Received: 8 August 2019 | Accepted: 17 September 2019

DOI: 10.1111/mec.15253

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics

Nicolas C. Rochette | Angel G. Rivera-Colón | Julian M. Catchen

Department of Evolution, Ecology, and Behavior, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Correspondence
Julian M. Catchen, Department of Evolution, Ecology, and Behavior, University of Illinois at Urbana-Champaign, Urbana, IL, USA.
Email: jcatchen@illinois.edu

Funding information
NSF, Grant/Award Number: 1645087

Abstract

For half a century population genetics studies have put type II restriction endonucleases to work. Now, coupled with massively-parallel, short-read sequencing, the family of RAD protocols that wields these enzymes has generated vast genetic knowledge from the natural world. Here, we describe the first software natively capable of using paired-end sequencing to derive short contigs from de novo RAD data. Stacks version 2 employs a de Bruijn graph assembler to build and connect contigs from forward and reverse reads for each de novo RAD locus, which it then uses as a reference for read alignments. The new architecture allows all the individuals in a metapopulation to be considered at the same time as each RAD locus is processed. This enables a Bayesian genotype caller to provide precise SNPs, and a robust algorithm to phase those SNPs into long haplotypes, generating RAD loci that are 400–800 bp in length. To prove its recall and precision, we tested the software with simulated data and compared reference-aligned and de novo analyses of three empirical data sets. Our study shows that the latest version of Stacks is highly accurate and outperforms other software in assembling and genotyping paired-end de novo data sets.

KEYWORDS
bioinformatics, conservation genetics, genotype calling, haplotype phasing, population genetics, restriction-site associated DNA sequencing

1 | INTRODUCTION

Type II restriction enzymes (Kelly & Smith, 1970; Smith & Welcox, 1970) remain one of the primary drivers in population genetics experiments. Starting with their first application in the mid-1970s (Botstein, White, Skolnick, & Davis, 1980; Grodzicker, Williams, Sharp, & Sambrook, 1974), restriction enzymes have been paired with advancing technologies, including the polymerase chain reaction coupled with polyacrylamide gel analysis (Blears, De Grandis, Lee, & Trevors, 1998), microarrays (Miller, Dunham, Amores, Cresko, & Johnson, 2007), and most recently, massively parallel, short-read sequencing to yield great insights into model and non-model organisms, laboratory and natural populations (Narum, Buerkle, Davey, Miller, & Hohenlohe, 2013; Schlötterer, 2004). While sequencing costs have decreased by orders of magnitude since the completion of the human genome project, the cost is still too high in the majority of ecologically-based, non-model studies for whole genome re-sequencing, leaving a wide niche for the set of Restriction-site Associated DNA sequencing (RADseq) protocols.

RADseq has grown into a family of protocols whose kin have been optimized to different criteria. The protocols vary on a few major axes; first, a protocol may employ one (GBS; Elshire et al., 2011; single-digest RAD, sdRAD; Baird et al., 2008; BestRAD; Ali et al., 2016) or two restriction enzymes (CRoPS; van Orsouw et al., 2007; double-digest RAD; Peterson, Weber, Kay, Fisher, & Hoekstra, 2012). Second, it may rely on the distance between cut sites to determine the length of DNA that is sampled (CRoPS, GBS, ddRAD), or it may employ sonication to create relatively uniform lengths of DNA (sdRAD, BestRAD). Third, protocols may use a size selection step to explicitly select the length range of DNA molecules to enrich, or they

may rely on PCR to enrich shorter sequences for the final sequencing library. Additional protocols are further specialized, focused on adapting more of the steps to off-the-shelf kits (ezRAD; Toonen et al., 2013). Others were designed to minimize primer-dimers (3RAD; Graham et al., 2015), or to use type IIb restriction enzymes (2bRAD; Wang, Meyer, McKay, & Matz, 2012), or to use biotinylated adapters to extract restriction site-associated DNA from other genomic DNA (BestRAD), as well as hybrids of the above approaches (for reviews see Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Davey et al., 2011).

While RADseq can produce orders of magnitude more genetic markers than earlier marker technologies, whole-genome sequencing (WGS) produces an order of magnitude greater still. The primary obstacle, however, remains cost. It is popular, but misguided, to relate changes in sequencing technologies to Moore's Law; Moore's Law requires a halving of cost every 18 months and has held for sixty years, while sequencing technology has instead reduced cost by orders of magnitude twice (first with 454 pyrosequencing, and later with Illumina sequencing-by-synthesis) (http://genome.gov/sequencingcosts), and while Illumina has continued to reduce costs incrementally, there is no clear path to any order-of-magnitude-changes on the horizon. In fact, long molecule sequencing is significantly more expensive than technology that came prior. For every 480 Mb threespine stickleback fish genome that is resequenced, far more than 100 individuals can be examined with RAD (sampling 3% of each genome with the rare-cutting *SbfI* enzyme) for the same sequencing cost and with a single library preparation (WGS requires a library per genome). For large studies, the resource advantage RAD provides appears to be stable for the near future.

The union of genome sampling protocols with massively parallel, short-read sequencing has produced an immensely successful research programme in population (Bassham, Catchen, Lescak, von Hippel, & Cresko, 2018), conservation (Dierickx, Shultz, Sato, Hiraoka, & Edwards, 2015) and landscape genomics (Bay et al., 2018), phylogenetics (Spriggs et al., 2019), and epigenetics (Trucchi et al., 2016), creating new experimental space for non-model organisms, and allowing, for example, ambitious sampling regimes in large geographical surveys (Dudaniec, Yong, Lancaster, Svensson, & Hansson, 2018), as well as wide-ranging taxon breadth in phylogenetic studies (Near et al., 2018). Regardless of the analytical approach, and in addition to any challenges of the experimental design, all RADseq strategies present two fundamental issues. First, the precision of the analytical results depends significantly on the quality and amount of DNA available (Casbon, Osborne, Brenner, & Lichtenstein, 2011). RADseq has expanded the scope of organisms that can be examined, but sampling many of these organisms from nature is a challenge and, in these cases, DNA may be degraded or available only in small quantities (Graham et al., 2015; Suchan et al., 2016). Second, restriction sites may not be conserved across all individuals in the experiment, depending on evolutionary distance between them, and the length of the restriction site(s) of interest. In both cases, the molecular library may not contain a sufficient number of template molecules from all of the alleles in the genome in each individual. Two additional processes will sample these template molecules in the library: PCR amplification will sample molecules to create additional copies, and the sequencer will select from the amplified molecules for inclusion on the sequencing flow cell. Having too few templates for amplification (Casbon et al., 2011), or selecting too few molecules to sequence (a low depth of coverage) can exacerbate the effects of allelic dropout.

Through our participation in a large number of studies conducted over the last decade, both our own, collaborations in the field, and by interacting with scores of scientists in the support of Stacks version 1 (v1), we have learned a lot. The quality of DNA, the success of library preparation, and the sequencing strategy – all contributing to differential allelic sampling – can separate the pathbreaking RADseq studies from the rest. Often, the differences between these studies generated substantial discussion in the community (Catchen et al., 2017; Lowry et al., 2017; McKinney, Larson, Seeb, & Seeb, 2017) and a lot of speculation as to the inherent limitations of reduced representation sequencing.

Our experience is that with the development of proper analytical protocols (Paris, Stevens, & Catchen, 2017; Rochette & Catchen, 2017), and with supportive software, we can close the performance gap between RAD studies and secure a successful experimental strategy for the next decade. We sought to design a software that could help identify poorly performing libraries and provide support in the design of new studies. We sought to maximize the amount of information we could extract from RAD protocols by focusing on Illumina paired-end, short-read sequencing and improving the analysis tools to provide the most polymorphic loci possible and the richest set of haplotypes to increase information yield significantly.

The collection of software to implement this strategy has resulted in the second major version of Stacks. Version 2 (v2) incorporates paired-end reads natively into the locus assembly algorithm providing for locus sizes >500 bp, increasing the number of polymorphic loci, significantly improving genotyping accuracy, and providing phased haplotype markers for those loci, in a massively scalable form. As we show, Stacks v2 outperforms every other RAD software for the purpose of analyzing de novo paired-end RADseq data. Finally, to vet and optimize the software, we designed and implemented an accurate RAD simulation system which shed light on basic processes, such as the effects of PCR duplicates and sequencing coverage while providing us with a road map to fully optimize our algorithms.

2 | MATERIALS AND METHODS

2.1 | Changes to the Stacks pipeline

Stacks v1 was designed to process individual samples independently, identifying polymorphic sites within each individual (USTACKS or PSTACKS), then connecting loci across samples, through the creation of a catalog, to provide a single view of the metapopulation (CSTACKS). Individuals could then be matched to the metapopulation data contained in the catalog with SSTACKS (Catchen, Amores, Hohenlohe,

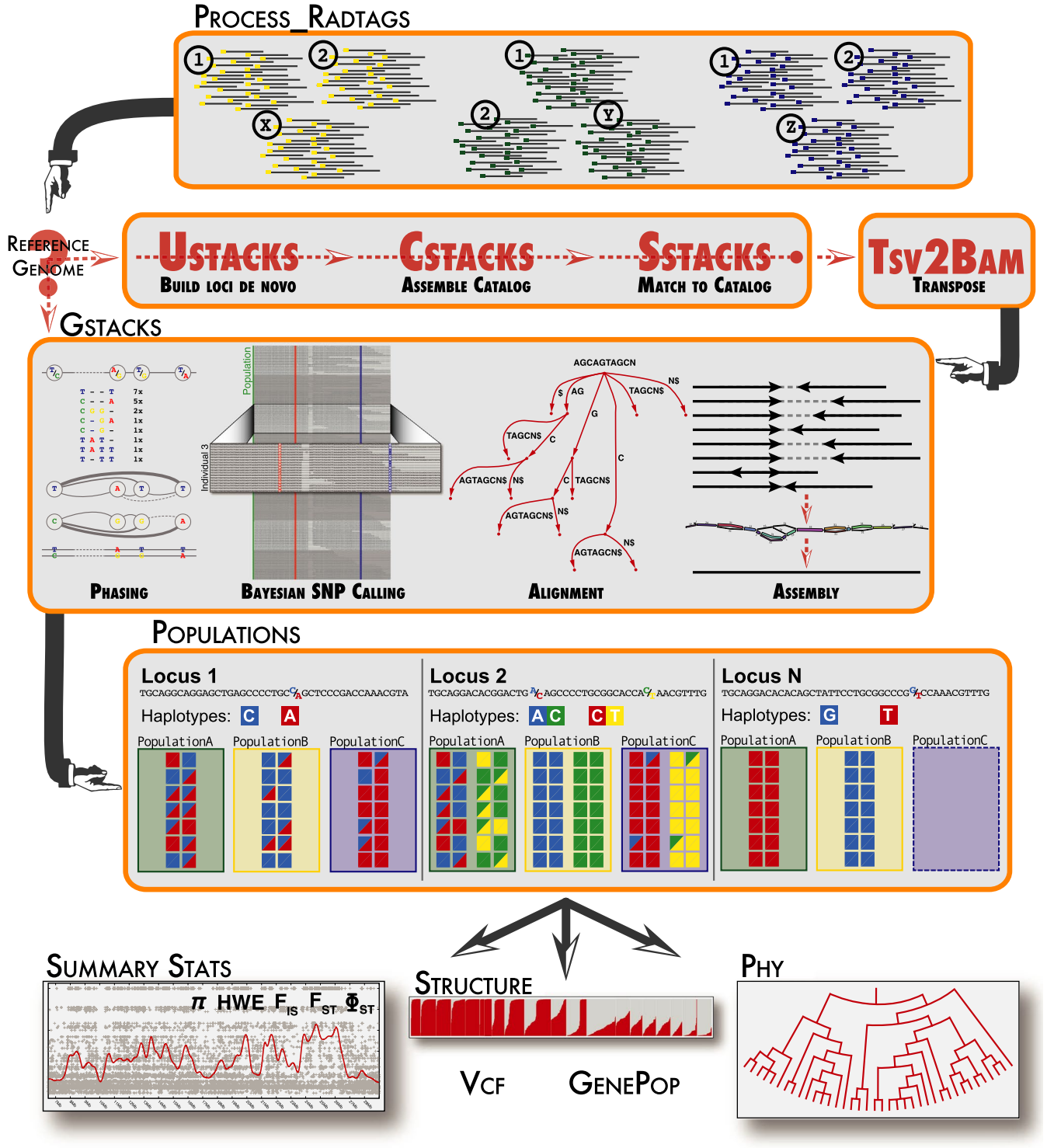

FIGURE 1 Stacks v2 pipeline overview. The software consists of four major components (shown graphically by pointed-finger icons). The USTACKS/CSTACKS/SSTACKS/TSV2BAM components are only used in de novo analyses. The GSTACKS component is new in Stacks v2 and is applied in both de novo and reference-based analyses. The POPULATIONS program applies a population genetics frame to the data and provides summary statistics and data exports

Cresko, & Postlethwait, 2011). This design was motivated by computational resource constraints as the pipeline needed to process potentially thousands of individual samples, each with millions of raw reads. However, the architecture limited how much information could be shared across individuals (e.g., for genotyping and phasing). Version 1.10 of Stacks (2013) patched this design weakness by incorporating the RXSTACKS program to share population-level information and make corrections retrospectively.

The central architectural change in v2 is the reorganization of individual-level data so that it is stored per locus instead of per

individual. This cosmetically simple change, implemented in the TSV2BAM program, enables large analytical gains downstream in the pipeline (Figure 1). De novo analyses start by clustering loci as in Stacks v1 using USTACKS, CSTACKS and SSTACKS, but then continue with TSV2BAM and additionally with the GSTACKS program, which now forms the analytical core of the v2 pipeline. The POPULATIONS program, which applies a population genetic-frame to the data, allowing for data filtering and data export, has also been rearchitected to process loci in a streaming fashion where previously all loci from all samples had to fit into computer memory at once, leading to a memory-bound program.

For reference-based analyses, the main Stacks pipeline now begins directly with the GSTACKS program (the v1 clustering pipeline is not employed) followed by a call to the POPULATIONS program.

2.2 | Improved locus clustering procedure

For de novo analyses, the core clustering algorithm (USTACKS-CSTACKS-SSTACKS), which builds loci out of the forward reads, remains as previously described (Catchen et al., 2011; Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013), but has received a number of improvements and parameter optimizations. Stacks has been capable of gapped assemblies since version 1.38 (2016), when Needleman–Wunsch comparisons between stacks sharing many k-mers were added, and in v2 this capability has become the default. In addition, when constructing the initial catalog (CSTACKS, version 2.0, 2018), the handling of loci differentially assembled across individuals has been changed from favouring under-merging to slightly favouring over-merging, so as to avoid dropping alleles due to pairwise similarity considerations.

2.3 | Transposing sequence data storage

The new TSV2BAM program, which concludes the locus clustering stage, sorts the reads (or read pairs) of each individual by catalog locus storing them as standard BAM files. Reverse reads can be incorporated by matching the set of forward read IDs in each locus. This is similar in principle to the sorting of alignment files by chromosome and coordinate, a strategy employed by most reference-based analysis pipelines, although as no locus reference sequence or alignments exist at this stage of the pipeline, reads are not sorted within loci. Importantly, this partial sorting step allows Stacks to stream the data locus by locus in subsequent analyses in GSTACKS, making it computationally feasible to work, at each locus, with the full sequencing information for all samples simultaneously.

2.4 | RAD-locus contig assembly

In de novo mode, given the set of clustered forward reads (produced by USTACKS-CSTACKS-SSTACKS), and the associated reverse reads (fetched by TSV2BAM), GSTACKS starts by assembling a contig for each locus. The reads of the locus (≤1,000 reads or read pairs sampled uniformly across individuals) are broken into k-mers and k-mers with a coverage of two or higher are inserted into a de Bruijn graph (Figure 2). By defureult, the method uses a k-mer size of 31, as for small single-locus de Bruijn graphs this value provides a good compromise between resolving repeats and maximizing useful k-mer coverage in presence of sequencing errors and polymorphisms. Stacks v2 then scores each connected subgraph for total coverage from forward and reverse reads (if any). The subgraphs with the highest total coverage for forward and for reverse reads are extracted (note that they may have connected into a single subgraph); other subgraphs are discarded.

These subgraphs are expected to be acyclic. If the subgraph for reverse reads contains cycles, reverse reads are discarded and the graph is recomputed using forward reads only. If the subgraph for forward reads contains cycles, the entire locus is discarded. Furthermore, to mitigate the effect of microsatellites, especially of long unbroken two-microsatellite repeats, we specifically screen for such events and attempt to remove the cycle by pruning one of the two involved k-mers (Figure S1).

The forward and reverse subgraph(s) are then converted into nucleotide sequence contig(s) by sorting them topologically and finding the path that maximizes the total coverage of the included k-mers, using conventional exact polynomial algorithms for directed acyclic graphs (DAGs). Using maximal cumulative coverage as the scoring criterion favours the inclusion of major alleles over minor ones, and marginally of insertions over deletions.

If the forward and reverse connected subgraphs are distinct, this procedure produces two separate contigs. A suffix tree is used to overlap them (see Section 2.5). If no matches were found using the suffix tree, the 3′ end of the forward-read locus is compared directly against the 5′ end of the reverse-read contig using a Needleman-Wunsch gapped alignment to check for any overlap that is below the minimum

FIGURE 2 Example de Bruijn graph built from the reverse reads for one locus of the *Pagothenia borchgrevinki* data set. Each rectangle represents one k-mer, with size scaling with coverage. In comparison with genomic de Bruijn graphs, our RAD-locus graphs are considerably smaller (this graph contains only 580 k-mers), but features more variable coverage and significantly more genetic diversity, visible as branching and loops, because the graph is built from the reads of 71 individuals. Identifying the path that has the highest total coverage (highlighted in blue and black) allows to derive a reference sequence for the locus

match length of the suffix tree. If there is no overlap, the two sequences are merged into a single scaffold with 10 'N' characters, to symbolize a gap of unknown length between the forward and reverse regions.

The algorithms described here also function well for paired-end data generated from ddRAD. Since both the forward and reverse reads are anchored to a restriction site, coverage will be uniform on both sides of the locus, and the length of the contig will reflect the distance between restriction sites. If the restriction sites are <2 read lengths apart, Stacks will overlap the two sides and yield a single continuous contig. There exist limit cases where the distance between restriction sites varies within the population due to structural variation, or to restriction site polymorphism and rescue (with ddRAD an allele that should be dropped due to a polymorphic restriction site may be rescued by another immediately adjacent restriction site). The algorithm naturally handles these cases as variable secondary restriction sites will all be captured by the de Bruijn graph which will collapse them into a single, linear sequence. These loci will often have atypical properties, such as being both continuous and longer than two read lengths.

2.5 | Read alignments

Once the full locus has been assembled, individual reads must be aligned in a per-sample context. A suffix tree is constructed from the locus consensus sequence using Ukkonen's algorithm (Gusfield, 1997), which can build the tree in linear time. Reads are then aligned against the suffix tree, with three possible outcomes. First, no alignments may be found, in which case the read is ignored. Second, a perfect alignment may have been found against the suffix tree, from which an alignment is calculated, or third, more than one maximal match was found against the suffix tree in which case the alignment fragments are ordered into a DAG, the consistent alignments from the DAG are used to populate a Needleman-Wunsch gapped alignment matrix, and a bounded, gapped alignment is conducted to connect the aligned fragments for the final alignment. A CIGAR string (Li et al., 2009) representing the alignment is recorded and the process is repeated until all reads from all samples have been aligned.

Once read alignments have been computed, the de novo pipeline continues with the same steps as in the reference-based pipeline, as if the locus contigs derived above formed an ad hoc reference genome.

2.6 | Reference-based analyses

In reference mode, Stacks v2 begins directly with the GSTACKS program (Figure 1) and relies on read alignments from an external alignment program (e.g., BWA-MEM or Bowtie; Langmead & Salzberg, 2012; Li, 2013; Li & Durbin, 2009), that are typically provided as one sorted BAM file per individual (Li et al., 2009). GSTACKS scans the genome for RAD loci using a sliding window method, reading all input BAM files simultaneously. Read pairs are clustered into RAD loci based on the alignment position of the partial restriction site at the 5′ end of forward reads (forward reads of which the 5′ end does not align are ignored). Locus starting positions (i.e., partial restriction sites) are considered by ascending coordinates and all read alignments mapping within 1 kb (by default) on either side of the locus are kept in memory. Within each individual, paired reads are matched by identifier. Once a locus has been built, the same filtering and genotyping methods are applied as in the de novo mode, then the program advances to the next locus.

2.7 | PCR duplicates removal

For paired-end data derived from shearing-based protocols, GSTACKS offers the option to filter out PCR duplicates, by identifying inserts (i.e., read pairs) that belong to the same individual and map to the same start and end coordinates. Filtering occurs by randomly discarding all but one pair of each set of reads. This approach relies on the randomness of the shearing process, which creates a mixture of inserts of diverse lengths that are then preserved through PCR amplification. While distinct source molecules will sometimes have identical insert sizes by chance, discarding one of them at random does not introduce bias, and these coincidences are rare in the typical medium-coverage configurations used in RADseq experiments (in contrast with e.g., RNAseq), so discarding inserts of identical sizes efficiently and selectively removes PCR duplicates.

This filtering method cannot be used with ddRAD protocols because insert lengths are determined by the distance between the two restriction sites and are invariant. Some protocols, however, incorporate random oligo-nucleotides into the barcodes of the molecular library (Graham et al., 2015); in this case, PCR duplicates may be removed prior to running the main pipeline by using the CLONE _ FILTER program of Stacks.

2.8 | Genotyping model

Stacks v1 employed the single nucleotide polymorphism (SNP) calling method described in Hohenlohe et al. (2010) which itself was modified from Lynch (2009). This model worked well; however, it required relatively high coverage for robust results, and it examined a single individual at a time instead of aggregating evidence for polymorphism across individuals. For Stacks v2, we again incorporated the work of Lynch and colleagues, who provide a tractable statistical framework in their Bayesian genotype caller (BGC; Maruki & Lynch, 2015, 2017). The GSTACKS program employs the BGC to identify the presence of biallelic SNPs within a locus by examining the read data from the entire metapopulation. Support for the presence of a SNP is then fed into the genotyping model as a Bayesian prior, and each individual is genotyped separately. Our implementation includes the following numerical stabilizations: (a) when computing the sequencing error rate under the assumption of polymorphism (equation 6 in Maruki & Lynch, 2015) we always assume at least 0.1 error nucleotides have been observed across the population, and (b) the genotype frequencies

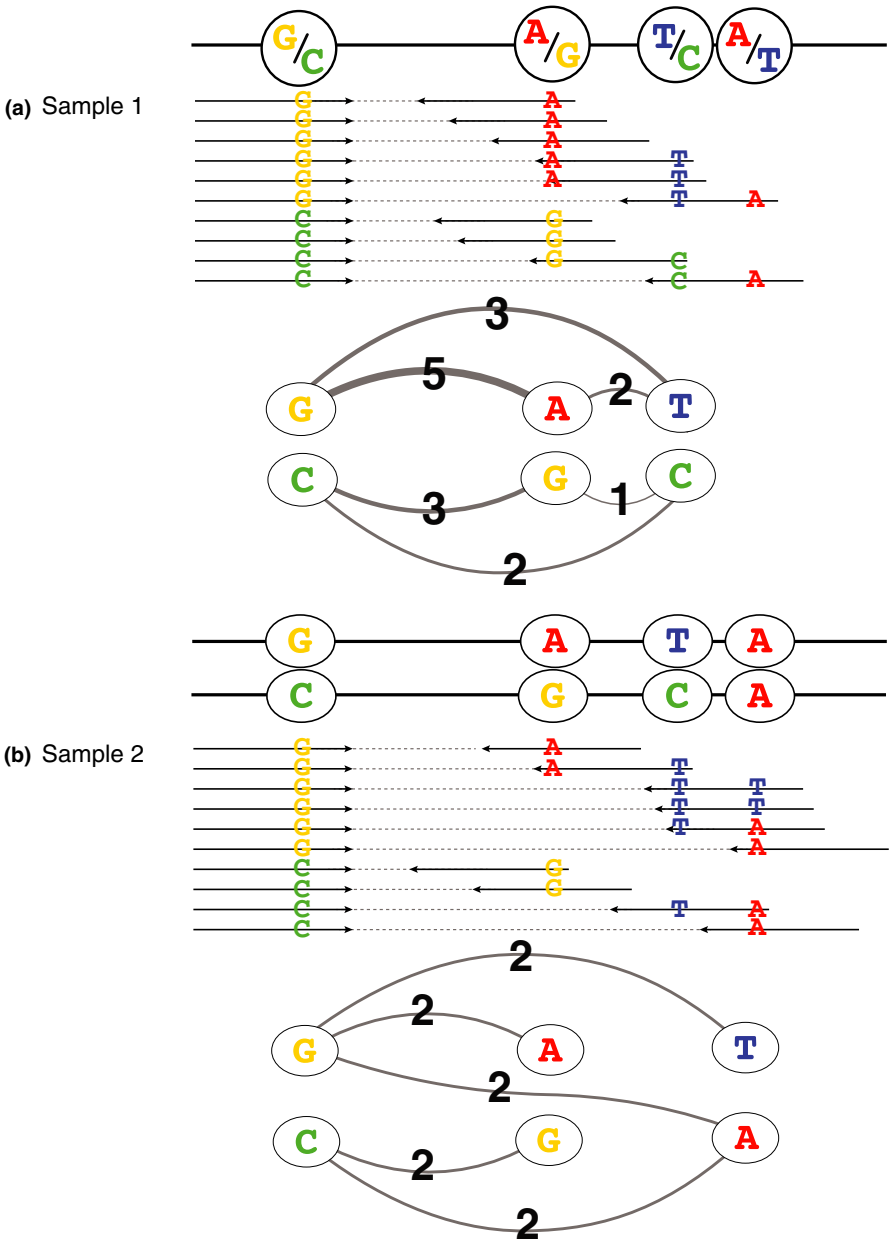

FIGURE 3 Algorithm for haplotyping heterozygotes. (a) Sample 1 is successfully phased at this locus as all alleles are found to co-occur on two or more read pairs producing two distinct subgraphs. (b) Sample 2 cannot be phased as at the fourth variable position the 'A' allele is observed on two different allelic backgrounds, confounding the haplotype graph

used in genotype likelihood computations are rescaled so as to always be greater or equal to 1 over the number of samples. This prevents the error rate and genotype frequencies estimates, respectively, from being zero.

Stacks v2 implements several alternative models to call SNPs and genotypes: BGC, along with the HGC (high-coverage genotype caller; Maruki & Lynch, 2017), and we still provide the method of Hohenlohe et al. (2010).

2.9 | Converting SNPs into phased haplotypes

SNP alleles that are observed on the same read or within the same read pair are part of the same haplotype, because the underlying sequence is a sample from a specific chromosome. We can take advantage of this natural phasing to provide sets of haplotypes from each RAD locus. While Stacks v1 provided short haplotypes, Stacks v2 extends this functionality by phasing heterozygotes using a graph-based algorithm that relies on sequence data, specifically on co-observations of alleles within a read (or read pair).

After genotyping, if a locus includes several SNPs and a (diploid) individual is heterozygous at two or more of them, GSTACKS reconstructs the combination of alleles that corresponds to the individual's two chromosomes (Figure 3). Stacks v2 implements a read-based phasing approach (as opposed to statistical phasing; Browning & Browning, 2011) that relies on the co-observation, in a given read (or read pair), of the alleles at several SNPs. A graph is built in which the nodes are the alternate alleles at heterozygous SNPs, and the edges are the co-observations of these alleles in reads. To limit the influence of sequencing errors, edges representing alleles seen together only once are ignored. If there are

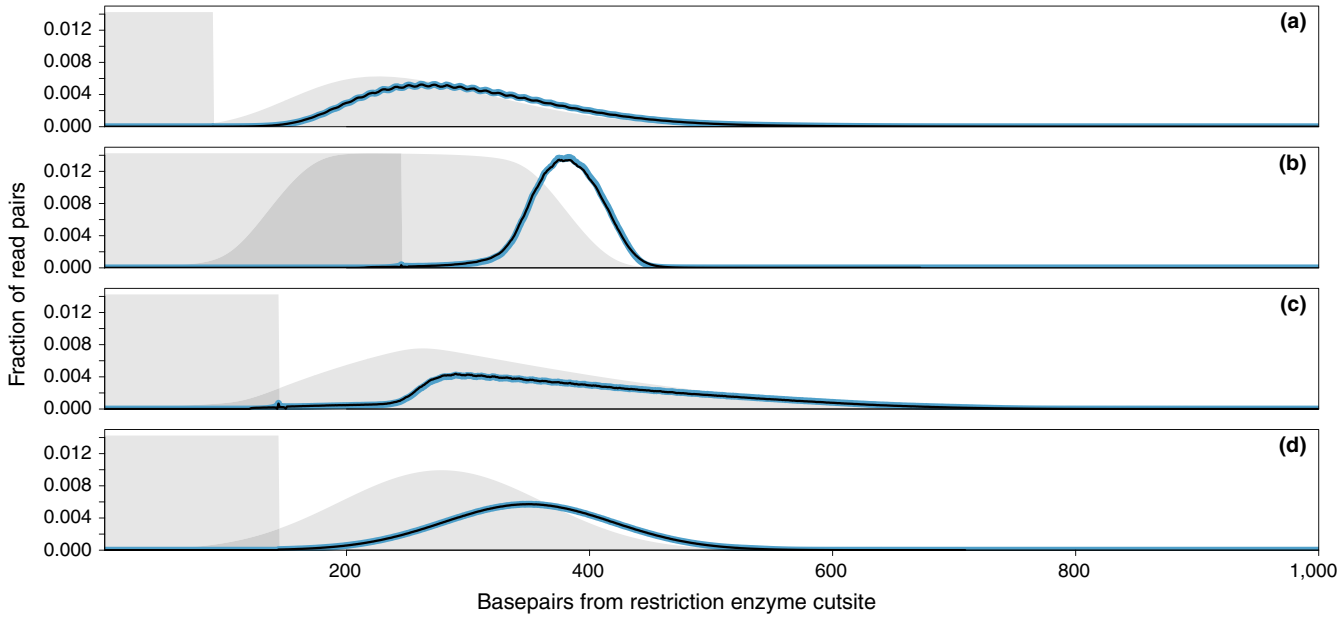

FIGURE 4 DNA library insert size distributions in the benchmark paired-end sdRAD data sets (a–d) as reconstructed by the reference-based (black lines) and de novo (heavy blue lines) approaches. The shaded areas represent the variation of the sequencing coverage along the length of RAD loci, as expected from the insert length distribution after all inserts are stacked by placing the first base of the restriction site at position 1, and accounting for read length in each data set (respectively 100, 250, 150 and 150; see Table 1). The left and right areas correspond to the forward and reverse reads respectively, with the overlap representing base pairs that have been sequenced in both directions. (a) Yellow warbler, (b) threespine stickleback, (c) bald notothen, (d) simulations, 20×. The periodic pattern of insert sizes apparent for the warbler data set, and to some extent for the bald notothen one, seems to be real (Figure S10)

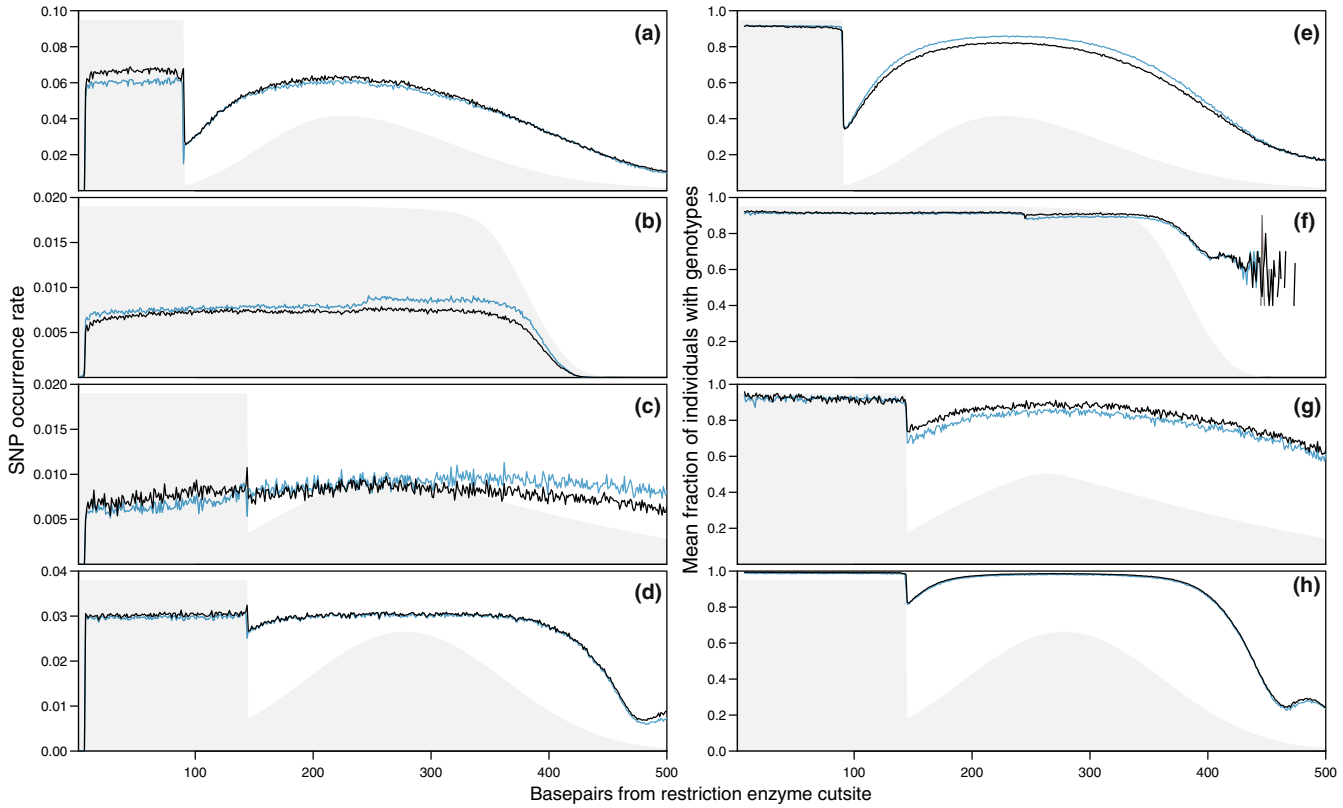

FIGURE 5 Variation of the average SNP (a–d) and genotype (e–h) call rates along the length RAD loci for each of the benchmark paired-end data sets (a, e: yellow warbler; b, f: threespine stickleback; c, g: bald notothen; d, h: simulations, 20×). The shared areas represent the sequencing coverage underlying the calls (see Figure 4). Consistent patterns are observed for the reference-based (black lines) and de novo (blue lines) approaches across the entire length of RAD loci, demonstrating that Stacks v2 appropriately handles the reverse reads for paired-end data from sdRAD libraries. Coverage is the main driver of the variation of call rates, with genotype calls being more sensitive to reduced coverage than SNP calls

TABLE 1 Statistics on inferred loci, single nucleotide polymorphisms and genotypes for all datasets and approaches

	Individuals	Useful read length (forward/reverse)		NS80 loci	Overlapped loci	Locus size	Library insert size (mean ± *SD*)	PCR duplicates	Coverage (read pairs)
Yellow Warbler[c]	241	90/100	Reference-based						
			Stacks	1,17,323	–	–	309 ± 84	34%	16.1×
			ANGSD	–	–	–	–	35%[b]	–
			De novo						
			Stacks	89,016	99.8%	597 ± 59	309 ± 81	33%	18.4×
Threespine Stickleback[c]	10	245/251	Reference-based						
			Stacks	4,17,427	–	–	378 ± 31	41%	7.8×
			ANGSD	–	–	–	–	41%[b]	–
			De novo						
			Stacks	3,60,292	99.9%	433 ± 14	380 ± 30	40%	8.3×
Bald Notothen	71	144/151	Reference-based						
			Stacks	39,620	–	–	387 ± 105	76%	21.1×
			ANGSD	–	–	–	–	76%[b]	–
			De novo						
			Stacks	34,460	99.9%	715 ± 47	392 ± 106	75%	22.8×
			DDOCENT	46,982	21.6%	328 ± 103	–	Not offered	110×
			RADASSEMBLER	34,350[d]	98.6%	676 ± 140	–	–	–
Simulations (20× run #0)	100	144/144	Reference-based						
			Stacks	26,576	–	–	350 ± 69	–	19.4×
			ANGSD	–	–	–	–	–	–
			De novo						
			Stacks	24,540	99.9%	550 ± 17	350 ± 70	–	19.9×
			DDOCENT	25,114	46.0%	344 ± 85	–	–	20.5×
			RADASSEMBLER	17,220[d]	99.9%	562 ± 26	–	–	–
Simulations (10× run #0)	100	144/144	Reference-based						
			Stacks	26,551	–	–	350 ± 69	–	9.7×
			ANGSD	–	–	–	–	–	–
			De novo						
			Stacks	24,262	99.7%	550 ± 18	350 ± 69	–	10.1×
			DDOCENT	25,005	42.1%	337 ± 82	–	–	10.3×
			RADASSEMBLER	17,856[d]	99.9%	561 ± 26	–	–	–

Abbreviations: NS80, loci for which at least 80% of individuals have one or more reads according to the method considered; TNS80, loci for which at least 80% of individuals have one or more BWA-MEM alignments (in contrast with NS80, this set is constant regardless of the analysis method considered); MAC3, SNPs with a minor allele count of three or more; GT80, SNPs with genotype calls in at least 80% of individuals; HAP80, remaining SNPs after filtering so that at least 80% of individuals have a fully resolved haplotype at each locus.

[a]We defined the reverse-read region as positions not covered by forward reads (e.g., for the warbler data set, positions >90).

[b]PCR duplicates were filtered from BWA-MEM alignments using SAMTOOLS.

[c]No data is shown for DDOCENT for the warbler and stickleback data sets because the pipeline did not complete. Similarly, RADASSEMBLER did not complete on the warbler data set and produced unstable results on the stickleback.

[d]Based on BWA read alignments to RADASSEMBLER loci.

no conflicts between reads, the haplotype graph will be composed of two distinct, connected subgraphs and the corresponding haplotypes can be extracted from each subgraph. Otherwise (i.e., if there is substantial conflict between reads) the phasing operation was probably confounded by sequencing errors, contamination, or over-merging. In this case, no phasing is provided, and the individual's SNP alleles are marked as unreliable, as they are probably affected by the same issues.

SNPs									
					GT80 & MAC3				
MAC3 (NS80 loci)	MAC3 (TNS80)	True positives	False positives	Precision	All	Reverse region[a]		HAP80 & MAC3 SNPs	HAP80 & MAC3 haplotype mean length
28,52,729					15,89,453	10,07,025	63%	11,23,081	9.8
–					17,70,794	–		–	–
20,58,649					12,43,449	8,22,289	66%	8,75,155	10.0
11,93,055					10,44,019	3,84,760	37%	10,09,300	3.3
–					12,40,121	–		–	–
9,92,691					8,74,740	3,33,804	38%	8,52,607	3.2
1,77,294					1,18,166	81,792	69%	1,06,836	3.3
–					1,48,536	–	–	–	–
1,68,456					99,942	71,827	72%	87,562	3.1
1,34,919					1,12,576	62,142	55%	–	–
–					–	–		–	–
3,59,977	3,60,080	3,57,858	2,222	0.994	3,17,874	2,07,724	65%	3,03,256	11.4
–	3,52,165	3,50,904	1,261	0.996	2,96,111	–		–	–
3,28,478	3,31,305	3,27,265	4,040	0.988	2,87,921	1,88,943	66%	2,73,890	11.3
2,19,124	2,16,817	2,08,223	8,594	0.960	2,10,309	1,25,994	60%	–	–
–	–	–	–	–	–	–		–	–
3,27,285	3,27,526	3,26,049	1,477	0.996	2,57,074	1,48,261	58%	2,21,151	8.3
–	2,16,817	2,08,223	8,594	0.960	2,10,689	–		–	–
2,96,499	3,01,652	2,97,420	4,232	0.986	2,29,213	1,32,455	58%	1,96,314	8.2
2,12,733	2,10,074	2,01,361	8,713	0.959	1,93,718	1,03,350	53%	–	–
–	–	–	–	–	–	–		–	–

2.10 | GSTACKS output files

The GSTACKS program proceeds through the data set one locus at a time and can be parallelized to run one locus per thread at a time. The program will produce a new catalog contained in two files: the consensus sequences for the catalog loci in a FASTA formatted file, and a custom file containing the SNP/haplotype calls for each locus and all individuals. In addition, GSTACKS provides a number of useful

data distributions, such as coverage, PCR duplicates and phasing rates, in its output.

2.11 | Improvements to the POPULATIONS program

The POPULATIONS program is designed to take the assembled data from GSTACKS and apply a population genetics frame to the data. The program can apply a population map to segment individuals in the metapopulation based on useful criteria (e.g., geography, phenotype, or sex). The program can filter data: keeping polymorphic sites that are found within a specified number of individuals or populations. A new option allows samples to be filtered according to haplotype presence. A user can ask to keep all haplotypes that are found (and complete) in, for example, more than 80% of individuals. Remarkably, even if all SNPs are present in 80% of individuals, it is possible that different SNPs will be missing in different individuals, resulting in <80% of complete haplotypes. In order to provide the longest possible haplotype at the required threshold, the POPULATIONS program will filter individual SNPs by their availability in the population until this criterion is met. This filtering will reduce the length of the haplotype at a particular locus, but ensure there is no missing data within the haplotypes.

After filtering, the POPULATIONS program will calculate a number of population genetic statistics, for both SNPs and haplotypes. These include π, heterozygosity, F_{IS} computed per-population and per-SNP, as well as F_{ST} computed for each pair of populations using an Analysis of Molecular Variance (AMOVA) approach (Excoffier, Smouse, & Quattro, 1992; Weir, 1996; Weir & Cockerham, 1984). For haplotypes, several versions of F_{ST} are calculated, including φ_{ST} (Bird, Karl, Mouse, & Toonen, 2011; Holsinger & Weir, 2009) and F'_{ST} (Meirmans, 2006), and we provide a haplotype-level calculation of D_{XY} (Cruickshank & Hahn, 2014; Nei, 1987). The POPULATIONS program also provides a measure of Hardy-Weinberg equilibrium for each SNP (Engels, 2009; Louis & Dempster, 1987) and each locus (Guo & Thompson, 1992). The two-allele SNP approach uses an exact test that is analogous to Fisher's exact test. For haplotypes, where multiple alleles can make an exact test computationally challenging, we employ a Markov chain approximation as described by Guo and Thompson (1992).

The POPULATIONS program is able to export data, after filtering, in a number of useful formats from VCF to FASTA, and for specialized programs such as STRUCTURE, GENEPOP, FINERADSTRUCTURE or PLINK.

For details on the real and simulated data sets examined, see File S1.

3 | RESULTS

3.1 | Calling SNPs, genotypes and haplotypes from paired-end RADseq data

We first surveyed the general properties of paired-end data derived from single-digest, shearing-based RADseq protocols. Using a reference genome approach, we reanalyzed three published data sets, including the Alaskan threespine stickleback (sdRAD, PstI enzyme, 2 × 250 bp reads; Nelson & Cresko, 2018), the North American yellow warbler (BestRAD, SbfI enzyme, 2 × 100 bp reads; Bay et al., 2018), plus one newly generated data set for a teleost fish, the Antarctic bald notothen (*Pagothenia borchgrevinki;* sdRAD, SbfI enzyme, 2 × 150 bp reads).

We aligned the paired reads of all individuals to their respective reference genomes using BWA-MEM (Li, 2013), then used Stacks to identify RAD loci, remove PCR duplicates and compute the distribution of insert lengths (Figure 4). We found the median insert sizes for warbler to be 309 bp, for stickleback to be 380 bp, and for *P. borchgrevinki* to be 387 bp. These distributions reflect the DNA shearing and size-selection steps of the library preparation. Accordingly, the distribution reconstructed for the *P. borchgrevinki* library matches the one estimated in vitro with a fragment analyzer (Figure S2).

Importantly, for paired-end RADseq data, the distribution of insert sizes has direct implications for the distribution of sequencing coverage within each locus. Because inserts are anchored on one side by the restriction site, sequencing both ends of each insert leads to the coverage patterns shown in Figure 4 and Figure S3: coverage on the restriction site end is constant over one read length, whereas coverage on the sheared end has a trapezium-like distribution (which appears bell-like if the width of the insert size distribution is larger than the read length). Figure 4 makes clear that three different size selection strategies were taken with the construction of the different libraries. While the stickleback library focused on a narrow size range while employing longer reads, the *P. borchgrevinki* library focused on a wider insert length, illustrating the trade-off between uniform coverage and longer contigs.

In turn, coverage affects polymorphism discovery and genotype calling over the length of each locus. SNP discovery remains possible even at low per-sample coverage (Figure 5a–c) because in Stacks v2 the existence of a polymorphism is a test on the population: evidence for alleles can be aggregated across individuals, so alternate alleles are visible as long as their total coverage in the population is substantial, independently of each individual's coverage (Maruki & Lynch, 2015). Genotype calls, in contrast, fundamentally relate to single individuals, and are therefore much more affected by coverage variations (Figure 5e–g), with calls becoming increasingly unreliable at depths under 7 (or 5–10 depending on the desired stringency level; Figure S4). For the stickleback, 1.04 million SNPs were found that could be genotyped in ≥80% of individuals (GT80), with 37% found in the reverse-read region (Table 1). In warbler, 1.59 million SNPs in the same class were found (63% in the reverse region), and in *P. borchgrevinki*, 0.12 million SNPs were likewise found, with 69% in the reverse region. The number of SNPs in the three data sets reflect the respective numbers of loci and the natural polymorphism rate, while their distribution across the forward and reverse regions reflect the insert lengths in the underlying RAD libraries. In the case of the bald notothen, there are relatively few RAD loci, but two-thirds of the SNPs were found with the addition of the reverse contig (in contrast, since the insert size distribution was wide, coverage was lower in the reverse region and genotype quality was reduced). These results demonstrate that paired-end sequencing very effectively increases the number of SNPs and genotypes generated by RADseq protocols.

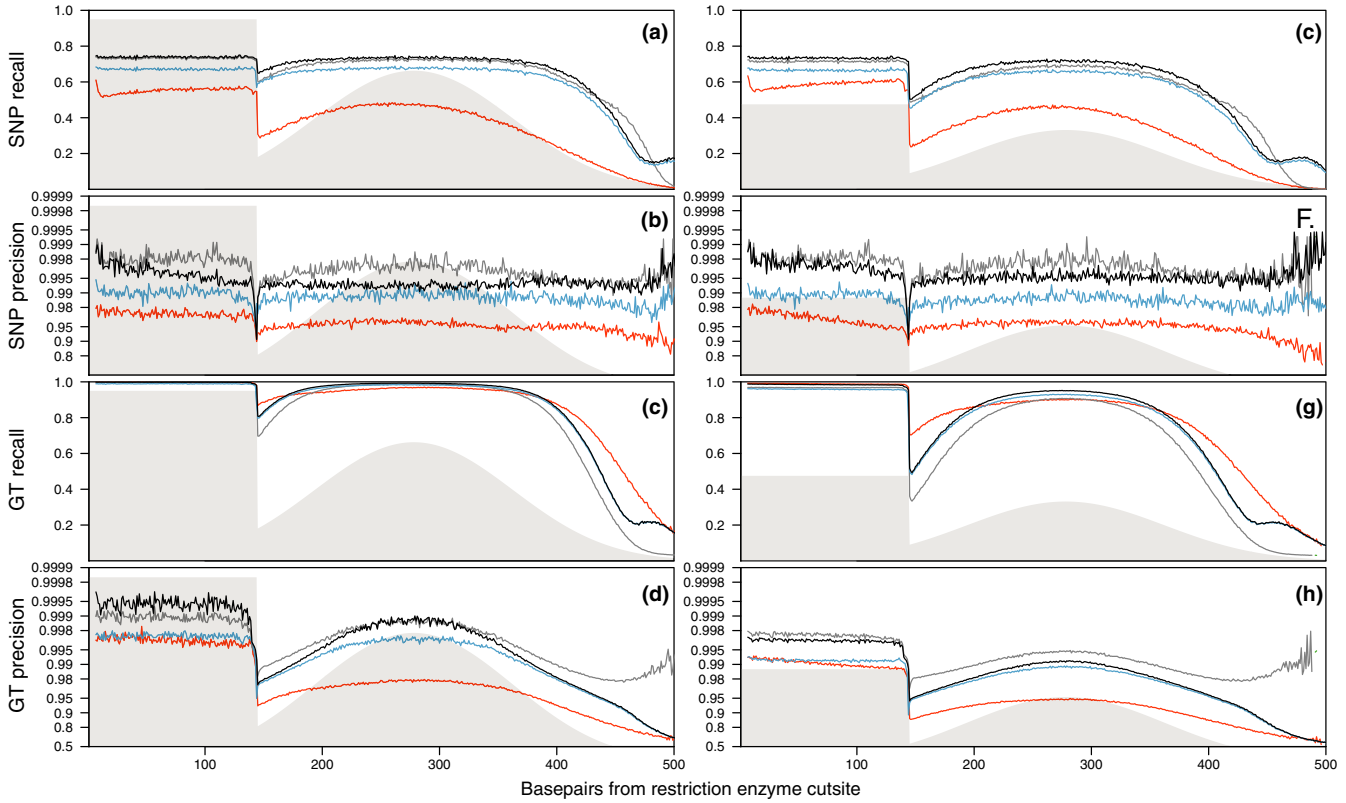

FIGURE 6 SNP and genotype (GT) precision and recall observed across RAD loci for the simulated data sets with 20× (left) and 10× (right) coverage for Stacks v2 using a reference-based approach (black), Stacks v2 using a de novo approach (blue), ANGSD (grey), and DDOCENT (orange)

3.2 | Stacks v2 accurately calls and phases genotypes in a reference-based context

The number of SNP and genotype calls may be inflated by erroneous calls, and thus cannot in itself be regarded as an indicator of the quality of these inferences. In order to assess the quality of the Stacks v2 SNP and genotype calls and their spatial variations within loci, we created simulation data sets with the RADINITIO package (Rivera-Colón, Rochette, & Catchen 2019), using the threespine stickleback reference genome as template (see File S1). We simulated two data sets, one high coverage (20×) and one low coverage (10×), each comprising 100 individuals split among four populations. Both data sets were based on the same population genetics parameters and had a nucleotide diversity (π) of 0.8%. For simulated data sets, the true loci, read alignments, positions of SNPs and genotypes are known, thus we could calculate the recall and precision of the SNPs and genotypes inferred by Stacks and a range of alternative methods (Figures 5d,h and 6).

We filtered results for a minor allele count (MAC) of ≥3, because that ensures an allele is seen in at least two diploid samples. Indeed, SNPs for which the minor allele has a low prevalence (expressed as MAC or equivalently as minor allele frequency, MAF) are more likely to be missed, or to be errors. Especially, precision and recall are markedly lower for SNPs with MAC = 1, even when using a reference-based approach on well-behaving simulated data sets (Figure S5).

The precision and recall of SNP and genotype calls vary across the length of RAD loci, in tight correlation with coverage, as is shown by our simulations at 20× coverage (Figure 6a–d) versus 10× (Figure 6e–h). As noted above, the relationship is more apparent with genotypes than with SNPs. SNP precision in the 10× data set (Figure 6f, black) was as high as in the 20× one (Figure 6b, black), and SNP recall was consistent across the locus. Unsurprisingly, genotype recall and precision were lower in the 10× data set (Figure 6h vs. Figure 6d, black), and deteriorated quickly in regions that are more rarely covered by reverse reads as a result of the insert size distribution.

Next, we compared the inferences of Stacks v2, which implements the BGC model (Maruki & Lynch, 2017), with those of ANGSD (Korneliussen, Albrechtsen, & Nielsen, 2014), using the same BWA-MEM read alignments for both methods. ANGSD is a package for SNP and genotype calling commonly used for whole genome data. ANGSD emphasizes the use of genotype likelihoods and probabilities rather than of explicit genotype calls, and is regarded as more accurate than GATK (Poplin et al., 2018) for the latter (Korneliussen et al., 2014; Maruki & Lynch, 2017). For SNP discovery, we found that Stacks had higher recall (Figure 6a,e; black vs. grey), but fractionally lower precision (Figure 6b,f; black vs. grey). For calling genotypes the two methods use different statistical approaches: ANGSD uses a posterior probability cutoff, whereas Stacks uses a *p*-value cutoff. As there is no natural equivalence

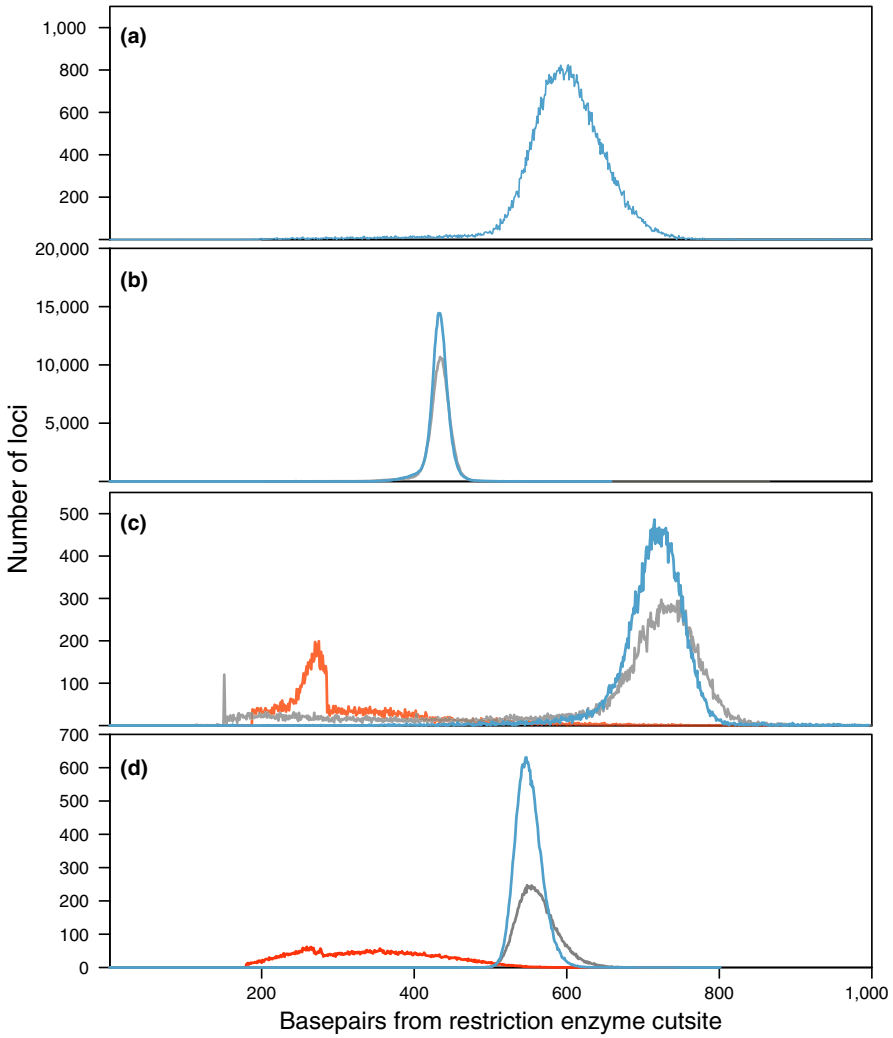

FIGURE 7 Observed lengths of the RAD loci reconstructed using de novo approaches. by Stacks v2 (blue), RADASSEMBLER (grey) and DDOCENT (orange) for the (a) yellow warbler, (b) threespine stickleback, (c) bald notothen, and (d) simulated data sets. Reconstructed locus sizes primarily depend on the shape of the distal tail of the insert size distribution (Figure 4). The number of loci represented by each curve includes only loci for which the forward and reverse regions could be combined in to a single contig. No data is shown for DDOCENT in (a) and (b) because the program did not allow for the analysis of these large data sets (see Section 2)

between the values for these thresholds, we chose the ANGSD cut-off value $p \geq 95\%$ so that the recall was approximately equal to that of Stacks. As a result, the recall was similar for both methods. Recall values for genotypes are higher than for SNPs because it makes more sense to measure genotype recall only for the subset of SNPs that were already correctly identified. For genotype precision, the comparison between the two methods depended on coverage. Stacks performed better than ANGSD at higher coverages, and worse at lower coverages (Figure 6d,h; black vs. grey).

3.3 | Stacks v2 can efficiently cluster RADseq loci, assemble small locus contigs from paired-end reads and map these reads back to loci

To evaluate the ability of Stacks v2 to process sheared, paired-end RADseq data sets when a genome is not available, we performed reanalyses of the data sets introduced above (threespine stickleback, yellow warbler, and bald notothen) using a de novo approach (see File S1). We reconstructed respectively 360,292, 89,016, and 34,460 loci present in ≥80% of individuals (NS80 loci, Table 1). Loci present in more than one or two individuals were generally present in most individuals (Figure S6).

Our de Bruijn graph method was able to assemble a consensus contig or a two-contig scaffold for 97.9%, 91.1% and 84.2% of these loci. Although GSTACKS can resolve simple microsatellite repeats in RAD loci, failure to assemble a consensus contig was frequently due to other repeated elements. The contig lengths (mean ± standard deviation) were 433 ± 14, 597 ± 59 and 715 ± 47 basepairs (Figure 7, Table 1).

The ability to provide a single, overlapped contig for the entire locus, rather than one contig for the restriction site end and another for the sheared end, depends primarily on the read length and distribution of insert sizes. Indeed, for the threespine stickleback, *P. borchgrevinki*, and yellow warbler data sets, which used 250 bp-, 150 bp-, and 100 bp-long reads and insert sizes of 380 bp, 309 bp, and 392 bp, respectively (see above), virtually all NS80 loci (>99%) were overlapped.

Once a consensus sequence has been determined for each locus, reads can be mapped back to these loci by our suffix tree-based alignment algorithm. Our method was able to map 96.1%, 99.3%, and 98.5% of the reads for the threespine stickleback, yellow warbler, and bald notothen data sets, respectively (for reads in NS80 loci). We validated our internal read alignment method by comparing the SNPs and genotypes they yielded to those obtained by extracting the consensus

locus sequences, aligning reads to them using BWA-MEM, and applying the rest of the Stacks v2 pipeline. We found that the results derived from the internal alignments and those derived from the BWA-MEM alignments were extremely close in their general statistics, such as the depth of coverage patterns or the number of SNPs and genotypes eventually called (Figure S7).

3.4 | The de novo approach performs comparably to the reference-based approach

We evaluated the global efficiency of the de novo approach for analyzing paired-end RADseq data by comparing its results with those of the reference-based approach. We analyzed threespine stickleback, yellow warbler, and bald notothen to gain a perspective under real conditions, and analyzed the high- and low-coverage simulated data sets for a performance comparison in absolute terms, since we knew the true underlying states.

For the three real and the two simulated data sets, the de novo analysis recovered on average 83% of the same NS80 loci found under the reference-based analysis (respectively 83%, 75%, and 87% for stickleback, warbler, and *P. borchgrevinki*, Table 1). These NS80 loci had comparable coverage, although loci from the de novo analysis comprised on average 110% as many aligned reads per individual as the reference-based one (respectively 8.3×, 18.4×, and 22.8× read coverage per locus per individual for stickleback, warbler, and *P. borchgrevinki*, Table 1). For all data sets, the observed insert size distributions were identical for the two approaches (Table 1).

Both approaches found similar densities of SNPs with similar spatial arrangements (Figure 5). Indeed, the de novo approach identified on average 83% of the same MAC ≥ 3 SNPs per locus (Table 1) that were found using the reference-based approach. The de novo approach found fewer SNPs, particularly in the forward read region, probably because low-coverage alleles may be classified as errors and missed during locus clustering (see Catchen et al., 2011 for details). Furthermore, the de novo approach called slightly fewer genotypes than the reference-based one (on average 91% as many at 20× coverage, 89% at 10×; Table 1). The reduction was due to a lower call rate in the distal region, despite coverage being equal with both approaches (Figure 5). Such a reduction in measured recall at the 3' end of the locus could be due to some whole RAD loci being discarded by the de novo algorithm.

In terms of absolute error rates in the simulated data sets, we found that the de novo approach was slightly less precise than the reference-based approach, but that the error rates of the two approaches were nevertheless comparable in magnitude, and low in both approaches. In the reference-based analyses, more than 99% of inferred MAC3 SNPs were true positives, while more than 98% were true positives in de novo analyses (Table 1). Importantly, precision was consistent across the length of loci, apart from the expected strong correlation between genotype precision and coverage (Figure 6). The reduction in precision between reference and de novo data is probably due to additional error generated in de novo assembly from very low coverage alleles being discarded and a small amount of locus over-merging both occurring in USTACKS.

3.5 | Stacks v2s methods for paired-end RADseq data performance compared to alternative packages

Next, we compared the results of Stacks on sheared, paired-end RADseq data with those obtained with alternate analysis packages, namely RADASSEMBLER (version 1.01; Li, Xue, Zhang, & Liu, 2018) and DDOCENT (version 2.2.20; Puritz, Hollenbeck, & Gold, 2014). Both packages differ from Stacks in their approach. RADASSEMBLER first clusters forward reads into loci using Stacks v1, then builds paired-end contigs with the CAP3 overlap-layout method (Huang & Madan, 1999). RADASSEMBLER focuses only on locus construction and does not identify SNPs or call genotypes. DDOCENT relies on CD-HIT (Li & Godzik, 2006), RAINBOW (Chong, Ruan, J., & Wu, 2012), and PEAR (Zhang, Kobert, Flouri, & Stamatakis, 2014) for locus and contig assembly and it applies the FREEBAYES (Garrison & Marth, 2012) software for SNP identification and genotyping.

In the simulated data sets, the number of NS80 loci RADASSEMBLER constructed was approximately 70% of the number of loci constructed in the same de novo Stacks analysis. For those loci constructed, the size and number that could be overlapped with the forward locus was very similar to Stacks. The DDOCENT software constructed ~105% of the loci that de novo Stacks did, however, those loci were 70% as long while only 40%–45% could be overlapped indicating assembly errors (Figure 7 and Table 1). In the empirical *P. borchgrevinki* data set, RADASSEMBLER performed similarly to Stacks overall, although the presence of a tail of markedly shorter contigs (Figure 7) suggests that a small subset of loci were not properly reconstructed. DDOCENT produced 136% as many loci, with a mean length only 46% as long as Stacks and the forward and reverse regions overlapped in only 21% of loci (Table 1). DDOCENT was not able to complete processing the other two empirical data sets.

3.6 | Stacks v2 is able to provide consistent and rich haplotypes

Producing useable haplotype markers involves two main processes. First, a consistent phasing of SNPs at each locus must be found in each individual. When considering a locus at the population level, however, the haplotype graphs between individuals may be different due to the presence or absence of particular SNPs in particular individuals (see Section 2). Stacks v2 is able to produce consistent, population-wide haplotypes that incorporate 95% of independently genotyped SNPs in the 20× simulated data set, and 86% given 10× sequencing. For the empirical data sets, the value ranged from 71%–97%. The haplotypes built were nearly 10 bp in mean length in the yellow warbler, vs. approximately 3 bp in stickleback and the bald notothen, suggesting the warbler genome may be more polymorphic, and also demonstrating the power of detecting rarer polymorphisms with deeper population sampling.

4 | DISCUSSION

4.1 | Long RADseq contigs

The appeal of using paired-end sequencing to derive short contigs from de novo sdRAD data has been promoted repeatedly (Amores, Catchen, Ferrara, Fontenot, & Postlethwait, 2011; Andrews et al., 2016; Etter, Preston, Bassham, Cresko, & Johnson, 2011; Hohenlohe et al., 2013; Li et al., 2018; Nelson & Cresko, 2018). However, the approach has remained elusive due to the technical difficulty, unreliability and inefficiency of generating these contigs in the absence of any software capable of performing this task natively. Thus, Etter et al. (2011) developed an approach combining Stacks and Velvet (Zerbino & Birney, 2008), while Hohenlohe et al. (2013) experimented with both Stacks and Velvet or Stacks and CAP3 (Huang & Madan, 1999), opting for the latter, and Nelson and Cresko (2018) used Stacks and Fastq-Join (Aronesty, 2011) together with a longer-read, high-overlap sequencing strategy. Finally, the recently published RADASSEMBLER method (Li et al., 2018) wraps around Stacks v1 and CAP3 in a similar manner to Hohenlohe et al. (2013). Each of these approaches applied one algorithmic strategy to the single-end reads, and a second, orthogonal strategy to the paired-end reads, partitioning the available information, which limited the accuracy of the results and the speed of the analysis.

The series of methodological developments that we introduce in Stacks v2 makes the analysis of paired-end sdRAD data efficient, reliable, and accessible to the majority of RADseq users. It yields more robust results and is considerably faster (Table S1) and easier to apply in comparison with previous approaches. Importantly, the length of the assembled paired-end contigs and the merging rate of the forward and reverse regions of the locus primarily depend on the distributions of insert sizes in the sequenced DNA library (see below). However, the average contig length can be expected to be in the 400–800 bp range, and we find that virtually all loci have a single contiguous contig provided that at least a small fraction of reads overlap.

In the de novo context, the availability of RAD locus contigs offers several clear advantages over the shorter loci obtained from single-end data, which are typically only as long as the individual reads. These loci can be more easily mapped to existing genomic resources for the purposes of providing functional annotation, conducting linkage-based genomic scans (Amores et al., 2011; Feulner & Seehausen, 2019), or designing capture baits (Ali et al., 2016). They allow one to make use of the paired-end reads derived from the sdRAD (Baird et al., 2008) and BestRAD (Ali et al., 2016) protocols, even when a de novo strategy is employed. As these protocols involve a random shearing step, this implies that PCR duplicates can be filtered natively based on the start position of the reverse read in the same way as can be done when paired alignments have been made to a reference genome (DePristo et al., 2011).

Finally, in addition to the qualitative advantages presented above, using paired-end reads increases the amount of genotype data that is produced. Although paired-end data do not change the number of RAD loci, they can more than double the average number of SNPs per locus for suitable library preparation and sequencing parameters (Table 1 and see below). On average, this results in more polymorphic loci, that are each more informative. In a de novo context, these results will benefit any study focused on species with low diversity, and in particular phylogenetics studies, where sparse genotype matrices are being constructed across many species (Near et al., 2018). When RAD loci are ordered onto chromosomes (i.e., using a native reference genome, or if de novo-assembled loci have been mapped to an external reference genome), a higher information density along the genome results, which should help resolve linkage patterns at a finer scale and identify evolutionary events that may have been missed with a sparser sampling of the genome (McKinney et al., 2017).

4.2 | Haplotypes

Expecting multiple SNPs at each locus opens the possibility of treating RAD loci as a set of haplotypes rather than as individual SNPs. Depending on the nucleotide diversity of the species, it not rare to see up to ten SNPs per locus in real data sets (Table 1). Since SNPs generally have two alleles, a given SNP only delineates two sets of individuals in the population. Within a small genomic region, choosing one SNP over another may result in a different evolutionary signal. By phasing these SNPs, Stacks v2 can instead provide multiallelic haplotypes which can encode a much larger amount of information regarding the provenance of the genomic region. These rich, physically phased haplotypes produced by Stacks v2 contain information on coalescence patterns that can be used to study population structure (Malinsky, Trucchi, Lawson, & Falush, 2018), demography (Trucchi et al., 2014), or to single out genomic islands of differentiation (Nelson & Cresko, 2018).

The haplotyping algorithm implemented in Stacks v2 relies strictly on the co-observation of alleles within a read (or read pair). This is in contrast with statistical phasing methods in which an individual's haplotypes are estimated in relation to a panel of haplotypes observed at the population level (Browning & Browning, 2011). Furthermore, our approach expects allele co-observations to appear consistent at a specified tolerance level (see Section 2). Other algorithms for read-based phasing are instead designed to find the haplotypes that are most consistent with the data, and subsequently accept them as the real ones (Patterson et al., 2015). While using a tolerance threshold makes our approach more sensitive to insufficient coverage, it allows one to identify and remove cases where the allelic observations for an individual at a particular locus are fundamentally at odds with diploidy. Such cases point to miscalled genotypes that can then be pruned out, or occasionally to contamination in the sequencing library (that makes an individual effectively non-diploid) as evidenced by the dramatically reduced phasing rates that may be observed for specific individuals (Figure S8). Over-merged loci, that collapse several paralogous genomic loci into a single polyploid one, also exhibit high phasing failure rates and can be filtered on this basis.

4.3 | Stacks v2 introduces improvements throughout the pipeline

Stacks v2 supports nearly all major protocols for reduced representation, marker-based experiments, including single- and double-digest RAD and DaRT (Kilian et al., 2012), using single- or paired-end sequencing, as well as 2bRAD and GBS using single-end sequencing. We confirm that it is not suitable for paired-end data derived from GBS, as forward and reverse reads contain essentially the same information. Applying our approach to paired-end GBS data would lead to assembling two half-coverage copies of each locus (one for each possible orientation).

One of the major advantages to the design of Stacks v2 is the vertically integrated nature of the software system which ensures that each stage of the analysis has access to all of the information collected. Most importantly, Stacks can take advantage of certain types of information that RADseq data provide, for example, the de Bruijn graph assembler knows that all reads in a RADseq analysis are sequenced in the same direction (and hence the de Bruijn graph only has to account for one strand). This is in contrast to the majority of alternative software pipelines available that incorporate pre-existing software as black box components (that is, how the software operates is opaque to the executing pipeline), and where such software was not designed for RAD data. In several cases, we could not run competing software (DDOCENT and RADASSEMBLER) because certain subcomponents failed on our data sets.

4.4 | Efficiently applying paired-end RADseq

Due to the particular nature of RADseq libraries, the distribution of insert sizes has significant consequences on coverage patterns in the reverse-read region (distal relative to the restriction site) of RAD loci, on whether (or how often) the forward and reverse regions can be overlapped, and on the maximal length of loci. The length of a reconstructed locus is primarily determined by the size of the longest inserts. Conversely, the overlapping of the forward and reverse regions depends on the existence of short inserts. However, a more spread out distribution of insert sizes (such as one having both short and long inserts) implies that reverse read positions, and in turn coverage in the reverse-read region, will themselves be more spread out. Consequently, the average coverage in the reverse-read region will be reduced relative to coverage in the forward-read region, which is constant (see Figure 4 and Figure S3). This is generally undesirable, and most RADseq users should find having smaller loci (e.g., 400–500 bp) and consistent coverage more beneficial than having longer loci (e.g., 700–800 bp) but unbalanced coverage.

For consistent coverage in the forward and reverse regions, the width of the size selection window during library construction should be <1 read length. In addition, overlapping of the forward and reverse regions into a single contig is only possible when there is overlapping within at least some read pairs (i.e., some inserts are shorter than two read lengths). In practice, high overlapping rates will be observed if at least 5% of read pairs overlap. Thus, we recommended a size selection window ranging from two to three read lengths, plus the length of the adapters. For 150 bp reads, an optimal insert length distribution could be obtained by size selecting for 300–400 bp inserts (that is, for 430–530 bp molecules if using 130 bp adapters).

4.5 | RADseq PCR duplicate rates must be monitored

Nearly all RAD protocols use PCR amplification as a method to enrich DNA adjacent to restriction enzyme cut sites. PCR amplification can create an illusion as to how much genetic information is available in a RAD library, particularly if the starting amount of template DNA was very small, and can result in bias when some alleles or loci are not sampled.

In our analysis, substantial PCR duplicate rates were observed in all data sets, ranging from 33% (yellow warbler) to 76% (bald notothen, Table 1). We do not doubt these figures as insert sizes were reconstructed robustly (see Section 3), and we could confirm them with alternative software in the case of reference-based analyses. The rates differed somewhat between individuals within data sets, but most of the differences occurred between data sets or possibly between libraries (Figure S9). The BestRAD (Ali et al., 2016) protocol seemed to yield fewer PCR duplicates (33%) than the original Baird et al. (2008) protocol (41%–76%). It is thus likely that the use of biotinylated adapters and bead capture in BestRAD improves restriction-site associated DNA recovery and library quality.

For ddRAD protocols, it is not usually possible to identify PCR duplicates, but methods based on degenerate adapters have been developed. Schweyen, Rozenberg, and Leese (2014) have reported PCR duplicate rates of 20%–45%, and Tin, Rheindt, Cross, and Mikheyev (2015) have reported rates of 30%–80%. While these figures represent varying systems, preservation states and protocols, they suggest that PCR duplicates are as prevalent in ddRAD as in shearing-based RAD protocols.

PCR duplicates can interfere with genotype calls, in particular by distorting the relative abundances of alleles in a heterozygote, thereby inflating the number of apparently homozygous samples, and possibly by introducing clonal errors that may artefactually appear as alleles and thus can bias downstream analyses. Furthermore, the PCR duplicate rate is often assumed to correlate with the quality of the starting DNA (Casbon et al., 2011). We note, however, that little effort has been made to accurately measure the impact of PCR duplicates. Such work would be useful both to confirm the theoretical argument that PCR duplicates alter genotype calls, and to estimate the extent of the bias that approaches that do not allow PCR duplicate removal putatively suffer from.

Nevertheless, our data imply that the rate of PCR duplicates coupled with sequencing coverage is the best measure of library quality and best predictor of analytical results. We therefore strongly encourage RADseq users to use experimental designs that allow them to characterize the rate of PCR duplicates, such as shearing-based

protocols coupled with paired-end sequencing (sdRAD; Baird et al., 2008; or BestRAD; Ali et al., 2016) or ddRAD protocols with random oligos, such as 3RAD (Graham et al., 2015), over protocols that mask the issue altogether. Users should also be careful to account for potentially significant PCR duplicate rates when planning the amount of sequencing necessary to meet the coverage target.

In conclusion, the family of RAD protocols developed over the past decade, coupled with commodity-priced, massively parallel, short-read sequencing, has found a valuable niche for conducting large population genomic and phylogenomic studies. We can optimize the accuracy and volume of data available to researchers employing reduced-representation strategies by providing a set of accurate, fast and accessible software. Stacks v2 provides the analytical tools to enhance RADseq when it is coupled with paired-end sequencing, assembling tens of thousands of loci that are 400–800 bp in length and can contain up to a dozen physically-phased SNPs on each locus. The software can scale to data sets with thousands of individual samples and we have shown here the effectiveness of the algorithms to assemble loci and genotype those individuals. By focusing software changes to provide robust genetic information in the current sequencing landscape we can fortify the utility of RAD sequencing for another decade.

ACKNOWLEDGEMENTS

The authors would like to thank Thom Nelson, Bill Cresko, and Susan Bassham for useful discussions and Rachel Bay for help accessing the yellow warbler data. We would also like to thank the users of Stacks for all of their input, bug reports, and early testing of the software. A.G.R.-C. and N.C.R. were supported by NSF grant 1645087.

AUTHOR CONTRIBUTIONS

J.M.C., and N.C.R. designed and implemented Stacks v2. J.M.C., N.C.R., and A.G.R.-C. designed and implemented the empirical experiments. J.M.C., N.C.R., and A.G.R.-C. designed the simulation experiments, N.C.R., and A.G.R.-C. implemented the simulation system. J.M.C., N.C.R., and A.G.R.-C. wrote the paper.

DATA AVAILABILITY STATEMENT

Stacks is released under the free software, GPLv3 license. Source code can be downloaded from the Catchen Lab website and is made available in a Bitbucket repository (https://bitbucket.org/jcatchen/stacks/). The threespine stickleback and yellow warbler data are publicly available; bald notothen data are in preparation for publication and are currently available by request from the author.

ORCID

Nicolas C. Rochette https://orcid.org/0000-0003-1899-1765

Angel G. Rivera-Colón https://orcid.org/0000-0001-9097-3241

Julian M. Catchen https://orcid.org/0000-0002-4798-660X

REFERENCES

Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). RAD capture (Rapture): Flexible and efficient sequence-based genotyping. *Genetics*, *202*(2), 389–400. https://doi.org/10.1534/genetics.115.183665

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*(4), 799–808. https://doi.org/10.1534/genetics.111.127324

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*(2), 81–92. https://doi.org/10.1038/nrg.2015.28

Aronesty, E. (2011). *ea-utils: Command-line tools for processing biological sequencing data*. Retrieved from https://github.com/ExpressionAnalysis/ea-utils

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Bassham, S., Catchen, J., Lescak, E., von Hippel, F. A., & Cresko, W. A. (2018). Repeated selection of alternatively adapted haplotypes creates sweeping genomic remodeling in stickleback. *Genetics*, *209*(3), 921–939. https://doi.org/10.1534/genetics.117.300610

Bay, R. A., Harrigan, R. J., Underwood, V. L., Gibbs, H. L., Smith, T. B., & Ruegg, K. (2018). Genomic signals of selection predict climate-driven population declines in a migratory bird. *Science*, *359*(6371), 83–86. https://doi.org/10.1126/science.aan4380

Bird, C. E., Karl, S. A., Mouse, P. E., & Toonen, R. J. (2011).Detecting and measuring genetic differentiation. In C. Held, S. Koenemann, & C. D. Schubart (Eds.), *Phylogeography and population genetics in Crustacea* (pp. 31–55). Boca Raton, FL: CRC Press.

Blears, M. J., De Grandis, S. A., Lee, H., & Trevors, J. T. (1998). Amplified fragment length polymorphism (AFLP): A review of the procedure and its applications. *Journal of Industrial Microbiology and Biotechnology*, *21*(3), 99–114. https://doi.org/10.1038/sj.jim.2900537

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, *32*(3), 314–331.

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: Existing methods and new developments. *Nature Reviews Genetics*, *12*(10), 703–714. https://doi.org/10.1038/nrg3054

Casbon, J. A., Osborne, R. J., Brenner, S., & Lichtenstein, C. P. (2011). A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Research*, *39*(12), e81. https://doi.org/10.1093/nar/gkr217

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: Building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda)*, *1*(3), 171–182. https://doi.org/10.1534/g3.111.000240

Catchen, J. M., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural

populations. *Molecular Ecology Resources*, *17*(3), 362–365. https://doi.org/10.1111/1755-0998.12669

Chong, Z., Ruan, J., & Wu, C.-I. (2012). RAINBOW: An integrated tool for efficient clustering and assembling RAD-seq reads. *Bioinformatics*, *28*(21), 2732–2737. https://doi.org/10.1093/bioinformatics/bts482

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, *23*(13), 3133–3157. https://doi.org/10.1111/mec.12796

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*(7), 499–510. https://doi.org/10.1038/nrg3012

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Dierickx, E. G., Shultz, A. J., Sato, F., Hiraoka, T., & Edwards, S. V. (2015). Morphological and genomic comparisons of Hawaiian and Japanese Black-footed Albatrosses (*Phoebastria nigripes*) using double digest RADseq: Implications for conservation. *Evolutionary Applications*, *8*(7), 662–678. https://doi.org/10.1111/eva.12274

Dudaniec, R. Y., Yong, C. J., Lancaster, L. T., Svensson, E. I., & Hansson, B. (2018). Signatures of local adaptation along environmental gradients in a range-expanding damselfly (*Ischnura elegans*). *Molecular Ecology*, *27*(11), 2576–2593. https://doi.org/10.1111/mec.14709

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

Engels, W. R. (2009). Exact tests for Hardy-Weinberg proportions. *Genetics*, *183*(4), 1431–1441. https://doi.org/10.1534/genetics.109.108977

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*(4), e18561. https://doi.org/10.1371/journal.pone.0018561

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics*, *131*(2), 479–491.

Feulner, P. G. D., & Seehausen, O. (2019). Genomic insights into the vulnerability of sympatric whitefish species flocks. *Molecular Ecology*, *28*(3), 615–629. https://doi.org/10.1111/mec.14977

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ArXiv*:1207.3907 [q-Bio]. Retrieved from http://arxiv.org/abs/1207.3907

Graham, C. F., Glenn, T. C., McArthur, A. G., Boreham, D. R., Kieran, T., Lance, S., … Somers, C. M. (2015). Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). *Molecular Ecology Resources*, *15*(6), 1304–1315. https://doi.org/10.1111/1755-0998.12404

Grodzicker, T., Williams, J., Sharp, P., & Sambrook, J. (1974). Physical mapping of temperature-sensitive mutations of adenoviruses. *Cold Spring Harbor Symposia on Quantitative Biology*, *39*, 439–446. https://doi.org/10.1101/SQB.1974.039.01.056

Guo, S. W., & Thompson, E. A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, *48*(2), 361–372. https://doi.org/10.2307/2532296

Gusfield, D. (1997). *Algorithms on strings, trees, and sequences: Computer science and computational biology*. New york, NY: Cambridge University Press.

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. *PLOS Genetics*, *6*(2), e1000862. https://doi.org/10.1371/journal.pgen.1000862

Hohenlohe, P. A., Day, M. D., Amish, S. J., Miller, M. R., Kamps-Hughes, N., Boyer, M. C., … Luikart, G. (2013). Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology*, *22*(11), 3002–3013. https://doi.org/10.1111/mec.12239

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: Defining, estimating and interpreting F(ST). *Nature Reviews Genetics*, *10*(9), 639–650. https://doi.org/10.1038/nrg2611

Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Research*, *9*(9), 868–877. https://doi.org/10.1101/gr.9.9.868

Kelly, T. J., & Smith, H. O. (1970). A restriction enzyme from *Hemophilus influenzae*. *Journal of Molecular Biology*, *51*(2), 393–409. https://doi.org/10.1016/0022-2836(70)90150-6

Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., … Uszynski, G. (2012). Diversity arrays technology: A generic genome profiling technology on open platforms. *Methods in Molecular Biology (Clifton, N.J.)*, *888*, 67–89. https://doi.org/10.1007/978-1-61779-870-2_5

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*, 356. https://doi.org/10.1186/s12859-014-0356-4

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv*, 1303.3997 [q-Bio]. Retrieved from http://arxiv.org/abs/1303.3997

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMTOOLS. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Li, Y.-L., Xue, D.-X., Zhang, B.-D., & Liu, J.-X. (2018). An optimized approach for local de novo assembly of overlapping paired-end RAD reads from multiple individuals. *Royal Society Open Science*, *5*(2), 171589. https://doi.org/10.1098/rsos.171589

Louis, E. J., & Dempster, E. R. (1987). An exact test for Hardy-Weinberg and multiple alleles. *Biometrics*, *43*(4), 805–811. https://doi.org/10.2307/2531534

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017). Responsible RAD: Striving for best practices in population genomic studies of adaptation. *Molecular Ecology Resources*, *17*(3), 366–369. https://doi.org/10.1111/1755-0998.12677

Lynch, M. (2009). Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics*, *182*(1), 295–301. https://doi.org/10.1534/genetics.109.100479

Malinsky, M., Trucchi, E., Lawson, D. J., & Falush, D. (2018). RADPAINTER and FINERADSTRUCTURE: Population inference from RADseq data. *Molecular Biology and Evolution*, *35*(5), 1284–1290. https://doi.org/10.1093/molbev/msy023

Maruki, T., & Lynch, M. (2015). Genotype-frequency estimation from high-throughput sequencing data. *Genetics*, *201*(2), 473–486. https://doi.org/10.1534/genetics.115.179077

Maruki, T., & Lynch, M. (2017). Genotype calling from population-genomic sequencing data. *G3: Genes, Genomes, Genetics*, *7*, 1393–1404. https://doi.org/10.1534/g3.117.039008

McKinney, G. J., Larson, W. A., Seeb, L. W., & Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: Comment on breaking RAD by Lowry et al. (2016). *Molecular Ecology Resources*, *17*(3), 356–361. https://doi.org/10.1111/1755-0998.12649

Meirmans, P. G. (2006). Using the AMOVA framework to estimate a standardized genetic differentiation measure. *Evolution; International Journal of Organic Evolution*, *60*(11), 2399–2402. https://doi.org/10.1554/05-631.1

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, *17*(2), 240–248. https://doi.org/10.1101/gr.5681207

Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., & Hohenlohe, P. A. (2013). Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, *22*(11), 2841–2847. https://doi.org/10.1111/mec.12350

Near, T. J., MacGuigan, D. J., Parker, E., Struthers, C. D., Jones, C. D., & Dornburg, A. (2018). Phylogenetic analysis of Antarctic notothenioids illuminates the utility of RADseq for resolving Cenozoic adaptive radiations. *Molecular Phylogenetics and Evolution*, *129*, 268–279. https://doi.org/10.1016/j.ympev.2018.09.001

Nei, M. (1987). *Molecular evolutionary genetics*. New York, NY: Columbia University Press.

Nelson, T. C., & Cresko, W. A. (2018). Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations. *Evolution Letters*, *2*(1), 9–21. https://doi.org/10.1002/evl3.37

Paris, J. R., Stevens, J. R., & Catchen, J. M. (2017). Lost in parameter space: A road map for stacks. *Methods in Ecology and Evolution*, *8*(10), 1360–1373. https://doi.org/10.1111/2041-210X.12775

Patterson, M., Marschall, T., Pisanti, N., van Iersel, L., Stougie, L., Klau, G. W., & Schönhuth, A. (2015). WhatsHap: Weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology*, *22*(6), 498–509. https://doi.org/10.1089/cmb.2014.0157

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. *PLoS ONE*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., der Auwera, G. A. V., … Banks, E. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, *201178*, https://doi.org/10.1101/201178

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. https://doi.org/10.7717/peerj.431

Rivera-Colón, A. G., Rochette, N. C., & Catchen, J. M. (2019). Simulation with RADinitio improves RADseq experimental design and sheds light on sources of missing data. *BioRxiv*, 775239. https://doi.org/10.1101/775239

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

Schlötterer, C. (2004). The evolution of molecular markers—Just a matter of fashion? *Nature Reviews Genetics*, *5*(1), 63–69. https://doi.org/10.1038/nrg1249

Schweyen, H., Rozenberg, A., & Leese, F. (2014). Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. *The Biological Bulletin*, *227*(2), 146–160. https://doi.org/10.1086/BBLv227n2p146

Smith, H. O., & Welcox, K. W. (1970). A Restriction enzyme from *Hemophilus influenzae*. *Journal of Molecular Biology*, *51*(2), 379–391. https://doi.org/10.1016/0022-2836(70)90149-X

Spriggs, E. L., Eaton, D. A. R., Sweeney, P. W., Schlutius, C., Edwards, E. J., & Donoghue, M. J. (2019). Restriction-site-associated DNA sequencing reveals a cryptic Viburnum species on the North American Coastal Plain. *Systematic Biology*, *68*(2), 187–203. https://doi.org/10.1093/sysbio/syy084

Suchan, T., Pitteloud, C., Gerasimova, N. S., Kostikova, A., Schmid, S., Arrigo, N., … Alvarez, N. (2016). Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *PLoS ONE*, *11*(3), e0151651. https://doi.org/10.1371/journal.pone.0151651

Tin, M. M. Y., Rheindt, F. E., Cros, E., & Mikheyev, A. S. (2015). Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. *Molecular Ecology Resources*, *15*(2), 329–336. https://doi.org/10.1111/1755-0998.12314

Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R., & Bird, C. E. (2013). ezRAD: A simplified method for genomic genotyping in non-model organisms. *PeerJ*, *1*, e203. https://doi.org/10.7717/peerj.203

Trucchi, E., Gratton, P., Whittington, J. D., Cristofari, R., Le Maho, Y., Stenseth, N. C., & Le Bohec, C. (2014). King penguin demography since the last glaciation inferred from genome-wide data. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1787), 20140528. https://doi.org/10.1098/rspb.2014.0528

Trucchi, E., Mazzarella, A. B., Gilfillan, G. D., Lorenzo, M. T., Schönswetter, P., & Paun, O. (2016). BsRADseq: Screening DNA methylation in natural populations of non-model species. *Molecular Ecology*, *25*(8), 1697–1713. https://doi.org/10.1111/mec.13550

vanOrsouw, N. J., Hogers, R. C. J., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., … vanEijk, M. J. T. (2007). Complexity reduction of polymorphic sequences (CRoPS): A novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE*, *2*(11), e1172. https://doi.org/10.1371/journal.pone.0001172

Wang, S., Meyer, E., McKay, J. K., & Matz, M. V. (2012). 2b-RAD: A simple and flexible method for genome-wide genotyping. *Nature Methods*, *9*(8), 808–810. https://doi.org/10.1038/nmeth.2023

Weir, B. S. (1996). *Genetic data analysis II: Methods for discrete population genetic data*. Sunderland, MA: Sinauer Associates.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution; International Journal of Organic Evolution*, *38*(6), 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. https://doi.org/10.1101/gr.074492.107

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620. https://doi.org/10.1093/bioinformatics/btt593

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Rochette NC, Rivera-Colón AG, Catchen JM. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Mol Ecol*. 2019;00:1–18. https://doi.org/10.1111/mec.15253