Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2016 doi: 10.1111/2041-210X.12700

Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference

Aaron B. A. Shafer†,[1,2], Claire R. Peart†,[1], Sergio Tusso[1], Inbar Maayan[1], Alan Brelsford[3], Christopher W. Wheat[4] and Jochen B. W. Wolf*,[1,5]

[1]*Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden;* [2]*Forensic Science and Environmental & Life Sciences, Trent University, 2014 East Bank Dr, K9J 7B8 Peterborough, Canada;* [3]*Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland;* [4]*Department of Zoology, Stockholm University, 106 91 Stockholm, Sweden; and* [5]*Division of Evolutionary Biology, Faculty of Biology, Ludwig-Maximilians University of Munich, Grosshaderner Str. 2, 82152 Planegg-Martinsried, Germany*

Summary

1. Restriction site-associated DNA sequencing (RAD-seq) provides high-resolution population genomic data at low cost, and has become an important component in ecological and evolutionary studies. As with all high-throughput technologies, analytic strategies require critical validation to ensure precise and unbiased interpretation.

2. To test the impact of bioinformatic data processing on downstream population genetic inferences, we analysed mammalian RAD-seq data (> 100 individuals) with 312 combinations of methodology (*de novo* vs. mapping to references of increasing divergence) and filtering criteria (missing data, HWE, F_{IS}, coverage, mapping and genotype quality). In an effort to identify commonalities and biases in all pipelines, we computed summary statistics (nr. loci, nr. SNP, π, Het_{obs}, F_{IS}, F_{ST}, N_e and m) and compared the results to independent null expectations (isolation-by-distance correlation, expected transition-to-transversion ratio T_s/T_v and Mendelian mismatch rates of known parent–offspring trios).

3. We observed large differences between reference-based and *de novo* approaches, the former generally calling more SNPs and reducing F_{IS} and T_s/T_v. Data completion levels showed little impact on most summary statistics, and F_{ST} estimates were robust across all pipelines. The site frequency spectrum was highly sensitive to the chosen approach as reflected in large variance of parameter estimates across demographic scenarios (single-population bottlenecks and isolation-with-migration model). Null expectations were best met by reference-based approaches, although contingent on the specific criteria.

4. We recommend that RAD-seq studies employ reference-based approaches to a closely related genome, and due to the high stochasticity associated with the pipeline advocate the use of multiple pipelines to ensure robust population genetic and demographic inferences.

Key-words: bioinformatics, GBS, genotyping by sequencing, population genetics, RAD-seq

Introduction

Advances in high-throughput sequencing technology have transformed ecological and evolutionary studies of non-model organisms (Ellegren 2014; Shafer *et al.* 2016a). While over 3000 eukaryotes have their entire genome sequenced (NCBI 2016), whole genome sequencing is still prohibitive for many species, especially those with large genomes, in terms of cost, bioinformatic resources and required DNA quality (Ekblom & Wolf 2014). These factors have led to a demand for alternative strategies. Leading the way is high-throughput sequencing of reduced representation libraries (i.e. a targeted subset of the genome), more specifically restriction site-associated DNA sequencing or RAD-seq (Andrews *et al.* 2016). Multiple terms and techniques have emerged, including RAD-seq, double-digest (dd) RAD-seq and genotyping by sequencing (GBS); here, we primarily use the term RAD-seq to highlight the generation of the library by restriction enzymes. The main advantages of RAD-seq are the relatively low costs, minimal required starting material and independence of a reference genome. Importantly, standard population genomic approaches like outlier scans (Kjeldsen *et al.* 2015), linkage mapping (Henning *et al.* 2014) and demographic analyses (Shafer *et al.* 2015a) can be conducted with RAD-seq data.

RAD-seq is not without caveats. The generation of reduced representation libraries generally relies on restriction enzymes and a PCR step. If a mutation has occurred in an enzyme cut-site in one individual, allelic dropout will occur and can bias biological inference (Arnold *et al.* 2013; Gautier *et al.* 2013).

*Correspondence author. E-mail: j.wolf@bio.lmu.de
†Shared first authorship.

Moreover, RAD-seq is particularly sensitive to PCR enrichment steps introducing artefacts (Davey *et al.* 2012). Alternative protocols using sonication have likewise been identified as problematic (Davey *et al.* 2012), and strand bias, i.e. different genotypes from forward and reverse reads, seems to be a general problem associated with short-read sequences (Guo *et al.* 2012). Not surprisingly, a debate has ensued as to which RAD-seq method is best (Andrews *et al.* 2014; Puritz *et al.* 2014). Although much attention has been focused on understanding the biases associated with different wet laboratory protocols, remarkably little attention has been directed towards the potential impact of bioinformatic pipelines on downstream biological inferences.

Studies comparing RAD-seq pipelines have generally been limited in focus to the number of loci and segregating variants or single nucleotide polymorphisms (SNPs) detected (Callicrate *et al.* 2014; Puritz, Hollenbeck & Gold 2014; Pante *et al.* 2015) – presumably under the premise that approaches yielding more SNPs are superior. What is most important, however, is not the number of SNPs a pipeline detects, but technical evaluations of how the called SNPs impact power, accuracy and precision of biological inferences. Given the importance of bioinformatic processing on biological inferences (Vijay *et al.* 2012; Chaisson, Wilson & Eichler 2015), comprehensive assessment of how different workflows influences RAD-seq data and resulting biological inferences is vital.

In this study we generated ddRAD data (Parchman *et al.* 2012; Peterson *et al.* 2012) for 106 individuals of the Galápagos sea lion (*Zalophus wollebaeki*), a top predator of the marine ecosystem of the Galápagos archipelago with evidence for spatially, behaviourally, morphologically and genetically structured ecotypes (Wolf *et al.* 2008; Jeglinski *et al.* 2015). With a mammalian-sized genome of approximately 3 Gb it constitutes a typical, yet – from a genomic perspective – challenging example of an ecoevolutionary model species (Ekblom & Wolf 2014). We analysed these data in a comprehensive framework to assess the potential impact of bioinformatic pipelines on the estimation of the site frequency spectrum (SFS), derived summary statistics and demographic inference (Fig. 1). In addition to descriptive statistics such as the number of loci or SNPs, we considered common population genetic summary statistics readily accessible from RAD-seq data such as nucleotide diversity (π), observed heterozygosity (Het_{obs}), measures of inbreeding (F_{IS}) and population differentiation (F_{ST}). We further explored pipeline effects in simple single-population and isolation-with-migration demographic models.

A key challenge facing population genomic studies of non-model organisms is to understand how bioinformatic pipelines impact inferences of biological processes when the *true biological pattern* remains unknown. In an effort to identify unbiased pipelines and suggest a best practice strategy we employed a two-pronged approach. We first assessed consistency and clustering of summary statistics among a total of 312 combinations of methods and filtering criteria (Fig. 1) providing insight into precision, commonalities and differences among pipelines. We then assessed potential biases comparing the computed summary statistics to independent null expectations including a previously published isolation-by-distance correlation (IBD; Wolf *et al.* 2008), the expected transition-to-transversion ratio (T_s/T_v; Li *et al.* 2010) and Mendelian mismatch rates derived from known parent–offspring trios (Pörschmann *et al.* 2010).

Materials and methods

SPECIES BACKGROUND AND SAMPLING STRATEGY

The Galápagos sea lion plays an integral role in the marine ecosystem of the Galápagos archipelago. The species lives in an isolated, small-scale marine environment with considerable ecological variation promoting two spatially, behaviourally, morphologically and genetically structured ecotypes (Wolf *et al.* 2008; Jeglinski *et al.* 2015). Sampling was conducted under permit PC 001-03/04 (export permit No. 099/04 SPNG) with support of the *Servicio Parque Nacional Galápagos* and the *Charles Darwin Research* Station; for DNA extraction and detailed sampling information, see Wolf *et al.* (2008). For this study we selected 94 individuals (51 males; 31 females; 12 unknown) from 10 different islands (see Data S1, Supporting Information). An additional 12 individuals from Caamaño Island constituting four known trios (mother–father–offspring) were also included (Pörschmann *et al.* 2010).

SEQUENCE DATA GENERATION

Double-digest restriction site-associated DNA sequencing (ddRAD) libraries were built using the 94 samples and the four trios following the protocol of Brelsford, Dufresnes & Perrin (2016), which was adapted from Parchman *et al.* (2012) and Peterson *et al.* (2012). Briefly, we used MseI and SbfI restriction enzymes and size-selected libraries of between 200 and 430 bp insert size. Paired-end sequencing (150 bp) was conducted on an Illumina HiSeq2500 machine. We also prepared paired-end libraries for whole-genome resequencing of 10 females (distributed across the Galápagos archipelago). We used the Illumina TruSeq Rapid SBS Kit according to the manufacturer's instructions (insert size: ~700 bp; range 300–1400 bp). Paired-end sequencing (150 bp) was conducted on an Illumina HiSeq2500 machine, yielding an average raw read sequence coverage of 5·7*x* (range 4·26*x*–7·57*x*). Sequence coverage estimation assumes a genome size of 3·08 Gb, as assessed from a C-value of 3·15 in the closely related California sea lion (Du & Wang 2006) and approximating the mean mass of a nucleotide pair as $1{\cdot}023 \times 10^{-9}$ pg/bp at 50% GC content (Doležel *et al.* 2003).

LOCUS ASSEMBLY AND SNP CALLING

Preliminary processing of ddRAD data

Due to the nature of the library preparation process, the majority of forward and reverse reads overlapped due to small fragment carryover (DaCosta & Sorenson 2014). Raw fastq files were therefore first treated with cutadapt v1.8 (Martin 2011) to remove adapter sequence; here, we included wildcards to account for the barcode and required a minimum read length of 50 bp after trimming. Reads were then demultiplexed using the *process_radtags* module of the STACKS pipeline (Catchen *et al.* 2013) with unmatched reads discarded. We merged all reads using PEAR v0.9.6 (Zhang *et al.* 2014) under default settings and a minimum length of 50 bp and proceeded to analyse the merged reads. Similar methods have detected between 10 000 and 50 000 loci (Torkamaneh,

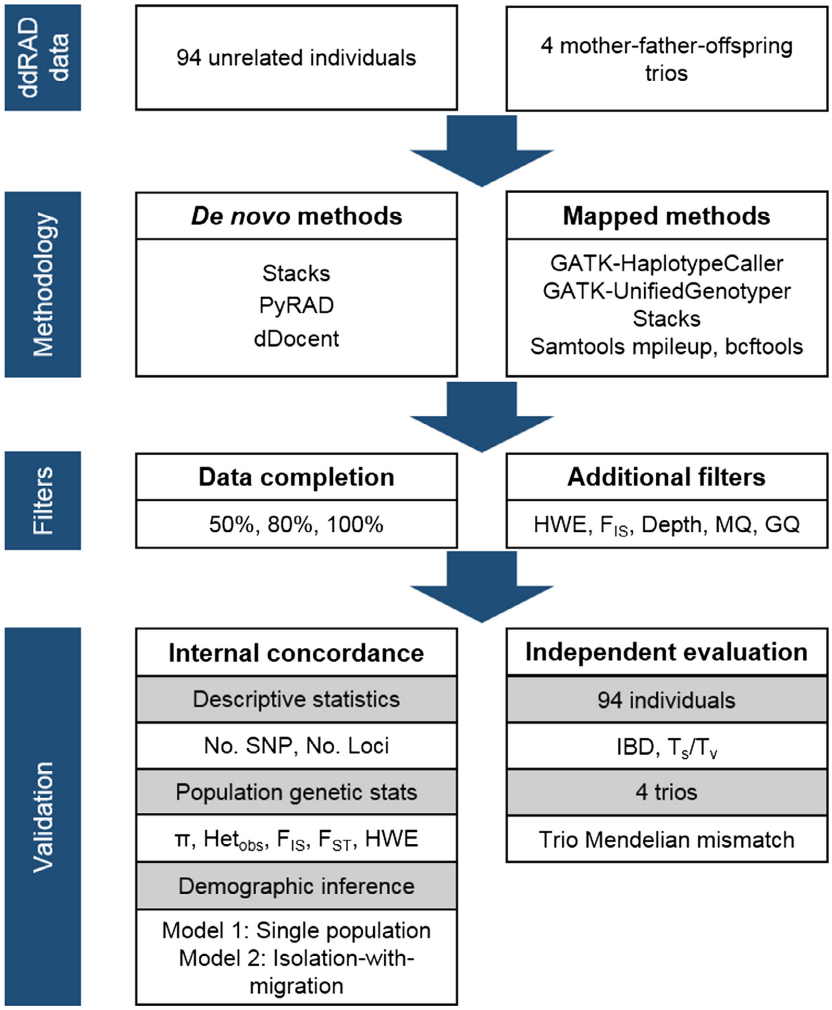

Fig. 1. Overview of this study. We generated double-digest restriction-associated DNA sequencing (ddRAD-seq) data for a total of 106 Galápagos sea lion individuals. The data were analysed using a series of *de novo* and reference-based methods including genome references of increasing sequence divergence (0, 0·01, 0·029 and 0·051) and a number of commonly used filters. To test for consistency among the resulting 312 combinations of method and filter (internal concordance), we calculated a series of common population genetic summary statistics and estimated demographic parameters. For external, independent validation, we used published data on isolation by distance, expected transition-transversion ratios and Medelian segregation as expected from known mother–father–offspring trios. Abbreviations (alphabetical order): F_{IS}, inbreeding coefficient; F_{ST}, genetic differentiation; GQ, genotyping quality; Het_{obs}, observed heterozygosity; HWE, the proportion of loci with significant deviations from Hardy–Weinberg equilibrium; IBD, Mantel correlation for isolation by distance; MQ, mapping quality; π, nucleotide diversity; T_s/T_v, transition-to-transversion ratio.

Laroche & Belzile 2016) which gives us an expected range as no genome is available for an *in silico* digest. Detailed bioinformatic steps and scripts are available in Data S1.

De novo methods (STACKS, PyRAD and dDocent)

All methods were executed with settings close to the default mode with additional pipeline parameters detailed in Table S1. We executed the STACKS modules ustacks, cstacks and sstacks (Catchen *et al.* 2013) on the merged reads under the following parameters: minimum stack depth (m) of 3 and 8, mismatch distance between loci within an individual (M) of 2 and number of mismatches between loci in the catalogue (n) of 1. These were selected through an optimization process similar to that found in Mastretta Yanes *et al.* (2015), but are close to the default (m = 2, M = 2, n = 0) commonly applied in the literature. As STACKS requires reads of equal length, we split the data set into two sets based on number of reads per length, one set with 5 quantiles, one set with 10 quantiles, respectively. Each quantile was then trimmed to its shortest length using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). The resulting quantile VCF files were then merged using VCFtools and treated as their own pipeline permutation.

The PyRAD pipeline (Eaton 2014) uses sequence similarity to cluster samples. This method is primarily used for phylogenetic analyses as it can incorporate insertion–deletion polymorphisms and lower levels of similarity between different species; however, high similarity values can be specified as clustering thresholds, allowing its use for population-level analyses. We used the PyRAD pipeline v.3.0.64 on the merged reads with two different clustering thresholds, 95% and 98%, and varied the minimum depth for a cluster to 3X, 6X and 8X.

Finally, the dDocent wrapper (Puritz, Hollenbeck & Gold 2014) that relies on linking other software was run using the FreeBayes SNP calling algorithm (Garrison & Marth 2012). Default parameters were used with the merged reads as input. dDocent used the a

mixture of programs for the *de novo* assembly (Rainbow: Chong, Ruan & Wu 2012 and CD-HiT: Fu *et al.* 2012) and mapping (BWA: Li & Durbin 2009).

Reference-based methods

Draft genomes are required for reference-based approaches. High-quality DNA – i.e. material with high molecular weight resulting in long-read inserts – was not available for the Galápagos sea lion. We therefore selected 10 females for resequencing totalling ~60X coverage and assembled a draft genome using the CLC workbench (QIAGEN Redwood City, Redwood City, CA, USA) consisting of 265 599 contigs. This number of contigs was too large for our analysis pipelines so the largest contigs were retained while the smaller contigs were artificially stitched together with strings of *N*s reducing the number to a manageable 5344 contigs. A second Galápagos sea lion genome draft was created by mapping the 10 females to the Antarctic fur seal (*Arctocephalus gazella)* genome assembly (Humble *et al.* 2016) using STAMPY v1.0.23 (Lunter & Goodson 2011) with a substitution rate of 0·01 and calling a consensus sequence using mpileup in SAMtools and bcftools (Li *et al.* 2009). While neither of these assembly strategies is optimal, both approaches are viable for many research groups, as high-quality draft genomes are still difficult and expensive to produce (Ekblom & Wolf 2014).

In many cases, researchers might choose to rely on published, but distantly related genomes. To investigate the effect of sequence divergence of the reference genome, we obtained three publicly available draft genomes from related pinniped species representing increasing divergence. These were as follows: Antarctic fur seal, Pacific walrus (*Odobenus rosmarus*) and Weddell seal (*Leptonychotes weddellii*). Samples were mapped to each draft genome using STAMPY v1.0.23 (Lunter & Goodson 2011). Default mapping parameters were used but with increasing substitution rates (0·00, 0·01, 0·029 and 0·051) for Galápagos sea lion, Antarctic fur seal, walrus and Weddell seal, respectively, based on divergence values among the Antarctic fur seal, Walrus and Weddell seal genomes (Humble *et al.* 2016). Following this the STACKS reference module for mapped reads was used with the following parameters: $m = 3$ and 8, $n = 1$. The mapped reads were sorted with SAMtools prior to SNP calling. SNPs were called with GATK (McKenna *et al.* 2010; DePristo *et al.* 2011), using both Haplotype Caller and Unified Genotyper and mpileup/bcftools (Li *et al.* 2009). The GATK pipeline was used to carry out local realignment for all samples combined and SNPs were called in the absence of any base quality score recalibration.

PIPELINE COMPARISONS

Locus filtering

All variable sites produced by each method were converted into VCF files, and we allowed for 50%, 20% and 0% missing data to give three levels of data completion. SNPs called using GATK Haplotype Caller, GATK Unified Genotyper and SAMtools/bcftools were further filtered using a mapping quality (MQ) threshold of 30, a minimum depth of 5, a genotype quality (GQ) of 30 and a maximum depth calculated as mean depth + $4\sqrt{\text{mean depth}}$ (Li 2014). All VCFs were subjected to a final filtering approach that retained only loci with F_{IS} [−0·05 to 0·05] and removed those deviating from Hardy–Weinberg equilibrium (HWE) at an α of 0·05. This latter step resulted in two classes of data (referred to as *filtered* and *unfiltered*) across the three levels of data completion. We note that all data were initially subjected to standard quality score checks and index verification.

Summary statistics

We calculated the following summary statistics in vcftools (Danecek *et al.* 2011): average π (per site), observed individual heterozygosity (Het_{obs}), F_{IS} (averaged across loci), proportion of loci deviating from HWE at an α of 0·05 and mean transition-to-transversion ratio (T_s/T_v). We report the number of SNPs and total number of loci; for the mapped methods that do not provide a summary of the number of loci, we defined a locus as continuously genotyped sites, allowing missing sites, with a gap of at least 200 bp between different loci (see Data S1 for script). F_{ST} (Weir & Cockerham 1984) was estimated between the east–west divide in the archipelago (Wolf *et al.* 2008) and IBD was estimated across all sampled rookeries using a Mantel test examining the relationship between geographic distance and $F_{ST}/1-F_{ST}$ (Rousset 1997). F_{ST} values between sampled rookeries were estimated using Genepop v4.4.3 (Rousset 2008) and Mantel tests (Mantel 1967; Smouse, Long & Sokal 1986) were performed using the R package *ecodist* v1.2.9 (Goslee & Urban 2007) with 10,000 permutations. For the four trios we calculated the proportion of SNPs displaying non-Mendelian inheritance patterns (mismatches) in vcftools.

Site frequency spectrum and demography

In the absence of an appropriate outgroup, as is commonly the case for RAD-seq data, the ancestral state of segregating variants cannot be determined. The following considerations are thus limited to the folded SFS. The shapes of the SFS were visually assessed by each method and filtering criteria. To account for missing data, all SFS were projected to half the sampled chromosomes using a hypergeometric distribution (Gutenkunst *et al.* 2009). To quantify the impact of the SFS on potential biological inferences, we estimated the parameters of two demographic models using a composite likelihood approach in ∂a∂i (Gutenkunst *et al.* 2009). The first model used the SFS to estimate range-wide changes in effective population size (timing and magnitude); wide parameter bounds were used to allow for a bottleneck or expansion detection. We further constructed the joint SFS of two demes to estimate the migration rates and split time between the east–west divide (Wolf *et al.* 2008). Confidence in parameters estimates was based on 100 bootstrap replicates (see Data S1 for code).

Evaluating the different methods and comparing to null expectations

We ran linear mixed-effect models with the summary statistics and trio mismatch score as response variables. Data completion, filtering (presence or absence), method (reference or *de novo*) and distance to reference genome were treated as fixed effects. The distance score for the *de novo* methods was entered as 0 and the SNP calling pipeline was treated as a random effect. The analyses were run using the *lmer* package in R.

To visualize the relationship among pipelines and fixed effects, the population summary statistics were first transformed to Z-scores. We then constructed a distance matrix using the *dist* function in R. A dendrogram was built using the *hclust* function available in the *R* cluster library (see Data S1 for code). For three statistics we had null expectations: we assume zero mismatches for true trios; IBDs similar to those obtained with microsatellites ($r = 0{\cdot}75$; Wolf *et al.* 2008) and a T_s/T_v of 2·1 similar to the panda genome, the sequenced genome most closely related to pinnipeds (Li *et al.* 2010). We calculated the Euclidean distances among the three summary statistics including the null expectation and constructed dendrograms as above. The congruence among dendrograms was assessed using a cophenetic correlation in R package *stats*.

Results

Population and trio ddRAD samples were sequenced on a HiSeq2500 resulting in 281 207 266 usable paired-end reads. In total, 624 VCF files were generated for the population and trio sequencing data (each consisting of 312 combinations of filtering, completion requirements and SNP calling pipelines – Fig. 1). The complete set of summary statistics is provided in Table S2.

VARIATION ACROSS SUMMARY STATISTICS AND THE SITE FREQUENCY SPECTRUM

The calculated summary statistics varied widely between the different methods and filtering options (Fig. 2a; Tables 1 and 2). The mean number of SNPs called was 13 984, but ranged from 16 to 159 451; GATK produces SNP counts on the tail ends of the distribution, with the primary difference being filter and completion criteria. Reference-based approaches called more SNPs before filtering than *de novo* methods, but there was large variation among methods (Fig. S1 and Table S2). The number of SNPs called per locus generally increased with distance to the reference genome. This is likely due to inappropriate alignment, also reflected in an artificial 'outbreeding signal' with strongly negative F_{IS} values for distant reference genomes (Table 1, Fig. S2).

Variation among methods was also large for per site π, Het_{obs} and F_{IS} estimates (Fig. 2a, Table 2). Linear mixed-effect models suggested that reference-based approaches produced higher values for per site π and individual Het_{obs}, but lower F_{IS} (Table 1). Filtering the data by F_{IS} and removing loci deviating from HWE generally reduced Het_{obs} and π estimates. $F_{ST,}$ and to a lesser extent isolation by distance were robust across pipelines (Table 1, Fig. 2). The level of missing data did not have a large effect on the summary statistics (Table 1); however, lower levels of missing data resulted in higher per site π values for GATK Unified Genotyper, but lower values for STACKS and mpileup (Table 2, Fig. S3).

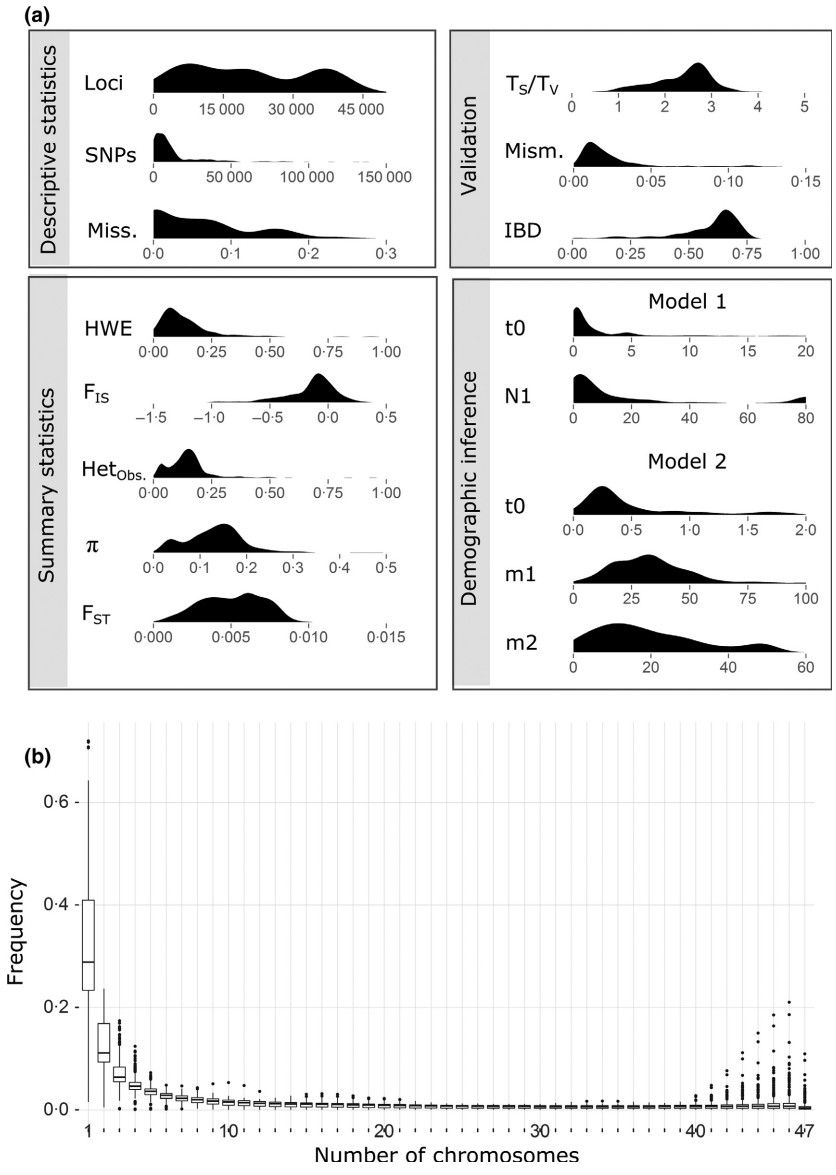

Fig. 2. Effect of bioinformatic processing on population genetic summary statistics. (a) Histogram depicting the range of parameter estimates from all 312 combinations of bioinformatic processing (Fig. 1) shown for *descriptive statistics*: the number of loci (Loci), the number single nucleotide polymorphisms (SNPs); proportion of loci with missing genotypes for any individual shown without 100% completion filter (Miss.) *summary statistics*: the proportion of loci with significant deviations from Hardy–Weinberg equilibrium (HWE), inbreeding coefficient (F_{IS}), observed heterozygosity (Het_{obs}), nucleotide diversity (π), genetic differentiation (F_{ST}); *external validation:* isolation-by-distance correlation (IBD), mismatches from expected Mendelian segregation (Mism.), transition-to-transversion ratio (T_s/T_v); *demographic inference:* split time (t0), change in population size (N1), asymmetric migration into population 1 (m1) or population 2 (m2). (b) Distribution of the folded SFS across all combinations shown as a boxplot depicting the frequency of the minor allele in all segregating sites (*y*-axis) as a function of the frequency of occurrence standardized to a maximum of 47 chromosomes (for the folded site frequency allowing for a maximum of 50% missing data, *x*-axis). Straight horizontal lines depict the median, box margins indicate the interquartile range between 25% and 75% quantiles and whiskers extend to 1·5 times the interquartile range.

Table 1. Coefficients and 95% confidence intervals in parentheses for the fixed effects in our mixed models

	De novo | Reference	Reference distance	Completion	Filtered | unfiltered	Residual variance
Summary statistics					
Heterozygosity	0·05 (0·02–0·09)	0·01 (0·00–0·01)	0·00 (0·00–0·00)	0·07 (0·05–0·09)	0·01 (0·08)
π	0·04 (0·02–0·06)	0·00 (0·00–0·00)	0·00 (0·00–0·00)	0·03 (0·03–0·04)	0·13 (0·36)
T_s/T_v	−0·17 (−0·34–0·01)	−0·06 (−0·09 to −0·04)	0·01 (0·01–0·01)	−0·16 (−0·24 to −0·08)	0·02 (0·12)
F_{ST}	0·00 (0·00–0·00)	0·00 (0·00–0·00)	0·00 (0·00–0·00)	0·00 (0·00–0·00)	1·49e−06 (0·00)
IBD	0·02 (−0·04–0·07)	−0·01 (−0·01–0·00)	0·00 (0·00–0·00)	0·00 (−0·03–0·03)	0·02 (0·13)
Mismatch	0·01 (−0·01–0·01)	0·00 (0·00–0·00)	0·00 (0·00–0·00)	0·00 (0·00–0·00)	6·25e−04 (0·03)
F_{IS}	−0·14 (−0·06 to –0·22)	−0·06 (−0·05 to −0·07)	0·00 (−0·01–0·00)	–	0·02 (0·13)
HW deviations	0·04 (−0·02–0·10)	0·04 (0·03–0·04)	0·00 (0·00–0·00)	–	0·01 (0·10)
Demographic parameters					
T_{Bottle}	−0·17 (−1·86–1·51)	−0·14 (−0·40–0·11)	0·01 (−0·01–0·02)	−0·11 (−0·98–0·76)	15·22 (3·90)
N_{Bottle}	−1·81 (−8·16–4·55)	0·41 (−0·44–1·23)	−0·01 (−0·08–0·07)	−2·45 (−5·40–0·44)	172·48 (13·13)
T_{Split}	−0·02 (−0·21–0·17)	0·00 (−0·04–0·02)	0·00 (0·00–0·00)	−0·10 (−0·20 to −0·01)	0·18 (0·42)
$M_{E\text{-}W}$	7·34 (1·00–13·69)	1·10 (0·06–2·16)	−0·03 (−0·12–0·05)	4·67 (1·07–8·28)	263·94 (16·25)
$M_{W\text{-}E}$	5·35 (0·14–5·36)	0·58 (−.23–1·38)	−0·07 (−0·14–0·00)	3·93 (1·15–6·70)	156·08 (12·49)

Response variables are shown in the leftmost column and the SNP calling pipeline is treated as a random effect. The residual variance and standard deviation (in parentheses) are also reported. Values for F_{IS} and deviation from Hardy–Weinberg proportions are before filtering.

Variation in estimates of the SFS across methods was large, in particular for low- and high-frequency alleles (Fig. 2b). *De novo* methods showed particularly high values for low-frequency alleles, in some cases over 60% of singletons. This increase in low-frequency alleles was mostly observed for the dDocent and pyRAD pipelines; SFS values resulting from the STACKS pipeline were more consistent and more similar to reference-based approaches (Fig. S4). The distribution of the SFS was strongly affected by distance to the reference. Additional peaks emerged with increasing distance to the reference, typically in high-frequency alleles (between 40 and 47 chromosomes in the SFS; Figs S5 and S6). Filtering for HWE and F_{IS} removed additional peaks while showing little effect on the distribution of low-frequency alleles (Fig. S6). Filtering by depth, MQ and GQ impacted the SFS, particularly for GATK Unified Genotyper and mpileup (Fig. S6).

VARIATION FOR DEMOGRAPHIC INFERENCE

Demographic inferences were highly variable, changing from predictions of a strong bottleneck to population expansion with large differences in estimates of the population size parameter N1 (Fig. 2a, Table S2). For the single-population demographic model, we found high variance but no strong systematic differences. Migration parameters in the isolation-with-migration model were influenced by pipeline and filtering; the split time parameter only by filtering (Table 1). The high variance in the demographic parameters obtained was mostly specific to each method or filter combination, as illustrated by the high residual variance of the statistical model (Table 1).

Based on results from the external validation (see below) we further investigated demographic inference within the confines of: a closely related reference genome (Galápagos sea lion or Antarctic fur seal) or STACKS *de novo*, a total SNP number between 2000 and 40 000 and a T_s/T_v ratio between 1·9 and 3·0. This substantially reduced the variance in demographic parameters in demographic model 1 (Fig. S7). The same exclusion criteria did not result in a large decrease in the variance of demographic parameters in the more complex demographic model 2. The split time estimate showed lower variance (c. 0·4), but migration rates were more variable. This suggests that the variance and consistency in demographic inference is not only contingent on bioinformatic processing, but also dependent on the demographic model under investigation.

VALIDATION USING EXTERNAL DATA

Variation in IBD was primarily impacted by the SNP calling pipeline (Tables 1 and 2). In addition, there was a significant reduction in the degree of correlation (Mantel test *r*-value) for methods with an extreme number of SNPs (lower than approximately 2000 or higher than 40 000). A high degree of missing data and distant reference genomes also resulted in lower Mantel correlations (Fig. S8). T_s/T_v was highly variable between methods and filtering strategies, ranging from 0·66 to 3·98 (mean 2·38). For the reference-based methods, distance from the reference negatively co-varied with T_s/T_v (Table 1, Fig. S9). Filtering SNPs by F_{IS} and HWE led to higher T_s/T_v (Table 1, Fig. S9). The GATK Unified Genotyper pipeline consistently had low T_s/T_v, as did mpileup methods when a high level of missing data was retained and there was no filtering for depth, MQ or GQ. The highest values were found using STACKS with a reference genome, resulting often in ratios over 3 (Table S2). The mean rate of trio mismatch based on Mendelian inheritance ranged from 0·02 to 0·03 for all four families. The linear mixed models showed no difference in mismatch rate after filtering between *de novo* and reference-based methods or with reference distance (Table 1). However, methods using GATK Haplotype Caller, mpileup and STACKS exhibited higher mismatch rates when using the most distant reference genomes, particularly when all missing data were removed (Fig. S11). Variance in mismatch rate appeared mainly due to differences between methods rather than between families (Fig. S10).

Table 2. Mean value and 95% confidence intervals in parentheses for the SNP calling pipelines

	STACKS	dDocent	PyRAD	GATK (Hap. caller)	GATK (Uni. genotyper)	mpileup
Summary statistics						
Heterozygosity	0·21 (0·18–0·22)	0·15 (0·12–0·18)	0·03 (0·03–0·03)	0·13 (0·12–0·14)	0·18 (0·14-0·22)	0·13 (0·11–0·14)
π	0·19 (0·17–0·20)	0·12 (0·11–0·13)	0·03 (0·03–0·03)	0·13 (0·11–0·14)	0·14 (0·12–0·16)	0·12 (0·11–0·13)
T_s/T_v	2·65 (2·52–2·76)	2·02 (2·01–2·04)	2·70 (2·67–2·73)	2·66 (2·60–2·73)	1·59 (1·48–1·70)	2·39 (2·26–2·52)
F_{ST}	0·01 (0·01–0·01)	0·01 (0·01–0·01)	0·00 (0·00–0·00)	0·01 (0·01–0·01)	0·00 (0·00–0·01)	0·01 (0·01–0·01)
IBD	0·64 (0·62–0·67)	0·62 (0·56–0·69)	0·56 (0·54–0·59)	0·58 (0·54–0·63)	0·49 (0·43–0·54)	0·62 (0·60–0·64)
Mismatch	0·02 (0·01–0·02)	0·02 (0·01–0·02)	0·03 (0·02–0·03)	0·03 (0·02–0·04)	0·03 (0·02–0·03)	0·03 (0·02–0·04)
F_{IS}	−0·17 (−0·23 to −0·11)	−0·29 (−0·29 to −0·28)	0·12 (0·08–0·16)	−0·04 (−0·01 to −0·08)	−0·31 (−0·21 to −0·42)	−0·16 (−0·20 to −0·11)
HW deviations	0·17 (0·13–0·21)	0·15 (0·15–0·15)	0·08 (0·05–0·11)	0·11 (0·09–0·14)	0·21 (0·14-0·30)	0·07 (0·05–0·09)
Demographic parameters						
T_{Bottle}	1·05 (0·60–1·50)	5·25 (0·70–9·80)	6·06 (4·75–7·39)	2·36 (1·25–3·47)	2·90 (1·63–4·16)	1·52 (0·66–2·40)
N_{Bottle}	2·25 (1·27–3·23)	39·72 (19·60–60·00)	74·00 (67·25–80·33)	10·92 (7·90–13. 95)	9·27 (6·02–12·52)	12·01 (8·73–15·30)
T_{Split}	0·86 (0·72–0·99)	0·16 (0·12–0·19)	0·29 (0·26–0·31)	0·51 (0·40–0·62)	0·37 (0·28–0·45)	0·38 (0·30–0·46)
$M_{E\text{-}W}$	37·20 (32·64–41·77)	30·53 (23·41–37·65)	16·75 (15·88–17·60)	33·45 (29·47–37·42)	36·80 (32·26–41·33)	37·00 (33·18–40·82)
$M_{W\text{-}E}$	17·15 (14·42–19·89)	20·53 (15·00–26·04)	10·38 (9·52–11·22)	16. 35 (12·89–19·82)	28·14 (24·54–31·75)	22·10 (18·18–26·02)

Note that these values were treated as a random effect in the mixed models. F_{IS} and Hardy–Weinberg deviation values are before filtering.

VISUALIZING PIPELINES AND IDENTIFYING THE LEAST-BIASED METHODS

A dendrogram summarizing the complex relationship among summary statistics as a function of the approach used is shown in Fig. 3a. In line with the aforementioned results, the dendrogram comprises several clusters that do not necessarily correspond to the different categories. The distribution of T_s/T_v values in Fig. 3b showed a variety of methods with close estimates to the null expectation. This included the *de novo* method dDocent and reference-based methods with a range of reference genomes. The *de novo* methods (often with low number of SNPs) showed the lowest trio mismatch (Fig. 3b); however, a variety of reference-based methods, particularly mpileup, had low levels of mismatch. Reference-based STACKS produced IBD correlations most similar to those of microsatellites (Fig. 3b). The clustering of dendrograms was not strongly correlated among the individual trio mismatch, IBD and T_s/T_v data sets: trio mismatch to T_s/T_v (r = 0·12); trio mismatch to IBD (r = 0·09); IBD to T_s/T_v (r = 0·15; dendrograms in Fig. S12).

Discussion

RAD-seq data are used extensively in population and ecological genomic studies, but there is no standard analysis pipeline with a variety of different alternatives commonly used (see Andrews *et al.* 2016). While there has been a considerable focus on detecting biases explicit to reduced representation sequencing (e.g. Davey *et al.* 2012; Arnold *et al.* 2013; Gautier *et al.* 2013), including the effect of experimental design and of library construction method (Andrews *et al.* 2014, 2016; Puritz *et al.* 2014), the effects of the bioinformatic pipeline on population genetic parameters have received little attention. Our study helps fill this knowledge gap by exploring the effects that standard bioinformatic pipelines for RAD-seq data have on population genetic inference. As we chose to work with an empirical example data set, certain aspects of the results will necessarily reflect unique features of the data. Studies simulating data (e.g. Vijay *et al.* 2012; Shafer *et al.* 2015a) are often better suited to assess precision and bias, yet they are bound to miss important variation introduced during wet laboratory data generation. In the absence of appropriate algorithms for simulating realistic RAD-seq data, an empirical approach is warranted and allows for broad conclusions on the impact of data processing.

Our key observation was a dramatic difference in summary statistics and demographic estimates dependent upon the bioinformatic pipeline, both when using *de novo* and reference-based approaches. Inference was particularly biased for demographic reconstruction, where some bioinformatic pipelines would imply outbreeding and population expansions, whereas others suggested a bottleneck and inbreeding. Relying on three summary statistics with independent null expectations, we found reference-based approaches of closely related genomes generally produced close to expected values, although no individual pipeline stood out as optimal. A key implication is this study calls into question previous population inferences made

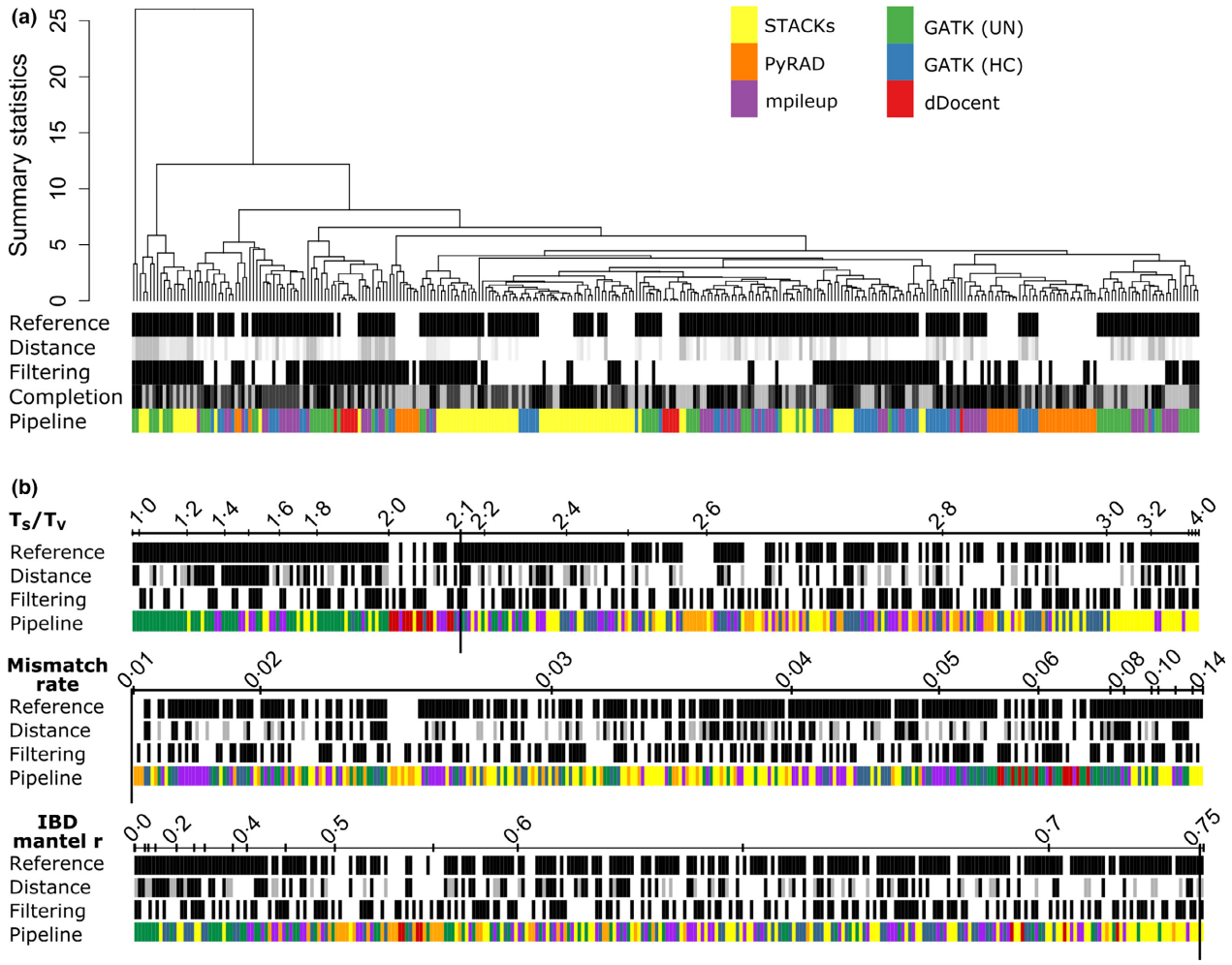

Fig. 3. Dendrogram summarizing the concordance of summary statistics as function of method and filtering for all 312 combinations (see Fig. 1). Categories are broken up by *reference: de novo* (white), reference based (black) further differentiated by *distance* to the reference genome (0·00, 0·01, 0·0029, 0·051 – gradient white to black), whether additional *filters* were applied (black) or not (white); by degree of *completion* (100%, 80%, 50% – black, dark grey, grey) and *method* used (see legend for colour coding). a) Dendrogram clustering approaches by similarity based on Euclidean distances integrated over the summary statistics π, Het_{obs}, F_{IS}, T_s/T_v, F_{ST}, Mantel correlation for isolation by distance (IBD), and proportion of loci deviating from expected Mendelian segregation (mismatch rate). (b) Combinations ordered linearly by T_s/T_v, mismatch rate and degree of IBD. Expected values obtained from external information are indicated with a vertical black bar.

using RAD-seq data and a single pipeline; importantly, our results imply that the allele frequency estimates produced from RAD-seq pipelines might not accurately reflect historical population processes. Multiple reference-based pipelines and independent null expectations should be employed to test for robustness and minimize the risk of extreme population genetic inferences.

SUMMARY STATISTICS

In the absence of independent estimates of summary statistics it is difficult to assess biases and select an appropriate bioinformatic method when it comes to RAD-seq data. Hohenlohe *et al.* (2010) reported reduced estimates of genetic diversity with RAD-seq and attributed it to conservative SNP calling. Our study adds another layer of complexity, as we found strong, often specific effects of bioinformatic pipeline on population genetic summary statistics. Importantly, this means many population genetic inferences based on RAD-seq data are likely to reflect bioinformatic processing as well as biological signal.

On a positive note, F_{ST} and to a lesser extent isolation-by-distance estimates were relatively robust across analyses (Tables 1 and 2), suggesting that some inferences into population differentiation can be made in spite of bias caused by the bioinformatic pipeline. However, there were often large differences among SFS and the remaining summary statistics between the *de novo* and reference-based methods. Generally, reference-based methods produced summary statistics most in line with our null expectations (Fig. 3). While the use of a reference genome is largely dictated by availability, even *ad hoc* draft genomes proved useful in our study. Reference genomes help by removing or masking repetitive regions and paralogous sequences, and permit analyses with greater statistical power (e.g. sliding window analyses, e.g. Hohenlohe *et al.* 2010; Martin *et al.* 2013). We also found that more SNPs were detected using reference-based approaches, even

with stringent filtering criteria, which might prove informative for studies of selection and adaptive divergence. However, the use of more distant genomes yielded unrealistically low T_s/T_v ratios, suggesting increasing miscalling. The results from several summary statistics suggest that the use of distantly related genomes should be avoided, although collectively reference-based approaches to moderately divergent genomes produced summary statistics most in line with null expectations (Fig. 3). Encouragingly, the two *ad hoc* Galápagos sea lion genome assemblies produced similar results to that of a closely related, quality-assembled genome, the Antarctic fur seal (Humble *et al.* 2016). Thus, when no closely related reference genome is available, the construction of a low-quality reference (10 individuals totalling roughly 60X coverage in our case) seems a viable option. Similar pseudo-reference genome approaches using RAD-seq data to construct a reference and then mapping to it (e.g. Hoffman *et al.* 2014; Mandeville *et al.* 2015) might be suitable alternatives. This option and further *de novo* pipelines that have not been investigated in this study (e.g. Sovic, Fries & Gibbs 2015) require further exploration.

RECONSTRUCTION OF POPULATION DEMOGRAPHY

Choice of bioinformatic pipeline showed a particularly dramatic impact on demographic reconstruction. While RAD-seq has not been used extensively in demographic analyses to date, it has great potential given the cost-effective information it yields on thousands of loci with independent coalescent histories. A study of simulated, short-read genomic data demonstrated that a variety of demographic models can in theory be reliably inferred (Shafer *et al.* 2015a). Accordingly, empirical genotyping-by-sequencing data were used to infer a demographic scenario consistent with the known historical biogeography of the reference species (Shafer *et al.* 2014). In our study, the consistency of parameter estimates appeared dependent on the demographic model, with migration and split times roughly consistent, while estimates of change in effective population size were highly variable (Table 2). These patterns are in line with findings from Shafer *et al.* (2015a) who showed that changes in effective population size (i.e. bottlenecks, expansions) are most difficult to accurately estimate, even in the case of simulated RAD-like genotyping data. These results illustrate the difficulty of correctly estimating the low-frequency variants of the SFS which hold key information regarding population demographic histories (Fig. 2b) – a problem also inherent in whole-genome resequencing (Pool *et al.* 2010). For our ddRAD data, the STACKS *de novo* pipeline and mapping to a close reference with moderate filters yielded the most consistent results (Fig. S7). Although additional analyses such as Approximate Bayesian Computation can be employed on these data, such methods also rely on summary statistics derived from the SFS and will be likewise impacted. Thus, in the absence of external references, such as a known date of a bottleneck or isolation (e.g. Nyman *et al.* 2014), researchers should assess the robustness of demographic parameters by screening a variety of bioinformatic pipelines.

FILTERING AND EXTERNAL VALIDATION

The trade-off between data quality and amount remains an open question for both RAD-seq data and for high-throughput sequencing data (Li 2014; Sims *et al.* 2014). The uncertainty in calling genotypes can often be incorporated into population genetic analyses through the use of genotype likelihoods (e.g. Li 2011; Korneliussen, Albrechtsen & Nielsen 2014) and has been applied to RAD-seq data in a *de novo* framework (Mandeville *et al.* 2015). Less stringent filtering which incorporates genotype uncertainty is thus a possible way forward. Interestingly, we observed no systematic effect of missing data on any of the summary statistics or demographic parameters, suggesting lower completion cut-offs might be appropriate for RAD-seq studies. In our study, strict filtering produced erratic SFS for the GATK Unified Genotyper pipeline, but showed no adverse effect on other reference-based methods (Fig. S6). We suspect the biggest differences between methods relates to locus definition (mapping and assembly errors) and the treatment of sequencing error. Sequencing errors in particular will result in an excess of singletons and a biased SFS (Johnson & Slatkin 2008); for lower coverage data sets the choice of SNP calling algorithm reflects a trade-off between SNP number and undercalling true heterozygotes (Nielsen *et al.* 2012). The variation within methods appears largely driven by completion (Fig. S5) which acts to shift the sample size and the lower bound of alleles that can be identified (Lynch 2009). Collectively, this illustrates the complex and at times unpredictable effects of data processing.

The best strategy for processing RAD-seq data will ultimately depend upon the aims and methodologies of each study and the level of uncertainty that can be tolerated. For example, the level of uncertainty surrounding demographic inferences reported here is likely not appropriate for conservation and management decisions (Shafer *et al.* 2015b), but might shed light on broad biogeographic hypotheses. There are modifications in laboratory protocols that can improve locus recovery (DaCosta & Sorenson 2014), and sample replicates can be used both to quantify errors and optimize bioinformatic methods (Mastretta Yanes *et al.* 2015); given the level of stochasticity presented here, we encourage the extra step of external validation through comparison of results to those derived from independent data.

Conclusions and best practice strategy

We showed a large effect of bioinformatic processing of RAD-seq data on standard population genetic inference. Despite complex implications specific to the applied combination of method and filter, we derive a set of recommendations for RAD-seq studies. (i) Use a closely related reference genome or consider an *ad hoc* genome assembly. If a *de novo* approach is the only option, our results suggest the industry standard STACKS is appropriate. (ii) Use multiple bioinformatic pipelines to assess the robustness of summary statistics. (iii) Use external information to guide the decision-making process. (iv) Replicate samples should be used to validate SNPs and

summary statistics (Mastretta Yanes *et al*. 2015) and selection of SNP pipelines should include genotype-likelihood approaches (e.g. Mandeville *et al*. 2015).

While the excitement surrounding data types like RAD-seq is warranted, the field has transitioned so rapidly that basic assessments of pipeline robustness and consistency have not been fully pursued. Studies assessing RAD-seq pipelines have begun to emerge (e.g. Callicrate *et al*. 2014; Puritz, Hollenbeck & Gold 2014; Pante *et al*. 2015) as the popularity in reduced genomic approaches continues to grow. The study presented here should be viewed as a cautionary tale advocating diligence in analysing and reporting RAD-seq data and calling for additional effort in the development of analysis pipelines.

Authors' contributions

Names listed within contributions are alphabetical. J.W. collected the samples and together with C.P. and A.S. conceived of the study; A.B., I.M. and A.S. generated the data; C.P., A.S. and S.T. analysed the data with help from I.M. and C.W.; all contributed to interpreting the data and writing the manuscript.

Acknowledgements

A.S. was funded by a Wenner-Gren and NSERC Banting Fellowship. Funding for the research came from the Swedish Royal Physiographic Society (A.S. and J.W.), Royal Swedish Academy of Science (A.S.) and The Swedish Research Council FORMAS (grant 231-2012-450 to J.W.).

Data accessibility

The raw ddRAD sequencing reads of 106 Galápagos sea lion individuals and whole-genome resequencing data of 10 individuals are available in the short-read archive (SRA) of the National Biotechnology Centre of Information (NCBI) under project number PRJNA349123. The two Galápagos sea lion assemblies derived from the resequencing data are available on the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.q14c1) (Shafer *et al*. 2016b). Three of pinniped genomes were previously available: Antarctic fur seal (Dryad Digital Repository http://dx.doi.org/10.5061/dryad.8kn8c.2) (Humble *et al*. 2016); Pacific walrus (NCBI Assembly: GCF_000321225.1) and Weddell seal (NCBI Assembly: GCF_000349705.1).

References

Andrews, K.R., Hohenlohe, P.A., Miller, M.R., Hand, B.K., Seeb, J.E. & Luikart, G. (2014). Trade-offs and utility of alternative RADseq methods: reply to Puritz et al. *Molecular Ecology*, **23**, 5943–5946.

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.

Brelsford, A., Dufresnes, C. & Perrin, N. (2016) High-density sex-specific linkage maps of a European tree frog (*Hyla arborea*) identify the sex chromosome without information on offspring sex. *Heredity*, **116**, 177–181.

Callicrate, T., Dikow, R., Thomas, J.W., Mullikin, J.C., Jarvis, E.D. & Fleischer, R.C. (2014) Genomic resources for the endangered Hawaiian honeycreepers. *BMC Genomics*, **15**, 1098.

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

Chaisson, M.J.P., Wilson, R.K. & Eichler, E.E. (2015) Genetic variation and the *de novo* assembly of human genomes. *Nature Reviews Genetics*, **16**, 627–640.

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

DaCosta, J.M. & Sorenson, M.D. (2014) Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PLoS ONE*, **9**, e106713.

Danecek, P., Auton, A., Abecasis, G. *et al*. & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, **27**, 2156–2158.

Davey, J.W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K. & Blaxter, M.L. (2012) Special features of RAD sequencing data: implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.

DePristo, M.A., Banks, E., Poplin, R. & Garimella, K.V. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Doležel, J., Bartoš, J., Voglmayr, H. & Greilhuber, J. (2003) Letter to the editor. *Cytometry Part A*, **51A**, 127–128.

Du, B. & Wang, D. (2006) C-values of seven marine mammal species determined by flow cytometry. *Zoological Science*, **23**, 1017–1020.

Eaton, D.A.R. (2014). PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses. *Bioinformatics (Oxford, England)*, **30**, 1844–1849.

Ekblom, R. & Wolf, J.B.W. (2014) A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, **7**, 1026–1042.

Ellegren, H. (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**, 51–63.

Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Garrison, E. & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907 [q-bio]*.

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.

Goslee, S.C. & Urban, D.L. (2007) The ecodist package for dissimilarity-based Analysis of Ecological Data | Goslee | Journal of Statistical Software. *Journal of Statistical Software*, **22**, 19.

Guo, Y., Li, J., Li, C.-I., Long, J., Samuels, D.C. & Shyr, Y. (2012) The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics*, **13**, 666.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D. (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data (ed. G. McVean). *PLoS Genetics*, **5**, e1000695–11.

Henning, F., Lee, H.J., Franchini, P. & Meyer, A. (2014) Genetic mapping of horizontal stripes in Lake Victoria cichlid fishes: benefits and pitfalls of using RAD markers for dense linkage mapping. *Molecular Ecology*, **23**, 5224–5240.

Hoffman, J.I., Simpson, F., David, P., Rijks, J.M., Kuiken, T., Thorne, M.A.S., Lacy, R.C. & Dasmahapatra, K.K. (2014) High-throughput sequencing reveals inbreeding depression in a natural population. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 3775–3780.

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A. & Cresko, W.A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags (P.A. Hohenlohe, S. Bassham, P.D. Etter, N. Stiffler, E.A. Johnson & W.A. Cresko, Eds.). *PLoS Genetics*, **6**, e1000862.

Humble, E., Martinez-Barrio, A., Forcada, J., Trathan, P.N., Thorne, M.a.S., Hoffmann, M., Wolf, J.B.W. & Hoffman, J.I. (2016) A draft fur seal genome provides insights into factors affecting SNP validation and how to mitigate them. *Molecular Ecology Resources*, **16**, 909–921.

Jeglinski, J.W.E., Wolf, J.B.W., Werner, C., Costa, D.P. & Trillmich, F. (2015) Differences in foraging ecology align with genetically divergent ecotypes of a highly mobile marine top predator. *Oecologia*, **179**, 1041–1052.

Johnson, P.L.F. & Slatkin, M. (2008) Accounting for bias from sequencing error in population genetic estimates. *Molecular Biology and Evolution*, **25**, 199–206.

Kjeldsen, S.R., Zenger, K.R., Leigh, K., Ellis, W., Tobey, J., Phalen, D., Melzer, A., FitzGibbon, S. & Raadsma, H.W. (2015) Genome-wide SNP loci reveal novel insights into koala (*Phascolarctos cinereus*) population variability across its range. *Conservation Genetics*, **17**, 337–353.

Korneliussen, T.S., Albrechtsen, A. & Nielsen, R. (2014) ANGSD: analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, **15**, 1.

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li, H. (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, **30**, 2843–2851.

Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. & 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, R., Fan, W., Tian, G. *et al*. (2010) The sequence and *de novo* assembly of the giant panda genome. *Nature*, **463**, 311–317.

Lunter, G. & Goodson, M. (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, **21**, 936–939.
Lynch, M. (2009) Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics*, **182**, 295–301.
Mandeville, E.G., Parchman, T.L., McDonald, D.B. & Buerkle, C.A. (2015) Highly variable reproductive isolation among pairs of Catostomusspecies. *Molecular Ecology*, **24**, 1856–1872.
Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.
Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.
Martin, S.H., Dasmahapatra, K.K., Nadeau, N.J. *et al*. (2013) Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Research*, gr.159426.113.
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.
McKenna, A., Hanna, M., Banks, E. *et al*. (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.
Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y. & Wang, J. (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data (ed. P. Awadalla). *PLoS ONE*, **7**, e37558.
Nyman, T., Valtonen, M., Aspi, J., Ruokonen, M., Kunnasranta, M. & Palo, J.U. (2014) Demographic histories and genetic diversities of Fennoscandian marine and landlocked ringed seal subspecies. *Ecology and Evolution*, **4**, 3420–3434.
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.
Parchman, T.L., Gompert, Z., Mudge, J., Schilkey, F.D., Benkman, C.W. & Buerkle, C.A. (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, **21**, 2991–3005.
Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S. & Hoekstra, H.E. (2012) Double digest RADseq: an inexpensive method for *De Novo* SNP discovery and genotyping in model and non-model species (ed. L. Orlando). *PLoS ONE*, **7**, e37135.
Pool, J.E., Hellmann, I., Jensen, J.D. & Nielsen, R. (2010) Population genetic inference from genomic sequence variation. *Genome Research*, **20**, 291–300.
Pörschmann, U., Trillmich, F., Mueller, B. & Wolf, J.B.W. (2010) Male reproductive success and its behavioural correlates in a polygynous mammal, the Galápagos sea lion (*Zalophus wollebaeki*). *Molecular Ecology*, **19**, 2574–2586.
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*, **2**, e431.
Puritz, J.B., Matz, M.V., Toonen, R.J., Weber, J.N., Bolnick, D.I. & Bird, C.E. (2014) Demystifying the RAD fad. *Molecular Ecology*, **23**, 5937–5942.
Rousset, F. (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, **145**, 1219–1228.
Rousset, F. (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.
Shafer, A.B., Davis, C.S., Coltman, D.W. & Stewart, R.E. (2014) Microsatellite assessment of walrus (*Odobenus rosmarus rosmarus*) stocks in Canada. *NAMMCO Scientific Publications*, **9**, 15–31.
Shafer, A.B.A., Gattepaille, L.M., Stewart, R.E.A. & Wolf, J.B.W. (2015a) Demographic inferences using short-read genomic data in an approximate Bayesian computation framework: *in silico* evaluation of power, biases and proof of concept in Atlantic walrus. *Molecular Ecology*, **24**, 328–345.
Shafer, A.B.A., Wolf, J.B.W., Alves, P.C. *et al*. (2015b). Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution*, **30**, 78–87.
Shafer, A.B.A., Northrup, J.M., Wikelski, M. & Wittemyer, G. (2016a) Forecasting ecological genomics: high-tech animal instrumentation meets high-throughput sequencing. *PLoS Biology*, **14**, e1002350.
Shafer, A.B.A., Peart, C.R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C.W. & Wolf, J.B.W. (2016b) Data from: Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Dryad Digital Repository*, http://dx.doi.org/10.5061/dryad.q14c1
Sims, D., Sudbery, I., Ilott, N.E., Heger, A. & Ponting, C.P. (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, **15**, 121–132.
Smouse, P.E., Long, J.C. & Sokal, R.R. (1986) Multiple regression and correlation extensions of the mantel test of matrix correspondence. *Systematic Biology*, **35**, 627–632.
Sovic, M.G., Fries, A.C. & Gibbs, H.L. (2015) AftrRAD: a pipeline for accurate and efficient *de novo* assembly of RADseq data. *Molecular Ecology Resources*, **15**, 1163–1171.
Torkamaneh, D., Laroche, J. & Belzile, F. (2016) Genome-wide SNP calling from genotyping by sequencing (GBS) data: a comparison of seven pipelines and two sequencing technologies. *PLoS ONE*, **11**, e0161333.
Vijay, N., Poelstra, J.W., Künstner, A. & Wolf, J.B.W. (2012) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silicoassessment of RNA-seq experiments. *Molecular Ecology*, **22**, 620–634.
Weir, B.S. & Cockerham, C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
Wolf, J.B.W., Harrod, C., Brunner, S., Salazar, S., Trillmich, F. & Tautz, D. (2008) Tracing early stages of species differentiation: ecological, morphological and genetic divergence of Galápagos sea lion populations. *BMC Evolutionary Biology*, **8**, 150.
Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, **30**, 614–620.

Received 6 September 2016; accepted 25 October 2016
Handling Editor: M. Gilbert

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Fig. S1. Plot of SNP and locus number.

Fig. S2. Boxplot of summary statistics grouped by reference genome.

Fig. S3. Boxplot of summary statistics grouped by reference genome and missing data.

Fig. S4. Boxplot of site frequency spectra.

Fig. S5. Site frequency spectra from data generated using the STACKs pipeline.

Fig. S6. Site frequency spectra from data generated using mpileup and GATK.

Fig. S7. Parameter estimates for demographic model 1.

Fig. S8. Relationship of SNP number and Mantel *r* (isolation by distance).

Fig. S9. Boxplot of T_s/T_v ratio by reference genome.

Fig. S10. Histogram of variance in mismatch rates.

Fig. S11. Boxplot of trio mismatches.

Fig. S12. Dendrograms of summary statistics.

Table S1. Additional pipeline parameters.

Table S2. Pipeline summary statistics.

Data S1. Scripts used for computational analyses.