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.

37 # Introduction

38

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

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

61 An alternative approach to capture probe design for non-model species involves the use of
62 a *de novo* assembled transcriptome for the species studied or a related species (e.g. Bi *et al.*, 2012;
63 Neves *et al.*, 2013). One of the limitations of this approach is that the intron-exon structure is
64 unknown, and probes based on transcript sequences spanning two consecutive exons therefore
65 hybridize only partially to genomic DNA, reducing capture efficiency (e.g. Neves *et al.*, 2013). One
66 way of addressing this issue is to use orthologous genes from closely related model species to
67 predict intron positions (Bi *et al.*, 2012; Stephens *et al.*, 2015; Bragg *et al.*, 2016). However, the use
68 of data from closely related model species biases target choice through the retention of genes
69 displaying the highest degree of conservation throughout evolution, which is potentially
70 problematic in population genomics studies. Even when orthologous genes are targeted,
71 transcriptome-based exon capture studies have been found to have a lower capture efficiency
72 close to intron-exon boundaries, because it is not possible to design capture probes spanning
73 exons and their (unknown) non-coding flanking regions (Bi *et al.*, 2012; Neves *et al.*, 2013; Portik *et*
74 *al.*, 2016).

75 Captured fragments are usually longer than probes. Transcriptome-based exon capture
76 therefore provides opportunities to capture the flanking regions of the exons, thereby expanding
77 gene models beyond the available coding sequence (Bi *et al.*, 2012; Neves *et al.*, 2013; Stephens *et*
78 *al.*, 2015; Bragg *et al.*, 2016; Portik *et al.*, 2016). This requires the construction, in a fragmented
79 way (e.g. long introns not fully captured), of the genomic sequences corresponding to the targeted
80 transcripts (e.g. Neves *et al.*, 2013). This assembly (generally referred to as an in-target assembly)
81 is used as a reference for the mapping of genomic reads or for phylogeny construction. This
82 assembly is obtained by cleaning the raw genomic sequences and assembling them with various k-
83 mer combinations. The genomic contigs obtained are then compared with the target sequences
84 for ordering and scaffolding with each other when they correspond to the same targets. Such a

85 complex approach is necessary for phylogeny construction when probes are used to capture
86 orthologous targets from related species, but may be dispensable in the context of population
87 genomics. Indeed, several studies have shown that the assembly process does not completely
88 rebuild all targets (e.g. Bragg *et al.*, 2016; Portik *et al.*, 2016), and this may decrease the likelihood
89 of detecting variants. Bragg et al. (2016) tried to attenuate this problem by mapping the genomic
90 reads of the species used for probe design directly onto the transcript target sequences, thereby
91 making it possible to cover exons that were not successfully assembled but were nevertheless
92 captured and sequenced.

93 Transcriptome-based exon capture has proved effective for the generation of thousands of
94 highly informative markers. Most studies to date have aimed to resolve the phylogenetic
95 relationships between various set of related species (Nicholls *et al.*, 2015; Stephens *et al.*, 2015;
96 Bragg *et al.*, 2016; Heyduk *et al.*, 2016; Ilves *et al.*, 2018), and applications in population genomics
97 remain rare (but see Bi *et al.*, 2013; Rellstab *et al.*, 2019). This is probably because population
98 studies traditionally analyze large numbers of individuals from multiple populations of non-model
99 species, which is complex, time-consuming and costly, even if sequencing capacity and costs are
100 optimized by indexing individuals for multiplexing (Shearer *et al.*, 2012). However, many research
101 questions in population genomics can be addressed by simply using the allele frequencies
102 calculated at population level. In this context, pooling individuals within the population before the
103 extraction of DNA may be an efficient, cheap and time-saving strategy. Pool-sequencing has been
104 shown to estimate allele frequencies within populations as accurately as individual genotyping,
105 but with much less effort for library construction and sequencing (e.g. Gautier *et al.*, 2013;
106 Schlötterer *et al.*, 2014). Allele frequencies have been estimated by a combination of targeted
107 capture methods and pool-sequencing strategies only in humans (Bansal *et al.*, 2011; Day-Williams

108 *et al.*, 2011; Ramos *et al.*, 2012; Ryu *et al.*, 2018) and, very recently, in pine trees (Rellstab *et al.*,
109 2019).

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

125

126

127 # Methods

128

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

130 We designed capture probes for ~5 Mb of randomly chosen coding exonic regions of *H. axyridis*.
131 Using a *de novo* transcriptome obtained by RNA-Seq on various tissues and conditions/stages of

132 the species (Vogel *et al.*, 2017), we sought putative peptide-coding sequences with FRAMEDP
133 (v1.2.2, Gouzy *et al.*, 2009), using bLASTX results (e-value ≤ $1e^{-7}$) of transcripts against the insect
134 proteomes of *Tribolium castaneum*, *Anoplophora glabripennis, Dendroctonus ponderosae,*
135 *Drosophila melanogaster* and the SwissProt database (v2016-02-17) for training. The
136 corresponding partial or complete coding sequences (CDS) were filtered to eliminate (i) *Wolbachia*
137 and other putative endosymbiont sequences, (ii) CDS with more than 1% missing data (*Ns*) or with
138 more than four consecutive *Ns*, (iii) CDS with a GC% below 25 or above 75 and (iv) CDS with
139 lengths below 400 bp (to avoid the capture of too great a proportion of flanking regions and to
140 allow better tiling) or above 3500 bp (to make it possible to target larger numbers of different
141 genes). We identified 28,326 putative CDS, 12,739 were retained after the application of the
142 filters. We then chose 5.5 Mb at random, corresponding to 5,736 CDS. We searched for sequence
143 similarity with complete proteins from the SwissProt database and the *D. melanogaster* and *T.*
144 *castaneum* proteomes with BlastX, and at least 895 CDS (15.6%) were found to be complete.
145 Overall, 5,120 CDS (89.3%) displayed hits with *T. castaneum* proteins (e-value ≤ $1e^{-7}$).

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

156 https://doi.org/10.5281/zenodo.2598388). This probe set was used for the preparation of a
157 SeqCap EZ Developer assay by Roche NimbleGen Technologies, which synthesized the probes as
158 biotinylated DNA oligomers. One capture reaction contained 2,100,000 overlapping probes of 50
159 to 99 bp in length, with a mean length of 74.71 ± 4.92 bp.
160
161 ***Individual and pooled genomic DNA library preparation***
162 Thirty-six individuals of *H. axyridis* from a wild population in China (Beijing; 40.057128°N,
163 116.53989°E) were collected in October 2015, immediately placed in RNAlater solution and stored
164 at -20°C. Based on the PIF tools proposed by Gautier *et al.* (2013), we would theoretically expect a
165 similar precision of allele frequency estimation for the sequencing of 23 individually indexed
166 individuals, each at 80X coverage, or a single pool of 36 individuals at 125X, assuming an
167 experimental error (i.e. ε, the coefficient of variation of the contribution of each individual to the
168 set of sequence reads obtained from the DNA pool) of 50%. We thus extracted DNA (i)
169 independently from 23 individuals, using half the body (without the hind legs), and (ii) from a pool
170 of 36 individuals, including the 23 individuals used for individual sequencing, using two hind legs
171 per individual. All DNA extractions were performed with the Qiagen DNeasy Blood & Tissue kit, in
172 accordance with the manufacturer's recommendations.
173 Genomic libraries were prepared with NimbleGen SeqCap EZ Library (v5.0). Briefly, for each
174 of the 24 samples (1 pool and 23 individuals), DNA (2 µg in 100 µl) was mechanically sheared to a
175 mean fragment length of 200 bp with a Covaris S2 E210 device (6x30s). Fragments were end-
176 repaired, A-tailed and indexed (one index per sample) with KAPA Library Preparation Kit Illumina
177 platforms. After the ligation of Illumina adaptors and indexes, only fragments with of 250 to 450
178 bp in length were retained. PCR was performed (7 cycles) with standard Illumina paired-end
179 primers and the amplicons were then purified with AMPure XP beads (Beckman). The length,

180 quality and concentration of the prepared DNA fragments were checked on a BioAnalyzer (Agilent
181 High Sensitivity DNA Assay) and a Qubit.
182
183 ***Sequence capture hybridization and sequencing***
184 The 23 indexed individuals were pooled at equimolar concentrations before hybridization. For
185 each capture (one with the pool and one with the indexed individuals), a total of 1 µg of amplified
186 DNA was used for exome enrichments, with the SeqCap EZ Developer probes described above,
187 strictly in accordance with the NimbleGen SeqCap EZ Library protocol (v5.0). For each capture, two
188 parallel post-capture PCRs (14 cycles) were performed on elution solution, and the PCR products
189 were then merged before purification with AMPure XP beads. The length, quality and
190 concentration of the final DNA fragments were checked on a BioAnalyzer (Agilent High Sensitivity
191 DNA Assay) and a Qubit. Each capture was sequenced on a half lane of an Illumina HiSeq3000
192 sequencer (on the same lane), in accordance with the manufacturer's instructions, in paired-end
193 mode, for 150 cycles. The data were then demultiplexed, exported as FastQ files and the 24
194 libraries were processed independently.
195
196 ***SNP calling by the mapping of reads to the CDS***
197 Sequence quality was checked with FastQC
198 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adapters were removed and low-
199 quality base pairs were trimmed with Trimmomatic (v0.35, Bolger *et al.*, 2014), with the following
200 parameters: ILLUMINACLIP:TruSeq-file.fa:2:30:10 LEADING:30 TRAILING:30
201 SLIDINGWINDOW:10:30 MINLEN:75, and FASTX (v0.0.14;
202 http://hannonlab.cshl.edu/fastx_toolkit/) to remove the first five bases.

203 We chose to align the reads directly with the targeted CDS sequences, because target
204 enrichment was performed on the same species as probe design, as suggested by Bragg *et al.*
205 (2016; see also Rellstab *et al.*, 2019). Doing so, we avoid the complex and time-consuming steps of
206 read assembly and scaffolding/ordering the genomic contigs obtained to generate the in-target
207 assemblies for read mapping (see introduction). For each library, cleaned reads were aligned with
208 the 5,717 targeted CDS, with BOWTIE2 (v2.2.5; Langmead & Salzberg, 2012), using the default
209 parameters and *local* option to allow for genomic reads not aligning with the targeted CDS
210 sequence over their entire length, whilst ensuring a minimum alignment length of 20 bp. Reads
211 were aligned as unpaired, to allow for the mapping of orphan reads. Reads with multiple
212 alignments, and those for which the mate-pair was aligned to a different target were discarded
213 after mapping. The aligned data in SAM format were converted into BAM format and sorted with
214 SAMtools (v0.1.19). Duplicate reads were estimated with picard tools (v1.113) MarkDuplicates,
215 using the default settings. Coverage at each target base was calculated with mosdepth (v0.2.3;
216 Pedersen & Quinlan, 2018).

217 Variant calling was performed with read alignments of more than 40 bases in length
218 containing fewer than 10 deletions, as deduced from the *cigar* format. SNP sites were called
219 separately for each of the 24 libraries with the SAMtools *mpileup* command, using option *-B*
220 followed by the varscan2 (v2.3; Koboldt *et al.*, 2012) *mpileup2snp* command with a minimum base
221 quality to count reads at a position of 35, with a *p*-value threshold of 0.01. For the other
222 parameters, we considered only positions with a minimum coverage of 40 reads for each
223 individual, and 250 reads for the pool (i.e. for these selected coverages, the same exome size was
224 analyzed: 90.3% and 93% of target bases were covered by at least 40 and 250 reads for the
225 individual and pooled exomes, respectively; see Fig. S1). There must also be at least 15 reads
226 supporting a variant for each individual, and three reads supporting the variant for the pool (for a

227 minimum coverage of 250X for the pool, we expected to have ~3 reads per haplotype if the 72
228 haplotypes in the pool were homogeneously captured). Moreover to account for the variation in
229 total read coverage across positions, we required a minimum variant allele frequency threshold of
230 0.25 for each individual and 0.01 for the pool. For individual libraries, an individual was considered
231 to be homozygous for the variant if the variant allele frequency f(v) was above 0.75, and
232 heterozygous if f(v) was between 0.25 and 0.75.

233 All 409,328 different SNP positions independently found in the 24 libraries were then
234 genotyped for each library with varscan2 *mpileup2cns,* using the same parameters as described
235 above. Several filters were then applied to all SNP positions, using either in-house PERL or R scripts
236 (R Development Core Team, 2015). We first retained only bi-allelic positions. We then retained
237 only positions genotyped for at least 20 individuals. Finally, we discarded SNPs with high levels of
238 coverage in the pool, which were likely to represent repetitive regions or paralogs of the genome:
239 maximum pool coverage was set at 6,960X, as calculated with the formula proposed by Li et al.
240 (2014): $c+4*\sqrt{c}$, where c is the mean coverage in the pool (c=6,634.8X). We retained 300,036 bi-
241 allelic SNPs after the application of all these filters.

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

248

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

250 Approaches based on the direct mapping of reads onto CDS have already been used to detect
251 SNPs (e.g. Rellstab *et al.*, 2019), but such approaches have never been evaluated. We used an *H.*
252 *axyridis* draft genome sequence (HaxR v1.0; Gautier *et al.*, 2018) to evaluate this approach and
253 that proposed below for the prediction of intron-exon boundaries (IEBs). We selected CDS targets
254 with genomic match, as proposed by Neves *et al.* (2013), with some adjustments. Briefly, CDS
255 target sequences were first aligned with genomic scaffolds with BLASTN (e-value cutoff of e^{-20} and
256 identity cutoff of 90%), and the best hit was retained for each target. The target sequences were
257 then realigned with the corresponding genomic scaffold, with both GMAP (version 2015-12-31;
258 Wu & Watanabe, 2005) and EXONERATE *est2genome* (v2.2.0;
259 http://www.genome.iastate.edu/bioinfo/resources/manuals/exonerate/). We isolated targets (i)
260 with genome matches over their entire length, (ii) for which intron/exon structure predictions
261 were the same for both alignment programs, (iii) which were covered by capture probes over their
262 entire length and (iv) for which we had access to at least 200 bp upstream and downstream from
263 the targeted CDS sequences. With these filters, we retained 3,161 CDS targets in total (3,220,718
264 bp, i.e. 60.23% of our targeted exome). These 3,161 full or partial CDS corresponded to 11,754
265 exons, with a mean of 3.72 exons per target (range: 1 to 17 exons) and a mean exon length of 274
266 bp (range: 12 to 3,417 bp). These 11,754 exons included 6,295 that were unambiguously
267 complete, because they were surrounded by two others, and 8,593 reliable IEBs in total were
268 identified between consecutive exons (Zenodo https://doi.org/10.5281/zenodo.2598388).

269 All genomic regions corresponding to these 3,161 CDS targets were then isolated, including
270 intron sequences and 200 bp upstream and downstream from the target sequences.
271 Independently for each library, all cleaned reads were aligned with these genomic regions, with
272 BOWTIE2 (v2.2.5), using the default parameters and the *end-to-end* option (i.e. reads must be
273 aligned along their full length). Reads with multiple alignments, and reads for which the mate-pair

274 was aligned to another target were discarded after mapping. Coverage calculation, SNP calling and
275 variant allele frequency estimation were performed as described above. We established the
276 correspondence between the position of a base in the genomic region and its position in the CDS
277 sequence, using the *vulgar* format of EXONERATE output. SNP positions for which there was a
278 deletion or an insertion between the CDS sequence and its associated genomic sequence (i.e. 559
279 positions) were excluded from the comparison.
280
281 ***Intron-exon boundary detection***
282 Our approach is based solely on the direct mapping of genomic reads to the CDS sequences, by
283 allowing reads to map over less than their full length (i.e. local alignment). Thus, a genomic read
284 with both part of an exon and its flanking sequences (intron or UTR) will start mapping to the
285 corresponding exon in the CDS sequence, and stop mapping at the exon end (Conklin *et al.*, 2013).
286 This will result in the generation of a particular signal if we focus, for a given position, on the
287 number of reads that either begin or end their local alignment at that position (hereafter referred
288 to as *nbBE*). Along a covered exon, *nbBE* should be much lower than the coverage rate, except at
289 the end of the exon, where we should observe a large increase in *nbBE* to a value theoretically
290 equal to the coverage (Fig. S2). This strong increase in *nbBE* should occur either at the base at the
291 exact end of the exon, or a few bases away if the start of the intron sequence is, by chance, highly
292 similar to the start of the next exon sequence on the CDS (Fig. S2). In this second case, it creates
293 an abnormally inflated estimate of coverage locally for a few bases at the boundary of the
294 following exon (Fig. S2 and Fig. S3 for an example). At an IEB (i.e. a junction of two exon ends in
295 the CDS), depending on whether a signal offset occurs for one or both exon ends, the size of the
296 area with abnormally inflated local coverage is variable, but generally less than 10 bp. In addition,
297 the presence of a SNP at some bases at the end of the exon may prevent reads with the variant

298 from aligning up to the IEB position, potentially generating another signal in addition to those of
299 the two exon ends of the IEB. Combinations of these different scenarios may also occur.

300 We developed and tested a method for predicting IEBs on the basis of the increase in *nbBE,*
301 by focusing on the 3,161 targeted transcripts for which genomic sequences were available, for
302 which intron-exon structure was known. We used the mapping output BAM file obtained with the
303 pool library. At each position in the transcript, *nbBE* was calculated from the start read alignment
304 position on the transcript and the *cigar* code, both present in the BAM file. The value of the *nbBE*
305 parameter at each position is a function of coverage (and read length). We took the possible effect
306 of coverage heterogeneity on *nbBE* into account by calculating the median *nbBE* over a window of
307 11 bases centered on the position concerned. A signal (i.e. a predicted IEB), was detected at a
308 position if *nbBE* was *X* times larger than the median, *X* being a tool parameter defined by the user.
309 For the definition of *X*, it may be necessary to identify a subset of transcripts with orthologs in a
310 close model species, and then to determine which value of *X* provides the best prediction of IEBs
311 for this subset. Signals at the exact ends of the CDS were not considered. We took the possible
312 signal deviation (Fig. S2) into account, by extending signal detection to the three bases upstream,
313 and three bases downstream from a position with measured signal (default value for parameter
314 *n*=3). A measured signal therefore predicts an IEB in a window of seven bases in total. An IEB,
315 composed of two exon ends, theoretically generates two close signals with overlapping windows.
316 If these two signals are found at the exact positions of the exon ends (no deviation), then the
317 predicted region measures 8 bp (Fig. S2). If these two signals overlap (e.g. because of the deviation
318 of one of the two signals) then the predicted region for the IEB is 7 bp long. If these signals deviate
319 by up to six bases from each other (i.e. cumulative deviation of eight bases, the deviation being in
320 opposite directions for the two signals) then the predicted region measures 14 bp (Fig. S2). More
321 rarely, if a third signal is detected in the same region, it increases the length of the region by a few

322 bases. If the two signals deviate by more than six bases from each other, then the tool predicts
323 two signal windows separated by a few bases for the IEB concerned. Thus, each transcript was
324 divided into short regions with prediction signals (+) alternating with regions without prediction
325 signals (-). Each region was then assessed as follows: true positive (TP) if the + region contained an
326 IEB; false positive (FP) if the + region did not contain an IEB; true negative (TN) if the - region did
327 not contain an IEB; and false negative (FN) if the - region contained an IEB. We evaluated the
328 performance of the method, by calculating the percentage of true IEBs correctly predicted (i.e.
329 sensitivity (SN)) as TP/(TP+FN) and the percentage of correct predictions (i.e. specificity (SP)) as
330 TP/(TP+FP) (Burset & Guigo, 1996). For a CDS with no introns (i.e. 862 of the 3,161 CDS), SN cannot
331 be evaluated; SP was 100 if no IEB was predicted (good prediction) and 0 otherwise. Scripts are
332 available from https://github.com/edeleury/IEB-finder.
333
334
335 **Results**
336
337 ***Capture data and filtration***
338 In the two target enrichment experiments, we obtained 186,998,614 and 185,457,659 *H. axyridis*
339 raw sequence read pairs for the pool library (SRA ERX3237193) and the 23 indexed individuals
340 (SRA ERX3237194 to ERX3237216), respectively. For each indexed individual library, we obtained a
341 mean of 8,063,376 raw read pairs (range: 6,912,561 to 9,176,292). After cleaning, 91.2% of the
342 raw read sequences were retained (see detailed information for all libraries in Table S1).
343
344 ***Capture efficiency, coverage and duplicates***

345 Rather than assembling the cleaned reads, we chose to map them directly onto the target CDS
346 sequences used for probe design. Mean capture specificity, defined as the percentage of cleaned
347 reads mapping correctly (i.e. meeting our mapping parameters) onto the 5,717 targeted CDS, was
348 83.2% (range over libraries: 80.6%-89.2%; Table S2). Aligned bases accounted for 74.6% (range
349 over all libraries: 72.6%-75.7%) of all cleaned read bases. The global median read alignment length
350 was 144 bp (range over libraries: 143-144 bp, see Table S2), with 62.7% of the mapped reads
351 (range over all libraries: 60.9%-64.3%) mapped over their entire length. A substantial proportion of
352 the mapped reads (22.2%; range over all libraries: 20%-25.9%) were 'orphaned' mates (i.e. only
353 one of the two mates aligned to the target sequences). This percentage strongly suggests that
354 flanking regions of targeted regions were also sequenced with our capture procedure (Yi *et al.*,
355 2010; Neves *et al.*, 2013), although quantification was not possible with this direct mapping
356 approach.

357 The mean percentage of the 5,347,461 targeted bases (5,717 target CDS) covered by at
358 least one read (i.e. the capture sensitivity) was 93.5% (range over all libraries: 93.4%-93.9%, see
359 Fig. S1 and Table S3). Overall, 93% of the targeted bases were covered in all 24 libraries, whereas
360 6% were covered by no read in any of the libraries (Table 1). We identified 491 full targets
361 (292,327 bp) with no read aligned in the 24 libraries. For these uncovered targets, 487 (99.2%) had
362 no match to the *H. axyridis* draft genome either, whereas, for the 5,226 covered targeted CDS,
363 only 74 (1.4%) had no match to the genome (here partial alignments of targets with genome were
364 also considered).

365 Considering all targeted bases, the median and mean base coverage for individual libraries
366 were 142X and 283X for the "23 individual" libraries, and 3,361X and 6,635X for the pool libraries
367 (see Table S3), respectively. A small number of targeted bases had very high coverage, but around
368 90% had a coverage <300X for individuals or <7,000X for the pool (Table S3). For individuals, 90.3%

369 of the targeted bases on average had a mean coverage greater than 40X. For the pool, 93% had a
370 coverage greater than 250X (Fig. S1).

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

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

389

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

391 SNP calling performed independently for the individuals and for the pool, following direct mapping
392 onto the targeted CDS sequences, led to the detection of 302,292 and 364,144 different putative

393 exonic SNPs, respectively, with 257,108 positions (62.8%) detected in both experiments. As
394 expected (because there were 13 more individuals in the pool), the number of SNPs called for the
395 pool was larger than that for the 23 individuals, with 26.2% of SNPs called for the pool only. The
396 vast majority (95.6%) of the 11% of SNPs called only for individuals had allele frequencies below
397 0.03.

398 All positions were genotyped in all libraries. After filtering, we finally selected a total of
399 300,036 exonic bi-allelic SNPs within 4,741 targeted CDS. Minor allele frequencies (MAFs) were
400 below 0.03 for 59.2% of these SNPs (Fig. 1); 92.75% were polymorphic in both the pool and the
401 individuals, 7.22% were polymorphic only in the pool, and 0.03% were polymorphic only in the
402 individuals. Only 1,007 of the total number of SNPs (i.e. 0.34%) displayed significant deviation
403 from Hardy-Weinberg equilibrium, as calculated with the individual genotypes. This number was
404 more than three times higher (i.e. 3,331 loci) before the removal of loci on the basis of the
405 maximum coverage filter applied to the pool.

406 When all 300,036 loci were considered, allele frequency estimates were strongly correlated
407 between the pool and individuals (r=0.99; Fig. 2A). These results were not dependent on the large
408 fraction of SNPs with low allele frequency estimates (r=0.98, calculated for 102,494 loci with MAF
409 ≥ 0.03).

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

412

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

414 We evaluated the ability of the method for mapping reads directly to CDS to call SNPs correctly, by
415 comparing the SNPs identified with this approach and those found by mapping reads to the
416 genome. We used only the subset of 3,161 targeted CDS with a match over their entire length to

417 sequences in the draft genome of *H. axyridis* for this comparison. All trimmed reads were also
418 mapped along their entire length onto the regions of the genome corresponding to this subset.
419 After application of the same procedure and filters for SNP calling as were used for direct mapping
420 onto CDS sequences, 290,620 filtered bi-allelic SNPs were genotyped within the subset, including
421 187,591 (64.5%) in exons and 103,029 (35.5%) in the flanking regions (intron or UTR) of the exons.
422 By mapping reads onto the genome, we demonstrated that allele frequency estimates were
423 strongly correlated between the pool and individuals, for both exonic and flanking SNPs (r=0.99;
424 Fig. 2B).

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

431 Some of the private bi-allelic SNPs were variants identified with both mapping approaches
432 but discarded during filtering in one of the approaches. For example, 1,467 bi-allelic SNPs private
433 to the CDS mapping process had a coverage above the maximum threshold, and 714 were tri-
434 allelic SNPs when mapped onto the genome. If we exclude this latter group, 8,853 and 8,333 were
435 found to be private to CDS mapping and mapping onto genomic regions, respectively. A non-
436 exhaustive analysis of these positions highlighted two possible origins of private SNPs for mapping
437 onto CDS. Some targeted regions (the full CDS or part of the CDS) have homologous copies in the
438 genome that we were unable to identify during target selection and for which the SNPs found
439 were not discarded with the "maximum coverage" filter. We have thus identified 43 CDS with
440 reads mapping onto their genomic regions in a non-unique manner, and these CDS with paralogs

441 displayed 3,141 SNPs, all private to CDS mapping. The second origin was false SNPs occurring
442 when the start of an intron could be aligned with the start of the next exon due to similarities
443 between the two sequences (Fig. S2 and S6, for example). Of the SNPs private to the CDS mapping
444 method, 6.9% (*N*=614) were positioned close to IEB positions (i.e. at the exact end of the exon or
445 on the three adjacent bases), whereas only 0.6% of the 174,307 SNPs common to both mapping
446 methods were found near IEBs.
447
448 ***Prediction of IEBs without read assembly***
449 The method we developed for IEB detection based on the mapping of genomic reads directly onto
450 CDS sequences was evaluated on the 3,161 targeted CDS for which we had genomic information,
451 and therefore knew the positions of the IEBs. An IEB signal was predicted at a particular position in
452 the CDS if the number of reads either beginning or ending their local alignment at that position
453 (*nbBE*) was 20 times greater (*X*=20) than the median *nbBE* value within a window of 11 bases
454 around the position concerned, and the signal measured at that position extended to a short
455 region of 7 bases (i.e. with a default parameter of *n*=3; see Methods section and Fig. S2). With
456 parameters *X*=20 and *n*=3, the number of 8,967 regions with lengths of 4 to 19 bp, were predicted
457 to be IEBs (Table 2). Regions of less than seven bases in length were next to non-covered regions
458 or regions with a measured signal close to the end of the CDS for which signal extension was not
459 possible. Of the 8,967 regions predicted to be IEBs, 8,575 actually corresponded to true IEBs (i.e.
460 method specificity = 95.63%). Of the 8,632 regions truly containing an IEB, 8,575 were accurately
461 predicted to be IEB regions (i.e. sensitivity = 99.34%; Table 2). Of the 862 CDS with no intron, 763
462 were correctly predicted to have no IEBs. As expected, the use of a lower detection threshold (i.e.
463 *X*=10) decreased the specificity of the method and increased its sensitivity (Table 2). The regions
464 predicted to have no IEB (i.e. the - regions) were between 1 and 3,417 bp in length, with a mean

465 length of 265.5 bp, entirely consistent with the known distribution of exon lengths (see Methods
466 section). The 153 very small (<10 bp) - regions attested to the possible cumulative deviation of
467 signals of more than eight bases for a small number of IEBs. Increasing the length of the signal
468 deviation (e.g. *n*=5 bases instead of 3) made it possible to take this situation into account in the
469 vast majority of cases. This slightly improved performance but increased the length, and thus, the
470 imprecision, of all predicted IEB positions (Table 2).
471
472
473 **Discussion**
474
475 Exome-targeted capture associated with next-generation sequencing has been mainly restricted to
476 model species for which genomic resources are available (Ng *et al.*, 2009). Transcriptome-based
477 exome capture has proved an efficient alternative for dealing with non-model species (e.g. Bi *et*
478 *al.*, 2012), but is mostly used in phylogeny studies. Here, we propose an extension of this method
479 with a high capture efficiency, providing accurate estimates of the allele frequencies of SNPs for
480 population genomics studies. Our methodology is original because (i) it combines transcriptome-
481 based exome capture with a pool-based approach (pooling the tissues of numerous individuals
482 before DNA extraction), (ii) it does not require a genome sequence for the species itself or another
483 closely related species, and (iii) it does not involve the *de novo* assembly of genomic reads. This
484 work opens up interesting prospects for population genomics studies on non-model species
485 because it is cheap and less labor-intensive than other methods.
486
487 ***Capture efficiency***

488 We found that capture efficiency was very high for the *H. axyridis* exome. The capture specificity
489 (83.2%) achieved was much higher than reported in more classical transcriptome-based studies
490 (range 25% - 73%), whereas capture sensitivity (93.5%) approached the highest reported values
491 (see Puritz & Lotterhos, 2018). Moreover, despite the differences in coverage observed between
492 targets and along the length of targets, reproducibility was high between libraries, whether in
493 terms of specificity, sensitivity or coverage. Our capture approach, based on a random choice of
494 targets from the *de novo* transcriptome of the species of interest itself, without the need to
495 identify orthologous genes in related model species, proved highly effective. This "blind" approach
496 had only a minor impact on capture performance: we only observed a slight decrease in coverage
497 for small exons and as towards the ends of the exons. Similar patterns have been observed in
498 classical transcriptome-based exome capture in which IEBs are predicted by orthology, due to the
499 use of less efficient probes designed from the sequences of two consecutive exons and the
500 absence of a probe straddling the end of the exon and the beginning of the intron (e.g. Bi *et al.*,
501 2012; Neves *et al.*, 2013).

502 The random choice of targets was probably the cause of the slightly lower capture
503 sensitivity in this study than in previous studies (Bi *et al.*, 2012; Puritz & Lotterhos, 2018). Targets
504 selected without looking for orthology in related species are necessarily only partly known, and
505 some may belong to other species (e.g. symbionts or parasites) or be chimeras present in the
506 transcriptome used. This is a small price to pay given the efficiency of the method, a feature of
507 considerable interest for population genomics studies. The absence of searches for orthologous
508 genes in related model species prevents the analysis from focusing purely on the most conserved
509 genes. Moreover, our method can be applied to any species for which no closely related model
510 species with genomic resources is available.

511

512 ***SNP calling and the pool-based approach***

513 With this method, we were able to genotype a very large number of SNPs (~300,000) over the
514 exome of *H. axyridis*. The study of the subset of CDS with complete genomic sequence matches
515 further confirmed that direct mapping onto the CDS has no major impact on SNP identification or
516 on the estimation of allele frequencies relative to mapping onto the genome. In addition to a few
517 residual paralogous sequences inherent to any approach dealing with non-model species, we
518 observed slightly larger numbers of SNPs called on the basis of mapping onto CDS in areas very
519 close to the IEBs, but the proportion of SNPs involved was minimal (0.33%).

520 For similar sequencing efforts, strong concordance was observed between the allele
521 frequencies for individual genotyping and those for the "pool-based" approach, whatever the
522 mapping method used. This finding demonstrates the feasibility of this approach, in situations in
523 which individual identification for many samples is practically impossible, technically difficult, too
524 costly and/or unnecessary. In population genomics studies, allele frequencies are sufficient for the
525 calculation of most summary statistics (e.g. nucleotide diversity, FST) and, thus, for addressing
526 most of the questions posed (Schlötterer *et al.*, 2014).

527 The use of a pool of individuals makes it impossible to remove SNPs that are not at Hardy-
528 Weinberg equilibrium. Such disequilibrium may be caused by copy number variations or
529 paralogous sequences resulting in false-positive SNPs and local over-coverage (Li, 2014). We show
530 here that the application of our "maximum coverage" filter to the pooled data made it possible to
531 discard almost 70% of the SNPs not at equilibrium, as shown in individuals. The extremely small
532 remaining proportion (0.34%) of SNPs not at Hardy-Weinberg equilibrium provides support for the
533 feasibility of our pooling-based method.

534

535 ***Method for predicting IEBs without read assembly***

536 We have developed a method for detecting IEBs that is particularly useful in the case of non-
537 model species (or species for which genome sequences are being assembled). The proposed tool
538 does not indicate the exact location of the IEBs, instead identifying an area of a few bases (about
539 eight, on average, in our case) predicted to contain an IEB. This method, which requires only the
540 genomic reads and the target CDS sequences, was effective when evaluated on a subset of CDS for
541 which we had a complete genomic match. Almost all the true IEBs were identified, whereas false-
542 positives accounted for less than 5% of the predicted IEBs. Montes *et al.* (2013) proposed a similar
543 approach, which they used to discard putative SNPs found in predicted IEB areas. They obtained
544 better genotyping results for the selected filtered SNPs than other similar studies on non-model
545 species. In this study, we went a step further by taking into account the possible offset of the
546 measured signal, and by evaluating the performance of the approach on a set of transcripts for
547 which the true intron-exon structure was known. In this study, we aimed to present evidence that
548 this approach is promising, although further improvement is required, together with evaluation on
549 larger dataset and at various level of coverage. A method for choosing an appropriate value for
550 parameter *X* for the correct identification of significant increases in *nbBE* depending on local
551 coverage must also be developed.
552
553 ***General recommendations, conclusion and perspectives***
554 The combination of approaches proposed here — i.e. exome transcriptome-based capture,
555 random choice of targets, and direct mapping of reads onto targeted CDS sequences, coupled with
556 a pool-seq approach before DNA extraction — significantly decreases costs, the laboratory time
557 required and the complexity of the sequencing analyses required to obtain a large number of
558 variants in exome-scale coding regions for non-model species. However, direct mapping onto
559 target sequences has two disadvantages. First, it provides no information on flanking regions

560 (including regulatory regions), and it is not, therefore, possible to genotype SNPs in these areas.
561 Nevertheless, detecting SNP polymorphism at exonic level would be sufficient in many cases,
562 particularly when the in-depth analysis of a large number of synonymous and non-synonymous
563 mutations is informative enough to address the question posed. It is still feasible to assemble the
564 genomic reads in situations in which flanking regions appear to be of interest. Second, direct
565 mapping onto CDS results in small biases in the detection of SNPs in the vicinity of IEBs. As a
566 means of limiting the risk of detecting false SNPs further, we therefore suggest using the IEB
567 detection method proposed here, followed by the elimination of any SNPs found in the predicted
568 IEB regions and in short regions (≤ 20 bp) surrounded by two predicted IEB regions. Using this
569 approach on our data with the parameters X=20 and n=3, we retained a final list of 296,736 SNPs.
570 A large proportion of the 3,300 SNPs excluded (408 SNPs, i.e. 12.36%) were not at Hardy Weinberg
571 equilibrium in individuals, confirming the relevance of this filter.

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

577 In conclusion, our study demonstrates the feasibility of obtaining, at low cost, for a non-
578 model species (i.e. without a reference genome), important and accurate information at
579 population level concerning (i) SNP polymorphism within the exome and (ii) the general structure
580 of the exome (intron-exon structure of the targeted genes). The results obtained with this method
581 applied to the Asian ladybird *Harmonia axyridis* were very promising, and we now need to
582 investigate other species with different genome sizes and complexities. Finally, the oversizing of

these experiments, resulting in very high levels of coverage, suggests that it should be possible to reduce costs further by capturing multiple indexed pools in a single reaction.

Acknowledgments

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

Author contributions

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

Competing interests

The authors have no competing interests to declare.

606

607

608 ## Data accessibility

609

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

615

616

617 ## References

618

619 Bansal, V. 2010. A statistical method for the detection of variants from next-generation
620 resequencing of DNA pools. *Bioinformatics* **26**: 318–324.

621 Bansal, V., Tewhey, R., LeProust, E.M. & Schork, N.J. 2011. Efficient and cost effective population
622 resequencing by pooling and in-solution hybridization. *PLoS One* **6**: 1–6.

623 Benjamini, Y. & Hochberg, Y. 1995. Controlling the false discovery rate - a practical and powerful
624 approach to multiple testing. *J. R. Stat. Soc. Ser. B-Methodological* **57**: 289–300.

625 Bi, K., Linderoth, T., Vanderpool, D., Good, J.M., Nielsen, R. & Moritz, C. 2013. Unlocking the vault:
626 Next-generation museum population genomics. *Mol. Ecol.* **22**: 6018–6032.

627 Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C. & Good, J.M. 2012. Transcriptome-based
628 exon capture enables highly cost-effective comparative genomic data collection at moderate

evolutionary scales. *BMC Genomics* **13**: 403.

Bolger, A.M., Lohse, M. & Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bossert, S. & Danforth, B.N. 2018. On the universality of target-enrichment baits for phylogenomic research. *Methods Ecol. Evol.* **9**: 1453–1460.

Bragg, J.G., Potter, S., Bi, K. & Moritz, C. 2016. Exon capture phylogenomics: efficacy across scales of divergence. *Mol. Ecol. Resour.* **16**: 1059–1068.

Burset, M. & Guigo, R. 1996. Evaluation of Gene Structure Prediction Programs. *Genomics* **34**: 353–367.

Castellano, S., Parra, G., Sanchez-Quinto, F.A., Racimo, F., Kuhlwilm, M., Kircher, M., *et al.* 2014. Patterns of coding variation in the complete exomes of three Neandertals. *Proc. Natl. Acad. Sci.* **111**: 6666–6671.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., *et al.* 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).* **6**: 80–92.

Conklin, D., Montes, I., Albaina, A. & Estonba, A. 2013. Improved conversion rates for SNP genotyping of nonmodel organisms. *Proc. Iwbbio 2013 Int. Work. Bioinforma. Biomed. Eng.* 127–134.

Cosart, T., Beja-Pereira, A., Chen, S., Ng, S.B., Shendure, J. & Luikart, G. 2011. Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics* **12**: 347. BioMed Central Ltd.

Day-Williams, A.G., McLay, K., Drury, E., Edkins, S., Coffey, A.J., Palotie, A., *et al.* 2011. An evaluation of different target enrichment methods in pooled sequencing designs for complex disease association studies. *PLoS One* **6**: e26279.

653 Förster, D.W., Bull, J.K., Lenz, D., Autenrieth, M., Paijmans, J.L.A., Kraus, R.H.S., *et al.* 2018.
654 Targeted resequencing of coding DNA sequences for SNP discovery in nonmodel species. *Mol.*
655 *Ecol. Resour.* 1356–1373.
656 Gautier, M., Foucaud, J., Gharbi, K., Cezard, T., Galan, M., Loiseau, A., *et al.* 2013. Estimation of
657 population allele frequencies from next-generation sequencing data: pool-versus individual-
658 based genotyping. *Mol. Ecol.* **22**: 3766–3779.
659 Gautier, M., Yamaguchi, J., Foucaud, J., Loiseau, A., Ausset, A., Facon, B., *et al.* 2018. The Genomic
660 Basis of Color Pattern Polymorphism in the Harlequin Ladybird. *Curr. Biol.* **28**: 3296–3302.
661 Gouzy, J., Carrere, S. & Schiex, T. 2009. FrameDP: Sensitive peptide detection on noisy matured
662 sequences. *Bioinformatics* **25**: 670–671.
663 Graffelman, J. 2015. Exploring Diallelic Genetic Markers: The HardyWeinberg Package. *J. Stat.*
664 *Softw.* **64**: 1–23.
665 Heyduk, K., Trapnell, D.W., Barrett, C.F. & Leebens-Mack, J. 2016. Phylogenomic analyses of
666 species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Biol. J.*
667 *Linn. Soc.* **117**: 106–120.
668 Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., *et al.* 2007. Genome-wide *in*
669 *situ* exon capture for selective resequencing. *Nat. Genet.* **39**: 1522–1527.
670 Ilves, K.L. & López-Fernández, H. 2014. A targeted next-generation sequencing toolkit for exon-
671 based cichlid phylogenomics. *Mol. Ecol. Resour.* **14**: 802–811.
672 Ilves, K.L., Torti, D. & López-Fernández, H. 2018. Exon-based phylogenomics strengthens the
673 phylogeny of Neotropical cichlids and identifies remaining conflicting clades (Cichliformes:
674 Cichlidae: Cichlinae). *Mol. Phylogenet. Evol.* **118**: 232–243.
675 Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., Mclellan, M.D., Lin, L., *et al.* 2012. VarScan 2:
676 Somatic mutation and copy number alteration discovery in cancer by exome sequencing.

677 *Genome Res.* **22**: 568–576.

678 Langmead, B. & Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**:
679 357–359.

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

682 Montes, I., Conklin, D., Albaina, A., Creer, S., Carvalho, G.R., Santos, M., *et al.* 2013. SNP Discovery
683 in European Anchovy (Engraulis encrasicolus, L) by High-Throughput Transcriptome and
684 Genome Sequencing. *PLoS One* **8**: e70051.

685 Neves, L.G., Davis, J.M., Barbazuk, W.B. & Kirst, M. 2013. Whole-exome targeted sequencing of the
686 uncharacterized pine genome. *Plant J.* **75**: 146–156.

687 Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., *et al.* 2009. Targeted
688 capture and massively parallel sequencing of 12 human exomes. *Nature* **461**: 272–276.
689 Nature Publishing Group.

690 Nicholls, J.A., Pennington, R.T., Koenen, E.J.M., Hughes, C.E., Hearn, J., Bunnefeld, L., *et al.* 2015.
691 Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the
692 neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Front. Plant Sci.* **6**: 1–20.

693 Ning, Z., Cox, A.J. & Mullikin, J.C. 2001. SSAHA: a fast search method for large DNA databases. -
694 Abstract - UK PubMed Central. 1725–1729.

695 Pedersen, B.S. & Quinlan, A.R. 2018. Mosdepth: Quick coverage calculation for genomes and
696 exomes. *Bioinformatics* **34**: 867–868.

697 Portik, D.M., Smith, L.L. & Bi, K. 2016. An evaluation of transcriptome-based exon capture for frog
698 phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Mol.*
699 *Ecol. Resour.* **16**: 1069–1083.

700 Puritz, J.B. & Lotterhos, K.E. 2018. Expressed exome capture sequencing: A method for cost-

701 effective exome sequencing for all organisms. *Mol. Ecol. Resour.* 1209–1222.

702 R Development Core Team. 2015. R: A language and environment for statistical computing. R

703 Foundation for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

704 Ramos, E., Levinson, B.T., Chasnoff, S., Hughes, A., Young, A.L., Thornton, K., *et al.* 2012.

705 Population-based rare variant detection via pooled exome or custom hybridization capture

706 with or without individual indexing. *BMC Genomics* **13**: 683.

707 Rellstab, C., Dauphin, B., Zoller, S., Brodbeck, S. & Gugerli, F. 2019. Using transcriptome

708 sequencing and pooled exome capture to study local adaptation in the giga-genome of *Pinus*

709 *cembra*. *Mol. Ecol. Resour.* **19**: 536–551.

710 Ryu, S., Han, J., Norden-Krichmar, T.M., Schork, N.J. & Suh, Y. 2018. Effective discovery of rare

711 variants by pooled target capture sequencing: A comparative analysis with individually

712 indexed target capture sequencing. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **809**: 24–31.

713 Elsevier.

714 Schlötterer, C., Tobler, R., Kofler, R. & Nolte, V. 2014. Sequencing pools of individuals — mining

715 genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15**: 749–763. Nature

716 Publishing Group.

717 Shearer, A.E., Hildebrand, M.S., Ravi, H., Joshi, S., Guiffre, A.C., Novak, B., *et al.* 2012. Pre-capture

718 multiplexing improves efficiency and cost-effectiveness of targeted genomic enrichment.

719 *BMC Genomics* **13**.

720 Stephens, J.D., Rogers, W.L., Heyduk, K., Cruse-Sanders, J.M., Determann, R.O., Glenn, T.C., *et al.*

721 2015. Resolving phylogenetic relationships of the recently radiated carnivorous plant genus

722 *Sarracenia* using target enrichment. *Mol. Phylogenet. Evol.* **85**: 76–87. Elsevier Inc.

723 Vogel, H., Schmidtberg, H. & Vilcinskas, A. 2017. Comparative transcriptomics in three ladybird

724 species supports a role for immunity in invasion biology. *Dev. Comp. Immunol.* **67**: 452–456.

725 Elsevier Ltd.

726 Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N. & Watson, M. 2015. Exome Sequencing:

727 Current and Future Perspectives. *Genes|Genomes|Genetics* **5**: 1543–1550.

728 Wu, T.D. & Watanabe, C.K. 2005. GMAP: A genomic mapping and alignment program for mRNA

729 and EST sequences. *Bioinformatics* **21**: 1859–1875.

730 Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., *et al.* 2010. Sequencing of Fifty

731 Human Exomes Reveals Adaptation to High Altitude. *Science (80-.).* **329**: 75–78.

732

733 # Figures and Tables

734

735

736

737

738

739

740

741

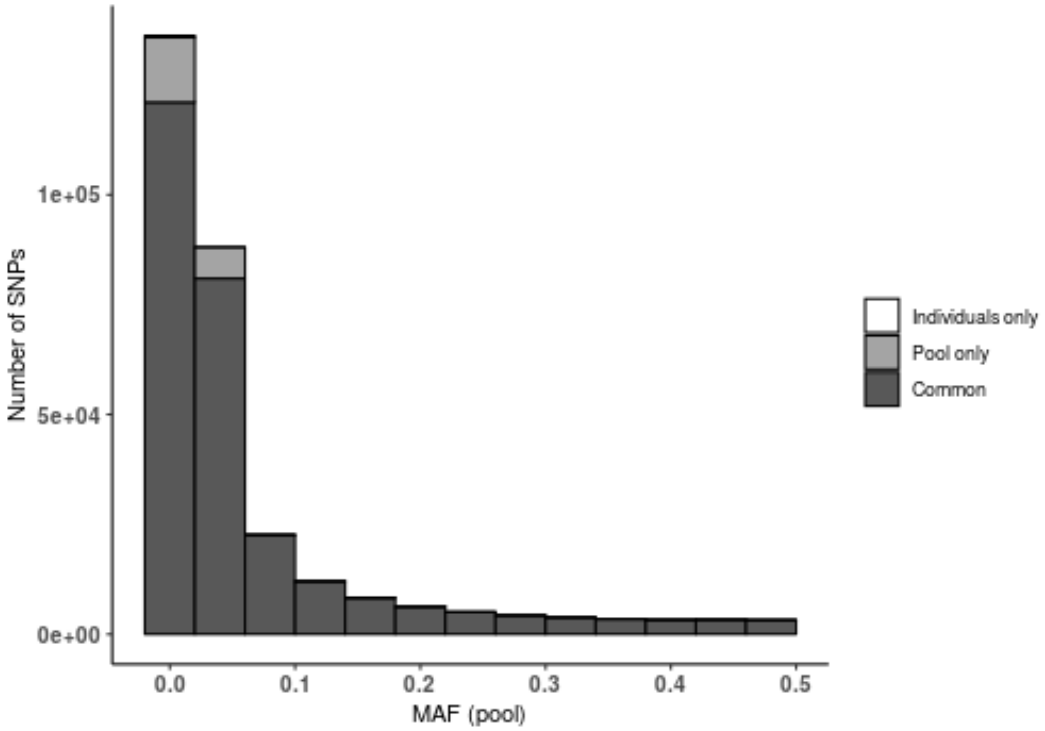
742

743

744

745

746



747

748

749 **Figure 1:** Folded-site frequency spectrum of the 300,036 filtered exonic bi-allelic SNPs (mapping onto

750 targeted CDS sequences) according to their polymorphism in the pool and in individuals. For SNPs

751 polymorphic in both the pool and individuals, the MAF shown is the MAF calculated for the pool.

752

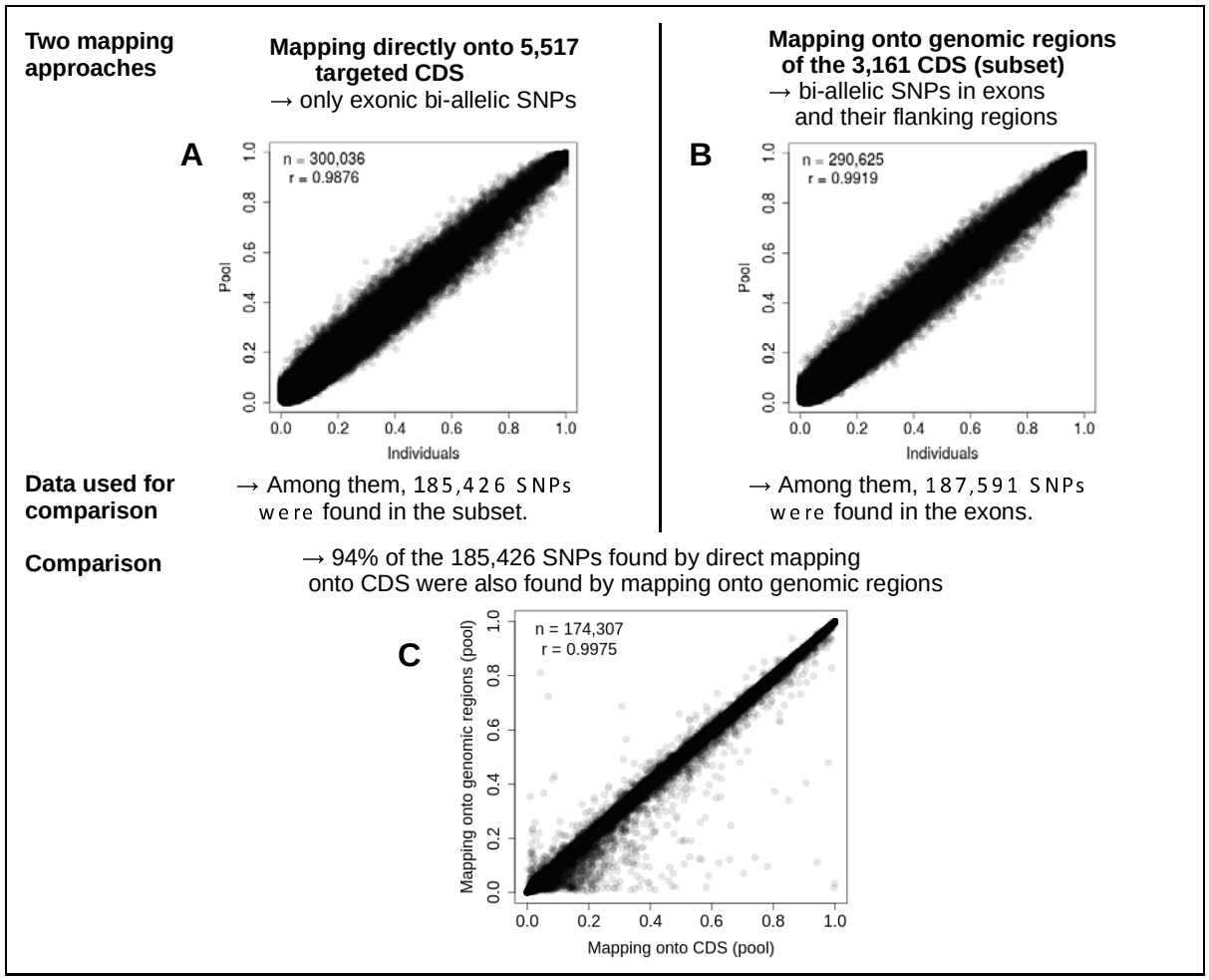


Figure 2: Correlation of the SNP allele frequencies estimated for the population pool and individuals by two mapping approaches: **(A)** mapping reads directly onto CDS and **(B)** mapping onto genomic regions available only for a subset of target CDS. For exonic bi-allelic SNPs common to both mapping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

759 **Table 1:** Reproducibility of capture sensitivity on the 24 libraries. 100% means that all 24 samples had at

760 least one read aligned with the target, whereas 0 % level indicates that no reads aligned with the target.

761

		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%)

762 **Table 2:** Evaluation of the IEB prediction method using pool data. The + region is a short region predicted to be an IEB. The - region is a region with no signal

763 (i.e. predicted to be an exon). Regions not covered by reads are regions without prediction. Sensitivity is the percentage of regions with true IEBs correctly

764 predicted to be + regions, and specificity is the percentage of + regions that are true IEBs.

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

765

766