

## ***dDocent*: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms**

Restriction-site associated DNA sequencing (RADseq) has become a powerful and useful approach for population genomics. Currently, no software exists that utilizes both paired-end reads from RADseq data to efficiently produce population-informative variant calls, especially for organisms with large effective population sizes and high levels of genetic polymorphism but for which no genomic resources exist. *dDocent* is an analysis pipeline with a user-friendly, command-line interface designed to process individually barcoded RADseq data (with double cut sites) into informative SNPs/Indels for population-level analyses. The pipeline, written in BASH, uses data reduction techniques and other stand-alone software packages to perform quality trimming and adapter removal, *de novo* assembly of RAD loci, read mapping, SNP and Indel calling, and baseline data filtering. Double-digest RAD data from population pairings of three different marine fishes were used to compare *dDocent* with *Stacks*, the first generally available, widely used pipeline for analysis of RADseq data. *dDocent* consistently identified more SNPs shared across greater numbers of individuals and with higher levels of coverage. This is most likely due to the fact that *dDocent* quality trims instead of filtering and incorporates both forward and reverse reads in assembly, mapping, and SNP calling, thus enabling use of reads with Indel polymorphisms. The pipeline and a comprehensive user guide can be found at (<http://dDocent.wordpress.com>).

1 ***dDocent*: a RADseq, variant-calling pipeline designed for population genomics of non-model**  
2 **organisms**

3 JONATHAN B. PURITZ<sup>†</sup>, CHRISTOPHER M. HOLLENBECK, AND JOHN R. GOLD

4 *Marine Genomics Laboratory, Harte Research Institute, Texas A&M University-Corpus Christi,*  
5 *6300 Ocean Drive, Corpus Christi, Texas 78412-5869*

6 <sup>†</sup>Author to whom correspondence should be addressed.

7 Email: jonathan.puritz@tamucc.edu Phone: 361-825-3343 Fax: 361-825-2050

## INTRODUCTION

8  
9 Next-generation sequencing (NGS) has transformed field of genetics into genomics by  
10 providing DNA sequence data at an ever increasing rate and reduced cost (Mardis, 2008). The  
11 nascent field of population genomics relies on NGS coupled with laboratory methods to  
12 reproducibly reduce genome complexity to a few thousand loci. The most common approach,  
13 restriction-site associated DNA sequencing (RADseq), uses restriction endonucleases to  
14 randomly sample the genome at locations adjacent to restriction-enzyme recognition sites that,  
15 when coupled with Illumina sequencing, produces high coverage of homologous SNP (Single  
16 Nucleotide Polymorphism) loci. As such, RADseq provides a powerful approach population  
17 level genomic studies (Ellegren, 2014; Narum et al., 2013; Rowe et al., 2011).

18 The original RADseq approach (Baird et al., 2008), and initial population genomic studies  
19 employing it (Hohenlohe et al., 2010), focused on SNP discovery and genotyping on the first  
20 (forward) read only. This is because the original RADseq method (Baird et al., 2008) utilized  
21 random shearing to produce RAD loci; paired-end reads were not of uniform length or coverage,  
22 making it problematic to find SNPs at high and uniform levels of coverage across a large  
23 proportion of individuals. As a result, the most comprehensive and widely used software package  
24 for analysis of RADseq data, *Stacks* (Catchen et al., 2013, 2011), provides SNP genotypes based  
25 only on first-read data. In contrast, RADseq approaches such as ddRAD (Peterson et al., 2012),  
26 2bRAD (Wang et al., 2012), and ezRAD (Toonen et al., 2013) rely on restriction enzymes to  
27 define both ends of a RAD locus, largely producing RAD loci of fixed length (fRAD). Paired-  
28 end Illumina sequencing of fRAD fragments provides an opportunity to significantly expand the  
29 number of SNPs that can be genotyped from a single RADseq library.

30 Here, the variant-calling pipeline *dDocent* is introduced as a tool for generating population  
31 genomic data; a brief methodological outline of the analysis pipeline also is presented. *dDocent*  
32 is a wrapper script designed to take raw RADseq data and produce population informative SNP

33 calls, taking full advantage of both paired-end reads. *dDocent* is configured for organism with  
34 high levels of nucleotide and INDEL polymorphisms, such as found in many marine organisms  
35 (Guo et al., 2012; Keever et al., 2009; Sodergren et al., 2006; Waples, 1998; Ward et al., 1994). As  
36 input, *dDocent* takes paired FASTQ files for individuals and outputs raw SNP and INDEL calls as  
37 well as filtered SNP calls in VCF format. The pipeline and a comprehensive online manual can  
38 be found at (<http://dDocent.wordpress.com>). Finally, results of pipeline analyses, using both  
39 *dDocent* and *Stacks*, of populations of three species of marine fishes are provided to demonstrate  
40 the utility of *dDocent* compared to *Stacks*, the first and most comprehensive existing software  
41 package for RAD population genomics.

## 42 METHODS

### 43 *Implementation and basic usage*

44 The *dDocent* pipeline is written in BASH and will run using most Unix-like operating  
45 systems. *dDocent* is largely dependent on other bioinformatics software packages, taking  
46 advantage of programs designed specifically for each task of the analysis and ensuring that each  
47 modular component can be updated separately. Proper implementation depends on the correct  
48 installation of each third-party packages/tools. A full list of dependencies can be found in the  
49 user manual at (<http://ddocent.wordpress.com/ddocent-pipeline-user-guide/>) and a sample script  
50 to automatically download and install the packages in a Linux environment can be found at the  
51 *dDocent* repository (<https://github.com/jpuritz/dDocent>).

52 *dDocent* is run by simply switching to a directory containing the input data and starting the  
53 program. There is no configuration file; *dDocent* will proceed through a short series of  
54 command-line prompts, allowing the user to set up analysis parameters. After all required  
55 variables are configured, including an e-mail address for a completion notification, *dDocent*  
56 provides instructions on how to move the program to the background and run, undisturbed, until  
57 completion. The pipeline is designed to take advantage of multiple processing core machines

58 and, whenever possible, processes should be invoked with multiple threads or occurrences. For  
59 most Linux distributions, the number of processing cores should be automatically detected. If  
60 *dDocent* cannot determine the number of processors, it will ask the user to input the value.

61 There are two distinct modules of *dDocent*: *dDocent.FB* and *dDocent.GATK*. *dDocent.FB*  
62 uses minimal, BAM-file preparation steps before calling SNPs and INDELS, simultaneously using  
63 FreeBayes (Garrison & Marth, 2012). *dDocent.GATK* uses GATK (McKenna et al., 2010) for  
64 INDEL realignment, SNP and INDEL genotyping (using HaplotypeCaller), and variant quality-  
65 score recalibration, largely following GATK Best Practices recommendations (Auwera &  
66 Carneiro, 2013;DePristo et al., 2011). The modules represent two different strategies for  
67 SNP/INDEL calling that are completely independent of one another. The remainder of this paper  
68 focuses on *dDocent.FB*; additional information on *dDocent.GATK* may be found in the user  
69 guide and results from *dDocent.GATK* can be found in Appendix S1.

#### 70 *Data input requirements*

71 *dDocent* requires demultiplexed forward and paired-end FASTQ files for every individual in  
72 the analysis. A simple naming convention (a single-word locality code/name and a single-word  
73 sample identifier separated by an underscore) must be followed for every sample; examples are  
74 *LOCA\_IND01.F.fq* and *LOCA\_IND01.R.fq*. A sample script for using a text file with barcodes  
75 and sample names and *process\_radtags* from *Stacks* (Catchen et al., 2013) to properly  
76 demultiplex samples and put them in the proper *dDocent* naming convention can be found at the  
77 *dDocent* repository (<https://github.com/jpuritz/dDocent>).

#### 78 *Quality trimming*

79 After *dDocent* checks that it is recognizing the proper number of samples in the current  
80 directory, it asks the user if s/he wishes to proceed with quality trimming of sequence data. If  
81 directed, *dDocent* can use the program *Trim Galore!*  
82 ([http://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)) to simultaneously remove

83 Illumina adapter sequences and trim ends of reads of low quality. By default, *Trim Galore!* looks  
84 for double-digest RAD adapters (Peterson et al., 2012) and trims bases with quality scores less  
85 than Phred 10. Typically, quality trimming only needs to be performed once on data, so the  
86 option exists to skip this step in subsequent *dDocent* analyses.

### 87 *De novo assembly*

88 Without reference material, population genomic analyses from RADseq depend on *de novo*  
89 assembly of a set of reference contigs. Inherently, not all RAD loci appear in all individuals due  
90 to stochastic processes inherent in library preparation and sequencing and to polymorphism in  
91 restriction-enzyme restriction sites (Catchen et al., 2011). Moreover, populations can contain  
92 large levels of within locus polymorphism, making generation of a reference sequence  
93 computationally difficult. *dDocent* minimizes the amount of data used for assembly by taking  
94 advantage of the fact that flRAD loci present in multiple individuals should have higher levels of  
95 exactly matching reads (forward and reverse) than loci that are only present in a few individuals.  
96 Caution is advised for unique reads with low levels of coverage throughout the data set as they  
97 likely represent sequencing errors or polymorphisms that are shared only by a few individuals.

98 During assembly, paired-end reads are reverse complemented and concatenated to forward  
99 reads. Unique paired reads are identified and their occurrences are counted in the entire data set.  
100 These data are tabulated into the number of unique reads per levels of 1X to 50X coverage; a  
101 graph is then generated and printed to the terminal. The distribution usually follows an  
102 asymptotic relationship (Figure 1), with a large proportion of reads only having one or two  
103 occurrences, meaning they likely will not be informative on a population scale. Highly  
104 polymorphic RAD loci still should have at least one allele present at the level of expected  
105 sequence coverage, so this can be used as a guide for informative data. The user chooses a cut-  
106 off level of coverage for reads to be used for assembly – note all reads are still used for  
107 subsequent steps of the pipeline.

108 After a cut-off level is chosen, remaining reads are returned in forward- and reverse-read  
109 files and then input directly into the RADseq assembly program *Rainbow* (Chong et al., 2012).  
110 The default parameters of *Rainbow* are used except that the maximum number of mismatches  
111 used in initial clustering should be changed from four to six. In short, *Rainbow* clusters forward  
112 reads based on similarity; clusters are then recursively divided, based on reverse reads, into  
113 groups representing single alleles. Reads in merged clusters are then assembled using a greedy  
114 algorithm (Pop & Salzberg, 2008). *dDocent* then selects the longest contig for each cluster as the  
115 representative reference sequence for that RAD locus. If the forward read does not overlap with  
116 the reverse read (almost always the case with flRAD), the forward read is concatenated to the  
117 reverse read with ten 'N' characters as padding. Finally, reference sequences are clustered based  
118 on overall sequence similarity (chosen by user, 90% by default), using the program *CD-HIT* (Fu  
119 et al., 2012; Li & Godzik, 2006). This final cluster step reduces the data set further, based on  
120 overall sequence identity after assembly. Alternatively, *de novo* assembly can be skipped and the  
121 user can provide a FASTA file with reference sequences.

## 122 *Read mapping*

123 *dDocent* uses the MEM algorithm (Li, 2013) of *BWA* (Li & Durbin, 2009, 2010) to map  
124 quality-trimmed reads to the reference contigs. Users can deploy the default values of *BWA* or  
125 set an alternative value for each mapping parameter (match score, mismatch score, and gap-  
126 opening penalty). The default settings are meant for mapping reads to the human genome, so  
127 users are encouraged to experiment with mapping parameters. *BWA* output is ported to  
128 *SAMtools* (Li et al., 2009), saving disk space, and alignments are saved to the disk as binary  
129 alignment/Map (BAM). BAM files are then sorted and indexed.

## 130 *SNP and INDEL discovery and genotyping*

131 *dDocent* uses a two-step process to optimize the computationally intensive task of  
132 SNP/INDEL calling. First, quality-trimmed forward and reverse reads are reduced to unique

133 reads. This data set is then mapped to all reference sequences using the previously entered  
134 mapping settings (see *Read Mapping* above). From this alignment, a set of intervals is created  
135 using BEDtools (Quinlan & Hall, 2010). The interval set saves computational time by directing  
136 the SNP-/INDEL-calling software to examine only reference sequences along contigs that have  
137 high quality mappings. Second, the interval list is then split into a single file for each processing  
138 core, allowing SNP/INDEL calling to be optimized with a scatter-gather technique. The program  
139 *FreeBayes* (Garrison & Marth, 2012) is then executed multiple times simultaneously (one  
140 execution per processor and genomic interval). *FreeBayes* is a Bayesian-based, variant-detection  
141 software that uses assembled haplotype sequences to simultaneously call SNPs, INDELS, multi-  
142 nucleotide polymorphisms (MNPs), and complex events (e.g., composite insertion and  
143 substitution events) from alignment files; *FreeBayes* has the added benefit for population  
144 genomics of using reads across multiple individuals to improve genotyping (Garrison & Marth,  
145 2012). *FreeBayes* is run with minimal changes to the default parameters; minimum mapping  
146 quality score and base quality score are set to PHRED 10. After all executions of *FreeBayes* are  
147 completed, raw SNP/INDEL calls are concatenated into a single variant call file (VCF), using  
148 VCFtools (Danecek et al., 2011).


#### 149 *Variant Filtering*

150 Final SNP data-set requirements are likely to be highly dependent on specific goals and aims  
151 of individual projects. To that end, *dDocent* uses *VCFtools* (Danecek et al., 2011) to provide only  
152 basic level filtering, mostly for run diagnostic purposes. *dDocent* produces a final VCF file that  
153 contains all SNPs, INDELS, MNPs, and complex events that are called in 90% of all individuals,  
154 with a minimum quality score of 30. Users are encouraged to use VCFtools and *vcflib* (part of  
155 the *FreeBayes* package; <https://github.com/ekg/vcflib>) to fully explore and filter data  
156 appropriately.

#### 157 *Comparison between dDocent and Stacks*



158 Two sample localities, each comprised of 20 individuals, were chosen randomly from  
159 unpublished RADseq data sets of three different, marine fish species: red snapper (*Lutjanus*  
160 *campechanus*), red drum (*Sciaenops ocellatus*), and silk snapper (*Lutjanus vivanus*). These three  
161 species are part of ongoing RADseq projects in our laboratory, and preliminary analyses  
162 indicated high levels of nucleotide polymorphisms across all populations. Double-digest RAD  
163 libraries were prepared, generally following Peterson *et al.* (2012). Individual DNA extractions  
164 were digested with *EcoRI* and *MspI*. A barcoded adapter was ligated to the *EcoRI* site of each  
165 fragment and a generic adapter was ligated to the *MspI* site. Samples were then equimolarly  
166 pooled and size-selected between 350 and 400 bp, using a Qiagen Gel Extraction Kit. Final  
167 library enhancement was completed using 12 cycles of PCR, simultaneously enhancing properly  
168 ligated fragments and adding an Illumina Index for additional barcoding. Libraries were  
169 sequenced on three separate lanes of an Illumina HiSeq 2000 at the University of Texas Genomic  
170 Sequencing and Analysis Facility.

171 Demultiplexed individual reads were analyzed with *dDocent*, using three different levels of  
172 final reference contig clustering (90%, 96%, and 99% similarity) in an attempt to alter the most  
173 comparable analysis variable in *dDocent* to match analysis variables of *Stacks*. The coverage cut-  
174 off for assembly was 12 for red snapper, 13 for red drum, and nine for silk snapper.  *dDocent*  
175 runs used mapping variables of one, three, and five for match-score value, mismatch score, and  
176 gap-opening penalty, respectively. For comparisons, complex variants were decomposed into  
177 canonical SNP and INDEL representation from the raw VCF files, using *vcfallelicprimitives* from  
178 *vcflib* (<https://github.com/ekg/vcflib>).

179 For *Stacks*, reads were demultiplexed and cleaned using *process\_radtags*, removing reads  
180 with 'N' calls and low-quality base scores. Because *dDocent* inherently uses both reads for  
181 SNP/INDEL genotyping, forward reads and reverse reads were processed separately with  
182 *denovo\_map.pl* (*Stacks* version 1.08), using three different sets of parameters. The first set had a

183 minimum depth of coverage of two to create a stack, a maximum distance of two between stacks,  
184 and a maximum distance of four between stacks from different individuals, with both the  
185 deleveraging algorithm and removal algorithms enabled. The second set had a minimum depth of  
186 coverage of three to create a stack, a maximum distance of four between stacks, and a maximum  
187 distance of eight between stacks from different individuals, with both the deleveraging algorithm  
188 and removal algorithms enabled. The third set had a minimum depth of coverage of three to  
189 create a stack, a maximum distance of four between stacks, and a maximum distance of 10  
190 between stacks from different individuals, with both the deleveraging algorithm and removal  
191 algorithms enabled. SNP calls were output in VCF format.

192 For both *dDocent* and *Stacks* runs, VCFtools was used to filter out INDELS and SNPs that had  
193 a minor allele count of less than five. SNP calls were then evaluated at different individual-  
194 coverage levels: the total number of SNPs; the number of SNPS called in 75%, 90%, and 99% of  
195 individuals at 3X coverage; the number of SNPS called in 75% and 90% of individuals at 5X  
196 coverage; the number of SNPS called in 75% and 90% of individuals at 10X coverage; and the  
197 number of SNPS called in 75% and 90% of individuals at 20X coverage. Overall coverage levels  
198 for red snapper were lower and likely impacted by a few low-quality individuals; consequently,  
199 the number of 5X and 10X SNPs shared among 90% of individuals (after removing the bottom  
200 10% of individuals in terms of coverage) were compared instead of SNP loci shared at 20X  
201 coverage. Results from two runs of *Stacks* (one using forward and one using reverse reads) were  
202 combined for comparison with *dDocent*, which inherently calls SNPs on both reads. All analyses  
203 and computations were performed on a 32-core Linux workstation with 128 GB of RAM.

## 204 RESULTS AND DISCUSSION

205 Results of SNP calling, including run times (in minutes) for each analysis (not including  
206 quality trimming), are presented in Table 1. Data from high coverage SNP calls, averaged over  
207 all runs for each pipeline, are presented in Figure 1. While *Stacks* called a larger number of low

208 coverage SNPs, limiting results to higher individual coverage and to higher individual call rates  
209 revealed that *dDocent* consistently called more high-quality SNPs. Run times were equivalent  
210 for both pipelines.

211 At almost all levels of coverage in three different data sets, *dDocent* called more SNPs across  
212 more individuals than *Stacks*. Two key differences between *dDocent* and *Stacks* likely contribute  
213 these discrepancies: (i) quality trimming instead of quality filtering, and (ii) simultaneous use of  
214 forward and reverse reads by *dDocent* in assembly, mapping, and genotyping, instead of  
215 clustering as employed by *Stacks*. As with any data analysis, quality of data output is directly  
216 linked to the quality of data input. Both *dDocent* and *Stacks* use procedures to ensure that only  
217 high-quality sequence data are retained; however, *Stacks* removes an entire read when a sliding  
218 window of bases drops below a preset quality score (PHRED 10, by default), while *dDocent* via  
219 *Trim Galore!* trims off low-quality bases, preserving high-quality bases of each read. Filtering  
220 instead of trimming results in fewer reads entering the *Stacks* analysis (between 65%-95% of the  
221 data compared to *dDocent*; data not shown), generating lower levels of coverage and fewer SNP  
222 calls than *dDocent*.

223 *dDocent* offers two advantages over *Stacks*: (i) it is specifically designed for paired-end data  
224 and utilizes both forward and reverse reads for *de novo* RAD loci assembly, read mapping,  
225 variant discovery, and genotyping; and (ii) it aligns reads to reference sequence instead of  
226 clustering by identity. Using both reads to cluster and assemble RAD loci helps to ensure that  
227 portions of the genome with complex mutational events, including INDELS or small repetitive  
228 regions, are properly assembled and clustered as homologous loci. Additionally, using *BWA* to  
229 map reads to reference loci enables *dDocent* to properly align reads with INDEL polymorphisms,  
230 increasing coverage and subsequent variant discovery and genotyping. Clustering methods  
231 employed by *Stacks*, whether clustering alleles within an individual or clustering loci between  
232 individuals, effectively remove reads, alleles, and loci with INDEL polymorphisms because the

233 associated frame shift effectively inflates the observed number of base-pair differences. For  
234 organisms with large effective population sizes and high levels of genetic diversity, such as many  
235 marine organisms (Waples, 1998; Ward et al., 1994), removing reads and loci with INDEL  
236 polymorphisms will result in a loss of shared loci and coverage.

237

### CONCLUSION

238 *dDocent* is an open-source, freely available population genomics pipeline configured for  
239 species with high levels of nucleotide and INDEL polymorphisms, such as many marine  
240 organisms. The *dDocent* pipeline reports more SNPs shared across greater numbers of  
241 individuals and with higher levels of coverage than current alternatives. The pipeline and a  
242 comprehensive online manual can be found at (<http://dDocent.wordpress.com>) and  
243 (<https://github.com/jpuritz/dDocent>).

244

### ACKNOWLEDGEMENTS

245 We thank T. Krabbenhoft for beta testing and use of his data, and C. Bird and D. Portnoy for  
246 useful discussions and comments on the manuscript.

- 248 Auwera G, Carneiro M. 2013. From FastQ Data to High-Confidence Variant Calls: The Genome  
249 Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*: 1–33.
- 250 Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA,  
251 Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD  
252 markers. *PLoS ONE* 3: e3376.
- 253 Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. 2011. Stacks: building and  
254 genotyping Loci de novo from short-read sequences. *G3 (Bethesda, Md.)* 1: 171–182.
- 255 Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set  
256 for population genomics. *Molecular ecology* 22: 3124–3140.
- 257 Chong Z, Ruan J, Wu C. 2012. Rainbow : an integrated tool for efficient clustering and  
258 assembling RAD-seq reads. : 1–6.
- 259 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G,  
260 Marth GT, Sherry ST, McVean G, Durbin R. 2011. The variant call format and VCFtools.  
261 *Bioinformatics (Oxford, England)* 27: 2156–2158.
- 262 DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, Angel G  
263 del, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernysky AM, Sivachenko AY,  
264 Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery  
265 and genotyping using next-generation DNA sequencing data. *Nature genetics* 43: 491–498.
- 266 Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. *Trends*  
267 *in ecology & evolution* 29: 51–63.
- 268 Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation  
269 sequencing data. *Bioinformatics (Oxford, England)* 28: 3150–3152.
- 270 Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. : 1–9.
- 271 Guo B, Zou M, Wagner A. 2012. Pervasive indels and their evolutionary dynamics after the fish-  
272 specific genome duplication. *Molecular biology and evolution* 29: 3005–3022.
- 273 Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population  
274 genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS*  
275 *genetics* 6: e1000862.
- 276 Keever CC, Sunday J, Puritz JB, Addison JA, Toonen RJ, Grosberg RK, Hart MW. 2009.  
277 Discordant distribution of populations and genetic variation in a sea star with high dispersal  
278 potential. *Evolution; international journal of organic evolution* 63: 3214–3227.
- 279 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.  
280 *Bioinformatics (Oxford, England)* 25: 1754–1760.
- 281 Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform.  
282 *Bioinformatics (Oxford, England)* 26: 589–595.
- 283 Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein  
284 or nucleotide sequences. *Bioinformatics (Oxford, England)* 22: 1658–1659.
- 285 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.  
286 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford,*  
287 *England)* 25: 2078–2079.
- 288 Li H. 2013. Aligning sequence reads , clone sequences and assembly contigs with BWA-MEM.  
289 00: 1–3.
- 290 Mardis ER. 2008. Next-generation DNA sequencing methods. *Annual review of genomics and*  
291 *human genetics* 9: 387–402.

- 292 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K,  
293 Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a  
294 MapReduce framework for analyzing next-generation DNA sequencing data. *Genome*  
295 *research* 20: 1297–1303.
- 296 Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA. 2013. Genotyping-by-sequencing  
297 in ecological and conservation genomics. *Molecular ecology* 22: 2841–2847.
- 298 Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: an  
299 inexpensive method for de novo SNP discovery and genotyping in model and non-model  
300 species. *PloS one* 7: e37135.
- 301 Pop M, Salzberg S. 2008. Bioinformatics challenges of new sequencing technology. *Trends in*  
302 *Genetics* 24: 142–149.
- 303 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic  
304 features. *Bioinformatics (Oxford, England)* 26: 841–842.
- 305 Rowe HC, Renaut S, Guggisberg A. 2011. RAD in the realm of next-generation sequencing  
306 technologies. *Molecular ecology* 20: 3499–3502.
- 307 Sodergren E, Weinstock GM, Davidson EH, Cameron RA, Gibbs RA, Angerer RC, Angerer LM,  
308 Arnone MI, Burgess DR, Burke RD, Coffman JA, Dean M, Elphick MR, Etensohn CA,  
309 Foltz KR, Hamdoun A, Hynes RO, Klein WH, Marzluff W, et al. 2006. The genome of the  
310 sea urchin *Strongylocentrotus purpuratus*. *Science (New York, N.Y.)* 314: 941–952.
- 311 Toonen RJ, Puritz JB, Forsman ZH, Whitney JL, Fernandez-Silva I, Andrews KR, Bird CE. 2013.  
312 ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ* 1:  
313 e203.
- 314 Wang S, Meyer E, McKay JK, Matz M V. 2012. 2b-RAD: a simple and flexible method for  
315 genome-wide genotyping. *Nature methods* 9: 808–810.
- 316 Waples RS. 1998. Separating the wheat from the chaff: patterns of genetic differentiation in high  
317 gene flow species. *Journal of Heredity* 89: 438–450.
- 318 Ward RD, Woodwark M, Skibinski DOF. 1994. A comparison of genetic diversity levels in  
319 marine, freshwater, and anadromous fishes. *Journal of Fish Biology* 44: 213–232.

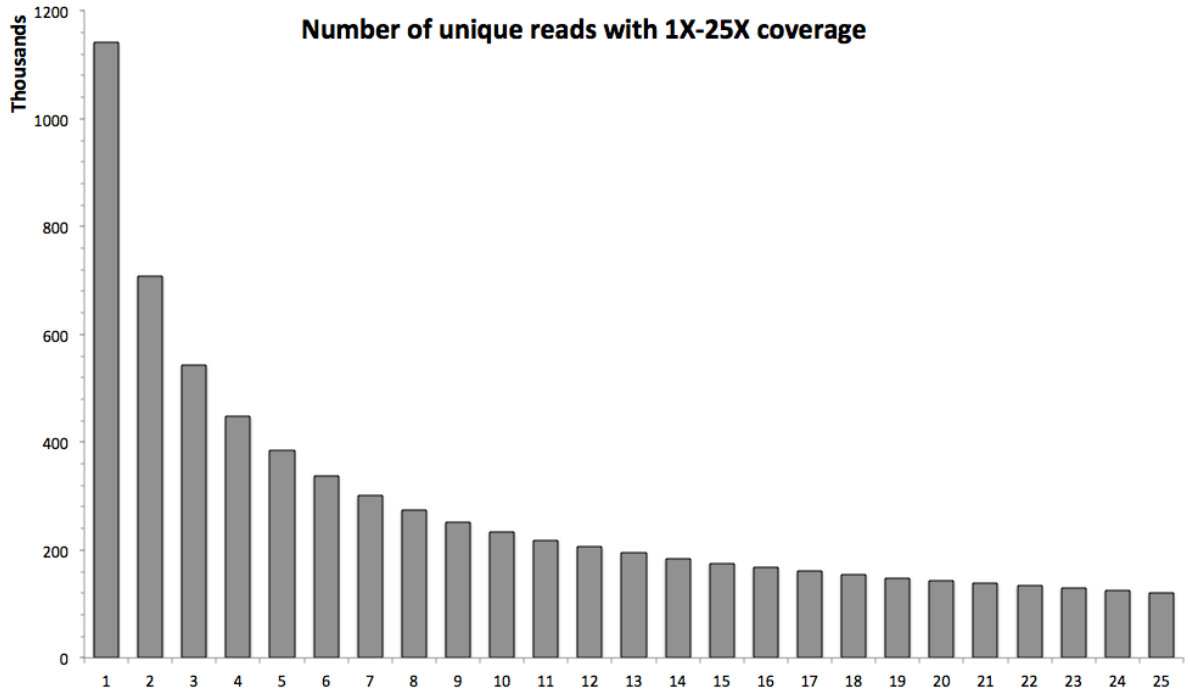
320 Table 1. Results from individual runs of *dDocent* and *Stacks*. *dDocent* runs varied in the level of  
 321 similarity used to cluster reference sequences: A (90%), B (96%), and C (99%). For *Stacks*,  
 322 forward reads and reverse reads were separately processed with *denovo\_map.pl* (*Stacks* version  
 323 1.08), using three different sets of parameters: A, minimum depth of coverage of two to create a  
 324 stack, a maximum distance of two between stacks, and a maximum distance of four between  
 325 stacks from different individuals; B, minimum depth of coverage of three to create a stack, a  
 326 maximum distance of four between stacks, and a maximum distance of eight between stacks from  
 327 different individuals; and C, minimum depth of coverage of three to create a stack, a maximum  
 328 distance of four between stacks, and a maximum distance of 10 between stacks from different  
 329 individuals. SNP calls were evaluated at different individual coverage levels: (i) total number of  
 330 SNPs; (ii) number of SNPS called in 75%, 90%, and 99% at 3X coverage; (iii) number of SNPS  
 331 called in 75% and 90% of individuals at 5X coverage; (iv) number of SNPS called in 75% and  
 332 90% of individuals at 10X coverage; and, (v) number of SNPS called in 75% and 90% of  
 333 individuals at 20X coverage. Results from forward and reverse reads of *Stacks* were combined  
 334 for comparison with *dDocent*, which inherently calls SNPs on both reads.

	<i>dDocent</i> A	<i>dDocent</i> B	<i>dDocent</i> C	<i>Stacks</i> A	<i>Stacks</i> B	<i>Stacks</i> C
	Red snapper					
Total 3X SNPS	30,130	30,043	29,907	28,817	33,479	34,459
75% 3X SNPs	12,507	12,249	12,012	4,150	5,735	5,728
90% 3X SNPs	5,368	5,187	5,039	675	987	983
99% 3X SNPs	52	25	5	0	0	0
75% 5X SNPs	8,144	7,946	7,793	2,632	4,351	4,324
90% 5X SNPs	2,775	2,696	2,606	179	579	574
75% 10X SNPs	4,151	4,017	3,914	783	1,618	1,579
90% 10X SNPS	785	729	682	7	48	47
90% IND 90% 5X	5,625	5,499	5,332	806	1,807	1,079
90% IND 90% 10x	2,403	2,298	2,196	129	441	434
Run time	59	58	57	70	47	53
	Red drum					
Total 3X SNPS	27,263	27,329	27,295	45,792	50,821	52,366

75% 3X SNPs	23,339	23,328	23,226	24,134	28,991	28,981
90% 3X SNPs	20,764	20,704	20,586	13,439	17,946	17,874
99% 3X SNPs	7,121	7,022	6,937	828	1,264	1,259
75% 5X SNPs	20,015	20,009	19,946	21,021	26,526	26,464
90% 5X SNPs	16,739	16,680	16,588	10,494	15,282	15,207
75% 10X SNPs	16,078	16,042	15,970	12,928	17,018	16,983
90% 10X SNPs	10,988	10,942	10,842	4,159	6,734	6,705
75% 20X SNPs	7,975	7,933	7,824	2,276	3,538	3,516
90% 20X SNPs	3,534	3,512	3,455	243	1,974	1,961
Run time	55	55	53	58	55	65
				Silk snapper		
Total 3X SNPs	35,763	35,645	35,509	48,742	55,505	58,352
75% 3X SNPs	17,518	17,244	16,992	7,596	9,705	9,696
90% 3X SNPs	8,586	8,353	8,157	2,007	3,439	3,433
99% 3X SNPs	2,552	2,380	2,276	132	527	523
75% 5X SNPs	10,775	10,547	10,385	4,789	7,290	7,274
90% 5X SNPs	4,936	4,725	4,606	1,225	2,573	2,570
75% 10X SNPs	5,252	5,018	4,876	2,094	3,547	3,546
90% 10X SNPs	2,191	2,058	1,938	489	1,224	1,223
75% 20X SNPs	2,220	2,098	1,984	703	1,415	1,411
90% 20X SNPs	801	721	675	136	417	418
Run time	98	100	60	93	89	204



335 Figure 1. Levels of coverage for each unique read in the red snapper data set. The horizontal  
336 axis represents the minimal level of coverage and the vertical axis represents the number of  
337 unique paired reads in thousands.



338 Figure 2. SNP results averaged across the three different run parameters for *dDocent* and *Stacks*.  
 339 (A) Red snapper, (B) Red drum, (C) Silk snapper (see Methods or Table 1 for SNP categories  
 340 description). Error bars represent standard error.

