



HAL
open science

DiscoSnp-RAD: de novo detection of small variants for RAD-Seq population genomics

Jérémy Gauthier, Charlotte Mouden, Tomasz Suchan, Nadir Alvarez, Nils Arrigo, Chloé Riou, Claire Lemaitre, Pierre Peterlongo

► **To cite this version:**

Jérémy Gauthier, Charlotte Mouden, Tomasz Suchan, Nadir Alvarez, Nils Arrigo, et al.. DiscoSnp-RAD: de novo detection of small variants for RAD-Seq population genomics. 2020. hal-01634232v3

HAL Id: hal-01634232

<https://inria.hal.science/hal-01634232v3>

Preprint submitted on 18 May 2020 (v3), last revised 10 Jun 2020 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DiscoSnp-RAD: de novo detection of small variants for RAD-Seq population genomics

Jérémy Gauthier¹, Charlotte Mouden^{1,2}, Tomasz Suchan³, Nadir Alvarez^{4,5}, Nils Arrigo⁴, Chloé Riou¹, Claire Lemaitre¹, and Pierre Peterlongo¹

¹Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France

²INRA, BIOGECO, UMR1202, Cestas, France

³W. Szafer Institute of Botany, Polish Academy of Sciences, Kraków, Poland

⁴Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

⁵Natural History Museum of Geneva, Geneva, Switzerland

Corresponding author:

Pierre Peterlongo

Email address: pierre.peterlongo@inria.fr

ABSTRACT

Restriction site Associated DNA Sequencing (RAD-Seq) is a technique characterized by the sequencing of specific loci along the genome, that is widely employed in the field of evolutionary biology since it allows to exploit variants (mainly Single Nucleotide Polymorphism - SNPs) information from entire populations at a reduced cost. Common RAD dedicated tools, such as *STACKS* or *IPyRAD*, are based on all-versus-all read alignments, which require consequent time and computing resources. We present an original method, *DiscoSnp-RAD*, that avoids this pitfall since variants are detected by exploiting specific parts of the assembly graph built from the reads, hence preventing all-versus-all read alignments. We tested the implementation on simulated datasets of increasing size, up to 1000 samples, and on real RAD-Seq data from 259 specimens of *Chiastocheta* flies, morphologically assigned to 7 species. All individuals were successfully assigned to their species using both STRUCTURE and Maximum Likelihood phylogenetic reconstruction. Moreover, identified variants succeeded to reveal a within-species genetic structure linked to the geographic distribution. Furthermore, our results show that *DiscoSnp-RAD* is significantly faster than state-of-the-art tools. The overall results show that *DiscoSnp-RAD* is suitable to identify variants from RAD-Seq data, it does not require time-consuming parameterization steps and it stands out from other tools due to its completely different principle, making it substantially faster, in particular on large datasets.

License: GNU Affero general public license

Availability: *DiscoSnp-RAD* belongs to the *DiscoSnp++* repository <https://github.com/GATB/DiscoSnp/>

1 INTRODUCTION

Next-generation sequencing and the ability to obtain genomic sequences for hundreds to thousands of individuals of the same species has opened new horizons in population genomics research. This has been made possible by the development of cost-efficient approaches to obtain sufficient homologous genomic regions, by reproducible genome complexity reduction and multiplexing several samples within a single sequencing run [1]. Among such methods, the most widely used over the last decade is “Restriction-site Associated DNA sequencing” (RAD-Seq). It uses restriction enzymes to digest DNA at specific genomic sites whose adjacent regions are then sequenced. This approach encompasses various methods with different intermediate steps to optimize the genome sampling, e.g. ddRAD [20], GBS [5], 2b-RAD [30], 3RAD/RADcap [11]. These methods share some basic steps: DNA digestion by one or more restriction enzymes, ligation of sequencing adapters and sample-specific barcodes, followed by optional fragmentation and fragment size selection, multiplexing samples bearing specific molecular tags, i.e. indices and barcodes, and finally sequencing. The sequencing output is thus composed of millions of

47 reads originating from all the targeted homologous loci. The usual bioinformatic steps consist in sample
48 demultiplexing, clustering sequences in loci and identifying informative homologous variations. If a
49 reference genome exists, the most widely used strategy is to align the reads to this reference genome
50 and to perform a classical variant calling, focusing on small variants, Single Nucleotide Polymorphisms
51 (SNPs) and small Insertion-Deletions (INDELs). However, RAD-Seq approaches are used on non-model
52 organisms for which a reference genome does not exist or is poorly assembled. The fact that all reads
53 sequenced from the same locus start and finish exactly at the same position makes it easy to compare
54 directly reads sequenced from a same locus. To *de novo* build homologous genomic loci and extract
55 informative variations, several methods have been developed, such as *STACKS* [2] and *PyRAD* [3], as well
56 as its derived rewritten version *IPyRAD* [4], being the most commonly used in the population genomics
57 community.

58 The main idea behind these approaches is to group reads by sequence similarity into clusters rep-
59 resenting each a distinct genomic locus. Since reads originating from the same locus start and end at
60 the same positions, they can be globally aligned, sequence variations can then be easily identified and a
61 consensus sequence is built for each locus. The key challenge is therefore the clustering part. To do so,
62 the classical approach relies on all-versus-all alignments. To reduce the number of alignments to compute,
63 the clustering is first performed within each sample independently, then sample consensus are compared
64 between samples. Nevertheless the number of alignments to perform remains very large in datasets
65 composed of many large read sets. Importantly, analysis of RAD-Seq data is highly dependent on the
66 chosen clustering method, the sequencing quality and the dataset composition, such as the presence of inter
67 and/or intra-specific specimens or the number of individuals. Thus, existing tools allow customization of
68 numerous parameters to fine-tune the analysis. Particularly, both methods have parameters controlling the
69 granularity of clustering: the number of mismatches allowed between sequences of a same locus within
70 and among samples for *STACKS* and the percentage of similarity for *PyRAD*. These can be arbitrarily
71 fixed by the user, but have a significant impact on downstream analyses [25].

72 We present here *DiscoSnp-RAD*, an utterly different approach to predict *de novo* small variants (SNPs
73 and indels) from large RAD-Seq datasets, without performing any read clustering, avoiding all-versus-all
74 read comparisons and without relying on a critical similarity threshold parameter. *DiscoSnp-RAD* takes
75 advantage of the *DiscoSnp++* approach [29, 19], that was initially designed for *de novo* prediction of
76 small variants, from shotgun sequencing reads, without the need of a reference genome. The basic idea
77 of the method is a careful analysis of the *de Bruijn graph* built from all the input read sets, to identify
78 topological motifs, often called *bubbles*, generated by polymorphisms. Notably, those bubbles arise
79 whatever the global similarity level between homologous reads, explaining why *DiscoSnp-RAD* is free of
80 similarity-related parameters. Note that *STACKS2* also uses a *de Bruijn graph* approach, but in a different
81 way, as it is used to build a so-called “*RAD-locus*” contig catalog on which reads are aligned for calling
82 SNPs [24].

83 After validation tests on simulated datasets of increasing size, we present an application of the
84 *DiscoSnp-RAD* implementation on double-digest RAD-Seq data (ddRAD) from a genus-wide sampling of
85 parasitic flies belonging to *Chiastocheta* genus. Using *DiscoSnp-RAD*, the 259 individuals analyzed could
86 be assigned to their respective species. Moreover, within-species analyses focused on one of these species,
87 identified variants revealing population structure congruent with sample geographic origins. Thus, the
88 information obtained from variants identified by *DiscoSnp-RAD* can be successfully used for population
89 genomic studies. The main notable difference between *DiscoSnp-RAD* and concurrent algorithms stands
90 in its easiness to use, in the fact that it does not require fine parameter tuning, and in its execution time, as
91 it is substantially faster than *STACKS* and *IPyRAD*.

92 2 MATERIAL AND METHODS

93 2.1 *DiscoSnp-RAD*: RAD-Seq adaptation of *DiscoSnp++*

94 Originally, *DiscoSnp++* was designed for finding variants from whole genome sequencing data. To
95 adapt to the RAD-Seq context, the core algorithm of *DiscoSnp++* was extended and modified as shown
96 Sections 2.1.1 and 2.1.2. Also, as presented Sections 2.1.3 and 2.1.4, specific features for post-processing
97 were added to the whole pipeline.

98 ***DiscoSnp++* basic algorithm.** We first recall the fundamentals of the *DiscoSnp++* algorithm, which
99 is based on the analysis of the *de Bruijn graph* (DBG) [21], which is a directed graph where the set of

100 vertices corresponds to the set of words of length k (k -mers) contained in the reads, and there is an oriented
 101 edge between two k -mers, say s and t , if they perfectly overlap on $k - 1$ nucleotides, that is to say if the
 102 last $k - 1$ suffix of s equals the first $k - 1$ prefix of t . In this case, we say that s can be *extended* by the last
 103 character of t , thus forming a word of size $k + 1$. A node that has more than one predecessor and/or more
 104 than one successor is called a branching node. Small variants, such as SNPs and INDELS, generate in the
 105 dBG recognizable patterns called “*bubbles*”. A bubble (Fig.1(a)) is defined by one *start* branching node that
 106 has two distinct successor nodes. From these two children nodes, two paths exist and merge in a *stop*
 107 branching node, which has two predecessors. The type of the variant, whether it is a single isolated SNP,
 108 several close SNPs (distant from one another by less than k nucleotides) or an INDEL, determines the
 109 length of each of the two paths of the bubble.

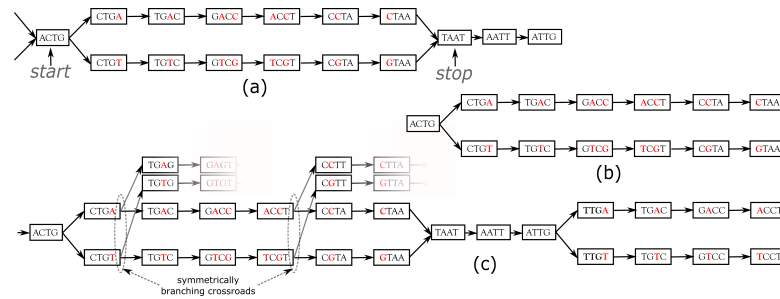


Figure 1. Examples of bubbles detected by SNPs in a toy de Bruijn graph, with $k = 4$. In (a) the bubble is complete: this corresponds to a bubble detected by *DiscoSnp++*. In (b), the bubble is symmetrically truncated: it is composed of a branching node (“ACTG”) whose two successors lead to two distinct paths that both have the same length and such that their last two nodes have no successor. Graph (c) shows an example of two bubbles from the same locus. The leftmost bubble contains two symmetrically branching crossroads.

110 *DiscoSnp++* first builds a dBG from all the input read samples combined, and then detects such
 111 bubbles. Sequencing errors or approximate repeats also generate bubbles, that can be avoided by filtering
 112 out kmers with a too low abundance in the read sets, and by limiting the type or number of branching
 113 nodes along the two paths. Detected bubbles are output as pairs of sequences in fasta format. The second
 114 main step of *DiscoSnp++* consists in mapping original reads from all samples on these sequences, in order
 115 to compute for each variant, its read depth per allele and per read set. From this coverage information,
 116 genotypes are inferred and variants are scored. The final output is a VCF file, where each variant is
 117 associated to a confidence score (the *rank*) and is genotyped in each read set, thanks to its allele coverages
 118 (see [19, 29]).

119 In *DiscoSnp-RAD*, these two main steps have been modified to adapt to the RAD-seq context and
 120 an additional third step has been developed in order to cluster the variants per locus and to output this
 121 information in the final VCF file. In short, *DiscoSnp-RAD 1/* constructs the de Bruijn graph and detects
 122 bubbles whose topology correspond to SNPs or indels, *2/* maps back reads on found bubble sequences,
 123 thus assessing the read coverage per allele and per read set, and *3/* performs clustering on predicted
 124 sequences. Those three steps are described in the three following sections.

125 2.1.1 Bubble detection with *DiscoSnp-RAD*

126 **A novel RAD-specific bubble model.** In *DiscoSnp++*, variants distant from less than k bp from a
 127 genomic extremity could not be detected, as associated bubbles do not open and/or close. This effect
 128 is negligible in the whole genome sequencing context, however, in the RAD-Seq context, sequenced
 129 genomic regions are limited to a hundred or to a few hundreds nucleotides (the read size), and thus a large
 130 amount of variants are likely to be located at the extremities of the loci. For instance, with reads of length
 131 100bp, and $k = 31$ (which is a usual k value), on average 62% of the variants are located in the first or last
 132 k nucleotides of a locus and cannot be detected by *DiscoSnp++*.

133 In the RAD-Seq context, all reads sequenced from the same locus start and end exactly at the same
 134 position. Thus, variants located less than k bp from loci extremities generate what we call *Symmetrically*
 135 *Truncated Bubbles* (Fig.1(b)). Such bubbles start with a node which diverges into two distinct paths that
 136 do not meet back, such that both of them cannot be extended because of absence of successor and both

137 paths have exactly the same length. Symmetrically, a variant located less than k bp apart from loci start
138 generates a bubble that is right closed, but that starts with two unconnected paths of the same length.

139 To further increase specificity of the truncated bubble model, we also constrain the last 3-mer of
140 both paths to be identical. Although this prevents the detection of variants as close as 3 bp from a
141 locus extremity, this enables to identify correctly the type of detected variant. Indeed, when the last
142 L nucleotides of two locus sequences are different, several mutation events could have taken place in
143 the genome resulting in the same observed differences: either an indel (of any size) or L successive
144 substitutions or a combination of the two types. When L is small, all events may be equally parsimonious
145 and we prefer to report none of these instead of a wrong one. Note that this does not prevent to detect
146 loci containing such variants as long as there is at least another variant detected in the locus. The value L
147 was set to 3 because it leads to a relatively low loss of recall (6% with reads of length 100), while the
148 probability of observing by chance three successive matches is low ($= \frac{1}{4^3} \approx 1.56\%$). Note that this issue
149 is also present in any mapping or clustering based approaches.

150 The core of the *DiscoSnp-RAD* algorithm SNP bubble detection is sketched in Algorithm 1. Al-
151 gorithm 1 is intentionally simplified and hides the process enabling to detect SNPs separated by less
152 than k nucleotides and INDELS. The full and detailed algorithm is proposed in supplementary materials.
153 Basically, after the graph construction, we loop over all its branching nodes (line 2), each branching
154 node is then considered as a potential bubble extremity. The pair of paths that can be generated from
155 this branching node are explored (lines 5 to the end). Notably, the two paths are created simultaneously
156 nucleotide by nucleotide. The extension stops 1/ if the extension is impossible (line 10, if there exists no
157 nucleotide α such that $kmer_1$ and $kmer_2$ can be extended with α); or 2/ if the bubble closes (line 11); or
158 3/ if the bubble is truncated (line 7).

159 **Dealing with entangled bubbles.** As RAD-Seq data often include a large number of individuals, this
160 is likely that many SNPs are close to each other (separated by less than k nucleotides), and that a large
161 number of distinct haplotypes co-exist. This situation generates bubbles that are imbricated in one another
162 and what we call “*Symmetrically Branching Crossroads*” (SBCs), as shown in Fig.1(c). SBCs appear
163 when more than one unique character may be used during extension. All possible extensions are explored
164 (line 12) in presence of SBCs. However, we limit the maximal number of traversals of SBCs per bubble
165 to 5 by default (line 14). This value has been chosen as larger values lead to longer computation time,
166 larger false positive calls (due to repetitive genomic regions), while not changing significantly recall, as
167 shown in the results. Depending on the user choice, we also propose a “high_precision” mode in which
168 bubbles containing one or more SBC(s) are not detected.

Algorithm 1 Simplified overview of the *DiscoSnp-RAD* SNP bubble detection (Indel bubble detection omitted)

```
1: Create a de Bruijn graph from all (any number  $\geq 1$ ) read set(s)
2: for Each right branching  $k$ -mer in the graph start do
3:   for each couple of successor  $kmer_1, kmer_2$  of  $k$ -mer start do
4:      $nb\_sym\_branching=0$ 
5:     while True do
6:       Extend  $kmer_1$  and  $kmer_2$  with  $\alpha \in \{A, C, G, T\}$ 
7:       if Both  $kmer_1$  and  $kmer_2$  have no successors then
8:         if last 3 characters from  $kmer_1$  and  $kmer_2$  are equal then
9:           Output bubble and break
10:        if Extension is impossible then break
11:        if  $kmer_1 = kmer_2$  then Output bubble and break
12:        if two or more possible extending nucleotides  $\alpha$  then
13:          Increase  $nb\_sym\_branching$ 
14:          if  $nb\_sym\_branching > 5$  then break
15:        else Explore recursively all possible extensions
```

169 2.1.2 Computing allele coverage and inferring genotypes

170 In this second step, original reads from all samples are mapped on all bubble sequences, in order to provide
171 the read coverage per allele and per read set. Importantly, this mapping step allows non-exact mapping,

172 allowing a high number of substitutions (up to 10 by default), except on the polymorphic positions of
173 the bubble. As shown in results, this choice enables to maximize the sensibility by allowing numerous
174 variations, while maintaining a high precision as no substitution is authorized on variant positions.

175 These coverage information enables to infer individual genotypes and to assign a score (called *rank*) to
176 each variant enabling to filter out potential false positive variants. Genotypes are inferred only if the total
177 coverage over both alleles is above a *min_depth* threshold (by default 3), using a maximum likelihood
178 strategy with a classical binomial model [19, 17], otherwise the genotype is indicated as missing ("./").
179 Variants with too many missing genotypes (by default more than 95 % of the samples) are filtered out.

180 Paralogous genomic regions represent a major issue in population genomic analyses as DNA sections
181 arising from duplication events can be aggregated in the same locus and thus, might encompass alleles
182 coming from non orthologous loci. Allele coverage information across many samples can be used to
183 filter out many of such paralog-induced variants. As the latter tend to occur in all the samples, their allele
184 frequency is thus non discriminant between samples. An efficient scoring scheme, called the *rank* value
185 in *DiscoSnp++*, reflects such discriminant power of variants. First, we define the Phi coefficient of a
186 given variant for a given pair of samples, as $\sqrt{\frac{\chi^2}{n}}$, with χ^2 being the chi-squared statistics computed
187 on the allele read counts contingency table for this pair of samples, and n being the sum of read counts
188 in this table. This is an association measure between two qualitative variables (here allele vs sample)
189 ranging between 0 (no association) and 1 (maximal association). Then, when more than two samples are
190 compared, the rank value is obtained by computing the Phi coefficient of all possible pairs of samples and
191 retaining the maximum value. We have shown in previous work [29, 19] that paralog-induced variants are
192 likely to generate bubbles in the dBG but with very low rank values (< 0.4) contrary to most real variants.
193 This filter is particularly effective when many samples are compared, as in the RAD-seq context. Thus, by
194 default, *DiscoSnp-RAD* discards all variants with such low rank values.

195 Noteworthy, some real variants can also harbour a low rank value: those which are heterozygous in
196 strictly all the samples. Such variants should be rare when hundreds of samples are considered. Moreover,
197 discarding such variants should not impact on most downstream analyses, such as deciphering population
198 structure or inferring phylogenies, as their results are mainly based on variants which can discriminate the
199 samples. Only the estimation of heterozygosity levels may be affected by removing such real variants:
200 they may be slightly under-estimated, but they would be more dramatically over-estimated without any
201 filtering of paralog-induced variants.

202 **2.1.3 Clustering variants per locus**

203 During the bubble detection phase, several independent bubbles can be predicted for the same RAD
204 locus. For instance, Fig.1(c) shows a toy example of a the dBG graph associated to a locus. In this
205 case, *DiscoSnp-RAD* detects two bubbles, that give no sign of physical proximity. In several population
206 genomics analyses, such proximity information can be useful, such as in population structure analyses,
207 where this is recommended to select only one variant per locus. In order to recover this information of
208 locus membership, we developed a post-processing method to cluster predicted variants per locus.

209 The method uses the fact that *DiscoSnp-RAD* is parameterized to output bubbles together with their
210 left and right contexts in the graph, which correspond to the paths starting from each extreme node and
211 ending at the first ambiguity (ie. a node with not exactly one successor). For instance, the leftmost
212 bubble of Fig.1.c is output as sequences ACTG**AC**CTAATg and ACTG**TC**GTAATg, where we represent
213 the context sequences in lower case, and rightmost bubble of the same figure is output as sequences
214 taATTG**AC**CT and taATTG**TC**CT.

215 By definition of these extensions, if a given locus contains several variants, each bubble of this locus
216 extended with its left and right contexts shares at least one $k - 1$ -mer with at least one other so extended
217 bubble of the same locus. For instance, the pairs of sequences of the two bubbles shown Fig.1.c share the
218 $k - 1$ -mer TAA (among others).

219 We exploit this property to group all bubbles per locus. For doing so, we create a graph in which a
220 node is a bubble (represented by its pair of sequences including the extensions), and there is an undirected
221 edge between two nodes N_i and N_j if any of the two sequences of N_i shares at least one $k - 1$ -mer with
222 any of the two sequences of N_j . Those edges are computed using *SRC_linker* [15].

223 Finally, we partition this graph by connected component. Each connected component contains all
224 bubbles for a given locus and this information is reported in the vcf file. By default, clusters containing
225 more than 150 variants are discarded, as they are likely to aggregate paralogous variants from repetitive

226 regions.

227 **2.1.4 Various optional filtering options**

228 The output of *DiscoSnp-RAD* is a VCF file containing predicted variants along with various information,
229 such as their genotypes and allele read counts in all samples, their *rank* value and the cluster ID (locus)
230 they belong to. This enables to apply custom filters at the locus level, as well as any variant level
231 classical RAD-Seq filters (such as the minimal read depth to call a genotype or the minimal minor
232 allele frequency to keep a variant). Several such RAD-seq filtering scripts are provided along with
233 the main program ([https://github.com/GATB/DiscoSnp/tree/master/discoSnpRAD/](https://github.com/GATB/DiscoSnp/tree/master/discoSnpRAD/post-processing_scripts)
234 `post-processing_scripts`).

235 **2.2 Testing environment**

236 The tests were performed on the GenOuest (genouest.org) cluster, on a node composed of 40 Intel
237 Xeon core processors with speed 2.6 GHz and 252 GB of RAM.

238 **2.3 Validation on simulated datasets**

239 Note that all scripts used for simulations and validations are publicly available [https://doi.org/](https://doi.org/10.5281/zenodo.3724518)
240 `10.5281/zenodo.3724518`.

241 **Simulation protocol.** RAD loci from *Drosophila melanogaster* genome (dm6) were simulated by
242 selecting 150 bp on both sides of 43,848 PstI restriction sites resulting in 87,696 loci. Several populations,
243 each composed of several diploid individuals were simulated as follows. For each simulated population,
244 SNPs were randomly generated at a rate of 1%. A first subset of them (70%) was introduced in all samples
245 from the population and represent shared polymorphism. The rest of these SNPs (30%) were distributed
246 between samples by a random picking of 10% of them and assigned to each sample. For each sample,
247 10% of the assigned SNPs, shared and sample specific, are introduced in only one of the homologous
248 chromosomes to simulate heterozygosity. This process was repeated to generate from 5 to 50 populations
249 each composed of 20 individuals. Finally, between 2,109,900 SNPs for 100 samples, and 2,547,337 SNPs
250 for 1,000 samples, were generated. Forward 150bp reads were simulated on right and left loci, with 1%
251 sequencing errors, with 20X coverage per individual (the complete pipeline is given in Supplementary
252 Figure 1).

253 **Evaluation protocol.** For estimating the result quality, predicted variants were localized on the *D.*
254 *melanogaster* genome and output in a vcf file. To do so, we used the standard protocol of *DiscoSnp++*
255 when a reference genome is provided, using BWA-mem [14]. The predicted vcf was compared to the
256 vcf storing simulated variant positions to compute the amount of common variants (true positive or TP),
257 predicted but not simulated variants (false positive or FP) and simulated but not predicted variants (false
258 negative or FN). Recall is then defined as $\frac{\#TP}{\#TP+\#FN}$, and precision as $\frac{\#TP}{\#TP+\#FP}$.

259 **Comparison with other tools.** For comparisons, *STACKS* v2.4 and *IPyRAD* v0.7.30 were run on the
260 simulated datasets. *STACKS* stacks were generated *de novo* (`denovo_map.pl`), with a minimum of 3
261 reads to consider a stack (-m 3). On the simulated dataset composed of 100 samples, five values of the
262 parameter -M governing stack merging (ie. 4, 6, 8, 10, and 12) were tested. On the remaining datasets, the
263 parameter -M was fixed to 6 following r80 method [18]. All other parameters were set to default values.
264 Similarly, *IPyRAD* was run using five values of clustering threshold on the dataset composed of 100
265 samples (ie. 0.75, 0.80, 0.85, 0.90 and 0.95) and then fixed to 0.80, following r80 method [18], for larger
266 datasets. The other parameters have been kept at the default values. Then, *de novo* tags from *STACKS* and
267 loci from *IPyRAD* were mapped to the *D. melanogaster* genome using BWA-mem and variant positions
268 were transposed on the genome positions with a custom script.

269 **2.4 Application to real data from *Chiastocheta* species**

270 **Data origin.** Tests on real data were performed on ddRAD reads previously obtained for the phylogenetic
271 study of seed parasitic pollinators from the genus *Chiastocheta* (Diptera: Anthomyiidae). The dataset
272 corresponds to the sequencing of 259 individuals sampled from 51 European localities generated by
273 Lausanne University, Switzerland [27] (<https://www.ebi.ac.uk/ena/data/view/PRJEB23593>). A total of
274 608,367,380 reads were used for the study with an average of 2.3 Million reads per individual.

275 **Variant prediction and filtering.** *DiscoSnp-RAD* was run with default parameters, searching for at most
276 five variants per bubble. For *IPyRAD* the same parameters as in the Suchan *et al.* study [27] have been
277 applied including a percentage of identity of 75% for the clustering and a minimum coverage of 6. For
278 *STACKS* we applied a minimum coverage by stack (-m) of 3, a maximum number of mismatches allowed
279 among sample (-M) of 8 and a maximum number of mismatches allowed between sample (-n) of 8. On
280 the output vcf from each tools, downstream classical filters were applied to follow as much as possible
281 the filters used in the Suchan *et al.* study [27]: a minimum genotype coverage of 6, a minimal minor
282 allele frequency of 0.01 and a minimum of 60% of the samples with a non missing genotype for each
283 variant. These filters remove less informative variants or those with an allele specific to a very small
284 subset of samples. These filters were also applied at the intra-specific level in one of the seven sampled
285 *Chiastocheta* species, i.e. *C. lophota*, on the same *DiscoSnp++* output.

286 **Population genomic analyses.** The species genetic structure was inferred using STRUCTURE [22]
287 v2.3.4. This approach requires unlinked markers, thus only one variant by locus, randomly selected, has
288 been kept. The STRUCTURE analysis was carried on the datasets generated by each tool. Simulations
289 were performed with genetic cluster number (*K*) set from 1 to 10. Best *K* was identified using Evanno's
290 method [7]. We used 20,000 MCMC iterations after a burn-in period of 10,000. The output is the posterior
291 probability of each sample to belong to each of the possible clusters. For *C. lophota* species, a multivariate
292 analysis were used to investigate intra-specific genetic structure using adegenet R package [12].

293 **Phylogenomic analyses.** Maximum likelihood (ML) phylogenetic reconstruction was performed
294 on a whole concatenated SNP dataset using GTRGAMMA model with the acquisition bias correc-
295 tion [13]. We applied rapid Bootstrap analysis with the extended majority-rule consensus tree stopping
296 criterion and search for best-scoring ML tree in one run, followed by ML search, as implemented in
297 RAxML v8.2.11 [26].

298 3 RESULTS

299 3.1 Results on simulated data

300 *DiscoSnp-RAD* was first run on several simulated RAD-Seq datasets composed of an increasing number of
301 samples (from 100 to 1,000) in order to validate the approach, to evaluate its speed and efficiency and to
302 compare it with the other clustering approaches. This experiment shows that *DiscoSnp-RAD* predictions
303 are accurate with a good compromise between recall and precision (see Figure 2).

304 On average, 84.6 % of the simulated variants are recovered with very few false positive calls, ie.
305 reaching a precision of 98.5 % on average. Importantly, these performances are not impacted by the
306 number of input samples in the dataset. For instance, recall varies from 84.6% to 83.3% between the
307 smallest and the largest datasets (100 vs 1.000 samples), and precision from 98.1 % to 98.5%.

308 By comparison with other tools, for each of the tested population sizes, recall and precision are
309 comparable between tools, with typically a recall lower than *STACKS* and *IPyRAD* and an intermediate
310 precision, lower than *IPyRAD* and higher than *STACKS*. The loss of recall may be explained by the fact
311 that *DiscoSnp-RAD* voluntarily does not detect the variants within 3 bp of each locus end (see Methods).
312 The amount of predicted loci are similar between all tools (Supplementary Table 2). The main differences
313 between the tools concern the run time and the disk space usage. These differences increase with the
314 number of samples in the dataset. For instance, on the largest dataset composed of 1,000 samples
315 *DiscoSnp-RAD* is more than 3 times faster than *STACKS* and more than 5 times faster than *IPyRAD*.
316 Moreover, if we consider the cumulative time required to test different parameters for *STACKS* and
317 *IPyRAD*, i.e. five sets of parameters for each tool, *DiscoSnp-RAD*, without parameter setting is more than
318 15 times faster than *STACKS* and more than 25 times faster than *IPyRAD*. Regarding the disk space used
319 by the tool during the process, *DiscoSnp-RAD* requires only a small amount of space compared to the
320 other tools. Full RAM memory, disk usage, and computation times of *DiscoSnp-RAD* are provided in
321 Supplementary Table 1.

322 **Robustness with respect to parameters.** In *DiscoSnp-RAD*, the main parameter is the size of *k*-mers,
323 used for building the DBG. As shown Figure 3, *DiscoSnp-RAD* results are robust with respect to *k*, the
324 main parameter, and its fine choice is thus not crucial. This figure also highlights the results robustness
325 with respect to other parameters such as the maximal number of predicted SNPs per bubble (5 by default),
326 the maximal number of substitutions authorized when mapping reads on bubble sequences (10 by default),

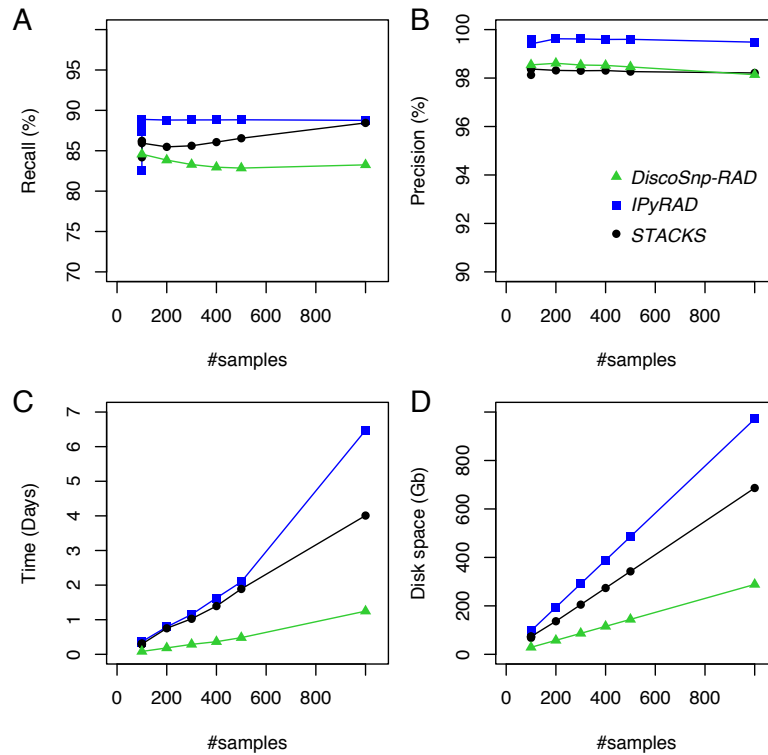


Figure 2. Recall (A), precision (B), time (C) and space (D) evolution on simulated data with different sampling sizes. For the sampling of 100 samples, five parameter sets were tested for *IPyRAD* and *STACKS* (see Material and Methods for details).

327 and the maximal number of symmetrically branching crossroads (also 5 by default). Concerning this last
 328 parameter, Figure 3 also shows the advantages of the “high_precision” mode which sets this parameter to
 329 zero, leading to a precision of nearly 100%.

330 We enriched results presented in this figure with two additional simulated datasets, following the
 331 protocol presented in the “Simulation protocol” section, but introducing SNPs at rates of 0.5% and 2%,
 332 instead of 1%. Results are presented in the Supplementary Figure 2. They also highlight the robustness of
 333 results and the rationale for the default parameter choices with two times more and two times less simulated
 334 diversity.

335 The robustness of *DiscoSnp-RAD* is an important point as other state-of-the-art tools are extremely
 336 sensible to their parameters, especially those directly linked to the expected sequence divergence, and
 337 require time consuming processes to set them properly [25].

338 3.2 Results on real data

339 In this section, we present an application of the *DiscoSnp-RAD* implementation on ddRAD sequences
 340 obtained from the anthomyiid flies from the *Chiastocheta* genus. In this genus, classical mitochondrial
 341 markers are not suitable for discriminating the morphologically described species [6]. Although RAD-
 342 sequencing dataset phylogenies supported the species assignment [27], the interspecific relationships
 343 between the taxa could not be resolved with high confidence due to high levels of incongruences in gene
 344 trees [9, 27]. The dataset is composed of 259 sequenced individuals from 7 species. Results obtained on
 345 *DiscoSnp-RAD* were compared to the prior work of Suchan and colleagues, based on *pyRAD* analysis [27].
 346 In addition, we provide a performance benchmark of *STACKS*, *IPyRAD* and *DiscoSnp-RAD* run on this
 347 dataset.

348 **Recovering all *Chiastocheta* species.** Variant calling was run on the 259 *Chiastocheta* samples with
 349 *DiscoSnp-RAD*. Before filtering, 115,920 SNPs were identified. After filtering, 4,364 SNPs, located in

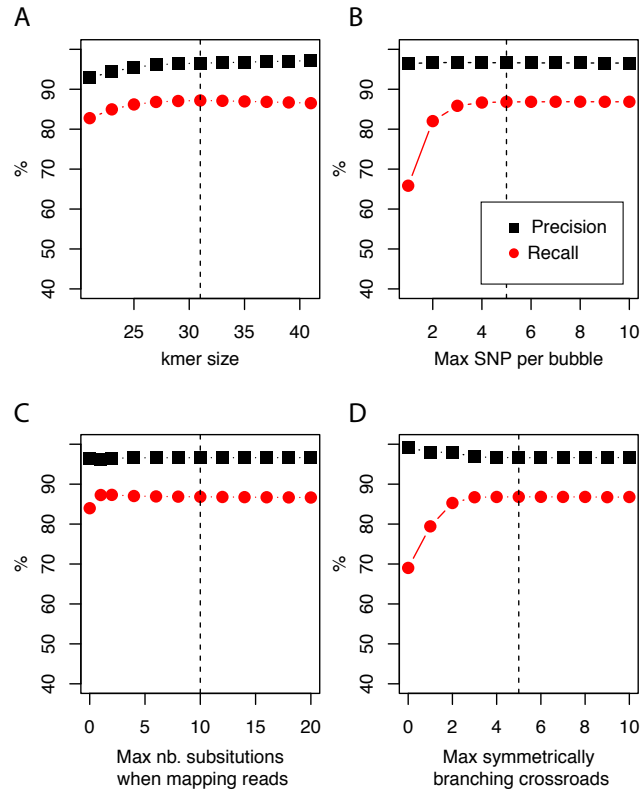


Figure 3. Recall and precision on simulated data of 100 samples using *DiscoSnp-RAD* with respect to **A.** k -mer sizes, **B.** maximal number of authorized SNP per bubble, **C.** maximal number of authorized substitutions while mapping reads on predicted variants sequences, and **D.** maximal number of symmetrically branching crossroads. Dashed vertical line represents on each plot the chosen default value.

350 1,970 clusters, were retained and used for population genomic analyses. The total number of clusters is
 351 coherent with the 1,672 loci from Suchan *et al.* [27].

352 Then, following the requirements of the STRUCTURE algorithm, only one variant per cluster was
 353 retained, resulting in a dataset composed of 1,970 SNPs. The most likely value of K is 7 (Supplementary
 354 Figure 3) and corresponds to the seven species described in [27]. STRUCTURE successfully assigned
 355 samples to the seven species, consistent with the morphological species assignment and previously
 356 published results [27] (Fig.4). The assignment values represent the probability with which STRUCTURE
 357 assigns a sample to a cluster, depending on the information carried by the variants. The assignment values
 358 are high with an average of 0.992 (sd 0.022) across samples and a minimum assignment of 0.810. These
 359 values are comparable to the assignment values obtained by Suchan *et al.* [27] with an average of 0.977
 360 (sd 0.042) and a minimum of 0.685. Genetic structure has also been investigated for the two other tools
 361 and give very similar population assignments (Supplementary Figure 4).

362 The phylogeny realized with RAxML on the 4,364 SNPs obtained after filtering, identifies clearly
 363 seven clusters representative of the seven species which are coherent with the clusters obtained by Suchan
 364 and colleagues [27] (Fig.4). The internal branches separating the seven species are well supported by high
 365 bootstrap values.

366 **Recovering phylogeographic patterns.** To assess the utility of *DiscoSnp-RAD* results for investigating
 367 the intra-specific genetic structure, we then focused the analysis on 40 samples of *C. lophota* species.
 368 From the same vcf file obtained with the 259 samples, the 40 *C. lophota* samples were extracted and the
 369 same filters, i.e. MAF, missing etc., were applied on this *C. lophota* dataset.

370 We obtained 1,306 SNPs by selecting one variant per locus extracted from 4,364 variants identified in
 371 this species. The multivariate analysis of this dataset identify three populations comprising respectively

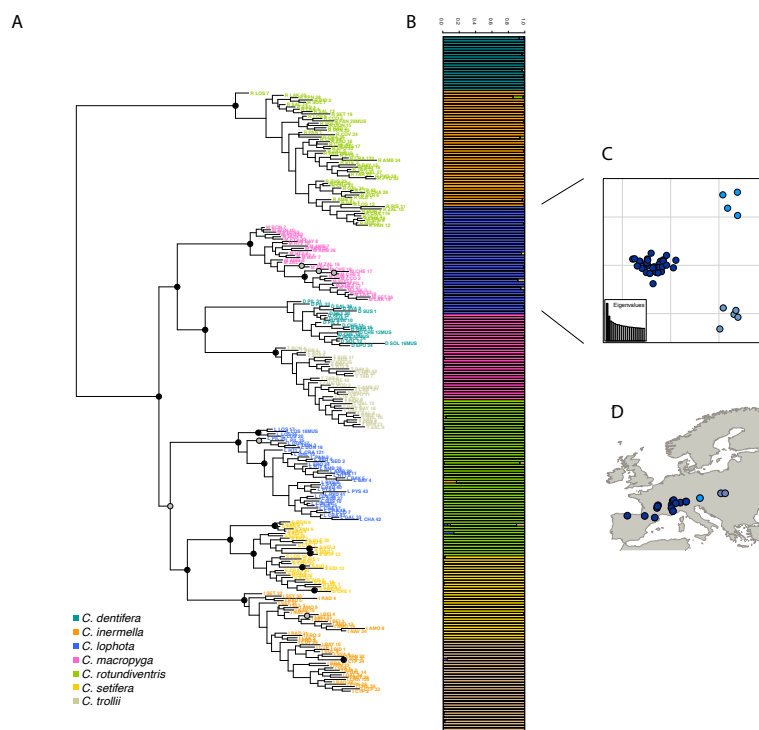


Figure 4. **A.** RAxML phylogeny realized on all variants predicted by *DiscoSnp-RAD*. Bootstrap node supports > 80 are shown denoted by gray dots, bootstrap node supports > 90 are shown denoted by black dots. **B.** STRUCTURE results obtained with SNP only and all variants on the seven *Chiastocheta* species. **C.** Plot of the two first PC from a multivariate analysis on *C. lophota* samples and, **D.** their geographic distribution (figure made with *Natural Earth* [16]).

372 31, 5 and 4 samples. (Fig.4). Notably, the genetic structure follows the geographic distribution of the
 373 samples, with samples from one population originating from western locations, another population from
 374 eastern locations and an intermediate population. Geographic structuring is the most frequent structuration
 375 factor observed in population genetics, pointing to the geographical isolation of divergent lineages. This
 376 clear geographic structuring is another hint that *DiscoSnp-RAD* recovers real biogeographic signal.

377 **Breakthrough in running time.** *DiscoSnp-RAD* run on the 259 *Chiastocheta* samples took about 30
 378 hours. This comprises the whole process from building the dBG to obtaining the final filtered vcf file
 379 with 1 SNP per locus. To compare the *DiscoSnp-RAD* performances with *STACKS* and *IPyRAD* on real
 380 data, we ran each of these tools using default parameters on the 259 *Chiastocheta* samples and measured
 381 running time and maximum memory usage. The difference is remarkable, *DiscoSnp-RAD* is more than
 382 4.65 times faster than *STACKS* (running time 138 hours) and 2.8 times faster than *IPyRAD* (running time
 383 82 hours) to perform the whole process. Moreover, contrary to *DiscoSnp-RAD*, *STACKS* and *IPyRAD*
 384 should be run several times to explore the parameters which represent a considerable amount of time and
 385 memory. For instance, in Suchan *et al.* [27], *IPyRAD* was run with 5 different combinations of parameter
 386 values, *DiscoSnp-RAD* being thus 14 times faster than *IPyRAD*.

387 4 DISCUSSION

388 ***DiscoSnp-RAD* efficiency.** *DiscoSnp-RAD* produced relevant results on ddRAD data from *Chiastocheta*
 389 species. SNPs identified allowed us to successfully i) distinguish the seven species based on the STRUC-
 390 TURE algorithm, and ii) reconstruct the phylogenetic tree of the genus, coherent with the phylogenies
 391 previously published [27]. Moreover, on the intra-specific scale, we obtained geographically meaningful
 392 results within *C. lophota* species. The variants identified by *DiscoSnp-RAD* can be used to study the
 393 species or population genetic structure and could be used to investigate deeper the mechanisms at the

394 origin of this structure such as potential gene flow between populations or their demographic histories. In
395 addition, *DiscoSnp-RAD* is also able to identify INDELS [19]. They were not used in this study but are
396 available for users.

397 Furthermore, the use of *DiscoSnp-RAD* presented considerable advantages in the run-time, and
398 parameters choice, compared to other common *de novo* RAD analysis tools, as described below.

399 **Run-time.** The use of *DiscoSnp-RAD* dramatically decreased the overall time for discovering and
400 selecting relevant variants, as compared to other tools. This is made possible thanks to the use of a unique
401 indexing data structure, the dBG built from all the reads. To build this graph, reads do not need to be
402 compared to each other. *DiscoSnp-RAD* speed depends on the graph size and at a lesser extend on the
403 number of reads. Importantly, it is not expected to increase quadratically with the dataset size. This can
404 likely be anticipated that with the drop of sequencing costs, RAD projects will grow in size, either by
405 using higher frequency cutting enzymes to obtain a dense genome screening, by increasing the sequencing
406 depth to compensate sequencing variation or by increasing the number of samples. In this context,
407 *DiscoSnp-RAD* will more easily scale to such very large datasets.

408 **Easy parameter choice.** Another substantial advantage of using *DiscoSnp-RAD* is the fact that param-
409 eters are not directly linked to the level of expected divergence of the compared samples. In fact, they
410 impact the number and type of detected variants, but are not related to the subsequent clustering step. As
411 a result, same parameters can be used whatever the type of analysis (for example, intra or inter-specific),
412 contrasting with classical tools in which parameters govern loci recovering. Indeed, in *STACKS*, the
413 parameters governing the merge of the stacks can compromise the detection of relevant variants if they are
414 not adapted to the studied dataset [25]. Therefore, the authors recommend to perform an exploration of
415 the parameter space before downstream analyses [18]. This is extremely time consuming, up to one month
416 as confessed by the authors [23], and may not always result in interpretable conclusions. In *IPyRAD*, the
417 similarity parameter for clustering also impacts variant detection, and usually several values have to be
418 tested to choose the best, as exemplified by Suchan and colleagues who tested five different values [27].

419 **By-locus assembly.** *DiscoSnp-RAD* output is a vcf file including pseudo-loci information, that allows
420 the application of standard variant filtering pipelines. One next objective is to recover loci consensus
421 sequences, that could be used for phylogenetic analysis based on full locus sequences. This could be
422 achieved by performing local assemblies per individual, from all bubbles contained in a cluster.

423 **Sequencing error rate.** Our tests on simulated data sets, performed with 1% error rate, show that, in
424 this worst case scenario, *DiscoSnp-RAD* can deal with a high error rate and can thus afford analyses
425 of data not generated with the most recent sequencing technologies. With a lower error rate, the good
426 performances of *DiscoSnp-RAD* are confirmed as shown by the results obtained on the real *Chiastocheta*
427 dataset. The breakthrough in running time with respect to the other approaches is slightly reduced with
428 the real dataset and this could be due to *STACKS* and *IPyRAD* being more impacted by sequencing errors
429 in the data.

430 **Potential applications.** *DiscoSnp-RAD* can handle all types of RAD data including original RAD-Seq,
431 GBS, ddRAD, etc. In addition it is able to use reads 2 from original RAD-Seq data that are often difficult
432 to analyse. These reads do not start and finish at the same position. Properly recovery of loci is therefore
433 not possible with read stacking approaches. This problem does not exist when using *DiscoSnp-RAD*,
434 and variants present in reads 2 can also be called. Indeed, the *DiscoSnp-RAD* method, being not based
435 on stacks of reads, is able to detect any variants that generate bubble motifs in the dBG, thus even if
436 present in reads whose starting positions differ. More precisely, if in a given locus some reads from reads
437 1 overlap some reads from reads 2 over at least $k - 1$ characters, then all variants from this locus are
438 clustered together and hence detected as belonging to the same locus. Conversely, variants detected from
439 reads 1 are not related to variants detected from reads 2.

440 This ability of *DiscoSnp-RAD* to handle reads that do not necessary start at the same genomic position
441 makes it particularly well suited to analyze the datasets produced by another group of genome-reduction
442 techniques, namely sequence capture approaches [10]. In these techniques, DNA shotgun libraries are
443 subject to enrichment using short commercially-synthesized [8] or in-house made [28] DNA or RNA
444 fragments acting as 'molecular baits', that hybridize and allow separation of homologous fragments
445 from genomic libraries. One of such promising approaches is HyRAD, a RAD approach combining
446 the molecular probes generated using ddRAD technique and targeted capture sequencing, designed for

447 studying old and/or poor quality DNA, likely to be too fragmented for RAD-sequencing [28]. In HyRAD,
448 capturing randomly fragmented DNA results in reads not strictly aligned and covering larger genomic
449 regions than RAD-Seq. Therefore RAD tools can not be used to reconstruct such loci, and the current
450 analysis consists in building loci consensus from reads, and then calling variants by mapping back the
451 reads on it. The use of *DiscoSnp-RAD* should simplify this process in a single *de novo* calling step, well
452 adapted to the specificities of data generated by reduction approaches: many compared samples, high
453 polymorphism and clustering by locus.

454 5 CONCLUSION

455 We propose *DiscoSnp-RAD*, an original method dedicated to the *de-novo* analyse of RAD-Seq data. We
456 have shown that on simulated data, the quality of the results is comparable to those obtained by state-of-the
457 art tools, *STACKS* and *IPYRAD*. On real data, *DiscoSnp-RAD* provides relevant results, enabling the
458 structuring at inter- and intra-level species, accurate enough for recovering the phylogeographic patterns.

459 Due to its methodological approach which is utterly different from existing methods, *DiscoSnp-RAD*
460 drastically reduces computation times and memory requirements. Another key difference stands in the
461 fact that *DiscoSnp-RAD* does not rely on fine tuning of parameters, contrary to existing methods that rely
462 on critical parameters, as those related to the input sequence similarity.

463 ACKNOWLEDGMENTS

464 Authors thank Camille Marchet for her precious help on the clustering implementation. Computations
465 have been made possible thanks to the resources of the Genouest infrastructures. This work was supported
466 by the French ANR-14-CE02-0011 SPECREP grant.

467 REFERENCES

- 468 [1] Andrews, K., Good, J., R Miller, M., Luikart, G., and A Hohenlohe, P. (2016). Harnessing the power
469 of radseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, 17:81–92.
- 470 [2] Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an analysis
471 tool set for population genomics. *Molecular Ecology*, 22(11):3124–3140.
- 472 [3] Eaton, D. A. R. (2014). Pyrad: assembly of de novo radseq loci for phylogenetic analyses. *Bioinfor-*
473 *matics*, 30(13):1844–1849.
- 474 [4] Eaton, D. A. R. and Overcast, I. (2020). ipyrad: Interactive assembly and analysis of RADseq datasets.
475 *Bioinformatics*. btz966.
- 476 [5] Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E.
477 (2011). A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. *PLOS*
478 *ONE*, 6(5):1–10.
- 479 [6] Espíndola, A., Buerki, S., and Alvarez, N. (2012). Ecological and historical drivers of diversification
480 in the fly genus *Chiastocheta pokorny*. *Molecular Phylogenetics and Evolution*, 63(2):466–474.
- 481 [7] EVANNO, G., REGNAUT, S., and GOUDET, J. (2005). Detecting the number of clusters of individuals
482 using the software structure: a simulation study. *Molecular Ecology*, 14(8):2611–2620.
- 483 [8] Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., and Glenn, T. C.
484 (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary
485 timescales. *Systematic Biology*, 61(5):717–726.
- 486 [9] Gori, K., Suchan, T., Alvarez, N., Goldman, N., and Dessimoz, C. (2016). Clustering genes of
487 common evolutionary history. *Molecular Biology and Evolution*, 33(6):1590–1605.
- 488 [10] Grover, C. E., Salmon, A., and Wendel, J. F. (2012). Targeted sequence capture as a powerful tool for
489 evolutionary analysis I. *American Journal of Botany*, 99(2):312–319.
- 490 [11] Hoffberg, S. L., Kieran, T. J., Catchen, J. M., Devault, A., Faircloth, B. C., Mauricio, R., and Glenn,
491 T. C. (2016). Radcap: sequence capture of dual-digest radseq libraries with identifiable duplicates and
492 reduced missing data. *Molecular Ecology Resources*, 16(5):1264–1278.
- 493 [12] Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinfor-*
494 *matics*, 24(11):1403–1405.
- 495 [13] Leaché, A. D., Banbury, B. L., Felsenstein, J., de Oca, A. N.-M., and Stamatakis, A. (2015). Short
496 tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring snp phylogenies.
497 *Systematic Biology*, 64(6):1032–1047.

- 498 [14] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform.
499 *Bioinformatics*, 25(14):1754–1760.
- 500 [15] Marchet, C., Limasset, A., Bittner, L., and Peterlongo, P. (2016). A resource-frugal probabilistic
501 dictionary and applications in (meta)genomics. *CoRR*, abs/1605.08319.
- 502 [16] Natural Earth contributors (2020). Natural earth data web page. <https://www.naturalearthdata.com/>. [Online; accessed 4-May-2020].
- 503 [17] Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from
504 next-generation sequencing data. *Nature Reviews Genetics*, 12:443–451.
- 505 [18] Paris, J. R., Stevens, J. R., and Catchen, J. M. (2017). Lost in parameter space: a road map for stacks.
506 *Methods in Ecology and Evolution*, 8(10):1360–1373.
- 507 [19] Peterlongo, P., Riou, C., Drezen, E., and Lemaitre, C. (2017). Discosnp++: de novo detection of
508 small variants from raw unassembled read set(s). *bioRxiv*.
- 509 [20] Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double digest
510 radseq: An inexpensive method for de novo snp discovery and genotyping in model and non-model
511 species. *PLOS ONE*, 7(5):1–11.
- 512 [21] Pevzner, P. A., Tang, H., and Tesler, G. (2004). De novo repeat classification and fragment assembly.
513 *Genome Research*, 14(9):1786–1796.
- 514 [22] Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using
515 multilocus genotype data. *Genetics*, 155(2):945–959.
- 516 [23] Rochette, N. C. and Catchen, J. M. (2017). Deriving genotypes from RAD-seq short-read data using
517 Stacks. *Nature Protocols*, 12(12):2640–2659.
- 518 [24] Rochette, N. C., Rivera-Colón, A. G., and Catchen, J. M. (2019). Stacks 2: Analytical methods for
519 paired-end sequencing improve radseq-based population genomics. *Molecular Ecology*, 28(21):4737–
520 4754.
- 521 [25] Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., and Wolf, J. B. W.
522 (2017). Bioinformatic processing of rad-seq data dramatically impacts downstream population genetic
523 inference. *Methods in Ecology and Evolution*, 8(8):907–917.
- 524 [26] Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large
525 phylogenies. *Bioinformatics*, 30(9):1312–1313.
- 526 [27] Suchan, T., Espíndola, A., Rutschmann, S., Emerson, B. C., Gori, K., Dessimoz, C., Arrigo, N.,
527 Ronikier, M., and Alvarez, N. (2017). Assessing the potential of rad-sequencing to resolve phylogenetic
528 relationships within species radiations: The fly genus *Chistocheta* (Diptera: Anthomyiidae) as a case
529 study. *Molecular Phylogenetics and Evolution*, 114:189–198.
- 530 [28] Suchan, T., Pitteloud, C., Gerasimova, N. S., Kostikova, A., Schmid, S., Arrigo, N., Pajkovic, M.,
531 Ronikier, M., and Alvarez, N. (2016). Hybridization capture using rad probes (hyrad), a new tool for
532 performing genomic analyses on collection specimens. *PLOS ONE*, 11(3):1–22.
- 533 [29] Uricaru, R., Rizk, G., Lacroix, V., Quillery, E., Plantard, O., Chikhi, R., Lemaitre, C., and Peterlongo,
534 P. (2015). Reference-free detection of isolated SNPs. *Nucleic Acids Research*, 43(2):1–11.
- 535 [30] Wang, S., Meyer, E., McKay, J., and Matz, M. (2012). 2b-rad: A simple and flexible method for
536 genome-wide genotyping. *Nature Methods*, 9:808–10.
- 537