An evaluation of pool-sequencing transcriptome-based exon capture for population genomics in non-model species

Emeline Deleury[1], Thomas Guillemaud[1], Aurélie Blin[1] & Eric Lombaert[1]

[1] INRA, CNRS, Université Côte d'Azur, ISA, France

Corresponding author:

Emeline Deleury

Institut Sophia Agrobiotech - 400 Route des Chappes - BP 167 – 06 903 Sophia Antipolis cedex - FRANCE

E-mail: emeline.deleury@inra.fr

Phone: +33 4 92 38 64 24

Keywords: target enrichment, non-model organism, population genomics, pool-sequencing, *Harmonia axyridis*, intron-exon boundary prediction

Abstract

Exon capture coupled to high-throughput sequencing constitutes a cost-effective technical solution for addressing specific questions in evolutionary biology by focusing on expressed regions of the genome preferentially targeted by selection. Transcriptome-based capture, a process that can be used to capture the exons of non-model species, is use in phylogenomics. However, its use in population genomics remains rare due to the high costs of sequencing large numbers of indexed individuals across multiple populations. We evaluated the feasibility of combining transcriptome-based capture and the pooling of tissues from numerous individuals for DNA extraction as a cost-effective, generic and robust approach to estimating the variant allele frequencies of any species at the population level. We designed capture probes for ~5 Mb of randomly chosen *de novo* transcripts from the Asian ladybird *Harmonia axyridis* (5,717 transcripts). We called ~300,000 bi-allelic SNPs for a pool of 36 non-indexed individuals. Capture efficiency was high, and pool-seq was as effective and accurate as individual-seq for detecting variants and estimating allele frequencies. Finally, we also evaluated an approach for simplifying bioinformatic analyses by mapping genomic reads directly to targeted transcript sequences to obtain coding variants. This approach is effective and does not affect the estimation of SNP allele frequencies, except for a small bias close to some exon ends. We demonstrate that this approach can also be used to predict the intron-exon boundaries of targeted *de novo* transcripts, making it possible to abolish genotyping biases near exon ends.

36 # Introduction

37

38 One of the core challenges in population genomics and phylogenetics is obtaining a reliable set of
39 orthologous loci from a sufficient number of individuals across multiple populations and from
40 species spanning a range of divergences, respectively. The targeted resequencing of a consistent
41 subset of genomic regions represents a reproducible and cost-effective technical solution, with
42 much lighter data handling than for whole-genome sequencing. Hybridization capture methods are
43 among the available reduced representation methods available for the enrichment of genomic DNA
44 samples in a preselected set of hundreds to thousands of genes or DNA fragments (Hodges *et al.*,
45 2007). The most common application of hybridization sequence capture is exon capture, in which
46 only coding regions of the genome (and their flanking regions) are sequenced (Warr *et al.*, 2015).
47 This approach provides focused information about gene function and adaptation, and, unlike
48 transcriptome sequencing, is not biased by variations of gene expression.

49 Exome capture requires a knowledge of the exon sequences of the species studied, for the
50 design of appropriate oligonucleotide probes. Consequently, its use has been largely limited to
51 model species (i.e. species for which a reference genome is available). Absolute probe specificity is
52 not required, however, and given the tendency of functional elements to be conserved even in after
53 millions of years of divergence, exons can be captured from a non-model species (i.e. a species
54 without a published genome sequence) with probes designed from the sequence of a related model
55 species (e.g. Cosart *et al.*, 2011; Förster *et al.*, 2018). This approach has been used for the successful
56 resolution of species phylogenies (e.g. Ilves & López-Fernández, 2014; Bossert & Danforth, 2018;
57 Ilves *et al.*, 2018). It is also an effective method for capturing and enriching degraded and
58 contaminated ancient DNA from extinct groups in paleontology (e.g. Castellano *et al.*, 2014).

59 An alternative approach to capture probe design for non-model species involves the use of
60 a *de novo* assembled transcriptome for the species studied or a related species (e.g. Bi *et al.*, 2012;
61 Neves *et al.*, 2013). One of the limitations of this approach is that the intron-exon structure is
62 unknown, and probes based on transcript sequences spanning two consecutive exons therefore
63 hybridize only partially to genomic DNA, reducing capture efficiency (e.g. Neves *et al.*, 2013). One
64 way of addressing this issue is to use orthologous genes from closely related model species to
65 predict intron positions (Bi *et al.*, 2012; Stephens *et al.*, 2015; Bragg *et al.*, 2016). However, the use
66 of data from closely related model species biases target choice through the retention of genes
67 displaying the highest degree of conservation throughout evolution, which is potentially
68 problematic in population genomics studies. Even when orthologous genes are targeted,
69 transcriptome-based exon capture studies have been found to have a lower capture efficiency close
70 to intron-exon boundaries, because it is not possible to design capture probes spanning exons and
71 their (unknown) non-coding flanking regions (Bi *et al.*, 2012; Neves *et al.*, 2013; Portik *et al.*, 2016).
72 Captured fragments are usually longer than probes. Transcriptome-based exon capture
73 therefore provides opportunities to capture the flanking regions of the exons, thereby expanding
74 gene models beyond the available coding sequence (Bi *et al.*, 2012; Neves *et al.*, 2013; Stephens *et*
75 *al.*, 2015; Bragg *et al.*, 2016; Portik *et al.*, 2016). This requires the construction, in a fragmented way
76 (e.g. long introns not fully captured), of the genomic sequences corresponding to the targeted
77 transcripts (e.g. Neves *et al.*, 2013). This assembly (generally referred to as an in-target assembly) is
78 used as a reference for the mapping of genomic reads or for phylogeny construction. This assembly
79 is obtained by cleaning the raw genomic sequences and assembling them with various k-mer
80 combinations. The genomic contigs obtained are then compared with the target sequences for
81 ordering and scaffolding with each other when they correspond to the same targets. Such a complex
82 approach is necessary for phylogeny construction when probes are used to capture orthologous

83 targets from related species, but may be dispensable in the context of population genomics. Indeed,
84 several studies have shown that the assembly process does not completely rebuild all targets (e.g.
85 Bragg *et al.*, 2016; Portik *et al.*, 2016), and this may decrease the likelihood of detecting variants.
86 Bragg et al. (2016) tried to attenuate this problem by mapping the genomic reads of the species
87 used for probe design directly onto the transcript target sequences, thereby making it possible to
88 cover exons that were not successfully assembled but were nevertheless captured and sequenced.
89 Transcriptome-based exon capture has proved effective for the generation of thousands of
90 highly informative markers. Most studies to date have aimed to resolve the phylogenetic
91 relationships between various set of related species (Nicholls *et al.*, 2015; Stephens *et al.*, 2015;
92 Bragg *et al.*, 2016; Heyduk *et al.*, 2016; Ilves *et al.*, 2018), and applications in population genomics
93 remain rare (but see Bi *et al.*, 2013; Rellstab *et al.*, 2019). This is probably because population studies
94 traditionally analyze large numbers of individuals from multiple populations of non-model species,
95 which is complex, time-consuming and costly, even if sequencing capacity and costs are optimized
96 by indexing individuals for multiplexing (Shearer *et al.*, 2012). However, many research questions in
97 population genomics can be addressed by simply using the allele frequencies calculated at
98 population level (e.g. Gutenkunst *et al.*, 2009). In this context, pooling individuals within the
99 population before the extraction of DNA may be an efficient, cheap and time-saving strategy. Pool-
100 sequencing has been shown to estimate allele frequencies within populations as accurately as
101 individual genotyping, but with much less effort for library construction and sequencing (e.g. Gautier
102 *et al.*, 2013; Schlötterer *et al.*, 2014). Allele frequencies have been estimated by a combination of
103 targeted capture methods and pool-sequencing strategies only in humans (Bansal *et al.*, 2011; Day-
104 Williams *et al.*, 2011; Ramos *et al.*, 2012; Ryu *et al.*, 2018) and, very recently, in pine trees (Rellstab
105 *et al.*, 2019).

106 We propose here an efficient way of quantifying the single-nucleotide polymorphism (SNP)
107 of exomes at the population level in non-model species for which no genome sequence is available.
108 This approach is based on a combination of transcriptome-based exon capture and the pooling of
109 tissues from numerous individuals before DNA extraction. We report the development of capture
110 probes for the Asian ladybird *Harmonia axyridis* from its *de novo* transcriptome, without the use of
111 the available draft genome (HaxR v1.0; Gautier *et al.*, 2018) and, thus, in the absence of known gene
112 structure, as in most non-model species. We used these probes for two target enrichment
113 experiments with the same individuals, the first based on capture with indexed multiplexed
114 individuals, and the second based on capture with a pool of non-indexed individuals. We evaluated
115 the efficiency of capture and the accuracy of the estimates of allele frequencies obtained by pool
116 sequencing in this context. We also evaluated a rapid approach for the processing of read data
117 without their assembly, based on the mapping of reads directly to targeted *de novo* transcript
118 sequences for the detection of variant loci (Bragg *et al.*, 2016; Rellstab *et al.*, 2019). This direct
119 mapping also makes it possible to effectively predict the locations of the intron-exon boundaries of
120 target sequences.
121
122

123 # Methods

124

125 ***Design of capture probes for target enrichment***

126 We designed capture probes for ~5 Mb of randomly chosen coding exonic regions of *H. axyridis*.
127 Using a *de novo* transcriptome obtained by RNA-Seq on various tissues and conditions/stages of the
128 species (Vogel *et al.*, 2017), we sought putative peptide-coding sequences with FRAMEDP (v1.2.2,
129 Gouzy *et al.*, 2009), using bLASTX results (e-value ≤ $1e^{-7}$) of transcripts against the insect proteomes

130 of *Tribolium castaneum*, *Anoplophora glabripennis, Dendroctonus ponderosae, Drosophila*
131 *melanogaster* and the SwissProt database (v2016-02-17) for training. The corresponding partial or
132 complete coding sequences (CDS) were filtered to eliminate (i) *Wolbachia* and other putative
133 endosymbiont sequences, (ii) CDS with more than 1% missing data (*Ns*) or with more than four
134 consecutive *Ns*, (iii) CDS with a GC% below 25 or above 75 and (iv) CDS with lengths below 400 bp
135 (to avoid the capture of too great a proportion of flanking regions and to allow better tiling) or above
136 3500 bp (to make it possible to target larger numbers of different genes). We identified 28,326
137 putative CDS, of which 12,739 were retained after the application of the filters. We then chose 5.5
138 Mb at random, corresponding to 5,736 CDS. We searched for sequence similarity with complete
139 proteins from the SwissProt database and the *D. melanogaster* and *T. castaneum* proteomes with
140 BlastX, and at least 895 CDS (15.6%) were found to be complete. Overall, 5,120 CDS (89.3%)
141 displayed hits with *T. castaneum* proteins (e-value ≤ $1e^{-7}$).

142 The final probes for the 5,736 selected CDS were designed and produced by NimbleGen. The
143 repetitiveness of the target sequences was assessed on the basis of the frequency of 15-mers in the
144 genomes of the coleopterans *T. castaneum* and *A. glabripennis*. Probes with more than one close
145 match, defined as no more than five single-base insertions, deletions or substitutions with the
146 SSAHA algorithm (Ning *et al.*, 2001), in the *H. axyridis de novo* transcriptome or in the draft genome
147 of *A. glabripennis* were discarded. Probes matching to sequences in the *H. axyridis* or *T. castaneum*
148 mitochondrial genome were also discarded. The few residual *Ns* in target sequences were replaced
149 with a random nucleotide to allow tile probing across a single unknown base. The final probe set
150 contained 6,400 regions of overlapping probes, corresponding to a targeted exome size of 5,347,461
151 bases distributed over 5,717 of our randomly selected CDS (Zenodo
152 https://doi.org/10.5281/zenodo.2598388). This probe set was used for the preparation of a SeqCap
153 EZ Developer assay by Roche NimbleGen Technologies, which synthesized the probes as biotinylated

154 DNA oligomers. One capture reaction contained 2,100,000 overlapping probes of 50 to 99 bp in
155 length, with a mean length of 74.71 ± 4.92 bp.
156
157 ***Individual and pooled genomic DNA library preparation***
158 Thirty-six individuals of *H. axyridis* from a wild population in China (Beijing; 40.057128°N,
159 116.53989°E) were collected in October 2015, immediately placed in RNAlater solution and stored
160 at -20°C. Based on the PIF tools proposed by Gautier *et al.* (2013), we would theoretically expect a
161 similar precision of allele frequency estimation for the sequencing of 23 individually indexed
162 individuals, each at 80X coverage, or a single pool of 36 individuals at 125X, assuming an
163 experimental error (i.e. ε, the coefficient of variation of the contribution of each individual to the
164 set of sequence reads obtained from the DNA pool) of 50%. We thus extracted DNA (i)
165 independently from 23 individuals, using half the body (without the hind legs), and (ii) from a pool
166 of 36 individuals, including the 23 individuals used for individual sequencing, using two hind legs per
167 individual. All DNA extractions were performed with the Qiagen DNeasy Blood & Tissue kit, in
168 accordance with the manufacturer's recommendations.
169 Genomic libraries were prepared with NimbleGen SeqCap EZ Library (v5.0). Briefly, for each
170 of the 24 samples (1 pool and 23 individuals), DNA (2 µg in 100 µl) was mechanically sheared to a
171 mean fragment length of 200 bp with a Covaris S2 E210 device (6x30s). Fragments were end-
172 repaired, A-tailed and indexed (one index per sample) with KAPA Library Preparation Kit Illumina
173 platforms. After the ligation of Illumina adaptors and indexes, only fragments with a length of 250
174 to 450 bp were retained. PCR was performed (7 cycles) with standard Illumina paired-end primers
175 and the amplicons were then purified with AMPure XP beads (Beckman). The length, quality and
176 concentration of the prepared DNA fragments were checked on a BioAnalyzer (Agilent High
177 Sensitivity DNA Assay) and a Qubit.

178

179 ***Sequence capture hybridization and sequencing***

180 The 23 indexed individuals were pooled at equimolar concentrations before hybridization. For each
181 capture (one with the pool and one with the indexed individuals), a total of 1 µg of amplified DNA
182 was used for exome enrichments, with the SeqCap EZ Developer probes described above, strictly in
183 accordance with the NimbleGen SeqCap EZ Library protocol (v5.0). For each capture, two parallel
184 post-capture PCRs (14 cycles) were performed on elution solution, and the PCR products were then
185 merged before purification with AMPure XP beads. The length, quality and concentration of the final
186 DNA fragments were checked on a BioAnalyzer (Agilent High Sensitivity DNA Assay) and a Qubit.
187 Each capture was sequenced on a half lane of an Illumina HiSeq3000 sequencer (on the same lane),
188 in accordance with the manufacturer's instructions, in paired-end mode, for 150 cycles. The data
189 were then demultiplexed and exported as FastQ files, and the 24 libraries were processed
190 independently.

191

192 ***SNP calling by the mapping of reads to the CDS***

193 Sequence quality was checked with FastQC
194 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adapters were removed and low-
195 quality base pairs were trimmed with Trimmomatic (v0.35, Bolger *et al.*, 2014), with the following
196 parameters: ILLUMINACLIP:TruSeq-file.fa:2:30:10 LEADING:30 TRAILING:30
197 SLIDINGWINDOW:10:30 MINLEN:75, and FASTX (v0.0.14; http://hannonlab.cshl.edu/fastx_toolkit/)
198 to remove the first five bases.

199 We chose to align the reads directly with the targeted CDS sequences, because target
200 enrichment was performed on the same species as probe design, as suggested by Bragg *et al.* (2016;
201 see also Rellstab *et al.*, 2019). Doing so, we avoid the complex and time-consuming steps of read

202 assembly and scaffolding/ordering the genomic contigs obtained to generate the in-target
203 assemblies for read mapping (see introduction). For each library, cleaned reads were aligned with
204 the 5,717 targeted CDS, with BOWTIE2 (v2.2.5; Langmead & Salzberg, 2012), using the default
205 parameters and *local* option to allow for genomic reads not aligning with the targeted CDS sequence
206 over their entire length, whilst ensuring a minimum alignment length of 20 bp. Reads were aligned
207 as unpaired, to allow for the mapping of orphan reads. Reads with multiple alignments, and those
208 for which the mate-pair was aligned to a different target were discarded after mapping. The aligned
209 data in SAM format were converted into BAM format and sorted with SAMtools (v0.1.19). Duplicate
210 reads were estimated with picard tools (v1.113) MarkDuplicates, using the default settings.
211 Coverage at each target base was calculated with mosdepth (v0.2.3; Pedersen & Quinlan, 2018).
212 Variant calling was performed with read alignments of more than 40 bases in length
213 containing fewer than 10 deletions, as deduced from the *cigar* format. SNP sites were called
214 separately for each of the 24 libraries with the SAMtools *mpileup* command, using option *-B*
215 followed by the varscan2 (v2.3; Koboldt *et al.*, 2012) *mpileup2snp* command with a minimum base
216 quality of 35 to count a read at a given position, and a *p*-value threshold of 0.01. For the other
217 parameters, we considered only positions with a minimum coverage of 40 reads for each individual,
218 and 250 reads for the pool (i.e. for these selected coverages, the same exome size was analyzed:
219 90.3% and 93% of target bases were covered by at least 40 and 250 reads for the individual and
220 pooled exomes, respectively; see Fig. S1). We also only included a variant if it was supported by 15
221 reads per individual, or by three reads for the single pool (given the pool's minimum coverage of
222 250X, we expected to have ~3 reads per haplotype if the 72 haplotypes in the pool were
223 homogeneously captured). Moreover to account for the variation in total read coverage across
224 positions, we required a minimum variant allele frequency threshold of 0.25 for each individual and
225 0.01 for the pool. For individual libraries, an individual was considered to be homozygous for the

226 variant if the variant allele frequency f(v) was above 0.75, and heterozygous if f(v) was between 0.25
227 and 0.75.

228 All 409,328 different SNP positions independently found in the 24 libraries were then
229 genotyped for each library with varscan2 *mpileup2cns,* using the same parameters as described
230 above. Several filters were then applied to all SNP positions, using either in-house PERL or R scripts
231 (R Development Core Team, 2015). We first retained only bi-allelic positions. We then retained only
232 positions genotyped for at least 20 individuals. Finally, we discarded SNPs with high levels of
233 coverage in the pool, which were likely to represent repetitive regions or paralogs of the genome:
234 maximum pool coverage was set at 6,960X, as calculated with the formula proposed by Li et al.
235 (2014): $c+4*\sqrt{c}$, where c is the mean coverage in the pool (c=6,634.8X). We retained 300,036 bi-
236 allelic SNPs after the application of all these filters.

237 At each site, and for each capture, variant allele frequency at population level was estimated
238 as the number of reads with the variant allele divided by total coverage. Individual genotypes were
239 used to calculate exact *p*-values for Hardy-Weinberg equilibrium tests with the "HardyWeinberg" R
240 package (Graffelman, 2015), with the correction of significance level by the false discovery rate
241 procedure (Benjamini & Hochberg, 1995). Finally, SNPs were annotated and categorized with the
242 SnpEff program (v4.3t; Cingolani *et al.*, 2012).

243

244 ***Evaluation of the approach based on read mapping directly onto CDS***

245 Approaches based on the direct mapping of reads onto CDS have already been used to detect SNPs
246 (e.g. Rellstab *et al.*, 2019), but they have never been evaluated. We used an *H. axyridis* draft genome
247 sequence (HaxR v1.0; Gautier *et al.*, 2018) to evaluate this approach and that proposed below for
248 the prediction of intron-exon boundaries (IEBs). We selected CDS targets with a genomic match, as
249 proposed by Neves *et al.* (2013), with some adjustments. Briefly, CDS target sequences were first

250 aligned with genomic scaffolds with BLASTN (e-value cutoff of e^{-20} and identity cutoff of 90%), and
251 the best hit was retained for each target. The target sequences were then realigned with the
252 corresponding genomic scaffold, with both GMAP (version 2015-12-31; Wu & Watanabe, 2005) and
253 EXONERATE *est2genome* (v2.2.0;
254 http://www.genome.iastate.edu/bioinfo/resources/manuals/exonerate/). We isolated targets (i)
255 with genome matches over their entire length, (ii) for which intron/exon structure predictions were
256 the same for both alignment programs, (iii) which were covered by capture probes over their entire
257 length and (iv) for which we had access to at least 200 bp upstream and downstream from the
258 targeted CDS sequences. With these filters, we retained 3,161 CDS targets in total (3,220,718 bp,
259 i.e. 60.23% of our targeted exome). These 3,161 full or partial CDS corresponded to 11,754 exons,
260 with a mean of 3.72 exons per target (range: 1 to 17 exons) and a mean exon length of 274 bp
261 (range: 12 to 3,417 bp). These 11,754 exons included 6,295 that were unambiguously complete,
262 because they were surrounded by two others, and 8,593 reliable IEBs in total were identified
263 between consecutive exons.

264 All genomic regions corresponding to these 3,161 CDS targets were then isolated, including
265 200 bp upstream and downstream from the target sequences. Independently for each library, all
266 cleaned reads were aligned with these genomic regions, with BOWTIE2 (v2.2.5), using the default
267 parameters and the *end-to-end* option (i.e. reads must be aligned along their full length). Reads with
268 multiple alignments, and reads for which the mate-pair was aligned to another target were
269 discarded after mapping. Coverage calculation, SNP calling and variant allele frequency estimation
270 were performed as described above. We established the correspondence between the position of
271 a base in the genomic region and its position in the CDS sequence, using the *vulgar* format of
272 EXONERATE output. SNP positions for which there was a deletion or an insertion between the CDS

273 sequence and its associated genomic sequence (i.e. 559 positions) were excluded from the
274 comparison.
275
276 ***Intron-exon boundary detection***
277 Our approach is based solely on the direct mapping of genomic reads to the CDS sequences, by
278 allowing reads to map over less than their full length (i.e. local alignment). Thus, a genomic read
279 with both part of an exon and its flanking sequences (intron or UTR) will start mapping to the
280 corresponding exon in the CDS sequence, and stop mapping at the exon end (Conklin *et al.*, 2013).
281 This will generate a signal of the intron-exon boundary in the number of reads that either begin or
282 end their local alignment at that position (hereafter referred to as *nbBE*). Along a covered exon,
283 *nbBE* should be much lower than the coverage rate, except at the end of the exon, where we should
284 observe a large increase in *nbBE* to a value theoretically equal to the coverage (Fig. S2). This strong
285 increase in *nbBE* should occur either at the base at the exact end of the exon, or a few bases away
286 if the start of the intron sequence is, by chance, highly similar to the start of the next exon sequence
287 on the CDS (Fig. S2). In this second case, it creates an abnormally inflated estimate of coverage
288 locally for a few bases at the boundary of the following exon (Fig. S2 and Fig. S3 for an example). At
289 an IEB (i.e. a junction of two exon ends in the CDS), depending on whether a signal offset occurs for
290 one or both exon ends, the size of the area with abnormally inflated local coverage is variable, but
291 generally less than 10 bp. In addition, the presence of a SNP at some bases at the end of the exon
292 may prevent reads with the variant from aligning up to the IEB position, potentially generating
293 another signal in addition to those of the two exon ends of the IEB. Combinations of these different
294 scenarios may also occur.
295 We developed and tested a method for predicting IEBs on the basis of the increase in *nbBE,*
296 by focusing on the 3,161 targeted transcripts for which genomic sequences were available, for which

297 intron-exon structure was known. We used the mapping output BAM file obtained with the pool
298 library. At each position in the transcript, *nbBE* was calculated from the start read alignment position
299 on the transcript and the *cigar* code, both present in the BAM file. The value of the *nbBE* parameter
300 at each position is a function of coverage (and read length). We took the possible effect of coverage
301 heterogeneity on *nbBE* into account by calculating the median *nbBE* over a window of 11 bases
302 centered on the position concerned. A signal (i.e. a predicted IEB), was detected at a position if *nbBE*
303 was *X* times larger than the median, *X* being a tool parameter defined by the user. For the definition
304 of *X*, it may be necessary to identify a subset of transcripts with orthologs in a close model species,
305 and then to determine which value of *X* provides the best prediction of IEBs for this subset. Signals
306 at the exact ends of the CDS were not considered. We took the possible signal deviation (i.e. a signal
307 that is slightly deported with respect to the IEB that generates it; Fig. S2) into account, by extending
308 signal detection to the three bases upstream, and three bases downstream from a position with
309 measured signal (default value for parameter *n*=3). A measured signal therefore predicts an IEB in a
310 window of seven bases in total. An IEB, composed of two exon ends, theoretically generates two
311 close signals with overlapping windows. If these two signals are found at the exact positions of the
312 exon ends (no deviation), then the predicted region measures 8 bp (Fig. S2). If these two signals
313 overlap (e.g. because of the deviation of one of the two signals) then the predicted region for the
314 IEB is 7 bp long. If these signals deviate by up to six bases from each other (i.e. cumulative deviation
315 of eight bases, the deviation being in opposite directions for the two signals) then the predicted
316 region measures 14 bp (Fig. S2). More rarely, if a third signal is detected in the same region, it
317 increases the length of the region by a few bases. If the two signals deviate by more than six bases
318 from each other, then the tool predicts two signal windows separated by a few bases for the IEB
319 concerned. Thus, each transcript was divided into short regions with prediction signals (+)
320 alternating with regions without prediction signals (-). Each region was then assessed as follows:

321 true positive (TP) if the + region contained an IEB; false positive (FP) if the + region did not contain
322 an IEB; true negative (TN) if the - region did not contain an IEB; and false negative (FN) if the - region
323 contained an IEB. We evaluated the performance of the method, by calculating the percentage of
324 true IEBs correctly predicted (i.e. sensitivity (SN)) as TP/(TP+FN) and the percentage of correct
325 predictions (i.e. specificity (SP)) as TP/(TP+FP) (Burset & Guigo, 1996). For a CDS with no introns (i.e.
326 862 of the 3,161 CDS), SN cannot be evaluated; SP was 100 if no IEB was predicted (good prediction)
327 and 0 otherwise. Scripts are available from https://github.com/edeleury/IEB-finder.
328
329
330 # Results
331
332 ***Capture data and filtration***
333 In the two target enrichment experiments, we obtained 186,998,614 and 185,457,659 *H. axyridis*
334 raw sequence read pairs for the pool library (SRA ERX3237193) and the 23 indexed individuals (SRA
335 ERX3237194 to ERX3237216), respectively. For each indexed individual library, we obtained a mean
336 of 8,063,376 raw read pairs (range: 6,912,561 to 9,176,292). After cleaning, 91.2% of the raw read
337 sequences were retained (see detailed information for all libraries in Table S1).
338
339 ***Capture efficiency, coverage and duplicates***
340 Rather than assembling the cleaned reads, we chose to map them directly onto the target CDS
341 sequences used for probe design. Mean capture specificity, defined as the percentage of cleaned
342 reads mapping correctly (i.e. meeting our mapping parameters) onto the 5,717 targeted CDS, was
343 83.2% (range over libraries: 80.6%-89.2%; Table S2). Aligned bases accounted for 74.6% (range over
344 all libraries: 72.6%-75.7%) of all cleaned read bases. The global median read alignment length was

345 144 bp (range over libraries: 143-144 bp, see Table S2), with 62.7% of the mapped reads (range over
346 all libraries: 60.9%-64.3%) mapped over their entire length. A substantial proportion of the mapped
347 reads (22.2%; range over all libraries: 20%-25.9%) were 'orphaned' mates (i.e. only one of the two
348 mates aligned to the target sequences). This percentage strongly suggests that flanking regions of
349 targeted regions were also sequenced with our capture procedure (Yi *et al.*, 2010; Neves *et al.*,
350 2013), although quantification was not possible with this direct mapping approach.

351 The mean percentage of the 5,347,461 targeted bases (5,717 target CDS) covered by at least
352 one read (i.e. the capture sensitivity) was 93.5% (range over all libraries: 93.4%-93.9%, see Fig. S1
353 and Table S3). Overall, 93% of the targeted bases were covered in all 24 libraries, whereas 6% were
354 covered by no read in any of the libraries (Table 1). We identified 491 full targets (292,327 bp) with
355 no read aligned in the 24 libraries. For these uncovered targets, 487 (99.2%) had no match to the *H.*
356 *axyridis* draft genome either, whereas, for the 5,226 covered targeted CDS, only 74 (1.4%) had no
357 match to the genome (here partial alignments of targets with genome were also considered).

358

		Percentage of samples with target sequence	
Target	Target size	100%	0%
CDS level	5 717	5 226 (91.4%)	491 (8.6%)
Base level	5 347 461	4 967 690 (92.9%)	323 418 (6.0%)

359 **Table 1:** Reproducibility of capture sensitivity on the 24 libraries. 100% means that all 24 samples had at least one read
360 aligned with the target, whereas 0 % level indicates that no reads aligned with the target.

361

362 Considering all targeted bases, the median and mean base coverage for individual libraries
363 were 142X and 283X for the "23 individuals" libraries, and 3,361X and 6,635X for the pool libraries
364 (see Table S3), respectively. A small number of targeted bases had very high coverage, but around
365 90% had a coverage <300X for individuals or <7,000X for the pool (Table S3). For individuals, 90.3%
366 of the targeted bases on average had a mean coverage greater than 40X. For the pool, 93% had a
367 coverage greater than 250X (Fig. S1).

368 Base coverage varied along the targeted CDS (Fig. S4A). Using the known complete exons (≥
369 20 bp) identified on targeted CDS with a match over their entire length with the *H. axyridis* draft
370 genome, we confirmed that mean coverage increased steadily with exon size (Fig. S4B), and that
371 coverage decreased from the middle of an exon to its ends (Fig. S4C). Alignments of less than 20 bp
372 between the reads and the genome sequence were not permitted with this mapping procedure, so
373 coverage was probably underestimated at exon ends and for short exons. For all the targeted CDS
374 considered together, mean coverage differed between targeted CDS but was highly reproducible
375 between libraries (Fig. S5): the correlation coefficient of the target coverage between individual
376 libraries ranged from 0.96 to 0.99, with a mean of 0.99, and the coefficient of correlation between
377 the two capture processes (i.e. between pool and individuals) was 0.999.

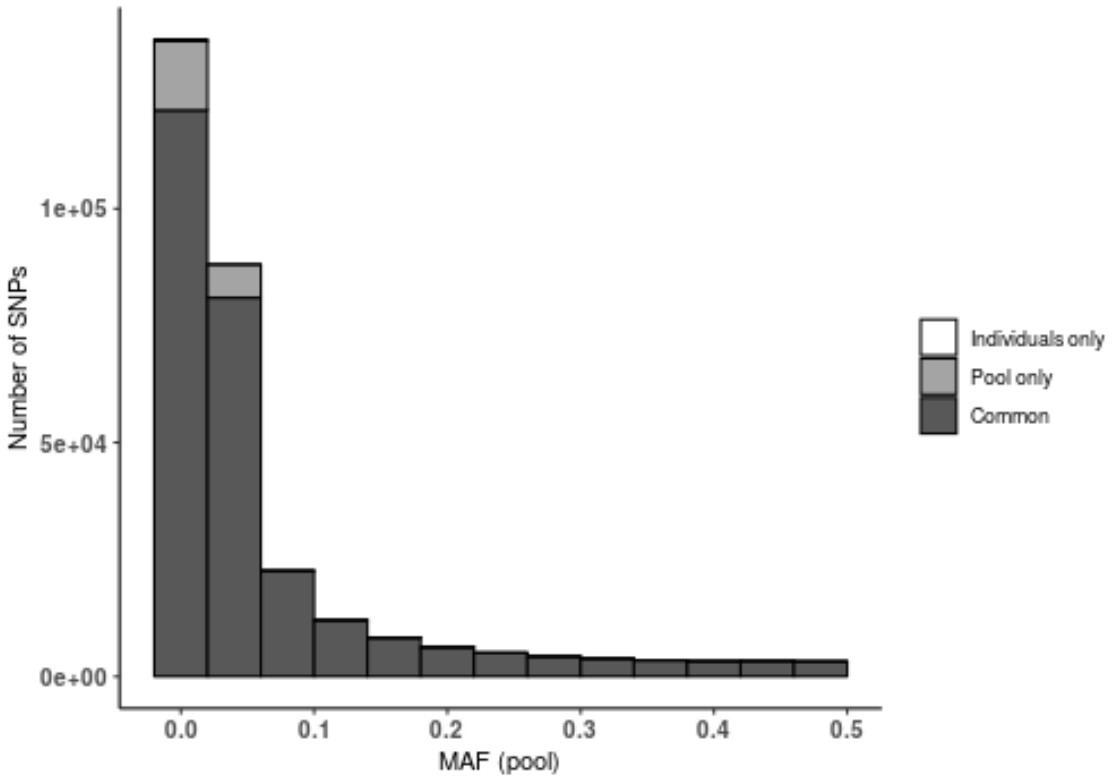
378 For the 23 individual libraries, a mean of 66.1% of reads were scored as duplicates (range
379 over libraries: 63.6%-68.9%; Table S4). This high proportion was due to the very strong sequencing
380 effort regarding the target region size. Indeed, pooling the reads of the 23 individual libraries
381 increased the proportion of duplicates detected to >96%, as for the pool (Table S4), suggesting that
382 the reads scored as duplicates in our experiment probably mostly come from different individuals.
383 We did not therefore discard duplicated reads before SNP calling, so as to avoid artificially biasing
384 the estimation of allele frequencies.

385

386 ***SNP calling and accuracy of allele frequency estimation***

387 SNP calling performed independently for the individuals and for the pool, following direct mapping
388 onto the targeted CDS sequences, led to the detection of 302,292 and 364,144 different putative
389 exonic SNPs, respectively, with 257,108 positions (62.8%) detected in both experiments. As
390 expected (because there were 13 more individuals in the pool), the number of SNPs called for the

391 pool was larger than that for the 23 individuals, with 26.2% of SNPs called for the pool only. The vast
392 majority (95.6%) of the 11% of SNPs called only for individuals had allele frequencies below 0.03.
393 All positions were genotyped in all libraries. After filtering, we finally selected a total of
394 300,036 exonic bi-allelic SNPs within 4,741 targeted CDS. Minor allele frequencies (MAFs) were
395 below 0.03 for 59.2% of these SNPs (Fig. 1); 92.75% were polymorphic in both the pool and the
396 individuals, 7.22% were polymorphic only in the pool, and 0.03% were polymorphic only in the
397 individuals. Only 1,007 of the total number of SNPs (i.e. 0.34%) displayed significant deviation from
398 Hardy-Weinberg equilibrium, as calculated with the individual genotypes. This number was more
399 than three times higher (i.e. 3,331 loci) before the removal of loci on the basis of the maximum
400 coverage filter applied to the pool.
401
402
403
404
405
406
407
408
409
410
411
412
413



414 **Figure 1:** Folded-site frequency spectrum of the 300,036 filtered exonic bi-allelic SNPs (mapping onto targeted CDS
415 sequences) according to their polymorphism in the pool and in individuals. For SNPs polymorphic in both the pool and
416 individuals, the MAF shown is the MAF calculated for the pool.

When all 300,036 loci were considered, allele frequency estimates were strongly correlated between the pool and individuals (*r*=0.99; Fig. 2A). These results were not dependent on the large fraction of SNPs with low allele frequency estimates (r=0.98, calculated for 102,494 loci with MAF ≥ 0.03).

SNP annotation revealed that 74.75% of SNPs were synonymous, whereas 24.92% were non-synonymous missense and 0.33% were non-synonymous nonsense SNPs.

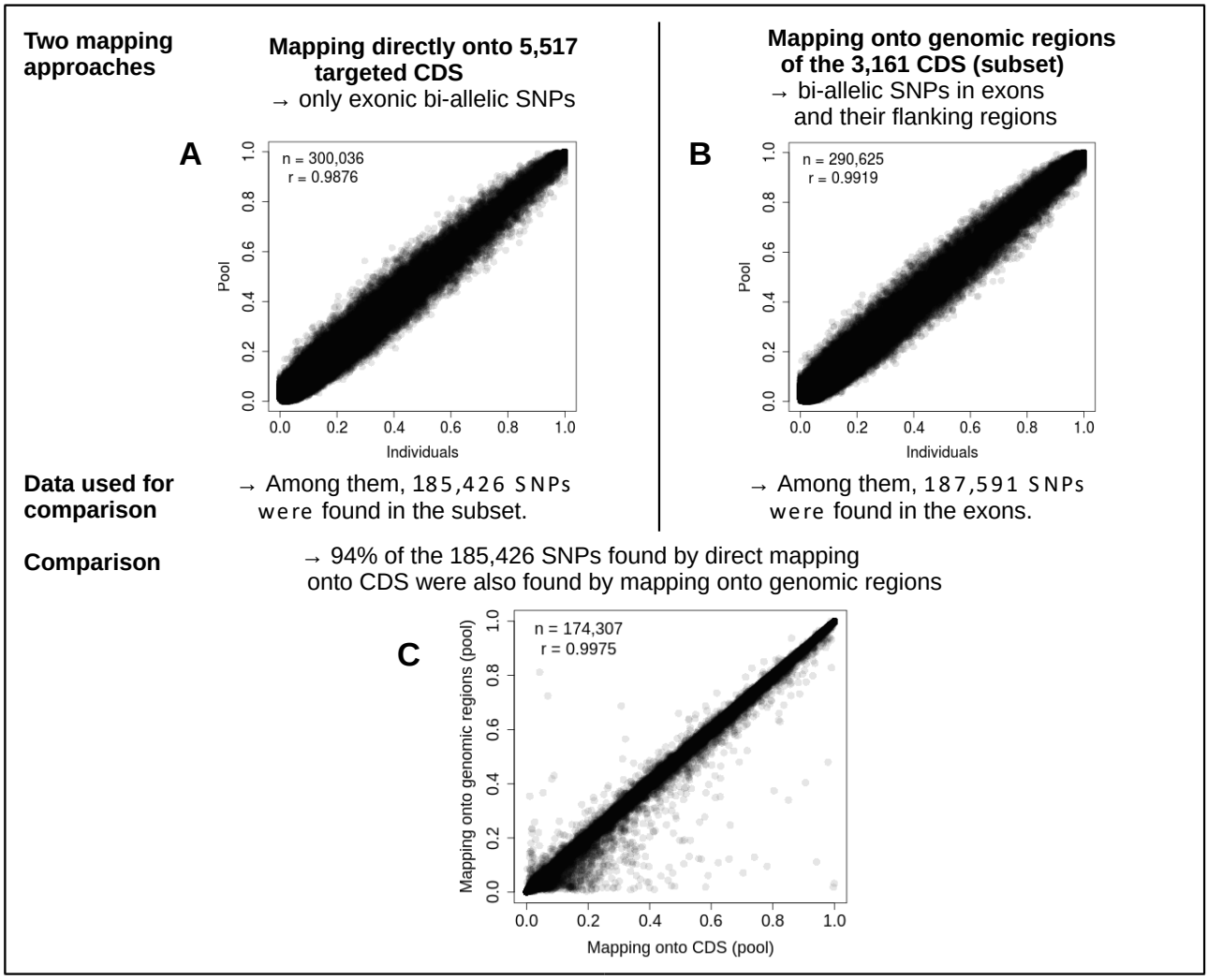


Figure 2: Correlation of the SNP allele frequencies estimated for the population pool and individuals by two mapping approaches: **(A)** mapping reads directly onto CDS and **(B)** mapping onto genomic regions available only for a subset of target CDS. For exonic bi-allelic SNPs common to both mapping approaches, we also show **(C)** the correlation between the allelic frequencies estimated for the pool by the two approaches.

429 ***Comparison of the SNPs detected with the two mapping approaches***

430 We evaluated the ability of the method for mapping reads directly to CDS to call SNPs correctly by
431 comparing the SNPs identified with this approach and those found by mapping reads to the genome.
432 We used only the subset of 3,161 targeted CDS with a match over their entire length to sequences
433 in the draft genome of *H. axyridis* for this comparison. All trimmed reads were also mapped along
434 their entire length onto the regions of the genome corresponding to this subset. After application
435 of the same procedure and filters for SNP calling as were used for direct mapping onto CDS
436 sequences, 290,620 filtered bi-allelic SNPs were genotyped within the subset, including 187,591
437 (64.5%) in exons and 103,029 (35.5%) in the flanking regions (intron or UTR) of the exons. By
438 mapping reads onto the genome, we demonstrated that allele frequency estimates were strongly
439 correlated between the pool and individuals, for both exonic and flanking SNPs (r=0.99; Fig. 2B).

440 Of the bi-allelic exonic SNPs found by direct mapping onto the targeted CDS (Fig. 2A),
441 185,426 mapped onto the subset of 3,161 targeted CDS, and 94% were also identified as bi-allelic
442 SNPs by mapping to genomic regions. For these 174,307 exonic bi-allelic SNPs found with both
443 mapping approaches (mapping onto the genome or directly onto transcripts), the allele frequency
444 estimates obtained with the two mapping methods were highly correlated both for the pool
445 (r=0.998; Fig. 2C) and for the individuals (r=0.998).

446 Some of the private bi-allelic SNPs were variants identified with both mapping approaches
447 but discarded during filtering in one of the approaches. For example, 1,467 bi-allelic SNPs private to
448 the CDS mapping process had a coverage above the maximum threshold, and 714 were tri-allelic
449 SNPs when mapped onto the genome. If we exclude this latter group, 8,853 and 8,333 were found
450 to be private to CDS mapping and mapping onto genomic regions, respectively. A non-exhaustive
451 analysis of these positions highlighted two possible origins of private SNPs for mapping onto CDS.
452 Some targeted regions (the full CDS or part of the CDS) have homologous copies in the genome that

453 we were unable to identify during target selection and for which the SNPs found were not discarded
454 with the "maximum coverage" filter. We have thus identified 43 CDS with reads mapping onto their
455 genomic regions in a non-unique manner, and these CDS with paralogs displayed 3,141 SNPs, all
456 private to CDS mapping. The second origin was false SNPs occurring when the start of an intron
457 could be aligned with the start of the next exon due to similarities between the two sequences (Fig.
458 S2 and S6, for example). Of the SNPs private to the CDS mapping method, 6.9% (*N*=614) were
459 positioned close to IEB positions (i.e. at the exact end of the exon or on the three adjacent bases),
460 whereas only 0.6% of the 174,307 SNPs common to both mapping methods were found near IEBs.
461
462 ***Prediction of IEBs without read assembly***
463 The method we developed for IEB detection based on the mapping of genomic reads directly onto
464 CDS sequences was evaluated on the 3,161 targeted CDS for which we had genomic information,
465 and therefore knew the positions of the IEBs. An IEB signal was predicted at a particular position in
466 the CDS if the number of reads either beginning or ending their local alignment at that position
467 (*nbBE*) was 20 times greater (*X*=20) than the median *nbBE* value within a window of 11 bases around
468 the position concerned, and the signal measured at that position extended to a short region of 7
469 bases (i.e. with a default parameter of *n*=3; see Methods section and Fig. S2). With parameters *X*=20
470 and *n*=3, 8,967 regions with lengths of 4 to 19 bp were predicted to be IEBs (Table 2). Regions of
471 less than seven bases in length were next to non-covered regions or regions with a measured signal
472 close to the end of the CDS for which signal extension was not possible. Of the 8,967 regions
473 predicted to be IEBs, 8,575 actually corresponded to true IEBs (i.e. method specificity = 95.63%). Of
474 the 8,632 regions truly containing an IEB, 8,575 were accurately predicted to be IEB regions (i.e.
475 sensitivity = 99.34%; Table 2). Of the 862 CDS with no intron, 763 (i.e. 88.51%) were correctly
476 predicted to have no IEBs. As expected, the use of a lower detection threshold (i.e. *X*=10) decreased

477 the specificity of the method and increased its sensitivity (Table 2). The regions predicted to have
478 no IEB (i.e. the - regions) were between 1 and 3,417 bp in length, with a mean length of 265.5 bp,
479 entirely consistent with the known distribution of exon lengths (see Methods section). The 153 very
480 small (<10 bp) - regions attested to the possible cumulative deviation of signals of more than eight
481 bases for a small number of IEBs. Increasing the length of the signal deviation (e.g. *n*=5 bases instead
482 of 3) made it possible to take this situation into account in the vast majority of cases. This slightly
483 improved performance but increased the length, and thus, the imprecision, of all predicted IEB
484 positions (Table 2).

485

Parameter		Regions +		Regions -		No prediction	Performance	
X	*n*	Number	Mean length [min;max] (bp)	Number	Mean length [min;max] (bp)	Number	Sensitivity	Specificity
20	3	8967	8.38 [4;19]	11849	265 [1;3417]	228	99.34	95.63
20	5	8856	12.44 [6;26]	11728	264.7 [1;3417]	228	99.88	97.01
10	3	9454	8.46 [4;26]	12254	255.8 [1;3417]	228	99.59	90.82
10	5	9306	12.55 [6;39]	12082	256.4 [1;3417]	228	99.94	92.34

486 **Table 2:** Evaluation of the IEB prediction method using pool data. The + region is a short region predicted to be an IEB.
487 The - region is a region with no signal (i.e. predicted to be an exon). Regions not covered by reads are regions without
488 prediction. Sensitivity is the percentage of regions with true IEBs correctly predicted to be + regions, and specificity is
489 the percentage of + regions that are true IEBs.

490

491

492 # Discussion

493

494 Exome-targeted capture associated with next-generation sequencing has been mainly restricted to
495 model species for which genomic resources are available (Ng *et al.*, 2009). Transcriptome-based
496 exome capture has proved an efficient alternative for dealing with non-model species (e.g. Bi *et al.*,
497 2012), but is mostly used in phylogeny studies. Here, we propose an extension of this method with

498 a high capture efficiency, providing accurate estimates of the allele frequencies of SNPs for
499 population genomics studies. Our methodology is original because (i) it combines transcriptome-
500 based exome capture with a pool-based approach (pooling the tissues of numerous individuals
501 before DNA extraction), (ii) it does not require a genome sequence for the species itself or another
502 closely related species, and (iii) it does not involve the *de novo* assembly of genomic reads. This work
503 opens up interesting prospects for population genomics studies on non-model species because it is
504 cheap and less labor-intensive than other methods.
505
506 ***Capture efficiency***
507 We found that capture efficiency was very high for the *H. axyridis* exome. The capture specificity
508 (83.2%) achieved was much higher than reported in more classical transcriptome-based studies
509 (range 25% - 73%), whereas capture sensitivity (93.5%) approached the highest reported values (see
510 Puritz & Lotterhos, 2018). Moreover, despite the differences in coverage observed between targets
511 and along the length of targets, reproducibility was high between libraries, whether in terms of
512 specificity, sensitivity or coverage. Our capture approach, based on a random choice of targets from
513 the *de novo* transcriptome of the species of interest itself, without the need to identify orthologous
514 genes in related model species, proved highly effective. This "blind" approach had only a minor
515 impact on capture performance: we only observed a slight decrease in coverage for small exons and
516 towards the ends of the exons. Similar patterns have been observed in classical transcriptome-based
517 exome capture in which IEBs are predicted by orthology, due to the use of less efficient probes
518 designed from the sequences of two consecutive exons and the absence of a probe straddling the
519 end of the exon and the beginning of the intron (e.g. Bi *et al.*, 2012; Neves *et al.*, 2013).
520 The random choice of targets was probably the cause of the slightly lower capture sensitivity
521 in this study than in previous studies (Bi *et al.*, 2012; Puritz & Lotterhos, 2018). Targets selected

522 without looking for orthology in related species are necessarily only partly known, and some may
523 belong to other species (e.g. symbionts or parasites) or be chimeras present in the transcriptome
524 used. This is a small price to pay given the efficiency of the method, a feature of considerable
525 interest for population genomics studies. The absence of searches for orthologous genes in related
526 model species prevents the analysis from focusing purely on the most conserved genes. Moreover,
527 our method can be applied to any species for which no closely related model species with genomic
528 resources is available.
529
530 ***SNP calling and the pool-based approach***
531 With this method, we were able to genotype a very large number of SNPs (~300,000) over the exome
532 of *H. axyridis*. The study of the subset of CDS with complete genomic sequence matches further
533 confirmed that direct mapping onto the CDS has no major impact on SNP identification or on the
534 estimation of allele frequencies relative to mapping onto the genome. In addition to a few residual
535 paralogous sequences inherent to any approach dealing with non-model species, we observed
536 slightly larger numbers of SNPs called on the basis of mapping onto CDS in areas very close to the
537 IEBs, but the proportion of SNPs involved was minimal (0.33%).
538 For similar sequencing efforts, strong concordance was observed between the allele
539 frequencies for individual genotyping and those for the "pool-based" approach, whatever the
540 mapping method used. This finding demonstrates the feasibility of this approach, in situations in
541 which individual identification for many samples is practically impossible, technically difficult, too
542 costly and/or unnecessary. In population genomics studies, allele frequencies are sufficient for the
543 calculation of most summary statistics (e.g. nucleotide diversity, F_{ST}, site frequency spectrum) and,
544 thus, for addressing most of the questions posed (Schlötterer *et al.*, 2014). In the specific case of
545 exome sequencing, allele frequencies can be used for large population-scale screening of protein

546 mutations, or to address questions related to adaptation or to the evolution of the genetic load (e.g.
547 Gutenkunst *et al.*, 2009; Yi *et al.*, 2010; Casals *et al.*, 2013; Popp *et al.*, 2017; Castellano *et al.*, 2019).
548 One of the limitations of pool-seq compared to individual-based approaches is that
549 "populations" must be carefully defined before sequencing, as it will be difficult to detect a Wahlund
550 effect or an unforeseen breeding system afterwards. It is therefore important to be particularly
551 careful during the sampling step. A second drawback of pool-seq is that it makes it impossible to
552 remove SNPs that are not at Hardy-Weinberg equilibrium. Such disequilibrium may be caused by
553 copy number variations or paralogous sequences resulting in false-positive SNPs and local over-
554 coverage (Li, 2014). We show here that the application of our "maximum coverage" filter to the
555 pooled data made it possible to discard almost 70% of the SNPs not at equilibrium, as shown in
556 individuals. The extremely small remaining proportion (0.34%) of SNPs not at Hardy-Weinberg
557 equilibrium provides support for the feasibility of our pooling-based method. Finally, pool-seq is not
558 optimal for studies in which analyses rely on linkage phase information. However, a prerequisite in
559 this context is the availability of a genetic (or physical) map, which is not the case in most non-model
560 organisms.
561 One of the limitations of pool-seq compared to individual-based approaches is that it is not
562 optimal for studies in which analyses rely on linkage phase information. However, a prerequisite in
563 this context is the availability of a genetic (or physical) map, which is not the case in most non-model
564 organisms. Another weakness of pool-seq is that "populations" must be carefully defined before
565 sequencing, as it will be difficult to detect a Wahlund effect or an unforeseen breeding system
566 afterwards. It is therefore important to be particularly careful during the sampling step, and to be
567 as familiar as possible with the biology of the species being studied. Finally, one more drawback of
568 pool-seq is that it makes it impossible to remove SNPs that are not at Hardy-Weinberg equilibrium.
569 Such disequilibrium may be caused by copy number variations or paralogous sequences resulting in

570 false-positive SNPs and local over-coverage (Li, 2014). We show here that the application of our
571 "maximum coverage" filter to the pooled data made it possible to discard almost 70% of the SNPs
572 not at equilibrium, as shown in individuals. The extremely small remaining proportion (0.34%) of
573 SNPs not at Hardy-Weinberg equilibrium provides support for the feasibility of our pooling-based
574 method.
575
576 ***Method for predicting IEBs without read assembly***
577 We have developed a method for detecting IEBs that is particularly useful in the case of non-model
578 species (or species for which genome sequences are being assembled). The proposed tool (available
579 from https://github.com/edeleury/IEB-finder) does not indicate the exact location of the IEBs,
580 instead identifying an area of a few bases (about eight, on average, in our case) predicted to contain
581 an IEB. This method, which requires only the genomic reads and the target CDS sequences, was
582 effective when evaluated on a subset of CDS for which we had a complete genomic match. Almost
583 all the true IEBs were identified, whereas false-positives accounted for less than 5% of the predicted
584 IEBs. Montes *et al.* (2013) proposed a similar approach, which they used to discard putative SNPs
585 found in predicted IEB areas. They obtained better genotyping results for the selected filtered SNPs
586 than other similar studies on non-model species. In this study, we went a step further by taking into
587 account the possible offset of the measured signal, and by evaluating the performance of the
588 approach on a set of transcripts for which the true intron-exon structure was known. We aimed to
589 present evidence that this approach is promising, although further improvement is required,
590 together with evaluation on larger dataset and at various level of coverage. A method for choosing
591 an appropriate value for parameter *X* for the correct identification of significant increases in *nbBE*
592 depending on local coverage must also be developed.
593

594 ***General recommendations, conclusion and perspectives***

595 The combination of approaches proposed here — i.e. exome transcriptome-based capture, random
596 choice of targets, and direct mapping of reads onto targeted CDS sequences, coupled with a pool-
597 seq approach before DNA extraction — significantly decreases costs, the laboratory time required
598 and the complexity of the sequencing analyses required to obtain a large number of variants in
599 exome-scale coding regions for non-model species. However, direct mapping onto target sequences
600 has two disadvantages. First, it provides no information on flanking regions (including regulatory
601 regions), and it is therefore not possible to genotype SNPs in these areas. Nevertheless, detecting
602 SNP polymorphism at exonic level would be sufficient in many cases, particularly when the in-depth
603 analysis of a large number of synonymous and non-synonymous mutations is informative enough
604 to address the question posed. It is still feasible to assemble the genomic reads in situations in which
605 flanking regions appear to be of interest. Second, direct mapping onto CDS results in small biases in
606 the detection of SNPs in the vicinity of IEBs. As a means of limiting the risk of detecting false SNPs
607 further, we therefore suggest using the IEB detection method proposed here, followed by the
608 elimination of any SNPs found in the predicted IEB regions and in short regions (≤ 20 bp) surrounded
609 by two predicted IEB regions. Using this approach on our data with the parameters X=20 and n=3,
610 we retained a final list of 296,736 SNPs. A large proportion of the 3,300 SNPs excluded (408 SNPs,
611 i.e. 12.36%) were not at Hardy Weinberg equilibrium in individuals, confirming the relevance of this
612 filter.

613 The allele frequencies measured for the pool with our method were consistent with those
614 obtained in the analysis of individuals. However, caution is nevertheless required with rare alleles
615 (i.e. SNPs with low MAFs), which may be confused with sequencing errors (Bansal, 2010). We
616 recommend the use of technical replicates. When possible, and depending on the biological
617 question addressed, the analysis could be restricted to SNPs with MAFs above a given threshold.

618 In conclusion, our study demonstrates the feasibility of obtaining, at low cost, for a non-
619 model species (i.e. without a reference genome), important and accurate information at population
620 level concerning (i) SNP polymorphism within the exome and (ii) the general structure of the exome
621 (intron-exon structure of the targeted genes). The results obtained with this method applied to the
622 Asian ladybird *Harmonia axyridis* were very promising, and we now need to investigate other
623 species with different genome sizes and complexities. Finally, the oversizing of these experiments,
624 resulting in very high levels of coverage, suggests that it should be possible to reduce costs further
625 by capturing multiple indexed pools in a single reaction.
626
627

628 Acknowledgments

629

630 We thank Stefan Toepfer for the *Harmonia axyridis* samples. We thank Thibaut Malausa, Arnaud
631 Estoup, Mathieu Gautier, Fabrice Legeai, and Bernhard Gschloessl for their helpful comments, which
632 helped us to improve this manuscript considerably. This work was funded by the INRA SPE
633 department and the INRA UMR Institut Sophia Agrobiotech. Sequencing was performed at the
634 GENOTOUL GeT platform (https://get.genotoul.fr/en/).
635
636

637 Author contributions

638

639 ED, EL and TG conceived the study. AB, ED and EL prepared the samples and the libraries. ED
640 developed the bioinformatics pipelines. ED and EL ran the analyses. ED, EL and TG wrote the paper.

641

642

643 ## Competing interests

644

645 The authors have no competing interests to declare.

646

647

648 ## Data accessibility

649

650 The exome sequence capture data reported here have been deposited in the European Nucleotide
651 Archive (accession no. PRJEB31592). The CDS sequences targeted and the description of the exon
652 positions on the subset of transcripts have been deposited at Zenodo
653 https://doi.org/10.5281/zenodo.2598388. The HaxR v1.0 genome sequence used in this study is
654 available from http://bipaa.genouest.org/sp/harmonia_axyridis/ (see Gautier *et al.*, 2018).

655

656

657 ## References

658

659 Bansal, V. 2010. A statistical method for the detection of variants from next-generation
660 resequencing of DNA pools. *Bioinformatics* **26**: 318–324.
661 Bansal, V., Tewhey, R., LeProust, E.M. & Schork, N.J. 2011. Efficient and cost effective population
662 resequencing by pooling and in-solution hybridization. *PLoS One* **6**: 1–6.
663 Benjamini, Y. & Hochberg, Y. 1995. Controlling the false discovery rate - a practical and powerful
664 approach to multiple testing. *J. R. Stat. Soc. Ser. B-Methodological* **57**: 289–300.
665 Bi, K., Linderoth, T., Vanderpool, D., Good, J.M., Nielsen, R. & Moritz, C. 2013. Unlocking the vault:
666 Next-generation museum population genomics. *Mol. Ecol.* **22**: 6018–6032.
667 Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C. & Good, J.M. 2012. Transcriptome-based
668 exon capture enables highly cost-effective comparative genomic data collection at moderate
669 evolutionary scales. *BMC Genomics* **13**: 403.
670 Bolger, A.M., Lohse, M. & Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence

671 data. *Bioinformatics* **30**: 2114–2120.
672 Bossert, S. & Danforth, B.N. 2018. On the universality of target-enrichment baits for phylogenomic
673 research. *Methods Ecol. Evol.* **9**: 1453–1460.
674 Bragg, J.G., Potter, S., Bi, K. & Moritz, C. 2016. Exon capture phylogenomics: efficacy across scales
675 of divergence. *Mol. Ecol. Resour.* **16**: 1059–1068.
676 Burset, M. & Guigo, R. 1996. Evaluation of Gene Structure Prediction Programs. *Genomics* **34**: 353–
677 367.
678 Casals, F., Hodgkinson, A., Hussin, J., Idaghdour, Y., Bruat, V., de Maillard, T., *et al.* 2013. Whole-
679 Exome Sequencing Reveals a Rapid Change in the Frequency of Rare Functional Variants in a
680 Founding Population of Humans. *PLoS Genet.* **9**.
681 Castellano, D., Maclà, M.C., Tataru, P., Bataillon, T. & Munch, K. 2019. Comparison of the full
682 distribution of fitness effects of new amino acid mutations across great apes. *Genetics* **213**:
683 953–966.
684 Castellano, S., Parra, G., Sanchez-Quinto, F.A., Racimo, F., Kuhlwilm, M., Kircher, M., *et al.* 2014.
685 Patterns of coding variation in the complete exomes of three Neandertals. *Proc. Natl. Acad.*
686 *Sci.* **111**: 6666–6671.
687 Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., *et al.* 2012. A program for
688 annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).*
689 **6**: 80–92.
690 Conklin, D., Montes, I., Albaina, A. & Estonba, A. 2013. Improved conversion rates for SNP
691 genotyping of nonmodel organisms. *Proc. Iwbbio 2013 Int. Work. Bioinforma. Biomed. Eng.*
692 127–134.
693 Cosart, T., Beja-Pereira, A., Chen, S., Ng, S.B., Shendure, J. & Luikart, G. 2011. Exome-wide DNA
694 capture and next generation sequencing in domestic and wild species. *BMC Genomics* **12**:
695 347. BioMed Central Ltd.
696 Day-Williams, A.G., McLay, K., Drury, E., Edkins, S., Coffey, A.J., Palotie, A., *et al.* 2011. An
697 evaluation of different target enrichment methods in pooled sequencing designs for complex
698 disease association studies. *PLoS One* **6**: e26279.
699 Förster, D.W., Bull, J.K., Lenz, D., Autenrieth, M., Paijmans, J.L.A., Kraus, R.H.S., *et al.* 2018.
700 Targeted resequencing of coding DNA sequences for SNP discovery in nonmodel species. *Mol.*
701 *Ecol. Resour.* 1356–1373.
702 Gautier, M., Foucaud, J., Gharbi, K., Cezard, T., Galan, M., Loiseau, A., *et al.* 2013. Estimation of
703 population allele frequencies from next-generation sequencing data: pool-versus individual-
704 based genotyping. *Mol. Ecol.* **22**: 3766–3779.
705 Gautier, M., Yamaguchi, J., Foucaud, J., Loiseau, A., Ausset, A., Facon, B., *et al.* 2018. The Genomic
706 Basis of Color Pattern Polymorphism in the Harlequin Ladybird. *Curr. Biol.* **28**: 3296–3302.
707 Gouzy, J., Carrere, S. & Schiex, T. 2009. FrameDP: Sensitive peptide detection on noisy matured
708 sequences. *Bioinformatics* **25**: 670–671.
709 Graffelman, J. 2015. Exploring Diallelic Genetic Markers: The HardyWeinberg Package. *J. Stat.*
710 *Softw.* **64**: 1–23.
711 Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D. 2009. Inferring the joint
712 demographic history of multiple populations from multidimensional SNP frequency data.
713 *PLoS Genet.* **5**.
714 Heyduk, K., Trapnell, D.W., Barrett, C.F. & Leebens-Mack, J. 2016. Phylogenomic analyses of
715 species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Biol. J.*
716 *Linn. Soc.* **117**: 106–120.
717 Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., *et al.* 2007. Genome-wide *in*

718 *situ* exon capture for selective resequencing. *Nat. Genet.* **39**: 1522–1527.
719 Ilves, K.L. & López-Fernández, H. 2014. A targeted next-generation sequencing toolkit for exon-
720 based cichlid phylogenomics. *Mol. Ecol. Resour.* **14**: 802–811.
721 Ilves, K.L., Torti, D. & López-Fernández, H. 2018. Exon-based phylogenomics strengthens the
722 phylogeny of Neotropical cichlids and identifies remaining conflicting clades (Cichliformes:
723 Cichlidae: Cichlinae). *Mol. Phylogenet. Evol.* **118**: 232–243.
724 Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., Mclellan, M.D., Lin, L., *et al.* 2012. VarScan 2 :
725 Somatic mutation and copy number alteration discovery in cancer by exome sequencing.
726 *Genome Res.* **22**: 568–576.
727 Langmead, B. & Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**:
728 357–359.
729 Li, H. 2014. Toward better understanding of artifacts in variant calling from high-coverage
730 samples. *Bioinformatics* **30**: 2843–2851.
731 Montes, I., Conklin, D., Albaina, A., Creer, S., Carvalho, G.R., Santos, M., *et al.* 2013. SNP Discovery
732 in European Anchovy (Engraulis encrasicolus, L) by High-Throughput Transcriptome and
733 Genome Sequencing. *PLoS One* **8**: e70051.
734 Neves, L.G., Davis, J.M., Barbazuk, W.B. & Kirst, M. 2013. Whole-exome targeted sequencing of the
735 uncharacterized pine genome. *Plant J.* **75**: 146–156.
736 Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., *et al.* 2009. Targeted
737 capture and massively parallel sequencing of 12 human exomes. *Nature* **461**: 272–276.
738 Nature Publishing Group.
739 Nicholls, J.A., Pennington, R.T., Koenen, E.J.M., Hughes, C.E., Hearn, J., Bunnefeld, L., *et al.* 2015.
740 Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the
741 neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Front. Plant Sci.* **6**: 1–20.
742 Ning, Z., Cox, A.J. & Mullikin, J.C. 2001. SSAHA: a fast search method for large DNA databases. -
743 Abstract - UK PubMed Central. 1725–1729.
744 Pedersen, B.S. & Quinlan, A.R. 2018. Mosdepth: Quick coverage calculation for genomes and
745 exomes. *Bioinformatics* **34**: 867–868.
746 Popp, B., Ekici, A.B., Thiel, C.T., Hoyer, J., Wiesener, A., Kraus, C., *et al.* 2017. Exome Pool-Seq in
747 neurodevelopmental disorders. *Eur. J. Hum. Genet.* **25**: 1364–1376. Springer US.
748 Portik, D.M., Smith, L.L. & Bi, K. 2016. An evaluation of transcriptome-based exon capture for frog
749 phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Mol.*
750 *Ecol. Resour.* **16**: 1069–1083.
751 Puritz, J.B. & Lotterhos, K.E. 2018. Expressed exome capture sequencing: A method for cost-
752 effective exome sequencing for all organisms. *Mol. Ecol. Resour.* 1209–1222.
753 R Development Core Team. 2015. R: A language and environment for statistical computing. R
754 Foundation for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
755 Ramos, E., Levinson, B.T., Chasnoff, S., Hughes, A., Young, A.L., Thornton, K., *et al.* 2012.
756 Population-based rare variant detection via pooled exome or custom hybridization capture
757 with or without individual indexing. *BMC Genomics* **13**: 683.
758 Rellstab, C., Dauphin, B., Zoller, S., Brodbeck, S. & Gugerli, F. 2019. Using transcriptome
759 sequencing and pooled exome capture to study local adaptation in the giga-genome of *Pinus*
760 *cembra*. *Mol. Ecol. Resour.* **19**: 536–551.
761 Ryu, S., Han, J., Norden-Krichmar, T.M., Schork, N.J. & Suh, Y. 2018. Effective discovery of rare
762 variants by pooled target capture sequencing: A comparative analysis with individually
763 indexed target capture sequencing. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **809**: 24–31.
764 Elsevier.

765 Schlötterer, C., Tobler, R., Kofler, R. & Nolte, V. 2014. Sequencing pools of individuals — mining
766 genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15**: 749–763. Nature
767 Publishing Group.
768 Shearer, A.E., Hildebrand, M.S., Ravi, H., Joshi, S., Guiffre, A.C., Novak, B., *et al.* 2012. Pre-capture
769 multiplexing improves efficiency and cost-effectiveness of targeted genomic enrichment.
770 *BMC Genomics* **13**.
771 Stephens, J.D., Rogers, W.L., Heyduk, K., Cruse-Sanders, J.M., Determann, R.O., Glenn, T.C., *et al.*
772 2015. Resolving phylogenetic relationships of the recently radiated carnivorous plant genus
773 *Sarracenia* using target enrichment. *Mol. Phylogenet. Evol.* **85**: 76–87. Elsevier Inc.
774 Vogel, H., Schmidtberg, H. & Vilcinskas, A. 2017. Comparative transcriptomics in three ladybird
775 species supports a role for immunity in invasion biology. *Dev. Comp. Immunol.* **67**: 452–456.
776 Elsevier Ltd.
777 Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N. & Watson, M. 2015. Exome Sequencing:
778 Current and Future Perspectives. *Genes|Genomes|Genetics* **5**: 1543–1550.
779 Wu, T.D. & Watanabe, C.K. 2005. GMAP: A genomic mapping and alignment program for mRNA
780 and EST sequences. *Bioinformatics* **21**: 1859–1875.
781 Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., *et al.* 2010. Sequencing of Fifty
782 Human Exomes Reveals Adaptation to High Altitude. *Science (80-.).* **329**: 75–78.
783
784 Bansal, V., Tewhey, R., LeProust, E.M. & Schork, N.J. 2011. Efficient and cost effective population
785 resequencing by pooling and in-solution hybridization. *PLoS One* **6**: 1–6.
786 Benjamini, Y. & Hochberg, Y. 1995. Controlling the false discovery rate - a practical and powerful
787 approach to multiple testing. *J. R. Stat. Soc. Ser. B-Methodological* **57**: 289–300.
788 Bi, K., Linderoth, T., Vanderpool, D., Good, J.M., Nielsen, R. & Moritz, C. 2013. Unlocking the vault:
789 Next-generation museum population genomics. *Mol. Ecol.* **22**: 6018–6032.
790 Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C. & Good, J.M. 2012. Transcriptome-based
791 exon capture enables highly cost-effective comparative genomic data collection at moderate
792 evolutionary scales. *BMC Genomics* **13**: 403.
793 Bolger, A.M., Lohse, M. & Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence
794 data. *Bioinformatics* **30**: 2114–2120.
795 Bossert, S. & Danforth, B.N. 2018. On the universality of target-enrichment baits for phylogenomic
796 research. *Methods Ecol. Evol.* **9**: 1453–1460.
797 Bragg, J.G., Potter, S., Bi, K. & Moritz, C. 2016. Exon capture phylogenomics: efficacy across scales
798 of divergence. *Mol. Ecol. Resour.* **16**: 1059–1068.
799 Burset, M. & Guigo, R. 1996. Evaluation of Gene Structure Prediction Programs. *Genomics* **34**: 353–
800 367.
801 Castellano, S., Parra, G., Sanchez-Quinto, F.A., Racimo, F., Kuhlwilm, M., Kircher, M., *et al.* 2014.
802 Patterns of coding variation in the complete exomes of three Neandertals. *Proc. Natl. Acad.*
803 *Sci.* **111**: 6666–6671.
804 Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., *et al.* 2012. A program for
805 annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).*
806 **6**: 80–92.
807 Conklin, D., Montes, I., Albaina, A. & Estonba, A. 2013. Improved conversion rates for SNP
808 genotyping of nonmodel organisms. *Proc. Iwbbio 2013 Int. Work. Bioinforma. Biomed. Eng.*
809 127–134.
810 Cosart, T., Beja-Pereira, A., Chen, S., Ng, S.B., Shendure, J. & Luikart, G. 2011. Exome-wide DNA
811 capture and next generation sequencing in domestic and wild species. *BMC Genomics* **12**:

812 347. BioMed Central Ltd.
813 Day-Williams, A.G., McLay, K., Drury, E., Edkins, S., Coffey, A.J., Palotie, A., *et al.* 2011. An
814 evaluation of different target enrichment methods in pooled sequencing designs for complex
815 disease association studies. *PLoS One* **6**: e26279.
816 Förster, D.W., Bull, J.K., Lenz, D., Autenrieth, M., Paijmans, J.L.A., Kraus, R.H.S., *et al.* 2018.
817 Targeted resequencing of coding DNA sequences for SNP discovery in nonmodel species. *Mol.*
818 *Ecol. Resour.* 1356–1373.
819 Gautier, M., Foucaud, J., Gharbi, K., Cezard, T., Galan, M., Loiseau, A., *et al.* 2013. Estimation of
820 population allele frequencies from next-generation sequencing data: pool-versus individual-
821 based genotyping. *Mol. Ecol.* **22**: 3766–3779.
822 Gautier, M., Yamaguchi, J., Foucaud, J., Loiseau, A., Ausset, A., Facon, B., *et al.* 2018. The Genomic
823 Basis of Color Pattern Polymorphism in the Harlequin Ladybird. *Curr. Biol.* **28**: 3296–3302.
824 Gouzy, J., Carrere, S. & Schiex, T. 2009. FrameDP: Sensitive peptide detection on noisy matured
825 sequences. *Bioinformatics* **25**: 670–671.
826 Graffelman, J. 2015. Exploring Diallelic Genetic Markers: The HardyWeinberg Package. *J. Stat.*
827 *Softw.* **64**: 1–23.
828 Heyduk, K., Trapnell, D.W., Barrett, C.F. & Leebens-Mack, J. 2016. Phylogenomic analyses of
829 species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Biol. J.*
830 *Linn. Soc.* **117**: 106–120.
831 Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., *et al.* 2007. Genome-wide *in*
832 *situ* exon capture for selective resequencing. *Nat. Genet.* **39**: 1522–1527.
833 Ilves, K.L. & López-Fernández, H. 2014. A targeted next-generation sequencing toolkit for exon-
834 based cichlid phylogenomics. *Mol. Ecol. Resour.* **14**: 802–811.
835 Ilves, K.L., Torti, D. & López-Fernández, H. 2018. Exon-based phylogenomics strengthens the
836 phylogeny of Neotropical cichlids and identifies remaining conflicting clades (Cichliformes:
837 Cichlidae: Cichlinae). *Mol. Phylogenet. Evol.* **118**: 232–243.
838 Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., Mclellan, M.D., Lin, L., *et al.* 2012. VarScan 2 :
839 Somatic mutation and copy number alteration discovery in cancer by exome sequencing.
840 *Genome Res.* **22**: 568–576.
841 Langmead, B. & Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**:
842 357–359.
843 Li, H. 2014. Toward better understanding of artifacts in variant calling from high-coverage
844 samples. *Bioinformatics* **30**: 2843–2851.
845 Montes, I., Conklin, D., Albaina, A., Creer, S., Carvalho, G.R., Santos, M., *et al.* 2013. SNP Discovery
846 in European Anchovy (Engraulis encrasicolus, L) by High-Throughput Transcriptome and
847 Genome Sequencing. *PLoS One* **8**: e70051.
848 Neves, L.G., Davis, J.M., Barbazuk, W.B. & Kirst, M. 2013. Whole-exome targeted sequencing of the
849 uncharacterized pine genome. *Plant J.* **75**: 146–156.
850 Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., *et al.* 2009. Targeted
851 capture and massively parallel sequencing of 12 human exomes. *Nature* **461**: 272–276.
852 Nature Publishing Group.
853 Nicholls, J.A., Pennington, R.T., Koenen, E.J.M., Hughes, C.E., Hearn, J., Bunnefeld, L., *et al.* 2015.
854 Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the
855 neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Front. Plant Sci.* **6**: 1–20.
856 Ning, Z., Cox, A.J. & Mullikin, J.C. 2001. SSAHA: a fast search method for large DNA databases. -
857 Abstract - UK PubMed Central. 1725–1729.
858 Pedersen, B.S. & Quinlan, A.R. 2018. Mosdepth: Quick coverage calculation for genomes and

859 exomes. *Bioinformatics* **34**: 867–868.
860 Portik, D.M., Smith, L.L. & Bi, K. 2016. An evaluation of transcriptome-based exon capture for frog
861 phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Mol.*
862 *Ecol. Resour.* **16**: 1069–1083.
863 Puritz, J.B. & Lotterhos, K.E. 2018. Expressed exome capture sequencing: A method for cost-
864 effective exome sequencing for all organisms. *Mol. Ecol. Resour.* 1209–1222.
865 R Development Core Team. 2015. R: A language and environment for statistical computing. R
866 Foundation for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
867 Ramos, E., Levinson, B.T., Chasnoff, S., Hughes, A., Young, A.L., Thornton, K., *et al.* 2012.
868 Population-based rare variant detection via pooled exome or custom hybridization capture
869 with or without individual indexing. *BMC Genomics* **13**: 683.
870 Rellstab, C., Dauphin, B., Zoller, S., Brodbeck, S. & Gugerli, F. 2019. Using transcriptome
871 sequencing and pooled exome capture to study local adaptation in the giga-genome of *Pinus*
872 *cembra*. *Mol. Ecol. Resour.* **19**: 536–551.
873 Ryu, S., Han, J., Norden-Krichmar, T.M., Schork, N.J. & Suh, Y. 2018. Effective discovery of rare
874 variants by pooled target capture sequencing: A comparative analysis with individually
875 indexed target capture sequencing. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **809**: 24–31.
876 Elsevier.
877 Schlötterer, C., Tobler, R., Kofler, R. & Nolte, V. 2014. Sequencing pools of individuals — mining
878 genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15**: 749–763. Nature
879 Publishing Group.
880 Shearer, A.E., Hildebrand, M.S., Ravi, H., Joshi, S., Guiffre, A.C., Novak, B., *et al.* 2012. Pre-capture
881 multiplexing improves efficiency and cost-effectiveness of targeted genomic enrichment.
882 *BMC Genomics* **13**.
883 Stephens, J.D., Rogers, W.L., Heyduk, K., Cruse-Sanders, J.M., Determann, R.O., Glenn, T.C., *et al.*
884 2015. Resolving phylogenetic relationships of the recently radiated carnivorous plant genus
885 *Sarracenia* using target enrichment. *Mol. Phylogenet. Evol.* **85**: 76–87. Elsevier Inc.
886 Vogel, H., Schmidtberg, H. & Vilcinskas, A. 2017. Comparative transcriptomics in three ladybird
887 species supports a role for immunity in invasion biology. *Dev. Comp. Immunol.* **67**: 452–456.
888 Elsevier Ltd.
889 Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N. & Watson, M. 2015. Exome Sequencing:
890 Current and Future Perspectives. *Genes|Genomes|Genetics* **5**: 1543–1550.
891 Wu, T.D. & Watanabe, C.K. 2005. GMAP: A genomic mapping and alignment program for mRNA
892 and EST sequences. *Bioinformatics* **21**: 1859–1875.
893 Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., *et al.* 2010. Sequencing of Fifty
894 Human Exomes Reveals Adaptation to High Altitude. *Science (80-.).* **329**: 75–78.