HM (with -r 0.2)	1,944	64	55	55
Den_OH[a]	2,086	44	44	37
Den_PAR[a]	636	39	39	30
Den_ROT (without -r)	6,317	71	–	–
Den_ROT (with -r 0.2)	3,299	57	47	44
Anthochaera phrygia				
ACT	937	38	20	24
ADL	428	34	29	31
NA	2,656	50	28	41
NNSW	361	33	25	27
NVIC	396	33	23	29
QLD	76	36	28	32
SVIC	467	40	23	31
BMTN	305	38	22	28

[a] This population was assessed without filters due to the lack of data using filters and the average values for all individuals below the threshold and all included individuals were taken without the filters

the aforementioned ranges, and we removed one individual from each with 60% and 68% missing data respectively (Table S1). In total, we removed six individuals from a total of 21 (i.e. 29% of the dataset was removed; Table S1). This improved the average missing data for each population except for Cutty Sark, where no individuals were excluded (Table 1). The strongest change could be observed for the Henningsvær population, which decreased from 53% to 28% missingness. Importantly, bad apples do not strictly correlate with sequencing coverage. For example, in Henningsvær and in Kristineberg, the samples with the highest coverage (80× and 75×) were identified as a bad apple.

Of all four datasets, the *Euhadra* dataset has the highest number of SNPs and the lowest missing data (Table 1). When running species separately, *Euhadra aomoriensis* has the highest average missingness, with 28% over 33,866 SNPs, while *E. quaesita* yielded the most SNPs (40,493) and an average of 23% missing data. *Euhadra senckenbergiana* had the lowest numbers in terms of variants and average missing data, with 19,126 SNPs and 10% respectively. In *E. aomoriensis*, we identified and removed two individuals with 44% and 34% missing data. In *E. quaesita*, we removed one individual with 30% missing data (threshold of 29%; Table S2). We also removed the individual of *E. senckenbergiana* with the highest degree of missingness (12%, Table S2), since we wanted to explore the effect of removing several individuals. In total, we removed four individuals from a total of 15 (27% of the dataset removed; Table S2). The average missingness for each population was decreased with the steepest decrease in *E. aomoriensis* (Table 1). In *Euhadra*, we found a correlation between missing data (and therefore the labelling of bad apple) and coverage (Table S2).

The *Dendrilla* dataset has the highest degrees of missingness of all four datasets. When running different populations separately, average missingness values range from 39% in the population Den_PAR to 77% in Den_FIL when no filtering is applied ('-r' flag, Table 1). The filter '-r 0.2' decreased missingness (Table 1), but the populations Den_OH and Den_PAR have no SNPs left. These are also the two populations with lowest number of SNPs (i.e. 2,086 and 636) when no filtering was applied. The Den_ROT population yielded the most SNPs, with 6,317 without filtering (decreasing to 3,299 SNPs after filtering). The difference between the obtained number of loci before and after filtering was less substantial in the Den_CIE population; in specific, there were 5,155 and 4,389 before and after filtering (Table 1). Using a threshold of 64% missing data, we removed at least one individual from each population (Table S3). The filtering was, in some cases, quite rigorous since it excluded a substantial number of individuals. For instance, in Den_CIE, we excluded as many as six out of nine individuals. We removed 31 individuals from a total of 62 (50% of the dataset; Table S3). Due to this rigorous removal of individuals, the average missingness decreased substantially in some populations (Table 1). For example, in Den_CIE, missing data decreased from 70% to 41%. Similar to the *Stygocapitella* dataset, there was no strict correlation between coverage and missing data, with multiple populations having the samples with most coverage removed (Table S3).

Finally, the *Anthochaera* dataset had the lowest number of SNPs across populations ranging from 76 in the QLD population to 2,656 in the NA population (Table 1). The extent of missing data is comparable to that observed in the *Stygocapitella* dataset, with the NA population when ran separately, having the most missing data (50%), and NNSW and NVIC having only 33% (Table 1). We removed at least one individual in each population, following a threshold of 39% (Table S4). Importantly, we kept some individuals above this threshold since the dataset included valuable individuals, namely historical specimens. We removed 41 individuals from a total of 230 (18% of the dataset was removed; Table S4), which led to a decrease in the missing data (Table 1). However, due to some of the retained individuals, which have substantially high values of missing data, the decrease is not substantial. For example, the decrease in the SVIC population is only from 40% to 31% instead of 23% (Table 1). In agreement with the *Stygocapitella* and *Dendrilla* datasets, we observed that bad apples do not correspond to samples with low coverage (Table S4).

3.2 | Improvement in the datasets after the removal of *bad apples*

In general, the clean and hybrid-clean datasets yielded substantially more loci, in some cases as much as 300% more than the unclean dataset (Figure 2). Notably, the relative difference increases with decreasing numbers of loci in the unclean dataset. This effect cannot merely be attributed to having a smaller number of samples, as discussed above, since the random datasets yield less loci than the cleaned and hybrid-cleaned, despite having the same number of samples removed. This therefore confirms that bad apples have an overall negative effect on the number of loci of a dataset. While these effects are noticeable in every dataset, they are least pronounced in *Euhadra*. This is because *Euhadra* was, relatively, the best of all the analysed datasets, since it had lower missing data and a higher number of loci. In contrast to the remaining three datasets, in *Euhadra*, all -R filtering had >100 loci (Table S5). However, even in an excellent dataset, such as *Euhadra*, the increase in loci in the clean and hybrid-clean datasets at lower numbers of loci is easily observed (Figure 2). This indicates that even in an excellent dataset there is room for improvement and removing bad apples can be performed. Another interesting observation is that only at the least restrictive settings (-R 0, -R 0.1), which result in the greatest number of loci, there is no negative effect of removing the bad apples on the number of loci across all datasets (Figure 2). Considering this, and the difference observed between the *Euhadra* dataset and the other three, it seems that in these cases (i.e. very low restrictive settings or very good dataset) the unclean dataset reaches a high number of loci. However, in all other cases, removing bad apples has a clear, sometimes substantial effect.

Except for the *Anthochaera* dataset, there is no obvious difference between the clean and hybrid-clean approach in terms of the

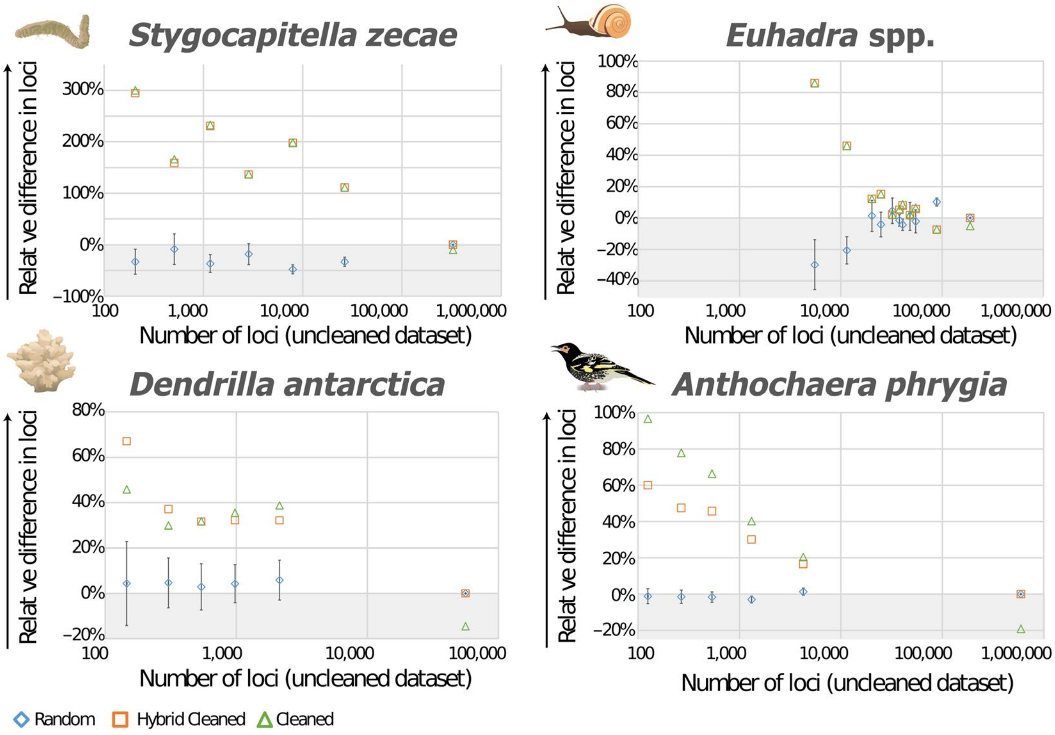

FIGURE 2 The relative difference in the number of loci for the random, hybrid-clean and clean datasets to the unclean dataset for *Stygocapitella*, *Euhadra*, *Dendrilla* and *Anthochaera* datasets in relation to the number of loci in the unclean dataset (the higher the number of loci, the lower the parameter R from the populations module). For the random datasets, the median and the standard deviations are given. The arrows indicate the directions of improvement and the grey zones the areas, where the unclean dataset is performing better

number of loci (Figure 2). In the *Stygocapitella* and *Euhadra* datasets, both these datasets retrieve nearly similar values. In the *Dendrilla* dataset, clean and hybrid-clean perform differently when varying -R values, with the clean performing slightly better at less restrictive settings. In the *Anthochaera* dataset, the clean approach performs much better than the hybrid-clean.

With respect to missing data, removing bad apples also improved the datasets. The only exceptions were two -R levels of the *Euhadra* dataset where there was no difference (0%) between the clean and unclean dataset (Figure 3). In the *Dendrilla* and *Anthochaera* datasets, the random datasets are clearly worse than the clean and hybrid-clean datasets. In the *Stygocapitella* dataset, there were different results following different filtering settings. In four cases, the clean and hybrid-clean approaches are clearly better than the random exclusion, while in three cases the random exclusion was good or even better than clean and hybrid-cleaned. In the *Euhadra* dataset, except for the two most restrictive settings, the cleaned and hybrid-cleaned are clearly better than the random exclusion. This may be explained by this being the best of all datasets, as presented above (Table 1, Table S5), and therefore there may be little room for improvement with respect to missingness at the most restrictive settings.

Considering the two cleaning approaches, clean and hybrid-clean, there is no difference in performance in the *Stygocapitella* and *Euhadra* datasets (Figure 3). In the *Dendrilla* dataset, the clean approach appears to perform slightly better at more restrictive settings (i.e. higher -R, lower number of loci) and the hybrid-clean at less restrictive settings. In the *Anthochaera* dataset, the clean outperforms hybrid-clean only in less restrictive settings.

Considering both the number of loci and the missingness, removing bad apples always improves the dataset, and the random datasets are clearly performing worse than the clean and hybrid-clean approaches (Figure 4). This difference is not as pronounced in the *Euhadra* dataset as in the other three. In the *Stygocapitella* and *Euhadra* datasets, the number of loci is increased and the missingness is reduced in the majority of the settings. However, the relationship between these two measurements is slightly negative. Hence, when there is little or no improvement in missingness, the number of loci increases and vice versa. The two cleaning approaches perform generally very similar in both datasets. In the remaining two datasets, the correlation is slightly different. In the *Dendrilla* dataset, both the number of loci and the missing data generally improve, but there seems to be no correlation between them. In contrast, in the *Anthochaera* dataset, using the hybrid-clean approach led to a generally positive correlation between both, while in the clean approach the improvement in missingness seems constant while the number of loci increases. In the *Dendrilla* dataset, the clean and hybrid-clean approaches perform slightly different, but no clear pattern being observed.

Removal of bad apples does not appear to bias the dataset against a particular class of loci (Figure 5). The removal of bad apples caused a removal of some loci, which had high missingness.

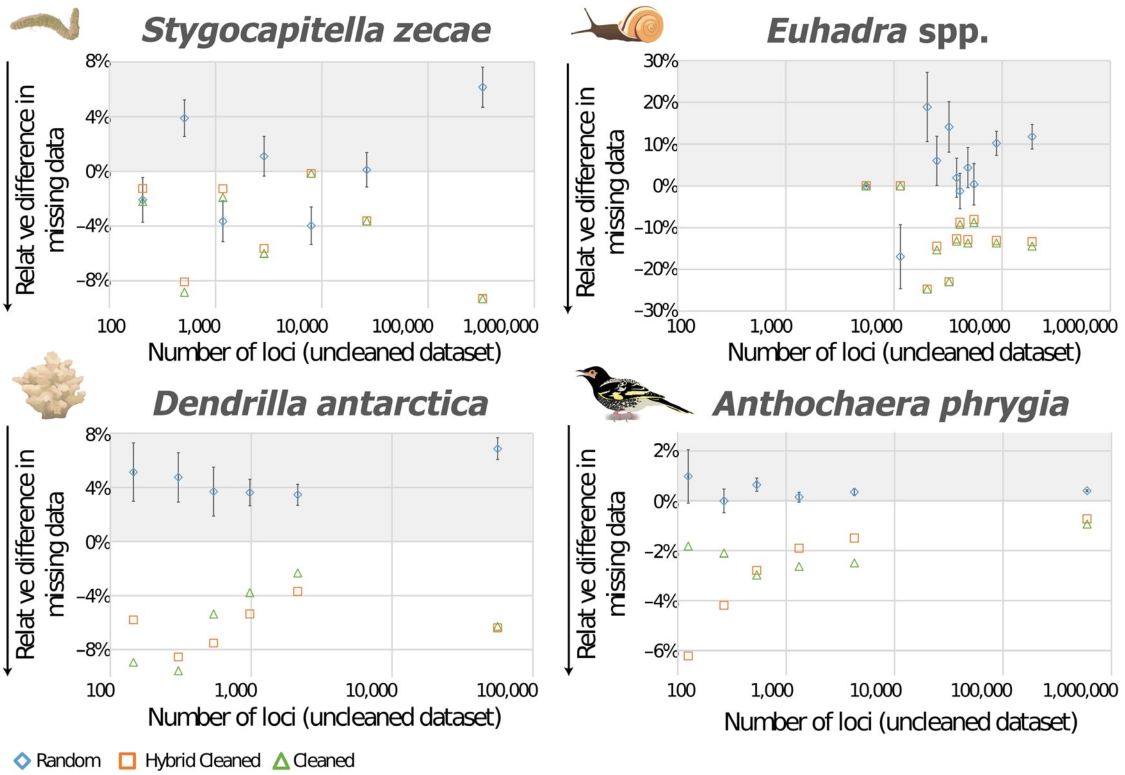

FIGURE 3 The relative difference in the average missingness for the random, hybrid-clean and clean datasets to the unclean dataset for *Stygocapitella*, *Euhadra*, *Dendrilla* and *Anthochaera* datasets in relation to the number of loci in the unclean dataset (the higher the number of loci, the lower the parameter R from the Populations module). For the random datasets, the median and the standard deviations are given. The arrows indicate the directions of improvement and the grey zones the areas, where the unclean dataset is performing better

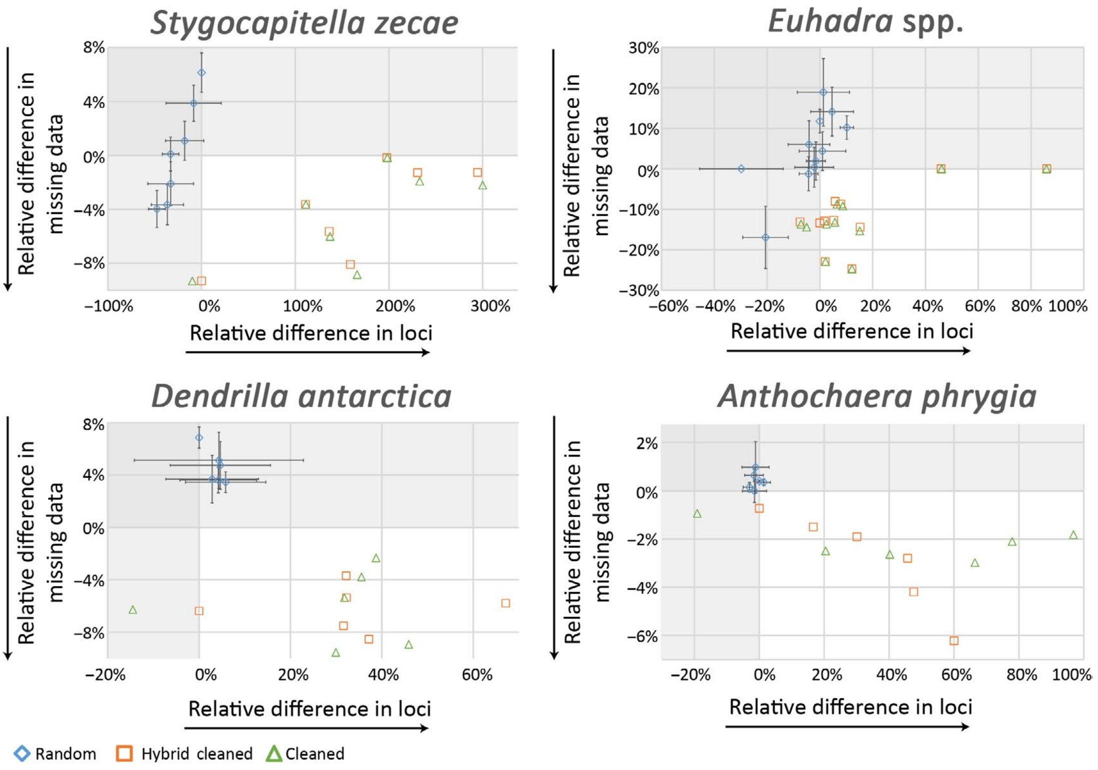

FIGURE 4 The relative difference in the number of loci in relation to the average missingness. For the random datasets, the median and the standard deviations are given. The arrow indicate the directions of improvement and the grey zones the areas, where the unclean dataset is performing better in one or both of the parameters

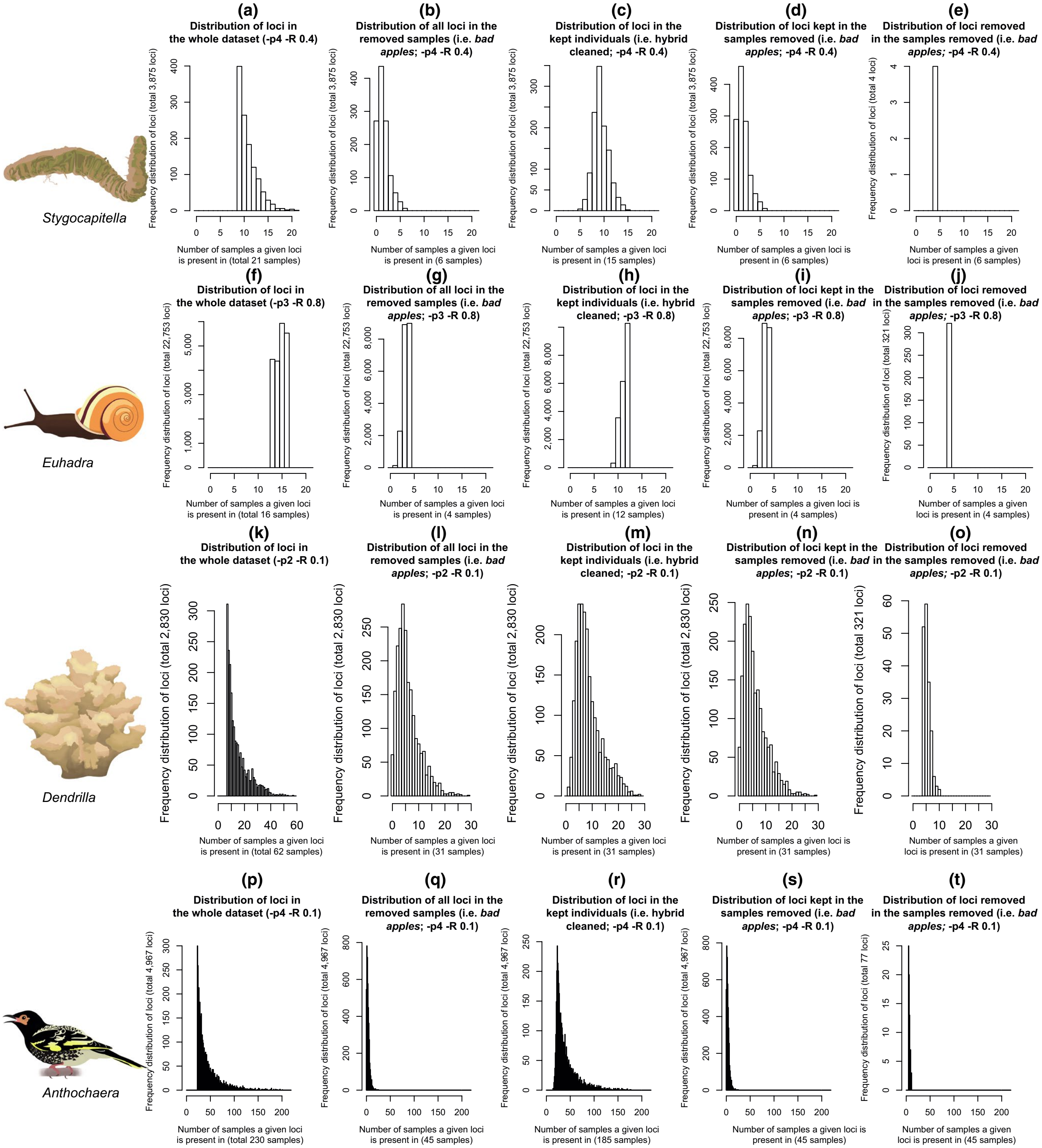

FIGURE 5 Distribution of loci in the whole dataset (a, f, k, p), in the removed samples or bad apples (b, g, l, q) and in the kept individuals (c, h, m, r). Distribution of only kept loci in the samples removed or bad apples (d, i, n, s), and the loci removed in the samples removed or bad apples (e, j, o, t). The four datasets *Stygocapitella* (a–e), *Euhadra* (f–j), *Dendrilla* (k–o) and *Anthochaera* (p–t) are presented. Notice different datasets had different cutting-thresholds

This was also evident from the very small number of removed loci: in the *Stygocapitella* dataset, four loci were removed from a total of 3,875; for the *Euhadra* dataset, 321 loci were removed from a total of 22,753; in *Dendrilla*, 321 loci were removed from a total of 21,140; in *Anthochaera* 77 loci were removed from a total of 4,967. This suggests that the removal of some loci after the exclusion of bad apples is driven by the inherent stochasticity of locus presence/absence in both good individuals and bad apples, rather than by systemic differences in the set of loci present in bad apples.

3.2.1 | No removal of outliers

The removal of samples generally did not remove outliers from the dataset. In a PCA, individuals with most missing data are pulled to the middle since the algorithm assumes average values for those individuals. This is most perceptible in populations with seven or more specimens in the dataset. For instance, in *Dendrilla* and *Anthochaera*, most bad apples represent individuals with central positions in the PCs (Figure S1). However, few samples in the extremities were also removed, as illustrated on DEN_OH, DEN_PAR (populations from *Dendrilla*), ACT, ADL and NA (populations from *Anthochaera*; Figure S1). In *Stygocapitella* and *Euhadra*, subgroup-runs had very few specimens, but generally results were similar to *Dendrilla* and *Anthochaera* (Figure S1). Changes in π and θ between datasets were not very pronounced between the unclean, hybrid-clean and clean datasets (Table S6). The hybrid-clean and clean generally performed similarly in these comparisons.

4 | DISCUSSION

When considering the number of loci and missing data together, removing bad apples (i.e. samples with high missing data) had a positive effect on the datasets, by increasing loci and/or decreasing missing data. This effect is not attributable to the removal of specimens, as the random dataset, where we removed a similar number of samples but chosen randomly, performed worse in comparison to hybrid-clean and clean approaches. Namely, the random retrieved less loci and generally retained more missing data. The hybrid-clean and clean approaches generally yielded similar performances, both in terms of missing data and number of loci (Figures 2–4). While some differences could be observed between clean and hybrid-clean, they do not seem to be predictable, consistent and depended on the dataset and parameter space investigated. This suggests that users should replicate the hybrid-clean approach as it is less resource and time consuming (i.e. Figure 1).

4.1 | Optimization of RADseq data

The identification and removal of the bad apples on the studied datasets yielded up to a threefold increase in the number of loci (Figure 2), at the cost of the removal of 18%–50% of the specimens. The retrieval of more loci allowed filtering the data more thoroughly, thus obtaining a 'high-quality' collection of variants (Paris et al., 2017). For instance, -R 0.6 in the hybrid-clean approach in *Stygocapitella*, a threshold considered for datasets consisting of highly diverged individuals (Paris et al., 2017), yielded more loci than -R 0.5 in the unclean dataset.

Despite these clear benefits, the identification and removal of bad apples should be conducted carefully. First, the principle behind bad apples requires that population and/or species are carefully determined as part of the experimental design. For most studies, the determination of populations and species is done a priori and is required by 'population maps', included as part of Stacks. The lack of precision, such as the inclusion of individuals from different populations together, may lead to the pruning of individuals from a minor/deviant genetic background. This may be particularly difficult for, for instance, marine populations where where the determination of population limits remains challenging (Cerca et al., 2018; Hellberg, 2009), or in cases where individuals from morphologically similar species (cryptic species) are overlooked (Struck & Cerca, 2019; Struck et al., 2018) and potentially considered as bad apples. However, these cases may be a minority in the landscape of RADseq studies. Second, some individuals may be of particular interest. For instance, hybridization or incomplete lineage sorting contributes to shifts in allelic frequencies (Sætre & Ravinet, 2019). If very divergent alleles lead to high missing data, and are restricted to only some samples, admixed individuals may then be wrongly pruned out. Third, historical specimens may be precious, as in the case of *A. phrygia*, even if yielding a high rate of allelic dropout. In practice, this translates that RADseq users need of carefully understand their data. Using a PCA, where samples which are genetically distinct will be easily identified, we found that removing bad apples generally targeted samples occupying intermediate positions in the PC ordinates. However, we also note that some exceptions to this pattern occurred, with some non-intermediate samples being removed. While these concerns may be applicable to particular datasets, we recommend researchers should always carefully analyse their data, either through simulation-based assessments (Rivera-Colón et al., 2020) or through approaches which decompose genetic variation such as principal components, so that users are guaranteed that they are not pruning outliers from their data (Figure S1).

4.2 | Mitigating allele dropout

Current strategies to mitigate dropout focus on improving laboratory practices and bioinformatics, however they may not work for every case. For instance, high quantities of high-quality DNA are desirable, but this is difficult to achieve for many non-model taxa. In the *Stygocapitella* dataset, a whole genome amplification was done to increase DNA concentration. While this can be a powerful approach for microscopic eukaryotes, it may, nonetheless, introduce biases in RADseq datasets (de Medeiros & Farrell, 2018). In the *Anthochaera* dataset, >100-year-old museum samples were included, thus yielding highly fragmented and low-concentration DNA. In these cases, optimization of libraries may be limited and, therefore, bioinformatic optimization may be needed. Attempts to mitigate dropout and its downstream issues include the removal of alleles below a certain coverage and identifying loci with high variance in read depth among individuals (O'Leary et al., 2018). Yet, these thresholds may not be applicable to the *Stygocapitella* and *Anthochaera* datasets. In the first case, it is because whole genome amplification may lead to differences in DNA coverage, as some strands of DNA may be over-represented after whole genome duplication, thus translating to differences among individuals. In the second case, coverage of historical samples may be biased due to their inherent properties. Thus, suggested methods based on depth filters may not work (O'Leary et al., 2018).

The proposed method allows distinguishing between the two sources of allele dropout, with clear benefits for RADseq population genetics inference. Allele dropout may stem from biological divergence (mutation in a restriction site; Ravinet et al., 2016) or from artefacts in library preparation or sequencing (O'Leary et al., 2018). By running Stacks at a subgroup level (population-by-population or species-by-species), we are able to set biological divergence aside during the optimization step of the pipeline, thereby isolating dropout stemming from artefacts in library preparation and sequencing. Removal of bad apples therefore targets poorly sequenced and prepared samples. Lowering the rates of allele dropout in RADseq inference is of significant importance since high rates of dropout can lead to bias in the estimation of various population-level statistics. For instance, high allele dropout (expressed by high levels of missing data) may lead to dramatically inflated estimates of F_{ST} and heterozygosity, and deflated rates of F_{IS}, as determined by comparisons between simulated data and empirical data (Arnold et al., 2013; Gautier et al., 2013; Hodel et al., 2017). Inflation of these metrics occurs because, with lower sample sizes, the extent of intra-population diversity is not represented and, for example, when comparing between populations, estimates of F_{ST} tend to be higher. Therefore, mitigation of allele dropout should be a priority when designing RADseq-based projects. In agreement with these works (Arnold et al., 2013; Gautier et al., 2013; Hodel et al., 2017), we report slight differences between π and θ in the clean and hybrid-clean when compared to the unclean dataset. While we cannot determine the exact drivers of these changes, it is likely they may be attributed to the decreasing of missing data in the cleaned and hybrid-cleaned datasets. We therefore conclude that the clean and hybrid-clean datasets may yield more correct estimates of π and θ.

Phylogenetic inference will also benefit from decreased rates of allele dropout. Best practices for the optimization of RADseq datasets for phylogenetic inference suggest that researchers pruning datasets for missing data cannot be too stringent or too permissive (Crotti et al., 2019) as, in either case, loci kept may have particular characteristics (Lee et al., 2018). On one hand, being conservative may exclude fast-evolving loci, thus jeopardizing the resolution of terminal branches (Eaton et al., 2017; Huang & Lacey Knowles, 2016; Lee et al., 2018). On the other hand, being too permissive may jeopardize phylogenetic inference as the signal to noise ratio will be blurred by missing data (Crotti et al., 2019). An important result from our approach is that excluded loci do not seem to differ from kept loci in any obvious way, suggesting that removing bad apples does not bias the final set of loci. In this way, recovery of more loci and reduction of missing data when building a phylogenetic data matrix may allow researchers to obtain improved inferences.

While we have optimized and applied currently suggested filtering method using the pipeline Stacks, we expect this approach to be applicable to other pipelines focusing on RADseq data analysis. This expectation results from other RADseq analysis pipelines, including ipyrad and dDocent (Eaton & Overcast, 2020; Puritz et al., 2014), analyse sequences in a similar framework as Stacks. Since these also output similar files (e.g. genotypes, vcf), implementing the bad apples approach is likely to yield similar results, and should be as easy to implement as it is for Stacks.

5 | CONCLUSIONS

The biggest advantage of genomics, that is the retrieval of a large amount of genetic data, is intimately coupled with its biggest hindrance, that is biases associated with big data. RADseq-based methods allow obtaining genomic-level data for phylogenetic and population genetic inference at affordable costs for organisms where reference genomes lack. However, optimizing de novo RADseq datasets still remains challenging, particularly when specimens are not closely related and when problems associated with library preparation and sequencing occur. Here, we suggest a simple procedure to mitigate some issues associated with allele dropout, which consists of the identification and removal of individuals with high degrees of missing data (bad apples) on a subgroup level (population-by-population level or species-by-species level). Comparisons of datasets with and without bad apples clearly suggest that removal of bad apples leads to an increase in the number of loci and/or lowering of missing data (Figure 4), while not removing outliers (Figure S1; Table S6). The more robust datasets obtained by removing bad apples are likely to improve phylogenetic and population genetic inferences. We recommend that users:

- Generate an unclean dataset and explore the level of missing data, as we did above, using vcftools (--missing-indv) and by exploring genetic variance-based clustering (principal component analysis).
- Determine bad apples by running Stacks at a reduced level of lineage divergence in the dataset (e.g. population-by-population or species-by-species, depending on the scale of the dataset). After running Stacks at the reduced level, users should obtain missing data using vcftools (--missing-indv) and determine bad apples as specimens by setting a by setting a missing data cut-off. For example, cutting specimens in which missing data is higher than the mean of missing data for that particular population.
- We advise against rerunning the whole pipeline of Stacks to generate a clean dataset, but only rerun the `populations` module, with bad apples removed, therefore reproducing the hybrid-clean dataset herein introduced. In this way, computational resources, including disc space and running time, will be saved.

ACKNOWLEDGEMENTS

The authors thank Ana Riesgo, Carlos Leiva Martinez, Sergio Taboada, George Olah, Angus Davidson and Ross Crates for kindly providing their data. We are grateful to Emma Falkeid Eriksen for the organismal illustrations provided. J.Ce. dedicates this paper to the vision, dedication and hard work of those who run Forbio—the

Norwegian Research School in Biosystematics. Without its workshops, courses and financial support, this paper would not have been possible. J.Ce. is also grateful to all those who fight for climate justice. J.Ce. was partially supported by an Godfrey Hewitt mobility award from the European Society for Evolutionary Biology (ESEB), and UiO:Life Sciences Internationalization fund—two initiates which permitted J.Ce. to visit J.Ca., N.C., A.R.-C. and N.R. in Urbana Champaign. Visiting UIUC greatly benefited J.Ce.'s bioinformatic and genomic skills. We thank thank two anonymous reviewers for their comments which have substantially improved our manuscript. This is NHM Genomics contribution number 23.

AUTHORS' CONTRIBUTIONS

J.Ce., T.H.S. and N.R. designed the study; J.Ce., M.F.M., T.H.S. and N.R. analysed the data with support, advice and infrastructure of A.R.-C., N.R. and J.Ca. J.Ce. drafted the manuscript. All the authors read and approved the draft.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13562.

DATA AVAILABILITY STATEMENT

With the exception of the *Stygocapitella* dataset, all data used here were downloaded from public repositories. The *Euhadra* dataset was generated and first analysed by Richards et al., (2017); the *Dendrilla antarctica* dataset was generated and first analysed by Leiva et al., (2019); and *Anthochaera phrygia* was generated and first analysed by Crates et al., (2019). The *Stygocapitella* dataset has been made public in the European Nucleotide Archive (ENA) under the project id PRJEB40223. Specimen-id can be easily cross-matched between the Table S1 of this work, and the name of the 'Submitted FTP' in ENA.

ORCID

José Cerca https://orcid.org/0000-0001-7788-4367
Nicolas C. Rochette https://orcid.org/0000-0003-1899-1765
Angel G. Rivera-Colón https://orcid.org/0000-0001-9097-3241
Julian M. Catchen https://orcid.org/0000-0002-4798-660X
Torsten H. Struck https://orcid.org/0000-0003-3280-6239

REFERENCES

Abalde, S., Tenorio, M. J., Uribe, J. E., & Zardoya, R. (2019). Conidae phylogenomics and evolution. *Zoologica Scripta*, *48*, 194–214. https://doi.org/10.1111/zsc.12329

Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, *22*, 3179–3190. https://doi.org/10.1111/mec.12276

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*, e3376. https://doi.org/10.1371/journal.pone.0003376

Birkeland, S., Gustafsson, A. L. S., Brysting, A. K., Brochmann, C., & Nowak, M. D. (2020). Multiple genetic trajectories to extreme abiotic stress adaptation in Arctic Brassicaceae. *Molecular Biology and Evolution*, *37*(7), 2052–2068. https://doi.org/10.1093/molbev/msaa068

Catchen, J., Amores, A., & Bassham, S. (2020). Chromonomer: A tool set for repairing and enhancing assembled genomes through integration of genetic maps and conserved synteny. *bioRxiv* 2020.02.04.934711. https://doi.org/10.1101/2020.02.04.934711

Cerca, J., Purschke, G., & Struck, T. H. (2018). Marine connectivity dynamics: Clarifying cosmopolitan distributions of marine interstitial invertebrates and the meiofauna paradox. *Marine Biology*, *165*, 123. https://doi.org/10.1007/s00227-018-3383-2

Crates, R., Olah, G., Adamski, M., Aitken, N., Banks, S., Ingwersen, D., Ranjard, L., Rayner, L., Stojanovic, D., Suchan, T., Von Takach Dukai, B., & Heinsohn, R. (2019). Genomic impact of severe population decline in a nomadic songbird. *PLoS ONE*, *14*, 1–19. https://doi.org/10.1371/journal.pone.0223953

Crotti, M., Barratt, C. D., Loader, S. P., Gower, D. J., & Streicher, J. W. (2019). Causes and analytical impacts of missing data in RADseq phylogenetics: Insights from an African frog (Afrixalus). *Zoologica Scripta*, *48*, 157–167. https://doi.org/10.1111/zsc.12335

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*, 499–510. https://doi.org/10.1038/nrg3012

de Medeiros, B. A. S., & Farrell, B. D. (2018). Whole genome amplification in double-digest RAD-seq results in adequate libraries but fewer sequenced loci. *PeerJ*, *6*, e5089. https://doi.org/10.7717/peerj.5089

Dincă, V., Lee, K. M., Vila, R., & Mutanen, M. (2019). The conundrum of species delimitation: A genomic perspective on a mitogenetically supervariable butterfly. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1911), 20191311. https://doi.org/10.1098/rspb.2019.1311

Eaton, D. A. R., & Overcast, I. (2020). Ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics*, *36*, 2592–2594. https://doi.org/10.1093/bioinformatics/btz966

Eaton, D. A. R., Spriggs, E. L., Park, B., & Donoghue, M. J. (2017). Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology*, *66*, 399–412. https://doi.org/10.1093/sysbio/syw092

Faria, R., Johannesson, K., Butlin, R. K., & Westram, A. M. (2019). Evolving inversions. *Trends in Ecology & Evolution*, *34*, 239–248. https://doi.org/10.1016/j.tree.2018.12.005

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., Cornuet, J. M., & Estoup, A. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, *22*, 3165–3178. https://doi.org/10.1111/mec.12089

Hellberg, M. E. (2009). Gene flow and isolation among populations of marine animals. *Annual Review of Ecology Evolution and Systematics*, *40*, 291–310. https://doi.org/10.1146/annurev.ecolsys.110308.120223

Hodel, R. G. J., Chen, S., Payton, A. C., McDaniel, S. F., Soltis, P., & Soltis, D. E. (2017). Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: Comparing microsatellites and RAD-Seq and investigating loci filtering. *Scientific Reports*, *7*, 1–14. https://doi.org/10.1038/s41598-017-16810-7

Huang, H., & Lacey Knowles, L. (2016). Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of rad sequences. *Systematic Biology*, *65*, 357–365. https://doi.org/10.1093/sysbio/syu046

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*, 3070–3071. https://doi.org/10.1093/bioinformatics/btr521

Knaus, B. J., & Grünwald, N. J. (2017). vcfr: A package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, *17*, 44–53. https://doi.org/10.1111/1755-0998.12549

Lee, K. M., Kivela, S. M., Ivanov, V., Hausmann, A., Kaila, L., Walberg, N., & Mutanen, M. (2018). Information dropout patterns in restriction site associated DNA phylogenomics and a comparison with multilocus sanger data in a species-rich moth genus. *Systematic Biology*, *67*(6), 925–939. https://doi.org/10.1093/sysbio/syy029

Leiva, C., Taboada, S., Kenny, N. J., Combosch, D., Giribet, G., Jombart, T., & Riesgo, A. (2019). Population substructure and signals of divergent adaptive selection despite admixture in the sponge *Dendrilla antarctica* from shallow waters surrounding the Antarctic Peninsula. *Molecular Ecology*, *28*, 3151–3170. https://doi.org/10.1111/mec.15135

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology*, *27*, 3193–3206. https://doi.org/10.1111/mec.14792

Paris, J. R., Stevens, J. R., & Catchen, J. M. (2017). Lost in parameter space: A road map for stacks. *Methods in Ecology and Evolution*, *8*, 1360–1373. https://doi.org/10.1111/2041-210X.12775

Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). dDocent: A RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, *2014*, e431. https://doi.org/10.7717/peerj.431

Ravinet, M., Elgvin, T. O., Trier, C., Aliabadian, M., Gavrilov, A., & Sætre, G. P. (2018). Signatures of human-commensalism in the house sparrow genome. *Proceedings of the Royal Society B-Biological Sciences*, *285*, 20181246. https://doi.org/10.1098/rspb.2018.1246

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M. A. F., Mehlig, B., & Westram, A. M. (2017). Interpreting the genomic landscape of speciation: A road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, *30*, 1450–1477. https://doi.org/10.1111/jeb.13047

Ravinet, M., Westram, A., Johannesson, K., Butlin, R., André, C., & Panova, M. (2016). Shared and nonshared genomic divergence in parallel ecotypes of *Littorina saxatilis* at a local scale. *Molecular Ecology*, *25*, 287–305. https://doi.org/10.1111/mec.13332

Richards, P. M., Morii, Y., Kimura, K., Hirano, T., Chiba, S., & Davison, A. (2017). Single-gene speciation: Mating and gene flow between mirror-image snails. *Evolution Letters*, *1*, 282–291. https://doi.org/10.1002/evl3.31

Rivera-Colón, A. G., Rochette, N. C., & Catchen, J. M. (2020). Simulation with RADinitio improves RADseq experimental design and sheds light on sources of missing data. *Molecular Ecology Resources*, *21*, 363–378.

Rochette, N. C., Brochier-Armanet, C., & Gouy, M. (2014). Phylogenomic test of the hypotheses for the evolutionary origin of eukaryotes. *Molecular Biology and Evolution*, *31*, 832–845. https://doi.org/10.1093/molbev/mst272

Rochette, N. C., & Catchen, J. M. (2017). Deriving genotypes from RAD-seq short-read data using Stacks. *Nature Protocols*, *12*, 2640–2659. https://doi.org/10.1038/nprot.2017.123

Rochette, N., Rivera-Colón, A., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21), 4737–4754. https://doi.org/10.1111/mec.15253

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sánchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution*, *34*, 3299–3302. https://doi.org/10.1093/molbev/msx248

Sætre, G.-P., & Ravinet, M. (2019). *Evolutionary genetics: Concepts, analysis, and practice* (1st ed.). Oxford University Press.

Struck, T. H., & Cerca, J. (2019). Cryptic species and their evolutionary significance. *eLS*, 1–9. https://doi.org/10.1002/9780470015902.a0028292

Struck, T. H., Feder, J. L., Bendiksby, M., Birkeland, S., Cerca, J., Gusarov, V. I., Kistenich, S., Larsson, K.-H., Liow, L. H., Nowak, M. D., Stedje, B., Bachmann, L., & Dimitrov, D. (2018). Finding evolutionary processes hidden in cryptic species. *Trends in Ecology & Evolution*, *33*(3), 153–163. https://doi.org/10.1016/j.tree.2017.11.007

Struck, T. H., Paul, C., Hill, N., Hartmann, S., Hösel, C., Kube, M., Lieb, B., Meyer, A., Tiedemann, R., Purschke, G., & Bleidorn, C. (2011). Phylogenomic analyses unravel annelid evolution. *Nature*, *471*, 95–98. https://doi.org/10.1038/nature09864

Weber, A.-A.-T., Stöhr, S., & Chenuil, A. (2019). Species delimitation in the presence of strong incomplete lineage sorting and hybridization. *Molecular Phylogenetics and Evolution*, *131*, 240218. https://doi.org/10.1101/240218

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Cerca J, Maurstad MF, Rochette NC, 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;12:805–817. https://doi.org/10.1111/2041-210X.13562