Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2017, **8,** 1360–1373 doi: 10.1111/2041-210X.12775

Lost in parameter space: a road map for STACKS

Josephine R. Paris[1], Jamie R. Stevens[1] and Julian M. Catchen*[,2]

[1]*Biosciences, College of Life and Environmental Sciences, University of Exeter, Exeter, UK; and* [2]*Department of Animal Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801, USA*

Summary

1. Restriction site-Associated DNA sequencing (RAD-seq) has become a widely adopted method for genotyping populations of model and non-model organisms. Generating a reliable set of loci for downstream analysis requires appropriate use of bioinformatics software, such as the program STACKS.

2. Using three empirical RAD-seq datasets, we demonstrate a method for optimising a *de novo* assembly of loci using STACKS. By iterating values of the program's main parameters and plotting resultant core metrics for visualisation, researchers can gain a much better understanding of their dataset and select an optimal set of parameters; we present the *80% rule* as a generally effective method to select the core parameters for STACKS.

3. Visualisation of the metrics plotted for the three RAD-seq datasets shows that they differ in the optimal parameters that should be used to maximise the amount of available biological information. We also demonstrate that building loci *de novo* and then integrating alignment positions is more effective than aligning raw reads directly to a reference genome.

4. Our methods will help the community in honing the analytical skills necessary to accurately assemble a RAD-seq dataset.

Key-words: alignment, *de novo* assembly, parameter optimisation, population genetics, RAD-seq, STACKS

Introduction

The last decade has been punctuated by innovations in the generation of genomic data for evolutionary and ecological science. The development of massively parallel, short-read sequencing, associated with lowered costs and open-access analysis tools, has enabled the genomic interrogation of a multitude of model and non-model species. Restriction site-Associated DNA sequencing (RAD-seq) has been proven to be an effective method for identifying and screening high-resolution polymorphism within and between populations (Lescak *et al.* 2015; Blanco-Bercial & Bucklin 2016), ecotypes (Hale *et al.* 2013; Pavey *et al.* 2015) and species (Nadeau *et al.* 2012; Wagner *et al.* 2013; Pante *et al.* 2015).

Restriction site-Associated DNA sequencing involves creating a reduced representation of a genome by isolating the DNA connected to a set of restriction enzyme cut sites. This cost-effective approach can be repeated in large numbers of samples to produce nearly the same reduced subset of the genome in each individual. After sequencing, the data are reassembled into loci, anchored by the presence of the restriction enzyme cut site (Baird *et al.* 2008; Etter *et al.* 2011), and subsequently SNPs are identified across those loci. Many different flavours of the original RAD protocol are now used (e.g. ddRAD, ezRAD, GBS, 2bRAD) and several analysis programs exist (reviewed in Andrews *et al.* 2016). Restriction site-Associated DNA sequencing provides a highly flexible experimental approach, which can be tuned by choosing restriction enzymes with different properties, such as cutting frequency, or by choosing combinations of enzymes; however, this flexibility also brings challenges, such as the quality of DNA required and ascertainment bias stemming from natural variation in restriction enzyme cut sites across a set of populations or species. A number of studies have outlined these aspects of experimental design and technical considerations (Davey *et al.* 2011, 2013; Rowe, Renaut & Guggisberg 2011; Arnold *et al.* 2013).

Restriction site-Associated DNA sequencing is facilitating a shift from using scores of genetic markers to make biological inferences (e.g. microsatellites), to using large-scale data obtained from tens of thousands of loci. The transition from familiar and well-established genotyping techniques into more complex genomic analyses remains a daunting task for many researchers. Thus, the ability to correctly handle the bioinformatic analysis of these vastly larger datasets is essential. Error quantification and hierarchical methods for filtering datasets to obtain biologically robust RAD-seq data do exist. However, these techniques require additional sequencing effort (Mastretta-Yanes *et al.* 2015) and particular RAD datasets for which testing the error is feasible (Fountain *et al.* 2016). Moreover, a RAD-seq analysis relies on competent bioinformatics knowledge, all of which are demanding, especially in a cost- or

*Correspondence author. E-mail: jcatchen@illinois.edu

Table 1. (Top) Three main parameters that control locus formation and polymorphism in STACKS, the default values, the STACKS component program that uses the parameter and a description of what part of the processes each parameter controls. (Bottom) Four additional parameters referenced in the paper (but not part of the optimisation process)

Parameter	Default value	STACKS component	Description
m	3	`ustacks`	Minimum number of raw reads required to form a stack (a putative allele)
M	2	`ustacks`	Number of mismatches allowed between stacks (putative alleles) to merge them into a putative locus
n	1	`cstacks`	Number of mismatches allowed between stacks (putative loci) during construction of the catalog
N	*M* + 2	`ustacks`	Number of mismatches allowed to align secondary reads (reads that did not form stacks) to assembled putative loci to increase locus depth
r	0	`populations`	Percentage of individuals that must possess a particular locus for it to be included in calculation of population-level statistics
max_obs_het	1	`populations`	For a particular locus the maximum number of heterozygous individuals that may be present
min_maf	0	`populations`	For a particular locus, alleles occurring below this frequency are discarded

time-limited framework. To this end, a fundamental protocol for developing an accurate and economical RAD-seq data interrogation strategy is currently lacking.

One of the most widely used programs for processing RAD-seq data is STACKS (Catchen *et al*. 2011, 2013a), a software pipeline designed to assemble loci from short-read sequences derived from restriction enzyme-based protocols, such as RAD-seq. STACKS can be used to assemble markers for genetic mapping analyses (Amores *et al*. 2011), or for population genomics (Epstein *et al*. 2016; Laporte *et al*. 2016), phylogeography (Emerson *et al*. 2010; Bryson *et al*. 2016), and phylogenomics (Jones *et al*. 2013; Díaz-Arce *et al*. 2016). The popularity of STACKS lies in its versatility and user-driven application. After cleaning and demultiplexing raw data (`process_radtags`), the researcher can proceed through one of two main pipelines depending on the availability of a reference genome, to build loci either *de novo* (`denovo_map.pl`; hereafter *de novo map*) or reference aligned (`ref_map.pl`; hereafter *ref map*). Throughout an analysis, values must be chosen for key parameters, which frequently have a significant effect on the building and quality of the resulting loci.

Stack assembly is controlled by several main parameters, the choice of which will depend on key features of a RAD-seq dataset: (i) Biological, such as the inherent polymorphism, level of ploidy and the biological hypothesis being tested; (ii) Study dependent, such as the number of individuals multiplexed, RAD flavour (e.g. RAD, ddRAD, 2bRAD) and restriction enzyme used, including the number of cut sites, the number and length of raw reads, coverage, sequencing platform and inherent error and (iii) Library development issues, for example, degraded DNA (Graham *et al*. 2015) and exogenous contamination (Trucchi *et al*. 2016). Given the uniqueness of each dataset, choosing which parameters are optimal for stack assembly can be difficult.

We tested and optimised the three main parameters within the *de novo map* pipeline (Table 1), which determine the number and the polymorphism of loci in a RAD-seq dataset. The first two affect how loci are built within each individual sample using the core component `ustacks`, where *m* is the minimum number of raw reads required to form a stack (or putative allele) and *M* is the number of mismatches allowed between stacks to merge them into a putative locus (Catchen *et al*. 2011, 2013a).

After the building of loci at an individual level, `cstacks` attempts to match loci across samples to build a *catalog*, which represents the homologous loci across all population samples. To accommodate fixed differences in loci between individuals, mismatches are also allowed during the construction of the *catalog* and the number allowed is controlled by the *n* parameter. Here, we outline a method where we demonstrate that by iterating a range of values for the main parameters, followed by plotting core assembly metrics gathered from the STACKS output files, the researcher can observe and make an informed choice of the best parameter sets for their data (Table 2).

An attractive attribute of RAD-seq is that it can be adopted in model and non-model organisms alike, by using either a reference genome or by constructing loci *de novo*. Both approaches require optimisation of the parameter space; in the case of a *de novo* assembly the parameters must be supplied to STACKS directly, whereas in a reference-aligned analysis the analogous parameters must be supplied to the chosen aligner. An advantage of a *de novo* assembly is that STACKS will identify putative alleles one after another and then merge them into putative loci – leveraging biological information – while an aligner will independently align each raw read. The reference genome also acts as a filter; for example, a draft genome will exclude loci not contained in the assembly and may fail to align reads that belong to loci captured in the reference more than once (e.g. as haplotypes). Alternatively, a reference genome of the study species may not be available, but the genome of a closely related species may be, providing positional information for loci at a reduced precision. We have developed a novel method to incorporate reference genome alignment information by building loci *de novo* and subsequently integrating the alignments of the consensus sequences into their *de novo* dataset (available in STACKS 1.42).

To extract meaningful biological information from a RAD-seq dataset, it is crucial to explore the parameter space and assess how the analysis software interacts with the biological signal. Below, we use empirical datasets from three different

Table 2. Decision framework for each main STACKS parameter that control locus formation and polymorphism in STACKS, the values that users should test and considerations when exploring the parameter space

Parameter	Test	Decision framework			Other considerations
m	3–7	↑ if coverage <15×	↑ if contamination	↑ if conducting phylogenetics	if >m6 disable use of secondary reads
M	1–8	↑ if high polymorphism	↑ if high genomic divergence	↓ if repetitive or polyploid genome	if *M* is too high, paralogous loci can be filtered in populations. Rescale parameters with increased read length (250 bp+)
n	= *M* = *M* + 1 = *M* − 1	↑ if high polymorphism	↓ if sampled from same population	↑ if conducting phylogenetics	plot to observe changes in SNP heterozygosity and fixation

species to demonstrate (i) a method of *de novo* parameter optimisation, and (ii) how integrating alignment information from a reference genome can be used to supplement alignment information to loci built *de novo*.

Materials and methods

RAD-SEQ DATASETS

We used three empirical RAD-seq datasets for analysis, representing a phylogenetically diverse group of organisms (Table 3). The first dataset (*TRT*) consists of data from brown trout (*Salmo trutta* L.) samples occupying two different environmental niches (Paris, King & Stevens 2015): clean (8 individuals) and metal-impacted (8 individuals) sites. The second dataset (*PGN*) is from the king penguin (*Aptenodytes patagonicus*), from two different colonies on different archipelagos (Cristofari *et al*. 2016a): KER (8 individuals) and PCM (8 individuals). The final dataset (*ETW*) contains 16 red earthworm (*Lumbricus rubellus*) individuals from a single population (OL2; Giska, Sechi & Babik 2015).

Analyses of the datasets were performed using STACKS version 1.42 (Catchen *et al*. 2011, 2013a). Data quality was first checked using FastQC (Andrews 2010). If required, reads were cleaned and demultiplexed using the `process_radtags` program. Using each dataset's respective reference genome, the restriction enzyme used and type of RAD library preparation, we estimated the expected number of cut sites and RAD loci *in silico* (Table 3). For *ETW*, two estimates were made; the first by searching for the *Mse*I cut site 200–400 bp downstream/upstream of *Sph*I (liberal estimate) and a second estimation accounting for a second (*Mse*I) cut site occurring within 0–200 bp downstream/upstream of the first cut site (representing allele dropout; conservative estimate). For each dataset, we used `kmer_filter` in STACKS to visualise the error profiles of the cleaned RAD-seq reads; resulting K-mer frequency distributions were plotted in Gnuplot (version 5.0, http://gnuplot.info).

DE NOVO MAP AND PARAMETER OPTIMISATION

For each dataset, we ran *de novo map* several times, varying just one parameter with each parse of the program. For the primary analysis, we varied the `ustacks` *m* parameter from 1 to 6 (m1–m6), the `ustacks` *M* parameter from 0 to 8 (M0–M8) and the `cstacks` *n* parameter from 0 to 10 (n0–n10), while keeping all other parameters consistent (m3, M2 and n0). For further validation, we repeated these same runs with the defaults set to (m6, M4/M6, n0).

We extracted and collated data on *de novo map* assembly metrics for each parameter iteration including: (i) the number of assembled loci; (ii) the number of polymorphic loci and (iii) the number of SNPs for the parameters *m* and *M*. For the *m* parameter, we also collected data on coverage. This information is reported by STACKS' component programs and captured in the log files of both *de novo map* and *ref map*. We extracted these data using simple shell scripts, which we provide (distributed with STACKS).

Differences in both natural polymorphism and read depths can vary across individuals (Davey *et al*. 2013), and so exploring discrepancies between the individuals across a dataset are important. For each parameter run for *m*, *M* and *n* we visualised the data across the population of samples using STACKS' `populations` module, varying the value for the *r* parameter so that a locus had to be present in a minimum of 40%, 60% and 80% of individuals (for *m* and *M*) and 80% of individuals for *n*. To further interrogate how *M* controls polymorphism, we assessed how many new polymorphic loci were identified across 80% of the population (*r80 loci*) for each increment of *M*.

To observe variation in changing the number of fixed differences allowed between loci across individuals when forming the *catalog* (the *n* parameter), we compared the SNPs from each individual with those in the *catalog* using a custom Python script (*count_fixed_catalog_snps.py*; distributed with STACKS). The script tabulates: (i) the number of heterozygous SNPs found in each individual; (ii) the variable sites identified across the whole population; (iii) the number of SNPs found in the *catalog*; and (iv) the number of SNPs found in the *catalog* but not found in any individual sample – that is, the fixed SNPs captured during the construction of the *catalog*. For increments of *n*, we also calculated the distribution of SNPs per *catalog* locus, as well as SNP distributions across polymorphic loci identified in 80% of the population (*r80 loci*).

Metrics for *m*, *M*, and *n* were plotted in GNUPLOT (version 5.0), visually assessed and optimal sets of parameters were chosen, specific to each dataset. The results generated by the optimal parameters were plotted against those obtained from an analysis using the STACKS default parameters (m3, M2, n1), and common population genetic statistics were calculated to assess how the optimal parameter runs compared to using the default values.

COMPARING LOCI BUILT *DE NOVO* TO REFERENCE-ALIGNED RAW READS

Our next objective was to compare loci derived from reference-aligned raw reads (*ref map*) with loci first assembled *de novo* and then aligned back to a genome – integrating the alignment position for each locus back into the STACKS output files (*integrated*).

Table 3. Details of the three Restriction site-Associated DNA sequencing datasets used for analysis

Dataset	Species	Estimated genome size	RAD flavour	Restriction enzyme	Reference genome	Reference estimation of cut sites	Reference estimation of RAD loci	*De novo* estimation of RAD loci
TRT	Brown trout (*Salmo trutta*)	3 Gb	Single digest (1 × 100 bp)	*Sbf*I	Atlantic salmon (*Salmo salar*)	58 197	116 394	257 322 (62 767–82 406)
PGN	King penguin (*Aptenodytes patagonicus*)	1·2 Gb	Single digest (2 × 100 bp)	*Sbf*I	Emperor penguin (*Aptenodytes forsteri*)	43 435	86 870	156 725 (20 664–71 122)
ETW	Red earthworm (*Lumbricus rubellus*)	420 Mb	Double digest (1 × 100 bp)	*Sph*I and *Mse*I	Red earthworm (*L. rubellus*)	127 402/278 544	143 380/399 731	832 898 (5884–56 666)

For each of the three species: the dataset abbreviation used; estimated genome size of the species; the type of RAD sequencing and restriction enzyme(s) used; the reference genome used for alignment analysis; *in silico* estimation of the number of RAD cut sites and resultant RAD loci using the reference genome and the number of RAD loci assembled *de novo* using optimal parameters. For *ETW*, two estimates of reference-related RAD cut sites and loci are presented to show both conservative and liberal estimates based on ddRAD digest sites and allele dropout. For *de novo* estimation, the total number assembled across the dataset is presented, and numbers in brackets represent those assembled across 40% and 80% of individuals within the dataset.

For the *ref map* analyses, clean raw reads were aligned to either a complementary reference genome of the same species or to that of a closely related sister species (Table 3). Raw reads were aligned using GSNAP (version 2015-12-31; Wu & Nacu 2010) specifying a maximum of five mismatches (-m 5), an indel penalty (-i 2) and turning off terminal alignments (–min-coverage = 0·95). Only reads that aligned uniquely (`unpaired_uniq`) were used. *TRT* was aligned to the closely related Atlantic salmon genome (*Salmo salar*, NCBI accession GCA_000233375.4), *PGN* was aligned against the emperor penguin (*Aptenodytes forsteri*, NCBI accession GCF_000699145.1), a close relative of the king penguin, and *ETW* was aligned against a draft genome of the same species (P. Kille and L. Cunha, Cardiff University). Resulting alignment files were run through *ref map*.

For the *integrated* analyses, we used GSNAP (using the same parameters as outlined above) to align the consensus sequence of the catalog loci from the optimal *de novo map* runs for each species against their respective genomes and used a Python script (*integrate_alignments.py*; distributed with STACKS) to integrate alignment information back into the original *de novo map* output files. We compared the number of uniquely mapping loci that were assembled using an alignment from *ref map,* to those aligned using *de novo map* consensus sequences – the *integrated* method.

Results

RAW DATA STATISTICS

Reads from the *TRT* samples were 95 bp in length with an average of 2 880 327 (± 290 590 SE) reads per sample (Table S1, Supporting Information); the *PGN* dataset was 96 bp with 3 312 159 (± 337 146 SE) reads per sample; the *ETW* dataset was 95 bp with 3 629 617 (± 409 701 SE) reads per sample. To get a sense of error and repeats, we generated the distribution of K-mer counts from our set of raw reads. Each true RAD locus can be described by a set of K-mers (subsequences of length *K*) and those K-mers will appear in the distribution in proportion to the number of times a RAD locus was sequenced. The mean of the K-mer distribution reflects the sequencing depth, rare K-mers represent sequencing or PCR error and frequent K-mers describe repetitive loci. Figure S1 demonstrates that *ETW* contains significantly more error than *TRT* and *PGN*.

CHOOSING A VALUE FOR THE MINIMUM NUMBER OF RAW READS (*M*) REQUIRES MEDIATING THE PROMOTION OF ERROR READS TO PUTATIVE ALLELES VS. EXCLUDING TRUE ALLELES

The first stage in the assembly of loci *de novo* (Catchen *et al.* 2011) is to collapse identical raw reads into STACKS and consider them putative alleles. The number of raw reads required to form an allele is governed by *m*. The general pattern for *m* is that using higher values increased average sample coverage (Fig. 1), but decreased the number of assembled loci (Fig. 2a) and the amount of polymorphism (Fig. 2b,c).

Figure 1 (green) shows that stack coverage improved with increasing values of *m*, and that after merging putative alleles into loci, coverage further increased across all datasets (Fig. 1, purple). Coverage between the datasets varied with *PGN*

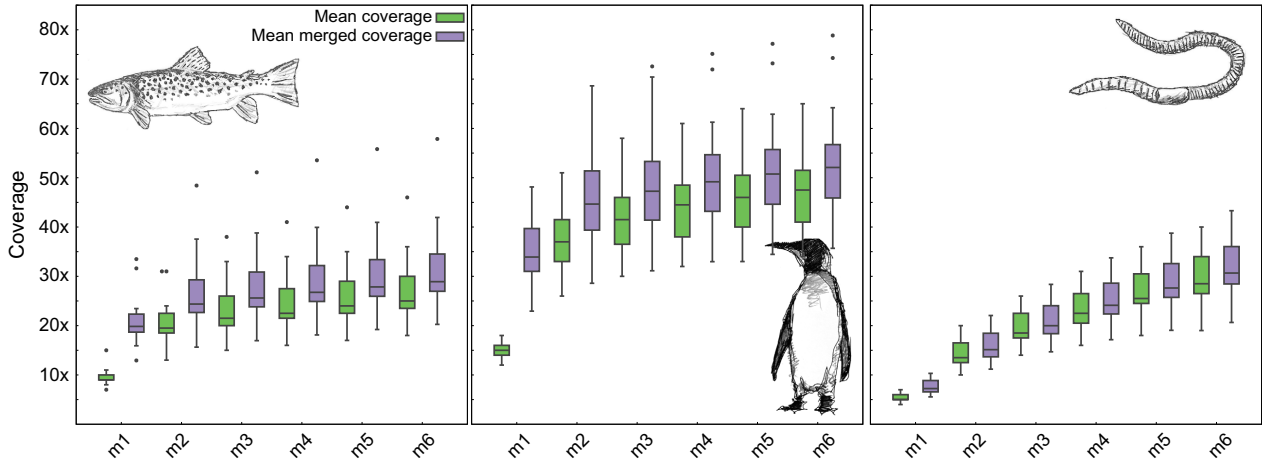

Fig. 1. Plots of mean coverage for *TRT, PGN* and *ETW*. Coverage for each of the 16 individuals is represented as a box plot, where mean coverage (in green) is the average mean coverage for primary reads and mean merged coverage (in purple) is the coverage after merging alleles.

showing the highest coverage, followed by *TRT* and finally *ETW* (Table S2). With the STACKS default value of m3; the average coverage for *TRT* was 23× (17–38×). Merging putative alleles into loci increased coverage further to 28× (16–51×). *PGN* had the highest coverage: 42× (30–58×); and a merged coverage of 49× (31–72×). *ETW* had the lowest coverage at 19× (14–23×), and only reached 21× (15–28×) after merging alleles.

At a value of m1, every raw read is treated as a putative allele and thus all datasets showed the most loci assembled for this value (Fig. 2a; Table S3a). Proportionately, we saw the fewest number of polymorphic loci and the fewest number of SNPs (Fig. 2b,c; Table S3b,c) for m1. Moreover, almost none of these SNPs was shared across the population.

When we increased m1 to m2, the average number of loci formed dropped dramatically. For *ETW*, ~50% more loci were assembled at m1 compared to m2, suggesting that *ETW* contains many unique reads, potentially indicating a large amount of PCR or sequencing error. This was confirmed in the K-mer distribution plots (Fig. S1). In all datasets, we observed a large drop in the number of loci assembled between m2 and m3, but after m3 the number of loci that were built stabilised.

The average number of assembled loci was significantly higher in *ETW* (138 905 at default m3) despite having the lowest coverage, moderate in *TRT* (92 626, default m3) and the fewest loci were assembled in *PGN* (58 876, default m3), despite its very high coverage. These results are consistent with our *in silico* estimates of the number of RAD loci (Table 3). The *TRT* dataset contained the lowest variance and also showed high consistency in the homologous loci being repeatedly assembled in 60% and 80% of the population. *PGN* showed a high, but consistent variance in the number of loci assembled, with many fewer loci shared across 60% and 80% of individuals in the population; however, the high coverage and consistent results with m3–m6 imply a true biological signal, and the distribution of assembled loci was likely bimodal. Indeed, considerably more loci were assembled in the KER population (m3: 74 646 ± 6934), compared to the PCM population (m3: 49 280 ± 5363), also representing higher levels of polymorphism (m3: KER: 7717 ± 1146; PCM: 2475 ± 427). Such patterns could be the result of a batch bias generated by sequencing the two populations on different lanes. Alternatively, this could be due to biological processes such as population bottlenecks, founder effects or stochastic demographic histories. Conversely, *ETW* showed very few loci assembled consistently across the population; combined with the low coverage, this again suggests a signal of error.

The average number of polymorphic loci in *TRT* and *PGN* (Fig. 2b) decreased by ~100 loci with each higher value of *m*, suggesting that increasing read depth excluded only a small amount of biological polymorphism. Given that *M* was fixed for these tests, we determined that *m* does not have a significant effect on polymorphism or SNP detection (Fig. 2c) and this was further confirmed in additional parameter tests (Fig. 2c; Tables S4 and S5). In *ETW*, several hundred loci were lost with each increasing value of *m*, corroborating the hypothesis for a high amount of error.

In a biologically unambiguous dataset with reasonable sequencing coverage (e.g. *TRT*), the *m* parameter converges after a value of 3 and on its own does not have a large impact on the detection of polymorphism. However, by varying *m*, we observed differences in a high coverage dataset that appeared to have different subsets of loci in different individuals (*PGN*) vs. a dataset with low coverage and apparently high error that does not show a strong biological signal (*ETW*).

SETTING VALUES FOR THE MAXIMUM NUMBER OF MISMATCHES (*M*) IS SPECIES SPECIFIC AND REQUIRES A BALANCE BETWEEN UNDERMERGING AND OVERMERGING LOCI

After putative alleles are formed, STACKS performs a search to match alleles together into putative loci. This search is governed by the *M* parameter, which controls for the maximum number of mismatches allowed between putative alleles. The general pattern for *M* was that higher values decreased the

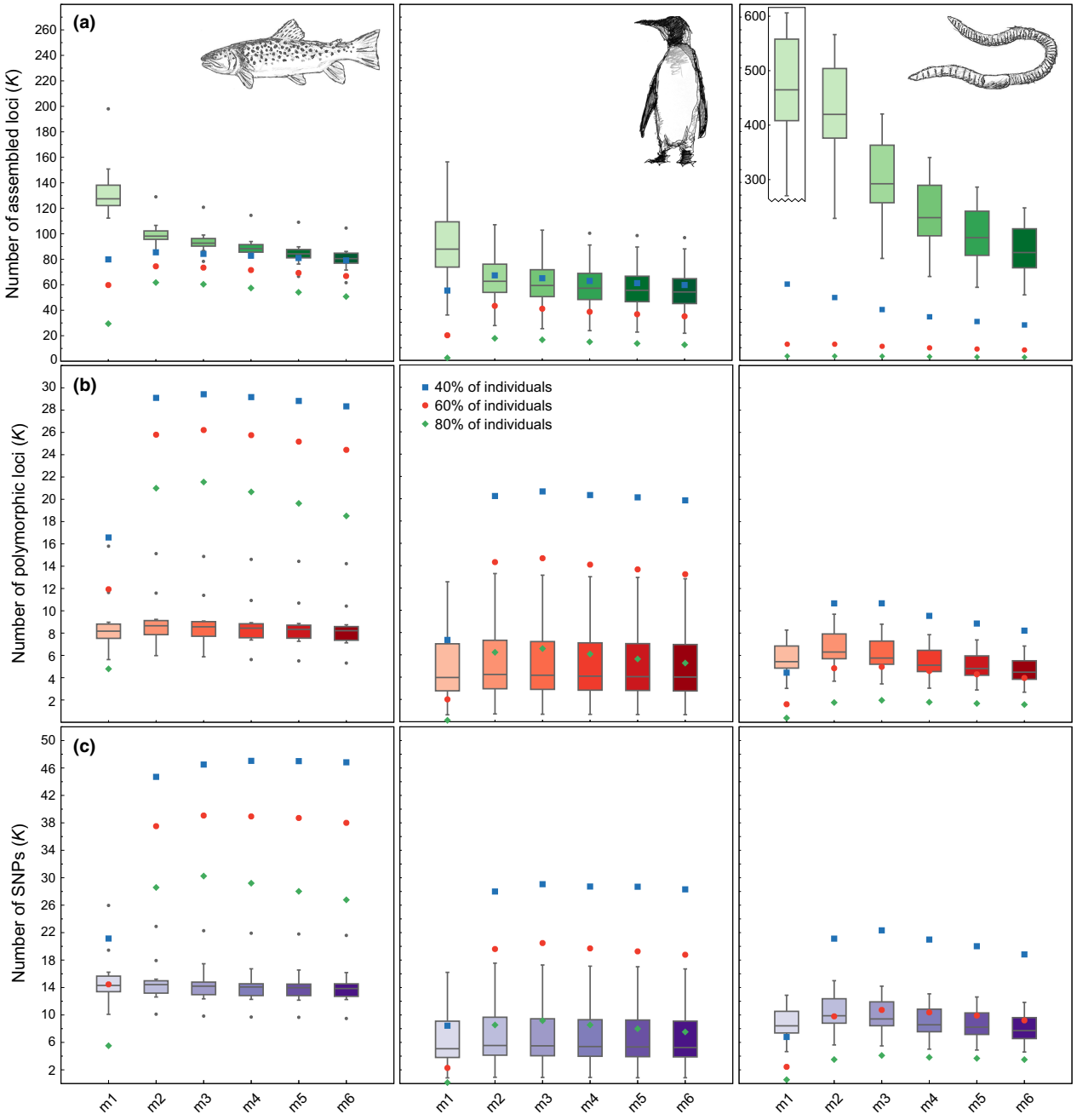

Fig. 2. Plots of iterating values the minimum number of raw reads required to form a stack (*m*) for the metrics: (a) the number of assembled loci; (b) the number of polymorphic loci and (c) the number of SNPs in *TRT, PGN* and *ETW*. Blue squares represent data found in at least 40% of the samples, red circles 60% and green diamonds 80% (*r80*).

number of assembled loci (Fig. 3a), but increased the number of polymorphic loci (Fig. 3b) and the number of SNPs (Fig. 3c).

Correctly setting *M* requires a balance – set it too low and alleles from the same locus will not collapse, set it too high and paralogous or repetitive loci will incorrectly merge together. When alleles from the same locus are undermerged, the software will incorrectly consider them as independent loci. When loci are overmerged because they happen to be close in sequence space, an errant locus with false polymorphism will result. Therefore, *M* is particularly dataset-specific because it depends on the natural levels of polymorphism in the species, as well as the amount of error generated during the preparation and sequencing of the RAD-seq libraries.

The number of assembled loci decreased with increasing values for *M* (Fig. 3a). For *TRT*, the proportion of putative alleles that collapsed levelled out between M3 and M4 (Fig. 3a; Table S6a). In *PGN* these patterns started at lower mismatch values, between M1 and M2. In *ETW*, the numbers that collapsed began to plateau between M2 and M3.

For the number of polymorphic loci (Fig. 3b; Table S6b) and the number of SNPs (Fig. 3c; Table S6c), obviously M0 is

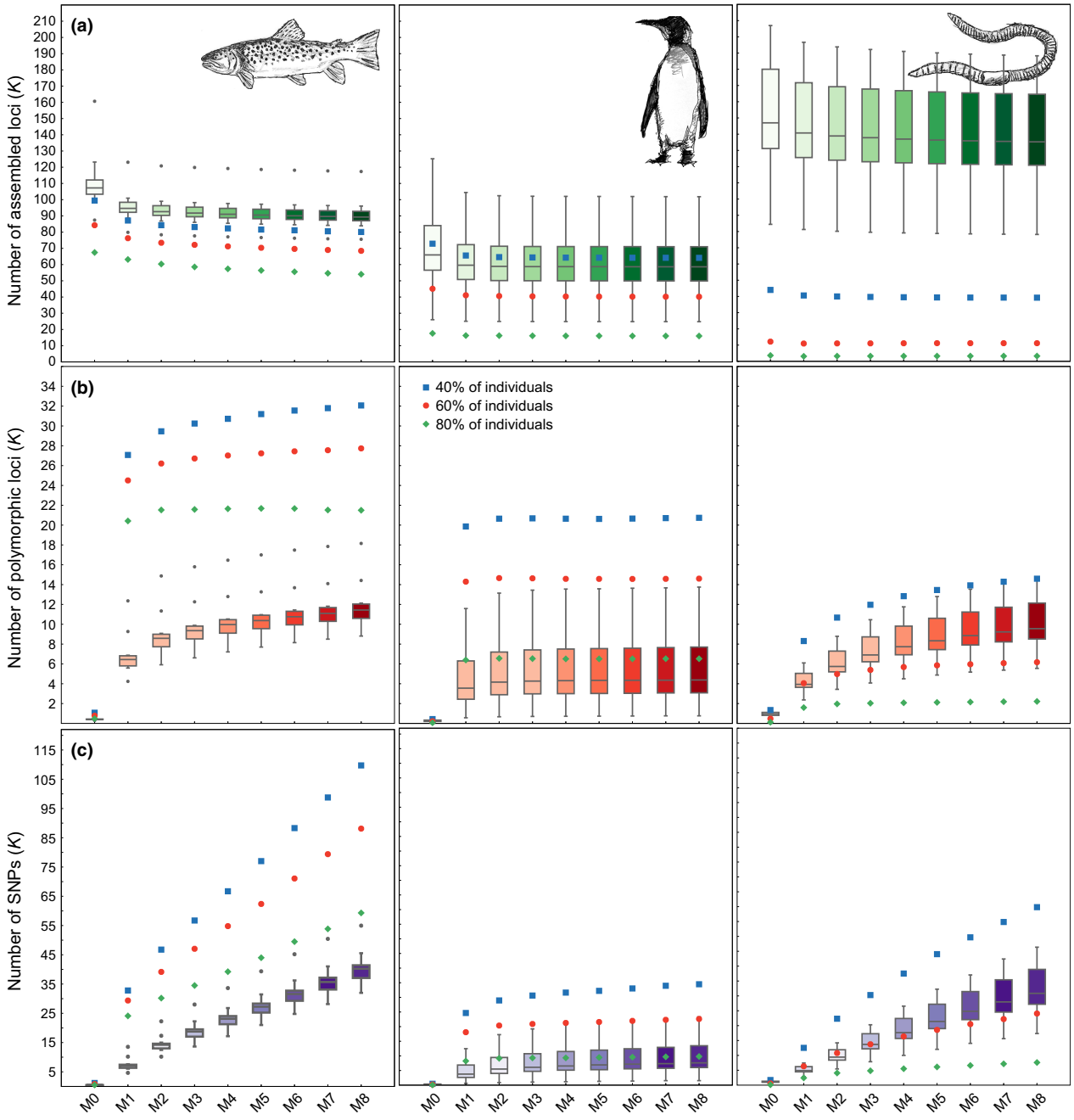

Fig. 3. Plots of iterating values for the distance allowed between two stacks (*M*), for the metrics: (a) the number of assembled loci; (b) the number of polymorphic loci and (c) the number of SNPs in *TRT*, *PGN* and *ETW*. Blue squares represent data found in at least 40% of the population samples, red circles 60% and green diamonds 80% (*r80*).

incorrect. A high number of loci was assembled (Fig. 3a) simply because putative alleles were not correctly merged into heterogeneous loci. Increasing *M* from 0 to 1 permitted alleles with a single polymorphism to merge and with increasing values of *M* no clear limit was seen in the average amount of polymorphism detected.

At M1, the majority of polymorphism and SNPs were already captured in *PGN*, and the amount of polymorphic loci was relatively uniform; the highest polymorphism detected across 80% of the population was at M2. In *TRT*, the average number of polymorphic loci identified within individuals began to flatten out at M5, and M5 also provided the highest amount of polymorphism across 80% of the population. Alternatively, in *ETW*, although polymorphic loci within individuals began to plateau at approximately M3, polymorphism at 80% continued to increase with incremental values of *M* (even up to unrealistically high levels of M11; Table S7), suggesting that the dataset contained a small number of loci with a high-density of SNPs.

In both *TRT* and *ETW* (Fig. 3c) we noticed a steep increase in the number of SNPs obtained with increasing *M*; however, a much smaller increase was detected in *PGN*. The steeper increase in SNPs in *TRT* reflects the higher value of *M* required

to detect the maximum amount of polymorphism across 80% of the population. Although *TRT* was quite polymorphic, a large proportion of these polymorphisms was not shared across the population, and many of the alleles were at low frequency.

When we observed the number of assembled loci (Fig. 3a) together with the number of polymorphic loci (Fig. 3b) and the number of SNPs called (Fig. 3c), we saw a small number of loci that contained many SNPs added with each increment of *M*. This is also reflected in the haplotype diversity in Fig. S2. With high coverage, these are likely to be true loci as the SNP calling algorithm can assess SNP calls with a higher likelihood. However, with low coverage these are more likely to be erroneously assembled loci.

Figure 4 describes how increasing values for *M* contributed new broadly shared, and therefore likely real, polymorphic loci (i.e. loci found in 80% of the population or, *r80 loci*). The number of novel polymorphic loci peaked at M3/M4 for *TRT*, M1/M2 for *PGN* and M2/M3 for *ETW*, and importantly, after a particular value for *M*, identification of new polymorphic loci appears to be robust to any further increases in *M* (Table S8). The data here corroborate the levels of polymorphism displayed in Fig. 3b,c in illustrating the relative homogeneity of *PGN*, compared to *TRT* and *ETW* which were more polymorphic and required higher *M* values.

SELECTING A VALUE FOR THE NUMBER OF MISMATCHES IN THE *CATALOG* (*N*) REQUIRES ASSESSING THE MAXIMUM AMOUNT OF 80% POLYMORPHIC LOCI (*R80*) AROUND *N* = *M*

After loci are assembled in each individual sample, they are compared across samples to match homologous loci into a single *catalog* locus for the population. The *n* parameter controls for the number of mismatches, or fixed differences, allowed during this process. The general pattern for *n* is that higher values increased the number of fixed differences found between samples (Fig. S3; Table S9). Choosing the value for *n* involves a trade-off between setting it too low and failing to find homologous loci in different samples that contain fixed differences, and setting *n* too high and thereby further collapsing loci close together in sequence space within and across samples.

At n0, some loci (and the variants they contained) had not been integrated into the *catalog*. However, once *n* was set higher than 0, the number of variant sites in the population stabilised. In all three datasets, increasing values of *n* provided a linear increase in the identification of more fixed differences between individuals, and increasing values of *n* did not cause the number of fixed differences to plateau (Fig. S3). There appears to be an unlimited number of loci that can be connected together in the *catalog* with increasing values of *n*, so we chose to focus again on optimising for broadly available loci, *r80 loci*.

By iterating over values for n0-n10 for M2, M4 and M6, we saw that at M2 all datasets contained the highest amount of *r80 loci* at $n = M$. At M4 and M6, the highest *r80* values were obtained for values of *n* one iteration either side of $n = M$, so that $n = M-1$ or $n = M + 1$ (Table S9). The optimal value of *n* for *TRT* (n4) provided a total of ~153K SNPs, of which ~24K (15%) were fixed (Table S10). In *PGN*, n2 provided ~62K SNPs, with ~8K being fixed (13%) and for *ETW* (n3) contributed ~156K SNPs and ~33K (21%) of these SNPs were fixed.

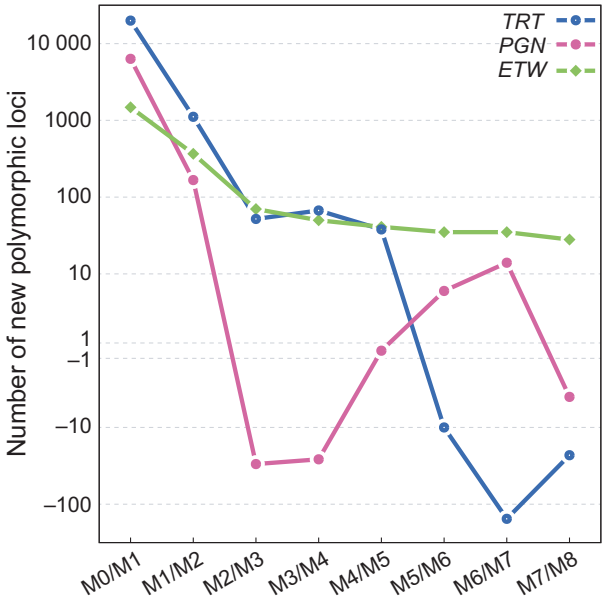

Fig. 4. Plots of the number of new polymorphic loci (*r80 loci*) added for each iteration of *M* (the distance between stacks) for the three datasets: *TRT*, *PGN* and *ETW*.

THE MAXIMS FOR SELECTING OPTIMAL PARAMETERS IN STACKS

Through collecting and plotting metrics of the iterations of *m*, *M* and *n* we have identified two general rules that allow for the identification of optimal parameters: (i) the 80% polymorphic (*r80*) loci rule, and (ii) $n = M$ plus or minus one iteration for linking loci together across samples. There are many possible metrics related to the *de novo* construction of loci we could focus on to optimise parameters. We chose to select a stable set of loci that are highly replicated across the population; these *r80* loci are unlikely to be derived from paralogous or repetitive sequence, or have a lot of sequencing error, and serve as a proxy for the true genome. Importantly, although we are using the *r80* loci as an optimisation target, we are not required to use only the *r80* loci for downstream analyses – we still have all subsets of the loci available to us, we have simply used the *r80* loci to optimise the assembly of all loci.

A value of m3 was optimal in providing the highest amount of polymorphism across all three datasets at 40%, 60% and most importantly 80% of the population. We observed that the *PGN* dataset was relatively monomorphic, and the highest amount of polymorphism identified at *r80* was detected at a value of M2. The *TRT* and *ETW* datasets showed higher polymorphism; indeed, the largest amount identified at *r80* was

found at M5 for *TRT*, making this the most polymorphic dataset. In *ETW*, the polymorphism at *r80* continued to increase with higher *M* values, but Fig. 4 showed that very few new polymorphic loci were added at *M* values greater than 3. Given the potential error in *ETW*, for this dataset, M3 also provides a balance between obtaining true polymorphism and introducing sequencing error.

All datasets continued to show more fixed SNPs with increasing values of *n* (Fig. S3). *PGN* showed a very gentle increase, *TRT* a moderate increase and *ETW* the steepest increase in fixed SNPs. This alone, however, does not provide enough information to choose a value for *n*. We found that the highest polymorphism of *r80 loci* resulted from $n = M$, $n = M-1$ or $n = M + 1$ (Table S9) and suggest this as the best method for obtaining the optimal value for *n*. The following optimal parameters were therefore chosen so that: *TRT* = m3, M5, n4; *PGN* = m3, M2, n2; *ETW*: m3, M3, n3.

Using these optimal parameter sets, in *TRT*, ~244K loci were identified across the population, 22% were polymorphic containing an average of 3·18 SNPs per locus; for *PGN*, ~155K loci were assembled, 32% were polymorphic, with 1·74 SNPs per locus; in *ETW*, ~865K loci were assembled, 13% were polymorphic, with 2·83 SNPs per locus. We suggest assessing and reporting these metrics in any RAD-seq analysis.

Figure 5 shows the main metrics plotted for the STACKS default parameters (m3, M2 and n1) and the optimal parameters for each dataset. Significantly, when compared to the defaults, the optimal parameters modified the number of constructed loci, the polymorphism levels and the numbers of SNPs (Fig. 5, Table S11). Moreover, the number of polymorphic loci discovered across 80% of the population (*r80*) increased with optimal parameters (Fig. 5, green dots). Although overall measures of population genetic statistics were not considerably altered (Table S12), changing the values for *M* considerably altered global SNP and haplotype diversity statistics across all three datasets (Fig. S2, Table S13).

DE NOVO ANALYSIS RECOVERS MORE LOCI COMPARED TO REFERENCE-ALIGNED RAW READS

It was apparent that in all three datasets the *de novo-integrated* approach recovered thousands more loci than *ref map* (Fig. S4; Table S14). In *TRT*, a total of ~116K *de novo* consensus loci aligned to the reference genome. Of this total, ~84K represented consensus loci that aligned to unique positions in the reference. This compares to just ~66K alignments using *ref map*. Similarly, in *PGN* over 117K consensus loci aligned using the *integrated* method, ~104K of which were aligned uniquely, in comparison to under 92K using *ref map*. The most obvious discrepancy was seen with the *ETW* dataset where ~253K loci aligned with the *integrated* method (~172K unique alignments) in comparison to just ~100K loci using *ref map*. In the large majority of cases, raw reads that were poorly aligned compared to well-integrated consensus *de novo* sequences were due to either insertions into the reference, or deletions in the sequence relative to the reference. Our comparison is of course dependent on the corresponding reference alignment parameters used; more promiscuous values may have closed the gap between *de novo map* and *ref map* a small amount.

Discussion

Restriction site-Associated DNA sequencing data interrogation is a non-trivial, yet necessary process for avoiding inaccurate building of loci, erroneous SNP calls and a loss of important biological information. Our aim was to develop a straightforward method for surveying and visualising the trends of assembly metrics from the output of STACKS, and we suggest that by doing so, researchers can accurately identify optimal parameter sets. We present two general maxims that can be used to optimise parameter space, but highlight that the *r80* rule, which uses a set of polymorphic loci repeatedly assembled across a set of samples, will provide an effective

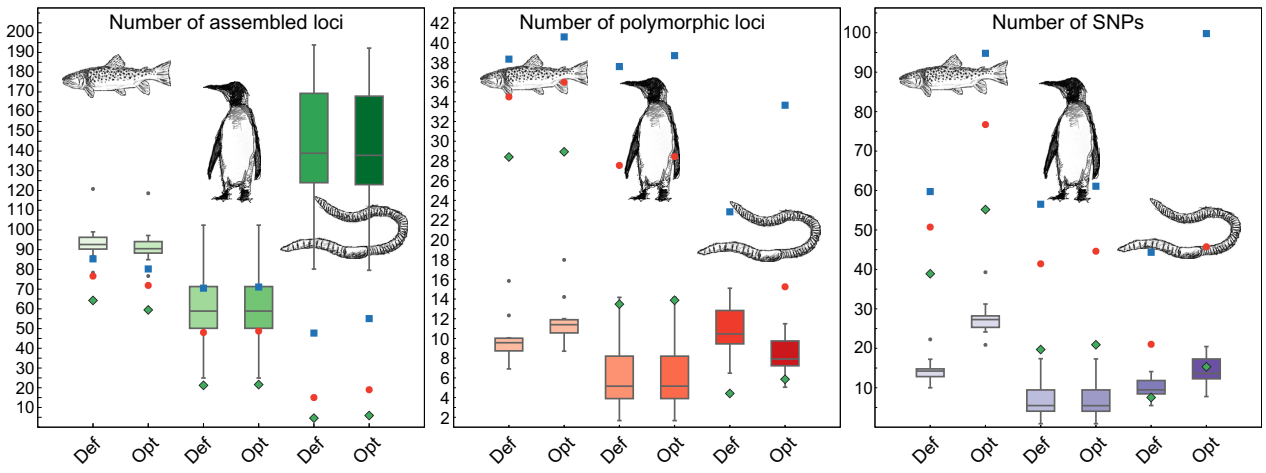

Fig. 5. Plots showing the differences between the default STACKS parameters (m3, M2, n1) and the optimal parameters selected for each dataset: *TRT:* m3, M5, n4; *PGN:* m3, M2, n2 and *ETW:* m3, M3, m3 for the number of assembled loci, the number of polymorphic loci and the number of SNPs. For every dataset, the optimal parameter sets gave the highest amount of polymorphic loci in 80% of the population (*r80 loci*, green diamonds).

optimisation target. Furthermore, we have shown that incorporating alignment information from a reference genome is an advantageous approach compared to using a standard method of alignment, particularly if using a draft reference genome or if the genome is phylogenetically distant from the focal species.

The STACKS *de novo* assembly algorithm proceeds in each individual sample in several stages (Catchen *et al*. 2011). First, exactly matching reads are collapsed into putative alleles, controlled by *m*. Sets of exactly matching reads with a size smaller than *m* are set aside and are referred to as secondary reads. Putative alleles are then compared against one another and collapsed into putative loci, controlled by *M*. Once the loci are formed, the secondary reads are aligned back against assembled loci with a relaxed mismatch limit (the `-N` parameter to `ustacks`) to increase locus depth. Next, the SNP model is executed and it evaluates the evidence at each nucleotide position for a homozygote or heterozygote. The model takes into account depth of coverage and the amount of error present (i.e. non-matching nucleotides at the current position) in a maximum likelihood statistical framework. Finally, after loci are assembled in each sample, homologous loci must be matched across samples to make a population-wide *catalog* of loci. Loci are compared across samples and sequences within a certain distance are collapsed into a single *catalog* (population-wide) locus, governed by *n* (`cstacks`).

We recommend scrutinising sample coverage as the first assembly metric when assessing the suitability of any given RAD-seq dataset and we suggest aiming for coverage thresholds greater than 25×. Fountain *et al*. (2016) showed that genotyping error rates were considerably higher when coverage was between 5 and 10×, and error rates were mostly robust to variation in sequence quality when coverage was ≥10. However, published datasets with coverage below 10× are not uncommon (e.g. Xu *et al*. 2014; Boehm *et al*. 2015; Ivy *et al*. 2016; Razkin *et al*. 2016).

As seen here, the main way the *de novo* assembly algorithm interacts with coverage is through the *m* parameter. If *m* is set to 1, each raw read becomes a putative allele and there are no secondary reads. However, as *m* is increased the status of a putative allele requires more evidence and the number of secondary reads grows. If *m* gets too large, alleles lacking coverage will not be recognised and will become secondary reads. These secondary reads will still be made available to the SNP calling model, as long as one of the two alleles was assembled, but if both alleles fall below a high *m* threshold the locus will be lost.

In the literature, wide ranges of values for the *m* parameter have been used. In some cases, the investigation was specifically addressing phylogenetic differences between divergent species, and hence underdetection of species-specific polymorphism when building loci was inconsequential (e.g. m50 – Keller *et al*. 2012; m125 – Wagner *et al*. 2013). Furthermore, a higher value of *m* may be required to exclude exogenous contamination (Trucchi *et al*. 2016). However, other studies have used inappropriately high values of *m* (Jezkova *et al*. 2015; Palaiokostas *et al*. 2015; Suyama & Matsuki 2015), presumably due to the assumption that higher values will reduce genotyping error. If *m* is set unrealistically high, secondary read incorporation must be disabled (`-N` option). This was demonstrated when comparing m3 and m6 with iterative values for *M*, where at M0 more SNPs were supposedly called at m6, compared to m3 (Table S5). However, these are not true SNPs, but simply represent sequencing error incorporated by secondary reads at higher read depths.

We recommend setting *m* high enough only to deny errant reads the status of a putative allele. We therefore do not suggest using a value for $m > 5$ and we have demonstrated here that the default value of m3 was favourable for all test datasets. Testing a range from 3 to 7 should allow the correct exploration of this parameter under most biological scenarios. If coverage is exceptionally high (i.e. >40×), or levels of polymorphism are exceptionally low (as in the *PGN* dataset), *m* can still be left at a moderate value, but secondary reads can be discarded entirely (setting `N` to 0), making for a clearer signal (e.g. Longo & Bernardi 2015). Despite controlling for the minimum read depth of alleles at the `ustacks` phase of the pipeline, many studies also incorporate a minimum stack depth required for individuals at a locus in the `populations` module of STACKS (e.g. Gaither *et al*. 2015; Ivy *et al*. 2016; Kjeldsen *et al*. 2016). Such a method is undesirable, as read depth has already been accounted for by the SNP model. Once the SNP model has made a determination, its evaluation should be trusted and using further non-statistically based limits on depth of coverage is ill advised and will result in the arbitrary dropping of loci.

Species with higher levels of polymorphism will require higher values of *M* (Campagna *et al*. 2015; Ravinet *et al*. 2016), as do studies assessing levels of genomic divergence (Jones *et al*. 2013; Lozier *et al*. 2016). If *M* is too small, alleles will not collapse into loci within individuals. When the *catalog* is constructed, different individuals will map alternative alleles together creating loci that appear homozygous, while the other alleles are excluded and may themselves form distinct loci. On the other hand, if *M* is too high, repetitive or paralogous loci will be erroneously merged together. These loci are noticeable, as they will appear as heterozygous in a large majority of individuals; they can be filtered out by the `populations` module after the main pipeline runs, based on their high heterozygosity (`--max_obs_het` option).

The three datasets examined here required different values for *M*. The *ETW* dataset is likely quite polymorphic; high levels of polymorphism have been shown in the *Lumbricus* genus (Kautenburger 2006; Shepeleva *et al*. 2008; Donnelly *et al*. 2014). However, the low coverage and potential error in the dataset means a definitive determination of true polymorphism could not be reliably assessed, but M3 provided the highest amount of *r80 loci*.

For the *TRT* dataset, M5 was the most suitable value for identifying polymorphism, and this is considerably higher than M2 as most commonly used in salmonid RAD-seq studies (Lemay & Russello 2015; Bernatchez *et al*. 2016) and in other teleost species (Catchen *et al*. 2013b; Martin & Feinstein 2014; Fowler & Buonaccorsi 2016). Observations of the *TRT* dataset suggested that a large proportion of the alleles were at a low

frequency; these could be filtered out after the main pipeline executes by implementing a minor allele frequency threshold (`--min_maf` option to `populations`).

On the other hand, the *PGN* dataset was relatively uniform and exhibited low levels of polymorphism. This corroborates with limited genetic variation and high gene flow that has been shown to exist in penguin colonies (Roeder *et al.* 2001; Nims *et al.* 2008; Freer *et al.* 2015). The bimodal distribution for the number of loci observed in *PGN* corroborates well with the known stochastic demographic history of the species (Cristofari *et al.* 2016a). Considerably fewer loci were observed in the PCM population, a colony sampled from an archipelago that was recolonised later after glaciation, that is, later than the archipelago home to the KER colony. While the length of the loci that were generated should be considered, with most short-read sequencing (~100 bp) surveying M1 to M8 should allow sufficient mismatch parameter space exploration for the large majority of RAD-seq datasets. Of course, if read lengths are longer (250 bp+), users should rescale the parameter, testing higher values of M as appropriate for the longer read length.

A main consideration when deciding on the number of mismatches allowed in the *catalog* (n) is how many fixed differences might be expected between individuals. In a closely related set of populations, it makes intuitive sense to set $n = M$, as M controls matching alleles within an individual, and n controls matching of the same set of alleles across individuals. If, by chance for a particular locus, only homozygous individuals were sampled, setting $n = M$ will allow the correct relationship for that locus to be recovered across individuals. If population sample sizes are large (so heterozygotes should have been found), and fixed differences are rare, it may make sense to set n to a value less than M (Barnard-Kubow, Debban & Galloway 2015; Saenz-Agudelo *et al.* 2015). Alternatively, if the samples originate from highly divergent individuals (Ravinet *et al.* 2016; Rougemont *et al.* 2016), or phylogenetic relationships are being explored between species (Combosch & Vollmer 2015; Tariel, Longo & Bernardi 2016), then a higher value of n may be required to detect these fixed polymorphisms. This being said, it may be difficult to derive a biological judgement of the known amount of differentiation between individuals – for example, if cryptic population structure exists. Therefore, it is essential to plot and explore how iterative values for this parameter affect how many fixed differences are detected.

Although the number of fixed SNPs continued to grow with incremental values for n for both *TRT* and *ETW*, the highest amount of *r80 loci* was equal to $n = M$ for *ETW*. High levels of intraspecific divergence are common in earthworm species (Klarica *et al.* 2012) and, in particular, natural populations of *L. rubellus* have been shown to consist of highly divergent mtDNA lineages representing a complex of cryptic species (King, Tibble & Symondson 2008). Thus, high divergence between individuals, even within the same population is very likely and validates the patterns we observed. The *TRT* dataset showed a moderate number of fixed SNPs, and the highest number of polymorphic *r80 loci* was recovered for $n = M-1$. This is in accordance with high levels of differentiation between geographically proximate populations of the species, occupying metal-contaminated and clean rivers (Paris, King & Stevens 2015). In king penguins, studies have shown low levels of genetic divergence both within (Cristofari *et al.* 2015) and between (Freer *et al.* 2015) colonies, and it has recently been suggested that Antarctic penguins should be considered single panmictic populations due to extreme levels of genetic homogeneity and low F_{ST} values between populations (Cristofari *et al.* 2016b). Coupled with the low levels of within-individual polymorphism, $n = M$ resolved the most polymorphic *r80 loci* across the populations for this dataset. Thus, the general maxims and visualisation of the RAD-seq plots created here and the subsequent parameters chosen corroborate well with the known biology of these species.

Our data clearly show (Fig. S5; Table S15) that in all three datasets there are a small number of loci containing high densities of SNPs. These loci appear in a *de novo* dataset with high values of M, and also appear in a reference-aligned dataset with promiscuous alignment parameters. Including these loci will affect standard population genetic statistics (Fig. S2; Table S13), naturally inflating them. Experiments that rely only on downstream population genetic measures to assess accuracy need to account for this (Rodríguez-Ezpeleta *et al.* 2016; Shafer *et al.* 2016) or they may produce misleading results.

A reference genome is a great asset in a RAD-seq analysis, enabling loci to be considered positionally and statistics to be computed across chromosomes. However, one aspect that may be both an asset and a liability is that the reference genome can act like a strict filter, and so what a researcher cannot observe as aligned on a chromosome is not included. In addition, alignment programs align each read independently and alignment algorithms can score different combinations of gaps and mismatches the same way, which can lead to alignment variation. Furthermore, RAD-seq is a key technology for non-model organisms, so, often a reference genome, if available, is a draft, or is a closely (or not so closely) related species, exacerbating the above problems. By building loci *de novo*, individual reads have already been merged into biologically plausible loci, thereby leveraging the maximum amount of information obtained from each individual read. In a straight alignment approach, each read is treated independently of every other read. Our method of building loci *de novo*, aligning the consensus sequences to a reference and then reintegrating alignment positions into the *de novo* data to compute chromosome-level statistics combines positive aspects of both *de novo* and reference analyses.

Other studies have used a version of the *de novo-integrated* approach (Jones *et al.* 2013; Wagner *et al.* 2013). In all datasets analysed here, the *de novo* and *integrated* method performed better than alignment and *ref map*. In both the brown trout (*TRT*) and king penguin (*PGN*) cases, the reference genome was not from the same species, but belonged to the same genus, suggesting that the genome of a closely related species can be utilised. Gene synteny is conserved between brown trout (*S. trutta*) and Atlantic salmon (*S.* salar) (Gharbi *et al.* 2006)

and the species are known to hybridise (Leaniz & Verspoor 1989; Elo *et al*. 1997). Similarly, phylogenetic relationships within the genus *Aptenodytes*, i.e. the king penguin (*A. patagonicus*) and emperor penguin (*A. forsteri*), are closely related (Baker *et al*. 2006; Ksepka, Bertelli & Giannini 2006). In both cases, the RAD-seq reads for the *TRT* and the *PGN* datasets aligned successfully to the reference genome, but this was improved by building loci *de novo*, most likely the result of constructing species-specific loci first, which accounts for the variance between the *reference* genome and the samples used.

In contrast, the genome used for the red earthworm *ETW*, *L. rubellus*, was from the same species. The highly divergent *L. rubellus* species consists of 11 lineages (Sechi 2013); the samples used for the RAD-seq analysis were identified as lineages A, C and E (Giska, Sechi & Babik 2015) and the reference genome was sequenced from lineage B. COI analysis has shown that other lineages are >12% different to lineage B and only 20% of RNA-seq reads map from lineage A to lineage B (Kille, pers. comm., unpubl. data). The straight alignment method using *ref map* was relatively unsuccessful. However, here the *integrated* method successfully aligned all of the *de novo* reads to the genome.

We have demonstrated how the STACKS software can be tailored to a researchers' RAD-seq dataset; we have provided a method of parameter optimisation and exhibited how to effectively implement and test our optimisation strategy using STACKS. These parameters should be adjusted to complement the biology of the species being studied, the biological hypotheses, the library construction and error inherent in the dataset. While we focused on STACKS in this work, other RAD-seq analysis software uses the same algorithmic strategy, governed by analogous parameters. Analyses in other domains also use similar strategies, such as assembling metagenomic loci, or during alignment for shotgun resequencing analyses; thus, this basic approach has broad relevance and can be widely applied.

Authors' contributions

J.P. and J.S. conceived of the brown trout experiments, J.P. implemented the brown trout molecular work. J.P. and J.C. conceived and implemented the analytical methods. J.P., J.C., and J.S. wrote the manuscript.

Acknowledgements

This work was co-funded by the Environment Agency, Westcountry Rivers Trust and the University of Exeter. Overseas collaboration for the project was made possible by funding from The Genetics Society, Santander and the University of Exeter. Thank you to many RAD-seq workshop participants for invaluable insight and new ideas. We thank Dr Nicolas Rochette for his insights into parameter analysis. Thanks also to Dr Andy King for assistance with the brown trout data molecular work and analysis, and Guy Freeman and Martin Young for the species illustrations. Prof Peter Kille and Dr Luis Cunha, Cardiff School of Biosciences, Cardiff University, kindly provided the reference genome of *L. rubellus*.

Data accessibility

The brown trout samples (TRT) are archived at the NCBI Sequence Read Archive, BioProject PRJNA379215, accession numbers SRR5344602–SRR5344617. The king penguin samples (PGN) are archived at EBI-EMBL, Project ID 308448; run accession numbers SRR3177636–SRR317765, and the red earthworm samples (ETW) are archived at EBI-EMBL Project ID 296755; run accession numbers SRR2962225–SRR2962229 and SRR2962234–SRR2962244.

References

Amores, A., Catchen, J., Ferrara, A., Fontenot, Q. & Postlethwait, J.H. (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics*, **188**, 799–808.

Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed 4 April 2017).

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G. & Hohenlohe, P.A. (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, **17**, 81–92.

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.

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.

Baker, A.J., Pereira, S.L., Haddrath, O.P. & Edge, K.-A. (2006) Multiple gene evidence for expansion of extant penguins out of Antarctica due to global cooling. *Proceedings Biological sciences/The Royal Society*, **273**, 11–17.

Barnard-Kubow, K.B., Debban, C.L. & Galloway, L.F. (2015) Multiple glacial refugia lead to genetic structuring and the potential for reproductive isolation in a herbaceous plant. *American Journal of Botany*, **102**, 1842–1853.

Bernatchez, S., Laporte, M., Perrier, C., Sirois, P. & Bernatchez, L. (2016) Investigating genomic and phenotypic parallelism between piscivorous and planktivorous Lake Trout (*Salvelinus namaycush*) ecotypes by means of RADseq and morphometrics analyses. *Molecular Ecology*, **25**, 4773–4792.

Blanco-Bercial, L. & Bucklin, A. (2016) New view of population genetics of zooplankton: RAD-seq analysis reveals population structure of the North Atlantic planktonic copepod *Centropages typicus*. *Molecular Ecology*, **25**, 1566–1580.

Boehm, J.T., Waldman, J., Robinson, J.D. & Hickerson, M.J. (2015) Population genomics reveals seahorses (*Hippocampus erectus*) of the western mid-Atlantic coast to be residents rather than vagrants. *PLoS ONE*, **10**, e0116219.

Bryson, R.W., Savary, W.E., Zellmer, A.J., Bury, R.B. & McCormack, J.E. (2016) Genomic data reveal ancient microendemism in forest scorpions across the California Floristic Province. *Molecular Ecology*, **25**, 3731–3751.

Campagna, L., Gronau, I., Silveira, L.F., Siepel, A. & Lovette, I.J. (2015) Distinguishing noise from signal in patterns of genomic divergence in a highly polymorphic avian radiation. *Molecular Ecology*, **24**, 4238–4251.

Catchen, J.M., Amores, A., Hohenlohe, P.A., Cresko, W.A. & Postlethwait, J.H. (2011) *Stacks*: building and genotyping Loci *de novo* from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171–182.

Catchen, J., Bassham, S., Wilson, T., Currey, M., O'Brien, C., Yeates, Q. & Cresko, W.A. (2013b) The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. *Molecular Ecology*, **22**, 2864–2883.

Catchen, J.M., Hohenlohe, P.A., Bassham, S., Amores, A. & Cresko, W.A. (2013a) *Stacks*: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Combosch, D.J. & Vollmer, S.V. (2015) Trans-Pacific RAD-Seq population genomics confirms introgressive hybridization in Eastern Pacific Pocillopora corals. *Molecular Phylogenetics and Evolution*, **88**, 154–162.

Cristofari, R., Bertorelle, G., Ancel, A. *et al*. (2016b) Full circumpolar migration ensures evolutionary unity in the Emperor penguin. *Nature Communications*, **7**, 11842.

Cristofari, R., Liu, X., Bonadonna, F. *et al*. (2016a) Climate-driven range shifts in fragmented ecosystems. *bioRxiv*, https://doi.org/10.1101/090852.

Cristofari, R., Trucchi, E., Whittington, J.D., Vigetta, S., Gachot-Neveu, H., Stenseth, N.C., Le Maho, Y. & Le Bohec, C. (2015) Spatial heterogeneity as a genetic mixing mechanism in highly philopatric colonial seabirds. *PLoS ONE*, **10**, e0117981.

Davey, J.W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K. & Blaxter, M.L. (2013) Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.

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.

Díaz-Arce, N., Arrizabalaga, H., Murua, H., Irigoien, X. & Rodríguez-Ezpeleta, N. (2016) RAD-seq derived genome-wide nuclear markers

resolve the phylogeny of tunas. *Molecular Phylogenetics and Evolution*, **102**, 202–207.

Donnelly, R.K., Harper, G.L., Morgan, A.J., Pinto-Juma, G.A. & Bruford, M.W. (2014) Mitochondrial DNA and morphological variation in the sentinel earthworm species *Lumbricus rubellus*. *European Journal of Soil Biology*, **64**, 23–29.

Elo, K., Ivanoff, S., Vuorinen, J.A. & Piironen, J. (1997) Inheritance of RAPD markers and detection of interspecific hybridization with brown trout and Atlantic salmon. *Aquaculture*, **152**, 55–65.

Emerson, K.J., Merz, C.R., Catchen, J.M., Hohenlohe, P.A., Cresko, W.A., Bradshaw, W.E. & Holzapfel, C.M. (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences*, **107**, 16196–16200.

Epstein, B., Jones, M., Hamede, R. *et al.* (2016) Rapid evolutionary response to a transmissible cancer in Tasmanian devils. *Nature Communications*, **7**, 12684.

Etter, P.D., Preston, J.L., Bassham, S., Cresko, W.A. & Johnson, E.A. (2011) Local *de novo* assembly of RAD paired-end contigs using short sequencing reads. *PLoS ONE*, **6**, e18561.

Fountain, E.D., Pauli, J.N., Reid, B.N., Palsbøll, P.J. & Peery, M.Z. (2016) Finding the right coverage: the impact of coverage and sequence quality on SNP genotyping error rates. *Molecular Ecology Resources*, **16**, 966–978.

Fowler, B.L.S. & Buonaccorsi, V.P. (2016) Genomic characterization of sex-identification markers in *Sebastes carnatus* and *Sebastes chrysomelas* rockfishes. *Molecular Ecology*, **25**, 2165–2175.

Freer, J.J., Mable, B.K., Clucas, G., Rogers, A.D., Polito, M.J., Dunn, M., Naveen, R., Levy, H. & Hart, T. (2015) Limited genetic differentiation among chinstrap penguin (*Pygoscelis antarctica*) colonies in the Scotia Arc and Western Antarctic Peninsula. *Polar Biology*, **38**, 1493–1502.

Gaither, M.R., Bernal, M.A., Coleman, R.R., Bowen, B.W., Jones, S.A., Simison, W.B. & Rocha, L.A. (2015) Genomic signatures of geographic isolation and natural selection in coral reef fishes. *Molecular Ecology*, **24**, 1543–1557.

Gharbi, K., Gautier, A., Danzmann, R.G. *et al.* (2006) A linkage map for brown trout (*Salmo trutta*): chromosome homeologies and comparative genome organization with other salmonid fish. *Genetics*, **172**, 2405–2419.

Giska, I., Sechi, P. & Babik, W. (2015) Deeply divergent sympatric mitochondrial lineages of the earthworm *Lumbricus rubellus* are not reproductively isolated. *BMC Evolutionary Biology*, **15**, 217.

Graham, C.F., Glenn, T.C., McArthur, A.G. *et al.* (2015) Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). *Molecular Ecology Resources*, **15**, 1304–1315.

Hale, M.C., Thrower, F.P., Berntson, E.A., Miller, M.R. & Nichols, K.M. (2013) Evaluating adaptive divergence between migratory and nonmigratory ecotypes of a salmonid fish, *Oncorhynchus mykiss*. *G3: Genes| Genomes| Genetics*, **G3**, 1273–1285.

Ivy, J.A., Putnam, A.S., Navarro, A.Y., Gurr, J. & Ryder, O.A. (2016) Applying SNP-derived molecular coancestry estimates to captive breeding programs. *Journal of Heredity*, **107**, 403–412.

Jezkova, T., Riddle, B.R., Card, D.C., Schield, D.R., Eckstut, M.E. & Castoe, T.A. (2015) Genetic consequences of postglacial range expansion in two codistributed rodents (genus *Dipodomys*) depend on ecology and genetic locus. *Molecular Ecology*, **24**, 83–97.

Jones, J.C., Fan, S., Franchini, P., Schartl, M. & Meyer, A. (2013) The evolutionary history of *Xiphophorus* fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. *Molecular Ecology*, **22**, 2986–3001.

Kautenburger, R. (2006) Genetic structure among earthworms (*Lumbricus terrestris* L.) from different sampling sites in western Germany based on random amplified polymorphic DNA. *Pedobiologia*, **50**, 257–266.

Keller, S., Levsen, N., Olson, M. & Tiffin, P. (2012) Local adaptation in the flowering-time gene network of balsam poplar, *Populus balsamifera* L. *Molecular Biology and Evolution*, **29**, 3143–3152.

King, R.A., Tibble, A.L. & Symondson, W.O.C. (2008) Opening a can of worms: unprecedented sympatric cryptic diversity within British lumbricid earthworms. *Molecular Ecology*, **17**, 4684–4698.

Kjeldsen, S.R., Zenger, K.R., Leigh, K., Ellis, W., Tobey, J., Phalen, D., Melzer, A., FitzGibbon, S. & Raadsma, H.W. (2016) Genome-wide SNP loci reveal novel insights into koala (*Phascolarctos cinereus*) population variability across its range. *Conservation Genetics*, **17**, 337–353.

Klarica, J., Kloss-Brandstätter, A., Traugott, M. & Juen, A. (2012) Comparing four mitochondrial genes in earthworms – Implications for identification, phylogenetics, and discovery of cryptic species. *Soil Biology and Biochemistry*, **45**, 23–30.

Ksepka, D.T., Bertelli, S. & Giannini, N.P. (2006) The phylogeny of the living and fossil Sphenisciformes (penguins). *Cladistics*, **22**, 412–441.

Laporte, M., Pavey, S.A., Rougeux, C. *et al.* (2016) RAD sequencing reveals within-generation polygenic selection in response to anthropogenic organic and metal contamination in North Atlantic Eels. *Molecular Ecology*, **25**, 219–237.

Leaniz, C.G. & Verspoor, E. (1989) Natural hybridization between Atlantic salmon, *Salmo salar*, and brown trout, *Salmo trutta*, in northern Spain. *Journal of Fish Biology*, **34**, 41–46.

Lemay, M.A. & Russello, M.A. (2015) Genetic evidence for ecological divergence in kokanee salmon. *Molecular Ecology*, **24**, 798–811.

Lescak, E.A., Bassham, S.L., Catchen, J., Gelmond, O., Sherbick, M.L., von Hippel, F.A. & Cresko, W.A. (2015) Evolution of stickleback in 50 years on earthquake-uplifted islands. *Proceedings of the National Academy of Sciences*, **112**, 201512020.

Longo, G. & Bernardi, G. (2015) The evolutionary history of the embiotocid surfperch radiation based on genome-wide RAD sequence data. *Molecular Phylogenetics and Evolution*, **88**, 55–63.

Lozier, J.D., Jackson, J.M., Dillon, M.E. & Strange, J.P. (2016) Population genomics of divergence among extreme and intermediate color forms in a polymorphic insect. *Ecology and Evolution*, **6**, 1075–1091.

Martin, C.H. & Feinstein, L.C. (2014) Novel trophic niches drive variable progress towards ecological speciation within an adaptive radiation of pupfishes. *Molecular Ecology*, **23**, 1846–1862.

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T.H., Piñero, D. & Emerson, B.C. (2015) Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Molecular Ecology Resources*, **15**, 28–41.

Nadeau, N.J., Whibley, A., Jones, R.T. *et al.* (2012) Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philosophical transactions of the Royal Society of London Series B, Biological Sciences*, **367**, 343–353.

Nims, B.D., Vargas, F.H., Merkel, J. & Parker, P.G. (2008) Low genetic diversity and lack of population structure in the endangered Galápagos penguin (*Spheniscus mendiculus*). *Conservation Genetics*, **9**, 1413–1420.

Palaiokostas, C., Bekaert, M., Taggart, J.B., Gharbi, K., McAndrew, B.J., Chatain, B., Penman, D.J. & Vandeputte, M. (2015) A new SNP-based vision of the genetics of sex determination in European sea bass (*Dicentrarchus labrax*). *Genetics Selection Evolution*, **47**, 68.

Pante, E., Abdelkrim, J., Viricel, A., Gey, D., France, S.C., Boisselier, M.C. & Samadi, S. (2015) Use of RAD sequencing for delimiting species. *Heredity*, **114**, 450–459.

Paris, J.R., King, R.A. & Stevens, J.R. (2015) Human mining activity across the ages determines the genetic structure of modern brown trout (*Salmo trutta* L.) populations. *Evolutionary Applications*, **8**, 573–585.

Pavey, S.A., Gaudin, J., Normandeau, E., Dionne, M., Castonguay, M., Audet, C. & Bernatchez, L. (2015) RAD sequencing highlights polygenic discrimination of habitat ecotypes in the panmictic American eel. *Current Biology*, **25**, 1666–1671.

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.

Razkin, O., Sonet, G., Breugelmans, K., Madeira, M.J., Gómez-Moliner, B.J. & Backeljau, T. (2016) Species limits, interspecific hybridization and phylogeny in the cryptic land snail complex *Pyramidula*: the power of RADseq data. *Molecular Phylogenetics and Evolution*, **101**, 267–278.

Rodríguez-Ezpeleta, N., Bradbury, I.R., Mendibil, I., Álvarez, P., Cotano, U. & Irigoien, X. (2016) Population structure of Atlantic Mackerel inferred from RAD-seq derived SNP markers: effects of sequence clustering parameters and hierarchical SNP selection. *Molecular Ecology Resources*, **16**, 991–1001.

Roeder, A.D., Marshall, R.K., Mitchelson, A.J. *et al.* (2001) Gene flow on the ice: genetic differentiation among Adélie penguin colonies around Antarctica. *Molecular Ecology*, **10**, 1645–1656.

Rougemont, Q., Gagnaire, P.-A., Perrier, C., Genthon, C., Besnard, A.-L., Launey, S. & Evanno, G. (2016) Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and nonparasitic lamprey ecotypes. *Molecular Ecology*, **26**, 142–162.

Rowe, H.C., Renaut, S. & Guggisberg, A. (2011) RAD in the realm of next-generation sequencing technologies. *Molecular Ecology*, **20**, 3499–3502.

Saenz-Agudelo, P., Dibattista, J.D., Piatek, M.J., Gaither, M.R., Harrison, H.B., Nanninga, G.B. & Berumen, M.L. (2015) Seascape genetics along environmental gradients in the Arabian Peninsula: insights from ddRAD sequencing of anemonefishes. *Molecular Ecology*, **24**, 6241–6255.

Sechi, P. (2013) An evolutionary history of the peregrine epigeic earthworm *Lumbricus rubellus*. PhD thesis, Cardiff University, Cardiff, UK.

Shafer, A.B.A., Peart, C.R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C.W. & Wolf, J.B.W. (2016) Bioinformatic processing of RAD-seq data dramatically

impacts downstream population genetic inference. *Methods in Ecology and Evolution*, doi: 10.1111/2041-210X.12700.

Shepeleva, O.A., Kodolova, O.P., Zhukovskaya, E.A. & Striganova, B.R. (2008) Genetic diversity of populations of the earthworm *Lumbricus rubellus*. *Biology Bulletin*, **35**, 170–177.

Suyama, Y. & Matsuki, Y. (2015) MIG-seq: an effective PCR-based method for genome-wide single-nucleotide polymorphism genotyping using the next-generation sequencing platform. *Scientific Reports*, **5**, 16963.

Tariel, J., Longo, G.C. & Bernardi, G. (2016) Tempo and mode of speciation in Holacanthus angelfishes based on RADseq markers. *Molecular Phylogenetics and Evolution*, **98**, 84–88.

Trucchi, E., Frajman, B., Haverkamp, T., Schoenswetter, P. & Paun, O. (2016) Genomic and metagenomic analyses reveal parallel ecological divergence in *Helisperma pusillusm* (Carophyllaceae). *bioRxiv*, https://doi.org/10.1101/044354.

Wagner, C.E., Keller, I., Wittwer, S., Selz, O.M., Mwaiko, S., Greuter, L., Sivasundar, A. & Seehausen, O. (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, **22**, 787–798.

Wu, T.D. & Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.

Xu, P., Xu, S., Wu, X., Tao, Y., Wang, B., Wang, S., Qin, D., Lu, Z. & Li, G. (2014) Population genomic analyses from low-coverage RAD-Seq data: a case study on the non-model cucurbit bottle gourd. *The Plant Journal*, **77**, 430–442.

Received 21 January 2017; accepted 13 March 2017
Handling Editor: Susan Johnston

Supporting Information

Details of electronic Supporting Information are provided below.

Appendix S1. Additional Figures S1–S5 and Tables S1–S15.