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 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 to obtain genotype calls for
39 a reliable set of orthologous loci from a sufficient number of individuals across multiple populations
40 and from species spanning a range of divergences, respectively. The targeted resequencing of a
41 consistent subset of genomic regions is a reproducible and cost-effective technical solution, with
42 much lighter data handling than for whole-genome sequencing. Hybridization capture is one of
43 these reduced representation methods that allows the enrichment of a preselected set of hundreds
44 to thousands of genes or DNA fragments from the genomic DNA (Hodges *et al.*, 2007). The most
45 common application of hybridization sequence capture is exon capture, in which only coding regions
46 of the genome (and their flanking regions) are sequenced (Warr *et al.*, 2015). This approach provides
47 focused information about gene function and adaptation, and, unlike transcriptome sequencing, is
48 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 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 for non-model species involves designing DNA capture probes on
60 the basis of sequence subset obtained from a *de novo* assembled transcriptome for the species
61 studied or a related species (e.g. Bi *et al.*, 2012; Neves *et al.*, 2013). One of the limitations of this
62 approach is that the intron-exon structure is unknown, and probes based on transcript sequences
63 spanning two consecutive exons therefore hybridize only partially to genomic DNA, reducing
64 capture efficiency (e.g. Neves *et al.*, 2013). One way of addressing this issue is to choose only targets
65 which have orthologous genes in closely related model species to predict intron positions (Bi *et al.*,
66 2012; Stephens *et al.*, 2015; Bragg *et al.*, 2016). However, the use of data from closely related model
67 species biases target choice toward the retention of genes displaying the highest degree of
68 conservation throughout evolution, which is potentially problematic in population genomics
69 studies. In addition, even when orthologous genes are targeted, transcriptome-based exon capture
70 studies have still been found to have a lower capture efficiency close to intron-exon boundaries,
71 because it is not possible to design capture probes spanning exons and their (unknown) non-coding
72 flanking regions (Bi *et al.*, 2012; Neves *et al.*, 2013; Portik *et al.*, 2016).
73 Captured fragments are usually longer than probes. Transcriptome-based exon capture
74 therefore provides opportunities to capture the flanking regions of the exons, thereby expanding
75 gene models beyond the available coding sequence (Bi *et al.*, 2012; Neves *et al.*, 2013; Stephens *et*
76 *al.*, 2015; Bragg *et al.*, 2016; Portik *et al.*, 2016). This requires the construction, in a fragmented way
77 (e.g. long introns not fully captured), of the genomic sequences corresponding to the targeted
78 transcripts (e.g. Neves *et al.*, 2013). This assembly (generally referred to as an in-target assembly) is
79 used as a reference for the mapping of genomic reads or for phylogeny construction. This assembly
80 is obtained by cleaning the raw genomic sequences and assembling them with various k-mer
81 combinations. The genomic contigs obtained are then compared with the target sequences for
82 ordering and scaffolding with each other when they correspond to the same targets. Such a complex

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

107 We propose here an efficient way of quantifying the single-nucleotide polymorphism (SNP)
108 of exomes at the population level in non-model species for which no genome sequence is available.
109 This approach is based on a combination of transcriptome-based exon capture and the pooling of
110 tissues from numerous individuals before DNA extraction. We report the development of capture
111 probes on a subset of the exome of the Asian ladybird *Harmonia axyridis* from its unannotated
112 available *de novo* transcriptome (Vilcinskas *et al.*, 2013; Vogel *et al.*, 2017), without the use of the
113 available draft genome (HaxR v1.0; Gautier *et al.*, 2018) and, thus, in the absence of known gene
114 structure, as in most non-model species. We used these probes for two target enrichment
115 experiments with the same population sample, the first based on capture with indexed multiplexed
116 individuals, and the second based on capture with a pool of non-indexed individuals. We evaluated
117 the efficiency of capture and the accuracy of the estimates of allele frequencies obtained by pool
118 sequencing in this context. We also evaluated a rapid approach for the processing of read data
119 without their assembly, based on the mapping of reads directly to targeted *de novo* transcript
120 sequences for the detection of variant loci (Bragg *et al.*, 2016; Rellstab *et al.*, 2019). This direct
121 mapping also makes it possible to effectively predict the locations of the intron-exon boundaries of
122 target sequences.
123
124

125 # Methods

126

127 The overall analysis workflow is summarized in Figure 1.

128

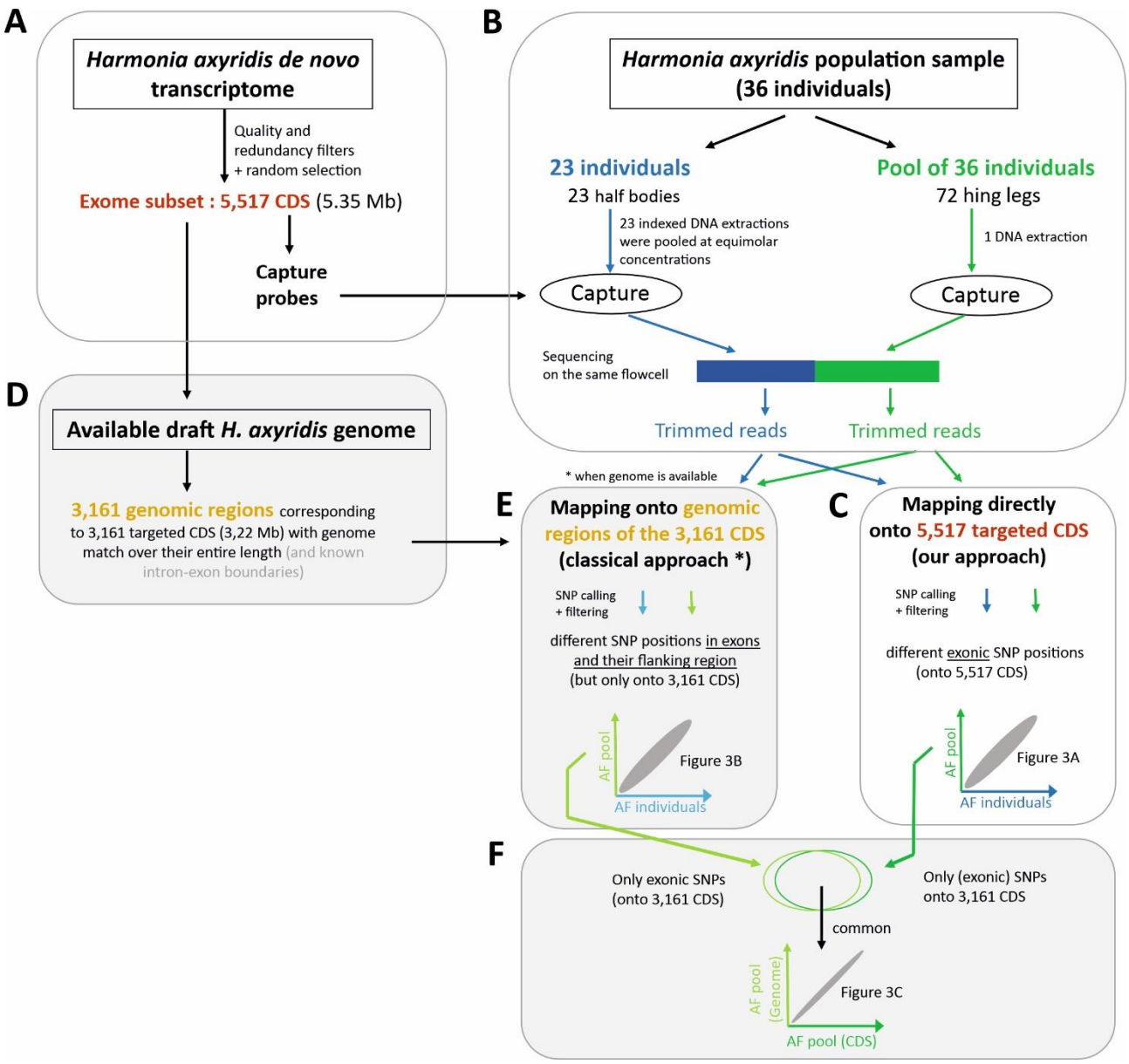

Figure 1: Overview the analysis workflow (see text for details).

Design of capture probes for target enrichment

Using an already available *de novo* transcriptome obtained by RNA-Seq on various tissues and conditions/stages of *H. axyridis* (Vilcinskas *et al.*, 2013; Vogel *et al.*, 2017), we sought putative peptide-coding sequences with FRAMEDP (v1.2.2, Gouzy *et al.*, 2009), using BLASTX results (e-value $\leq 1e^{-7}$) of transcripts against the insect proteomes of *Tribolium castaneum*, *Anoplophora glabripennis, Dendroctonus ponderosae, Drosophila melanogaster* and the SwissProt database (v2016-02-17) for training. We have chosen to work on an exome subset, particularly with the future

139 goal of working simultaneously on a large number of population samples. To maximize the
140 effectiveness of our probes, target sequences were selected as follows. The corresponding partial
141 or complete coding sequences (CDS) were filtered to eliminate (i) *Wolbachia* and other putative
142 endosymbiont sequences, (ii) CDS with more than 1% missing data (*Ns*) or with more than four
143 consecutive *Ns*, (iii) CDS with a GC% below 25 or above 75 and (iv) CDS with lengths below 400 bp
144 (to avoid the capture of too great a proportion of flanking regions and to allow better tiling) or above
145 3500 bp (to make it possible to target larger numbers of different genes). We identified 28,326
146 putative CDS, of which 12,739 were retained after the application of the filters. We then chose a
147 subset of 5.5 Mb at random: the index (1 to 12,739) of the CDS was drawn from a discrete uniform
148 distribution U[1;12,739] without replacement until the desired subset exome size is reached.

149 The probes based on the selected CDS were designed and produced by NimbleGen. The
150 repetitiveness of the probe sequences was assessed on the basis of the frequency of 15-mers in the
151 genomes of the coleopterans *T. castaneum* and *A. glabripennis*. Probes with more than one close
152 match, defined as no more than five single-base insertions, deletions or substitutions with the
153 SSAHA algorithm (Ning *et al.*, 2001), in the *H. axyridis de novo* transcriptome or in the draft genome
154 of *A. glabripennis* were discarded. Probes matching to sequences in the *H. axyridis* or *T. castaneum*
155 mitochondrial genome were also discarded. The few residual *Ns* in target sequences were replaced
156 with a random nucleotide to allow tile probing across a single unknown base. The final probe set
157 contained 6,400 regions of overlapping probes, corresponding to a targeted exome size of 5,347,461
158 bases distributed over 5,717 of our selected CDS (Fig. 1A; Zenodo
159 https://doi.org/10.5281/zenodo.2598388). This final probe set was used for the preparation of a
160 SeqCap EZ Developer assay by Roche NimbleGen Technologies, which synthesized the probes as
161 biotinylated DNA oligomers. One capture reaction contained 2,100,000 overlapping probes of 50 to
162 99 bp in length, with a mean length of 74.71 ± 4.92 bp.

163

164 ***Individual and pooled genomic DNA library preparation***

165 Thirty-six individuals of *H. axyridis* from a wild population in China (Beijing; 40.057128°N,
166 116.53989°E) were collected in October 2015, immediately placed in RNAlater solution and stored
167 at -20°C. Based on the PIFs ("Pool Individual Frequency SNP") tools proposed by Gautier *et al.* (2013),
168 we would theoretically expect a similar precision of allele frequency estimation for the sequencing
169 of 23 individually indexed individuals, each at a sequencing depth of 80X, or a single pool of 36
170 individuals at 125X, assuming an experimental error (i.e. ε, the coefficient of variation of the
171 contribution of each individual to the set of sequence reads obtained from the DNA pool) of 50%.
172 We thus extracted DNA (i) independently from 23 individuals, using half the body (without the hind
173 legs), and (ii) from a pool of 36 individuals, including the 23 individuals used for individual
174 sequencing, using two hind legs per individual (Fig. 1B). All DNA extractions were performed with
175 the Qiagen DNeasy Blood & Tissue kit, in accordance with the manufacturer's recommendations.

176 Genomic libraries were prepared with NimbleGen SeqCap EZ Library (v5.0). Briefly, for each
177 of the 24 samples (1 pool and 23 individuals), DNA (2 µg in 100 µl) was mechanically sheared to a
178 mean fragment length of 200 bp with a Covaris S2 E210 device (6x30s). Fragments were end-
179 repaired, A-tailed and indexed (one index per sample) with KAPA Library Preparation Kit Illumina
180 platforms. After the ligation of Illumina adaptors and indexes, only fragments with a length of 250
181 to 450 bp were retained. PCR was performed (7 cycles) with standard Illumina paired-end primers
182 and the amplicons were then purified with AMPure XP beads (Beckman). The length, quality and
183 concentration of the prepared DNA fragments were checked on a BioAnalyzer (Agilent High
184 Sensitivity DNA Assay) and a Qubit.

185

186 ***Sequence capture hybridization and sequencing***

187 The 23 indexed individuals were pooled at equimolar concentrations before hybridization. For each
188 capture (one with the pool and one with the indexed individuals), a total of 1 µg of amplified DNA
189 was used for exome enrichments, with the SeqCap EZ Developer probes described above, strictly in
190 accordance with the NimbleGen SeqCap EZ Library protocol (v5.0). For each capture, two parallel
191 post-capture PCRs (14 cycles) were performed on elution solution, and the PCR products were then
192 merged before purification with AMPure XP beads. The length, quality and concentration of the final
193 DNA fragments were checked on a BioAnalyzer (Agilent High Sensitivity DNA Assay) and a Qubit.
194 Each capture was sequenced on a half lane of an Illumina HiSeq3000 sequencer (on the same lane;
195 Fig. 1B), in accordance with the manufacturer's instructions, in paired-end mode, for 150 cycles. The
196 data were then demultiplexed and exported as FastQ files, and the 24 libraries were processed
197 independently.
198
199 ***SNP calling by the mapping of reads to the CDS***
200 Sequence quality was checked with FastQC
201 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adapters were removed and low-
202 quality base pairs were trimmed with Trimmomatic (v0.35, Bolger *et al.*, 2014), with the following
203 parameters: ILLUMINACLIP:TruSeq-file.fa:2:30:10 LEADING:30 TRAILING:30
204 SLIDINGWINDOW:10:30 MINLEN:75, and FASTX (v0.0.14; http://hannonlab.cshl.edu/fastx_toolkit/)
205 to remove the first five bases.
206 We chose to align the reads directly with the targeted CDS sequences, because target
207 enrichment was performed on the same species as probe design, as suggested by Bragg *et al.* (2016;
208 see also Rellstab *et al.*, 2019) (Fig. 1C). Doing so, we avoid the complex and time-consuming steps
209 of read assembly and scaffolding/ordering the genomic contigs obtained to generate the in-target
210 assemblies for read mapping (see introduction). For each library, cleaned reads were aligned with

211 the 5,717 targeted CDS, with BOWTIE2 (v2.2.5; Langmead & Salzberg, 2012), using the default
212 parameters and *local* option to allow for genomic reads not aligning with the targeted CDS sequence
213 over their entire length, whilst ensuring a minimum alignment length of 20 bp. Reads were aligned
214 as unpaired, to allow for the mapping of orphan reads. Reads with multiple alignments, and those
215 for which the mate-pair was aligned to a different target were discarded after mapping. The aligned
216 data in SAM format were converted into BAM format and sorted with SAMtools (v0.1.19). Duplicate
217 reads were estimated with picard tools (v1.113) MarkDuplicates, using the default settings.
218 Coverage at each target base was calculated with mosdepth (v0.2.3; Pedersen & Quinlan, 2018).

219 Variant calling was performed with read alignments of more than 40 bases in length
220 containing fewer than 10 deletions, as deduced from the *cigar* format. SNP sites were called
221 separately for each of the 24 libraries with the SAMtools *mpileup* command, using option *-B*
222 followed by the varscan2 (v2.3; Koboldt *et al.*, 2012) *mpileup2snp* command with a minimum base
223 quality of 35 to count a read at a given position, and a *p*-value threshold of 0.01. For the other
224 parameters, we considered only positions with a minimum sequencing depth of 40 reads for each
225 individual, and 250 reads for the pool (i.e. for these selected depths, the same exome size was
226 analyzed: 90.3% and 93% of target bases were supported by at least 40 and 250 reads for the
227 individual and pooled exomes, respectively; see Fig. S1). We also only included a variant if it was
228 supported by 15 reads per individual (to take into account the variance between both haplotypes
229 sequencing depths), or by three reads for the single pool (given the pool's minimum sequencing
230 depth of 250X, we expected to have ~3 reads per haplotype if the 72 haplotypes in the pool were
231 homogeneously captured). Moreover, to account for the variation in total read sequencing depth
232 across positions, we required a minimum variant allele frequency threshold of 0.25 for each
233 individual and 0.01 for the pool. For individual libraries, an individual was considered to be
234 heterozygous if the variant allele frequency f(v) was between 0.25 and 0.75.

235 All 409,328 different SNP positions independently found in the 24 libraries were then
236 genotyped for each library with varscan2 *mpileup2cns,* using the same parameters as described
237 above, but without imposing a minimum number of reads with variant for individuals. Several filters
238 were then applied to all SNP positions, using either in-house PERL or R scripts (R Development Core
239 Team, 2015). We first retained only bi-allelic positions. In order to properly estimate allelic
240 frequencies with individuals, we retained only positions genotyped for at least 20 individuals.
241 Finally, we discarded SNPs with high levels of sequencing depth in the pool, which were likely to
242 represent repetitive regions or paralogs of the genome: maximum pool sequencing depth was set
243 at 6,960X, as calculated with the formula proposed by Li et al. (2014): $c+4*\sqrt{c}$, where c is the mean
244 sequencing depth in the pool (c=6,634.8X). We retained 300,036 bi-allelic SNPs (73.3%) after the
245 application of all these filters (Fig. 1C).
246 At each site, and for each capture, variant allele frequency at population level was estimated
247 as the number of reads with the variant allele divided by total sequencing depth. Individual
248 genotypes were used to calculate exact *p*-values for Hardy-Weinberg equilibrium tests with the
249 "HardyWeinberg" R package (Graffelman, 2015), with the correction of significance level by the false
250 discovery rate procedure (Benjamini & Hochberg, 1995). Finally, SNPs were annotated and
251 categorized with the SnpEff program (v4.3t; Cingolani *et al.*, 2012).
252
253 ***Evaluation of the approach based on read mapping directly onto CDS***
254 Approaches based on the direct mapping of reads onto CDS have already been used to detect SNPs
255 (e.g. Rellstab *et al.*, 2019), but they have never been evaluated. We used an *H. axyridis* draft genome
256 sequence (HaxR v1.0; Gautier *et al.*, 2018) to evaluate this approach and that proposed below for
257 the prediction of intron-exon boundaries (IEBs), but only for a subset of our targets, i.e. those with
258 a genomic match (Fig. 1D). We selected CDS targets with a genomic match, as proposed by Neves

259 *et al.* (2013), with some adjustments. Briefly, CDS target sequences were first aligned with genomic
260 scaffolds with BLASTN (e-value cutoff of e^{-20} and identity cutoff of 90%), and the best hit was
261 retained for each target. The target sequences were then realigned with the corresponding genomic
262 scaffold, with both GMAP (version 2015-12-31; Wu & Watanabe, 2005) and EXONERATE
263 *est2genome* (v2.2.0; http://www.genome.iastate.edu/bioinfo/resources/manuals/exonerate/). We
264 isolated targets (i) with genome matches over their entire length, (ii) for which intron/exon structure
265 predictions were the same for both alignment programs, (iii) which were covered by capture probes
266 over their entire length and (iv) for which we had access to at least 200 bp upstream and
267 downstream from the targeted CDS sequences. With these filters, we retained 3,161 CDS targets in
268 total (3,220,718 bp, i.e. 60.23% of our targeted exome subset). These 3,161 full or partial CDS
269 corresponded to 11,754 exons, with a mean of 3.72 exons per target (range: 1 to 17 exons) and a
270 mean exon length of 274 bp (range: 12 to 3,417 bp). These 11,754 exons included 6,295 (53.6%)
271 that were unambiguously complete, because they were surrounded by two others, and 8,593
272 reliable IEBs in total were identified between consecutive exons.

273 All genomic regions corresponding to these 3,161 CDS targets were then isolated, including
274 200 bp upstream and downstream from the target sequences (Fig. 1D). Independently for each
275 library, all cleaned reads were aligned with these genomic regions, with BOWTIE2 (v2.2.5), using the
276 default parameters and the *end-to-end* option (i.e. reads must be aligned along their full length).
277 Reads with multiple alignments, and reads for which the mate-pair was aligned to another target
278 were discarded after mapping. Coverage calculation, SNP calling and variant allele frequency
279 estimation were performed as described above (Fig. 1E). We established the correspondence
280 between the position of a base in the genomic region and its position in the CDS sequence, using
281 the *vulgar* format of EXONERATE output (Fig. 1F). SNP positions for which there was a deletion or

282 an insertion between the CDS sequence and its associated genomic sequence (i.e. 559 positions)
283 were excluded from the comparison.

284 In order to compare quantitatively the allelic frequencies obtained with the different
285 approaches presented here (i.e. pool vs. individuals; CDS mapping vs genome mapping), we
286 calculated in each case the Pearson correlation coefficient and the Bland-Altman average bias with
287 its standard deviation which provide information upon agreement between measurements (Altman
288 & Bland, 1983).

289

290 ***Intron-exon boundary detection***

291 Our approach is based solely on the direct mapping of genomic reads to the CDS sequences, by
292 allowing reads to map over less than their full length (i.e. local alignment). Thus, a genomic read
293 with both part of an exon and its flanking sequences (intron or UTR) will start mapping to the
294 corresponding exon in the CDS sequence, and stop mapping at the exon end (Conklin *et al.*, 2013).
295 This will generate a signal of the intron-exon boundary in the number of reads that either begin or
296 end their local alignment at that position (hereafter referred to as *nbBE*). Along a covered exon,
297 *nbBE* should be much lower than the sequencing depth rate, except at the end of the exon, where
298 we should observe a large increase in *nbBE* to a value theoretically equal to the sequencing depth
299 (Fig. S2A). This strong increase in *nbBE* should occur either at the base at the exact end of the exon,
300 or a few bases away if the start of the intron sequence is, by chance, highly similar to the start of
301 the next exon sequence on the CDS (Fig. S2B). In this second case, it creates an abnormally inflated
302 estimate of sequencing depth locally for a few bases at the boundary of the following exon (Fig. S2B
303 and Fig. S3 for an example). At an IEB (i.e. a junction of two exon ends in the CDS), depending on
304 whether a signal offset occurs for one or both exon ends, the size of the area with abnormally
305 inflated local sequencing depth is variable, but generally less than 10 bp. In addition, the presence

306 of a SNP at some bases at the end of the exon may prevent reads with the variant from aligning up
307 to the IEB position, potentially generating another signal in addition to those of the two exon ends
308 of the IEB. Combinations of these different scenarios may also occur.

309 We developed and tested a method for predicting IEBs on the basis of the increase in *nbBE,*
310 by focusing on the 3,161 targeted transcripts for which genomic sequences were available (Fig. 1D).
311 We used the mapping output BAM file obtained with the pool library (Fig. 1C). At each position in
312 the transcript, *nbBE* was calculated from the start read alignment position on the transcript and the
313 *cigar* code, both present in the BAM file. The value of the *nbBE* parameter at each position is a
314 function of sequencing depth (and read length). We took the possible effect of sequencing depth
315 heterogeneity on *nbBE* into account by calculating the median *nbBE* over a window of 11 bases
316 centered on the position concerned. A signal (i.e. a predicted IEB), was detected at a position if *nbBE*
317 was *X* times larger than the median, *X* being a tool parameter defined by the user. For the definition
318 of *X*, it may be necessary to identify a subset of transcripts with orthologs in a close model species,
319 and then to determine which value of *X* provides the best prediction of IEBs for this subset. Signals
320 at the exact ends of the CDS were not considered. We took the possible signal deviation (i.e. a signal
321 that is slightly deported with respect to the IEB that generates it; Fig. S2) into account, by extending
322 signal detection to the three bases upstream, and three bases downstream from a position with
323 measured signal (default value for parameter *n*=3). A measured signal therefore predicts an IEB in a
324 window of seven bases in total. An IEB, composed of two exon ends, theoretically generates two
325 close signals with overlapping windows. If these two signals are found at the exact positions of the
326 exon ends (no deviation), then the predicted region measures 8 bp (Fig. S2A). If these two signals
327 overlap (e.g. because of the deviation of one of the two signals) then the predicted region for the
328 IEB is 7 bp long. It can be less than 7 bp if the signal is detected close to non-covered regions or
329 regions with measured signals close to the end of the CDS for which signal extension is not possible.

330 If these signals deviate by up to six bases from each other (i.e. cumulative deviation of eight bases,
331 the deviation being in opposite directions for the two signals) then the predicted region measures
332 14 bp (Fig. S2). More rarely, if a third signal is detected in the same region, it increases the length of
333 the region by a few bases. If the two signals deviate by more than six bases from each other, then
334 the tool predicts two signal windows separated by a few bases for the IEB concerned. Thus, each
335 transcript was divided into short regions with prediction signals (+) alternating with regions without
336 prediction signals (-). Each region was then assessed as follows: true positive (TP) if the + region
337 contained an IEB; false positive (FP) if the + region did not contain an IEB; true negative (TN) if the -
338 region did not contain an IEB; and false negative (FN) if the - region contained an IEB. We evaluated
339 the performance of the method, by calculating the percentage of true IEBs correctly predicted (i.e.
340 sensitivity (SN)) as TP/(TP+FN) and the percentage of correct predictions (i.e. specificity (SP)) as
341 TP/(TP+FP) (Burset & Guigo, 1996). For a CDS with no introns (862, i.e. 27.3% of the 3,161 CDS), SN
342 cannot be evaluated; SP was 100 if no IEB was predicted (good prediction) and 0 otherwise. Scripts
343 are available from https://github.com/edeleury/IEB-finder.
344
345

346 # Results

347

348 ***Capture data and filtration***

349 In the two target enrichment experiments, we obtained 186,998,614 and 185,457,659 *H. axyridis*
350 raw sequence read pairs for the pool library (SRA ERX3237193) and the 23 indexed individuals (SRA
351 ERX3237194 to ERX3237216), respectively. For each indexed individual library, we obtained a mean
352 of 8,063,376 raw read pairs (range: 6,912,561 to 9,176,292). After cleaning, 91.2% of the raw read
353 sequences were retained (see detailed information for all libraries in Table S1).

354

355 ***Capture efficiency, sequencing depth and duplicates***

356 Rather than assembling the cleaned reads, we chose to map them directly onto the target CDS
357 sequences used for probe design (Fig. 1C). Mean capture specificity, defined as the percentage of
358 cleaned reads mapping correctly (i.e. meeting our mapping parameters) onto the 5,717 targeted
359 CDS, was 83.2% (range over libraries: 80.6%-89.2%; Table S2). Aligned bases accounted for 74.6%
360 (range over all libraries: 72.6%-75.7%) of all cleaned read bases. The global median read alignment
361 length was 144 bp (range over libraries: 143-144 bp, see Table S2), with 62.7% of the mapped reads
362 (range over all libraries: 60.9%-64.3%) mapped over their entire length. A substantial proportion of
363 the mapped reads (22.2%; range over all libraries: 20%-25.9%) were 'orphaned' mates (i.e. only one
364 of the two mates aligned to the target sequences). This percentage strongly suggests that flanking
365 regions of targeted regions were also sequenced with our capture procedure (Yi *et al.*, 2010; Neves
366 *et al.*, 2013), although quantification was not possible with this direct mapping approach.

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

374

		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%)

375 **Table 1:** Reproducibility of capture sensitivity on the 24 libraries. 100% means that all 24 samples had at least one read
376 aligned with the target, whereas 0 % level indicates that no reads aligned with the target.
377

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

384 Base sequencing depth varied along the targeted CDS (Fig. S4A). Using the known complete
385 exons (≥ 20 bp) identified on targeted CDS with a match over their entire length with the *H. axyridis*
386 draft genome, we confirmed that mean sequencing depth increased steadily with exon size (Fig.
387 S4B), and that base sequencing depth decreased from the middle of an exon to its ends (Fig. S4C).
388 Alignments of less than 20 bp between the reads and the CDS sequence were not permitted with
389 this mapping procedure, so sequencing depth was probably underestimated at exon ends and for
390 short exons. For all the targeted CDS considered together, mean sequencing depth differed between
391 targeted CDS but was highly reproducible between libraries (Fig. S5): the Pearson correlation
392 coefficient of the target sequencing depth between individual libraries ranged from 0.96 to 0.99,
393 with a mean of 0.99, and the Pearson coefficient of correlation between the two capture processes
394 (i.e. between pool and individuals) was 0.999.

395 For the 23 individual libraries, a mean of 66.1% of reads were scored as duplicates (range
396 over libraries: 63.6%-68.9%; Table S4). This high proportion was due to the very strong sequencing
397 effort regarding the target region size. Indeed, pooling the reads of the 23 individual libraries
398 increased the proportion of duplicates detected to >96%, as for the pool (Table S4), suggesting that
399 the reads scored as duplicates in our experiment probably mostly come from different individuals.

400 We did therefore not discard duplicated reads before SNP calling, so as to avoid artificially biasing
401 the estimation of allele frequencies.
402
403 ***SNP calling and accuracy of allele frequency estimation***
404 SNP calling performed independently for the individuals and for the pool, following direct mapping
405 onto the targeted CDS sequences, led to the detection of 302,292 and 364,144 different putative
406 exonic SNPs, respectively. Among the total 409,328 SNPs, 257,108 positions (62.8%) were detected
407 in both experiments, 45,184 (11%) only in individuals and 107,036 (26.2%) only in the pool. As
408 expected (because there were 13 more individuals in the pool), the number of SNPs called for the
409 pool was larger than that for the 23 individuals, with 26.2% of SNPs called for the pool only. The vast
410 majority (95.6%) of the SNPs called only for individuals had allele frequencies below 0.03.
411 All positions were genotyped in all libraries. After filtering, we finally selected a total of
412 300,036 exonic bi-allelic SNPs within 4,741 targeted CDS (Fig. 1C). Minor allele frequencies (MAFs)
413 were below 0.03 for 59.2% of these SNPs (Fig. 2); 92.75% were polymorphic in both the pool and
414 the individuals, 7.22% were polymorphic only in the pool, and 0.03% were polymorphic only in the
415 individuals. Only 1,007 of the 300,036 SNPs (i.e. 0.34%) displayed significant deviation from Hardy-
416 Weinberg equilibrium, as calculated with the individual genotypes. This number was more than
417 three times higher (i.e. 3,331 loci) before the removal of loci on the basis of the maximum
418 sequencing depth filter applied to the pool.
419

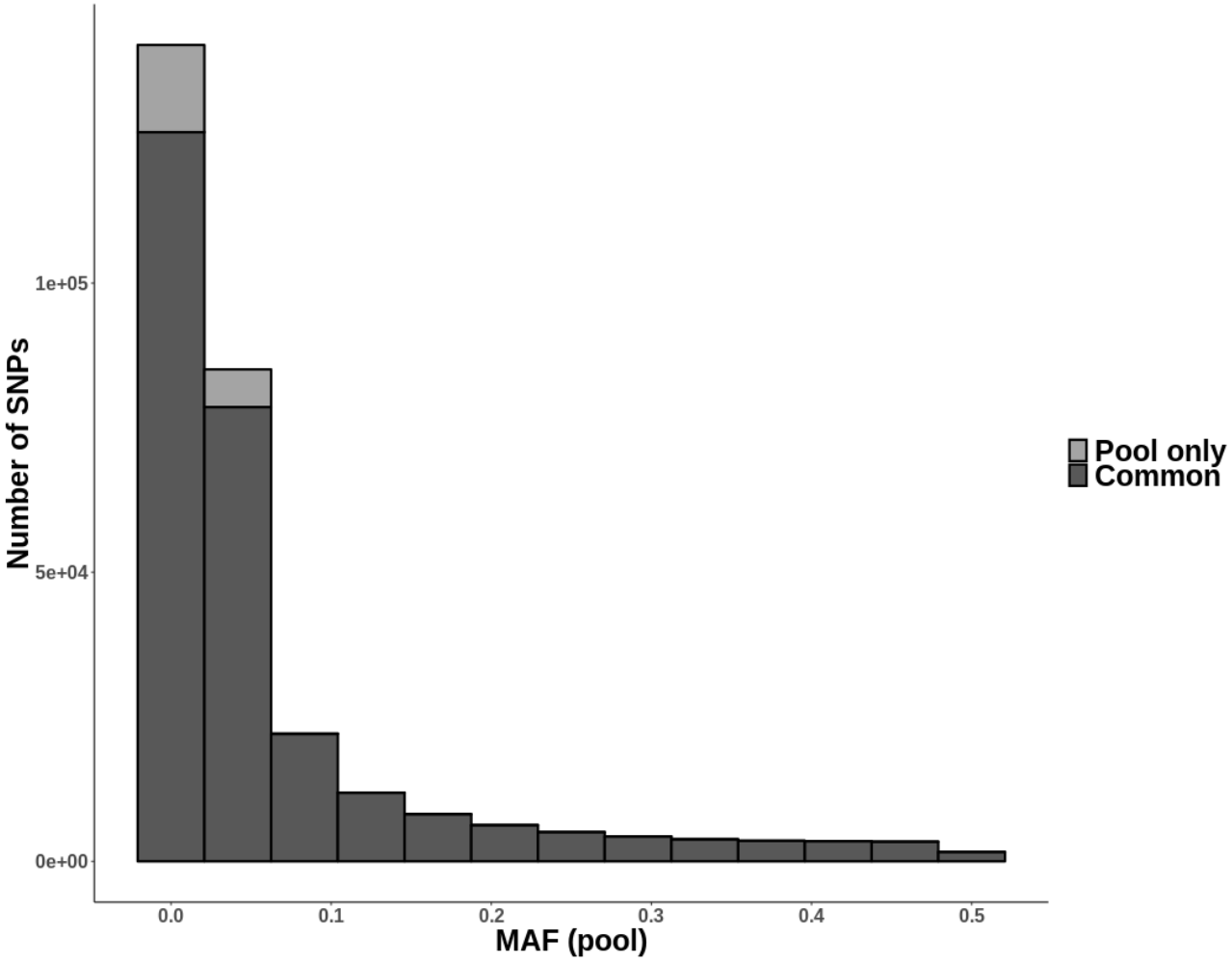

Figure 2: Folded-site frequency spectrum of the 300,036 filtered exonic bi-allelic SNPs (mapping onto targeted CDS sequences) according to their polymorphism in the pool and in individuals. For SNPs polymorphic in both the pool and individuals, the MAF shown is based on the pool. 86 SNPs (0.03%) that were polymorphic only in the individuals are not represented here, all have MAF<0.05.

When all 300,036 loci were considered, allele frequency estimates were strongly correlated between the pool and individuals (Pearson's $r = 0.99$; Fig. 3A), with a good level of agreement (average bias = $9.5E10^{-5}$; bias SD = 0.025).

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.

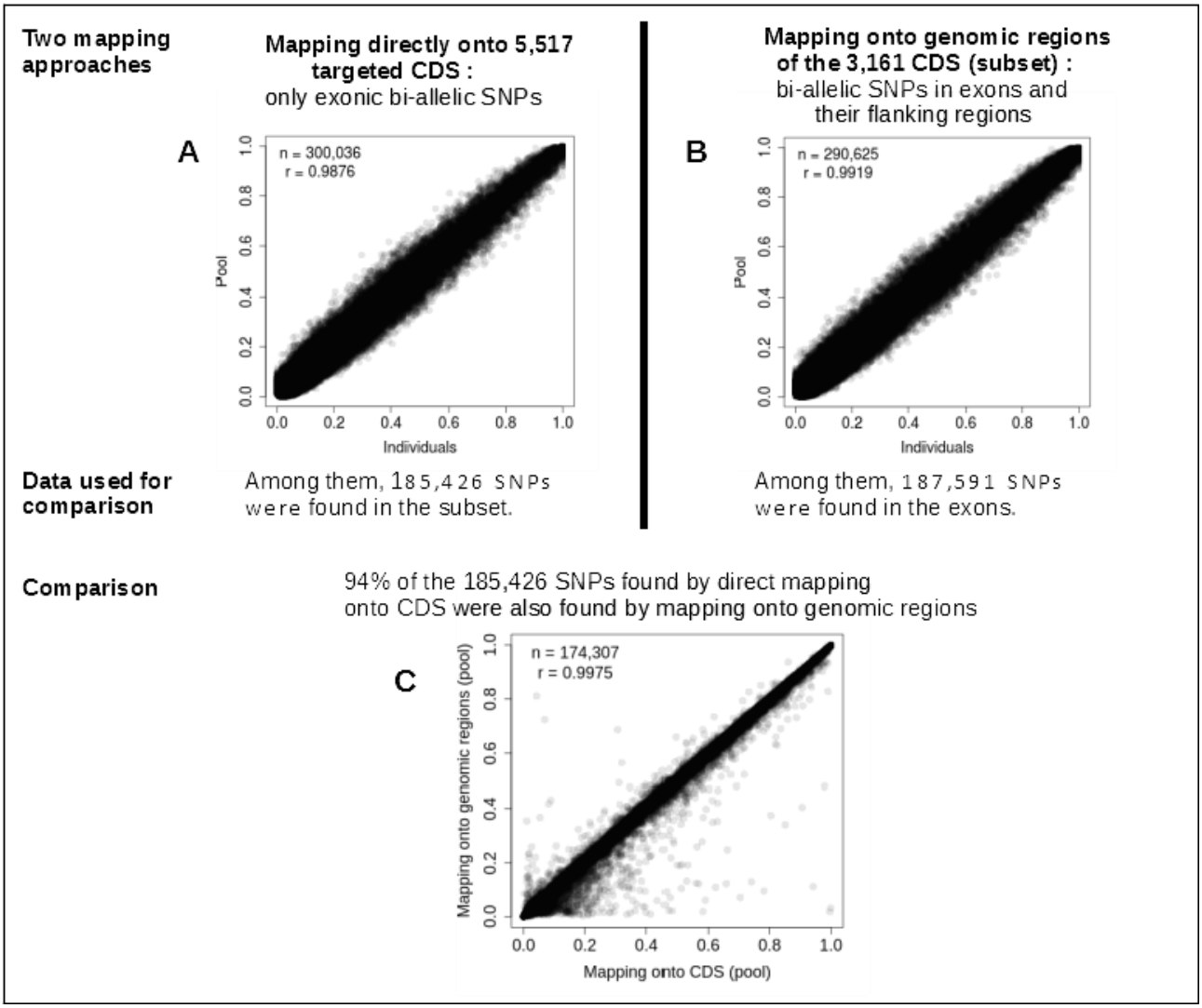

Figure 3: Allele frequency estimates from the population pool and individuals by two mapping approaches: **(A)** mapping reads directly onto CDS and **(B)** mapping onto genomic regions available only for a subset of target CDS. For exonic bi-allelic SNPs common to both mapping approaches, we also show **(C)** the allele frequency estimates from the pool by the two approaches. Note: Pearson correlation coefficient (*r*) are displayed.

Comparison of the SNPs detected with the two mapping approaches

We evaluated the ability of the method for mapping reads directly to CDS (Fig. 1C) to call SNPs correctly by comparing the SNPs identified with this approach and those found by mapping reads to the genome (Fig. 1E). We used only the subset of 3,161 targeted CDS with a match over their entire length to sequences in the draft genome of *H. axyridis* for this comparison. For each library (pool and individuals), trimmed reads were also mapped along their entire length onto the regions of the

445 genome corresponding to this subset (Fig. 1E). After application of the same procedure and filters
446 for SNP calling as were used for direct mapping onto CDS sequences, 290,620 filtered bi-allelic SNPs
447 were genotyped within the subset, including 187,591 (64.5%) in exons and 103,029 (35.5%) in the
448 flanking regions (intron or UTR) of the exons. By mapping reads onto the genome, we demonstrated
449 that allele frequency estimates were strongly correlated, with a good level of agreement, between
450 the pool and individuals, for both exonic and flanking SNPs (Pearson's *r*=0.99; average bias = -
451 $7.3E10^{-4}$; bias SD = 0.025; Fig. 3B).

452 Of the bi-allelic exonic SNPs found by direct mapping onto the targeted CDS (Fig. 3A),
453 185,426 mapped onto the subset of 3,161 targeted CDS, and 94% were also identified as bi-allelic
454 SNPs by mapping to genomic regions (Fig. 1F). For these 174,307 exonic bi-allelic SNPs found with
455 both mapping approaches (mapping onto the genome or directly onto transcripts), the allele
456 frequency estimates obtained with the two mapping methods were highly correlated, with good
457 level of agreement, both for the pool (Pearson's *r*=0.998; average bias = $6.8E10^{-4}$; bias SD = 0.013;
458 Fig. 3C) and for the individuals (Pearson's *r*=0.998; average bias = $1.0E10^{-3}$; bias SD = 0.012).

459 Some of the private bi-allelic SNPs (i.e. SNPs identified in a single mapping approach) were
460 variants initially identified with both mapping approaches but discarded during filtering in one of
461 the approaches. For example, 1,467 bi-allelic SNPs private to the CDS mapping process had a
462 sequencing depth above the maximum threshold, and 714 were tri-allelic SNPs when mapped onto
463 the genome. If we exclude this latter group, 8,853 and 8,333 were found to be private to CDS
464 mapping and mapping onto genomic regions, respectively. A non-exhaustive analysis of these
465 positions highlighted two possible origins of private SNPs for mapping onto CDS. Some targeted
466 regions (the full CDS or part of the CDS) have homologous copies in the genome that we were unable
467 to identify during target selection and for which the SNPs found were not discarded with the
468 "maximum sequencing depth" filter. We have thus identified 43 CDS (0.7%) with reads mapping

469 onto their genomic regions in a non-unique manner, and these CDS with paralogs displayed 3,141
470 SNPs, all private to CDS mapping. The second origin was false SNPs occurring when the start of an
471 intron could be aligned with the start of the next exon due to similarities between the two
472 sequences (Fig. S2 and S6, for example). Of the SNPs private to the CDS mapping method, 6.9%
473 (*N*=614) were positioned close to intron-exon boundary (IEB) positions (i.e. at the exact end of the
474 exon or on the three adjacent bases), whereas only 0.6% of the 174,307 SNPs common to both
475 mapping methods were found near IEBs.
476
477 ***Prediction of intron-exon boundaries without read assembly***
478 The method we developed for IEB detection based on the mapping of genomic reads directly onto
479 CDS sequences (Fig. 1C) was evaluated on the 3,161 targeted CDS for which we had genomic
480 information, and therefore knew the positions of the IEBs (Fig. 1D). An IEB signal was predicted at a
481 particular position in the CDS if the number of reads either beginning or ending their local alignment
482 at that position (*nbBE*) was 20 times greater (*X*=20) than the median *nbBE* value within a window of
483 11 bases around the position concerned, and the signal measured at that position extended to a
484 short region of 7 bases (i.e. with a default parameter of *n*=3; see Methods section and Fig. S2). With
485 parameters *X*=20 and *n*=3, 8,967 regions with lengths of 4 to 19 bp were predicted to be IEBs (Table
486 2). Of the 8,967 regions predicted to be IEBs, 8,575 actually corresponded to true IEBs (i.e. method
487 specificity = 95.63%). Of the 8,632 regions truly containing an IEB, 8,575 were accurately predicted
488 to be IEB regions (i.e. sensitivity = 99.34%; Table 2). Of the 862 CDS with no intron, 763 (i.e. 88.51%)
489 were correctly predicted to have no IEBs. As expected, the use of a lower detection threshold (i.e.
490 *X*=10) decreased the specificity of the method and increased its sensitivity (Table 2). The regions
491 predicted to have no IEB (i.e. the - regions) were between 1 and 3,417 bp in length, with a mean
492 length of 265.5 bp, entirely consistent with the known distribution of exon lengths (see Methods

493 section). We counted 153 very small (<10 bp) - regions with *n=3* and only 51 with *n=5*, which attested
494 to the possible cumulative deviation of signals of more than eight bases for a small number of IEBs.
495 Increasing the length of the signal deviation (e.g. *n*=5 bases instead of 3) made it possible to take
496 this situation into account in the vast majority of cases. This slightly improved performance but
497 increased the length, and thus, the imprecision, of all predicted IEB positions (Table 2).

498

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

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

503

504

505 # Discussion

506

507 Exome-targeted capture associated with next-generation sequencing has been mainly restricted to
508 model species for which genomic resources are available (Ng *et al.*, 2009). Transcriptome-based
509 exome capture has proved an efficient alternative for dealing with non-model species (e.g. Bi *et al.*,
510 2012), but is mostly used in phylogeny studies. Here, we propose an extension of this method with
511 a high capture efficiency, providing accurate estimates of the allele frequencies of SNPs for
512 population genomics studies. Our methodology is original because (i) it combines transcriptome-
513 based exome capture with a pool-based approach (pooling the tissues of numerous individuals

514 before DNA extraction), (ii) it does not require a genome sequence for the species itself or another
515 closely related species, and (iii) it does not involve the *de novo* assembly of genomic reads. This work
516 opens up interesting prospects for population genomics studies on non-model species because it is
517 cheap and less labor-intensive than other methods.
518
519 ***Capture efficiency***
520 We found that capture efficiency was very high for the *H. axyridis* exome subset we targeted. The
521 capture specificity (83.2%) achieved was much higher than reported in more classical transcriptome-
522 based studies (range 25% - 73%), whereas capture sensitivity (93.5%) approached the highest
523 reported values (see Puritz & Lotterhos, 2018). Moreover, despite the differences in sequencing
524 depth observed between targets and along the length of targets, reproducibility was high between
525 libraries, whether in terms of specificity, sensitivity or sequencing depth. Our capture approach,
526 based on the choice of targets from the *de novo* transcriptome of the species of interest itself,
527 without the need to identify orthologous genes in related model species, proved highly effective.
528 This "blind" approach had only a minor impact on capture performance: we only observed a slight
529 decrease in sequencing depth for small exons and towards the ends of the exons. Similar patterns
530 have been observed in classical transcriptome-based exome capture in which IEBs are predicted by
531 orthology. In both cases, this is due to the use of less efficient probes designed from the sequences
532 of two consecutive exons and the absence of a probe straddling the end of the exon and the
533 beginning of the intron (e.g. Bi *et al.*, 2012; Neves *et al.*, 2013).
534 The choice of targets without looking for orthology in related species was probably the cause
535 of the slightly lower capture sensitivity in this study than in previous studies (Bi *et al.*, 2012; Puritz
536 & Lotterhos, 2018). Targets selected this way are necessarily only partly known, and some may
537 belong to other species (e.g. symbionts or parasites) or be chimeras present in the transcriptome

538 used. This is a small price to pay given the efficiency of the method, a feature of considerable
539 interest for population genomics studies. The absence of searches for orthologous genes in related
540 model species prevents the analysis from focusing purely on the most conserved genes. Moreover,
541 our method can be applied to any species for which no closely related model species with genomic
542 resources is available.
543
544 ***SNP calling and the pool-based approach***
545 With this method, we were able to genotype a very large number of SNPs (~300,000) over a random
546 subset of the exome of *H. axyridis*. The study of the subset of CDS with complete genomic sequence
547 matches further supported the idea that direct mapping onto the CDS has probably no major impact
548 on SNP identification or on the estimation of allele frequencies relative to mapping onto the
549 genome. In addition to a few residual paralogous sequences inherent to any approach dealing with
550 non-model species, we observed slightly larger numbers of SNPs called on the basis of mapping onto
551 CDS in areas very close to the IEBs, but the proportion of SNPs involved was minimal (0.33%).
552 For similar sequencing efforts, strong concordance was observed between the allele
553 frequencies for individual genotyping and those for the "pool-based" approach, whatever the
554 mapping method used. This finding demonstrates the feasibility of this approach, in situations in
555 which individual identification for many samples is practically impossible, technically difficult, too
556 costly and/or unnecessary. In population genomics studies, allele frequencies are sufficient for the
557 calculation of most summary statistics (e.g. nucleotide diversity, F_{ST}, site frequency spectrum) and,
558 thus, for addressing most of the questions posed (Schlötterer *et al.*, 2014). In the specific case of
559 exome sequencing, allele frequencies can be used for large population-scale screening of protein
560 mutations, or to address questions related to adaptation or to the evolution of the genetic load (e.g.
561 Gutenkunst *et al.*, 2009; Yi *et al.*, 2010; Casals *et al.*, 2013; Popp *et al.*, 2017; Castellano *et al.*, 2019).

562 One of the limitations of pool-seq compared to individual-based approaches is that it is not
563 optimal for studies in which analyses rely on linkage phase information. However, a prerequisite in
564 this context is the availability of a genetic (or physical) map, which is not the case in most non-model
565 organisms. Another weakness of pool-seq is that "populations" must be carefully defined before
566 sequencing, as it will be difficult to detect a Wahlund effect or an unforeseen breeding system
567 afterwards. It is therefore important to be particularly careful during the sampling step, and to be
568 as familiar as possible with the biology of the species being studied. Finally, one more drawback of
569 pool-seq is that it makes it impossible to remove SNPs that are not at Hardy-Weinberg equilibrium.
570 Such disequilibrium may be caused by copy number variations or paralogous sequences resulting in
571 false-positive SNPs and local large sequencing depth (Li, 2014). We show here that the application
572 of our "maximum sequencing depth" filter to the pooled data made it possible to discard almost
573 70% of the SNPs not at equilibrium, as shown in individuals. The extremely small remaining
574 proportion (0.34%) of SNPs not at Hardy-Weinberg equilibrium provides support for the feasibility
575 of our pooling-based method.
576
577 ***Method for predicting IEBs without read assembly***
578 We have developed a method for detecting IEBs that is particularly useful in the case of non-model
579 species (or species for which genome sequences are being assembled). The proposed tool (available
580 from https://github.com/edeleury/IEB-finder) does not indicate the exact location of the IEBs,
581 instead it identifies an area of a few bases (about eight, on average, in our case) predicted to contain
582 an IEB. This method, which requires only the genomic reads and the target CDS sequences, was
583 effective when evaluated on a subset of CDS for which we had a complete genomic match. Almost
584 all the true IEBs were identified, whereas false-positives accounted for less than 5% of the predicted
585 IEBs. Montes *et al.* (2013) proposed a similar approach, which they used to discard putative SNPs

586 found in predicted IEB areas. They obtained better genotyping results for the selected filtered SNPs
587 than other similar studies on non-model species. In this study, we went a step further by taking into
588 account the possible offset of the measured signal, and by evaluating the performance of the
589 approach on a set of transcripts for which the true intron-exon structure was known. We aimed to
590 present evidence that this approach is promising, although further improvement is required,
591 together with evaluation on larger dataset and at various sequencing depth levels. A method for
592 choosing an appropriate value for parameter *X* for the correct identification of significant increases
593 in *nbBE* depending on local sequencing depth must also be developed.
594
595 ***General recommendations, conclusion and perspectives***
596 The combination of approaches proposed here — i.e. exome transcriptome-based capture and
597 direct mapping of reads onto targeted CDS sequences, coupled with a pool-seq approach before
598 DNA extraction — significantly decreases costs, the laboratory time required and the complexity of
599 the sequencing analyses required to obtain a large number of variants in exome-scale coding regions
600 for non-model species. However, direct mapping onto target sequences has two disadvantages.
601 First, it provides no information on flanking regions (including regulatory regions), and it is therefore
602 not possible to genotype SNPs in these areas. Nevertheless, detecting SNP polymorphisms at exonic
603 level would be sufficient in many cases, particularly when the in-depth analysis of a large number
604 of synonymous and non-synonymous mutations is informative enough to address the question
605 posed. It is still feasible to assemble the genomic reads in situations in which flanking regions appear
606 to be of interest. Second, direct mapping onto CDS results in the detection of some false SNPs in the
607 vicinity of IEBs. As a means of limiting this risk further, we therefore suggest using the IEB detection
608 method proposed here, followed by the elimination of any SNPs found in the predicted IEB regions
609 and in short regions (≤ 20 bp) surrounded by two predicted IEB regions. Using this approach on our

610 data with the parameters *X*=20 and *n*=3, we retained a final list of 296,736 SNPs. A large proportion
611 of the 3,300 SNPs excluded (408 SNPs, i.e. 12.36%) were not at Hardy Weinberg equilibrium in
612 individuals, which is an additional evidence of the relevance of this filter.

613 The allele frequencies measured for the pool with our method were in good agreement with
614 those obtained in the analysis of individuals. However, caution is nevertheless required with rare
615 alleles (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 sequencing depths, 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

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.

Conflict of interest disclosure

The authors of this preprint declare that they have no financial conflict of interest with the content of this article. Thomas Guillemaud and Eric Lombaert are both recommenders at PCI Evolutionary Biology, PCI Ecology and PCI Entomology.

Data accessibility

The exome sequence capture data reported here have been deposited in the European Nucleotide Archive (accession no. PRJEB31592). The CDS sequences targeted and the description of the exon positions on the subset of transcripts have been deposited at Zenodo

655 https://doi.org/10.5281/zenodo.2598388. The HaxR v1.0 genome sequence used in this study is
656 available from http://bipaa.genouest.org/sp/harmonia_axyridis/ (see Gautier *et al.*, 2018).
657
658

659 # References

660

661 Altman, D.G. & Bland, J.M. 1983. Measurement in Medicine: The Analysis of Method Comparison
662 Studies. *Stat.* **32**: 307.
663 Bansal, V. 2010. A statistical method for the detection of variants from next-generation
664 resequencing of DNA pools. *Bioinformatics* **26**: 318–324.
665 Bansal, V., Tewhey, R., LeProust, E.M. & Schork, N.J. 2011. Efficient and cost effective population
666 resequencing by pooling and in-solution hybridization. *PLoS One* **6**: 1–6.
667 Benjamini, Y. & Hochberg, Y. 1995. Controlling the false discovery rate - a practical and powerful
668 approach to multiple testing. *J. R. Stat. Soc. Ser. B-Methodological* **57**: 289–300.
669 Bi, K., Linderoth, T., Vanderpool, D., Good, J.M., Nielsen, R. & Moritz, C. 2013. Unlocking the vault:
670 Next-generation museum population genomics. *Mol. Ecol.* **22**: 6018–6032.
671 Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C. & Good, J.M. 2012. Transcriptome-based
672 exon capture enables highly cost-effective comparative genomic data collection at moderate
673 evolutionary scales. *BMC Genomics* **13**: 403.
674 Bolger, A.M., Lohse, M. & Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence
675 data. *Bioinformatics* **30**: 2114–2120.
676 Bossert, S. & Danforth, B.N. 2018. On the universality of target-enrichment baits for phylogenomic
677 research. *Methods Ecol. Evol.* **9**: 1453–1460.
678 Bragg, J.G., Potter, S., Bi, K. & Moritz, C. 2016. Exon capture phylogenomics: efficacy across scales
679 of divergence. *Mol. Ecol. Resour.* **16**: 1059–1068.
680 Burset, M. & Guigo, R. 1996. Evaluation of Gene Structure Prediction Programs. *Genomics* **34**: 353–
681 367.
682 Casals, F., Hodgkinson, A., Hussin, J., Idaghdour, Y., Bruat, V., de Maillard, T., *et al.* 2013. Whole-
683 Exome Sequencing Reveals a Rapid Change in the Frequency of Rare Functional Variants in a
684 Founding Population of Humans. *PLoS Genet.* **9**.
685 Castellano, D., Maclà, M.C., Tataru, P., Bataillon, T. & Munch, K. 2019. Comparison of the full
686 distribution of fitness effects of new amino acid mutations across great apes. *Genetics* **213**:
687 953–966.
688 Castellano, S., Parra, G., Sanchez-Quinto, F.A., Racimo, F., Kuhlwilm, M., Kircher, M., *et al.* 2014.
689 Patterns of coding variation in the complete exomes of three Neandertals. *Proc. Natl. Acad.*
690 *Sci.* **111**: 6666–6671.
691 Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., *et al.* 2012. A program for
692 annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).*
693 **6**: 80–92.
694 Conklin, D., Montes, I., Albaina, A. & Estonba, A. 2013. Improved conversion rates for SNP
695 genotyping of nonmodel organisms. *Proc. Iwbbio 2013 Int. Work. Bioinforma. Biomed. Eng.*
696 127–134.
697 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.

Förster, D.W., Bull, J.K., Lenz, D., Autenrieth, M., Paijmans, J.L.A., Kraus, R.H.S., *et al.* 2018. Targeted resequencing of coding DNA sequences for SNP discovery in nonmodel species. *Mol. Ecol. Resour.* 1356–1373.

Gautier, M., Foucaud, J., Gharbi, K., Cezard, T., Galan, M., Loiseau, A., *et al.* 2013. Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Mol. Ecol.* **22**: 3766–3779.

Gautier, M., Yamaguchi, J., Foucaud, J., Loiseau, A., Ausset, A., Facon, B., *et al.* 2018. The Genomic Basis of Color Pattern Polymorphism in the Harlequin Ladybird. *Curr. Biol.* **28**: 3296–3302.

Gouzy, J., Carrere, S. & Schiex, T. 2009. FrameDP: Sensitive peptide detection on noisy matured sequences. *Bioinformatics* **25**: 670–671.

Graffelman, J. 2015. Exploring Diallelic Genetic Markers: The HardyWeinberg Package. *J. Stat. Softw.* **64**: 1–23.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**.

Heyduk, K., Trapnell, D.W., Barrett, C.F. & Leebens-Mack, J. 2016. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Biol. J. Linn. Soc.* **117**: 106–120.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., *et al.* 2007. Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* **39**: 1522–1527.

Ilves, K.L. & López-Fernández, H. 2014. A targeted next-generation sequencing toolkit for exon-based cichlid phylogenomics. *Mol. Ecol. Resour.* **14**: 802–811.

Ilves, K.L., Torti, D. & López-Fernández, H. 2018. Exon-based phylogenomics strengthens the phylogeny of Neotropical cichlids and identifies remaining conflicting clades (Cichliformes: Cichlidae: Cichlinae). *Mol. Phylogenet. Evol.* **118**: 232–243.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., Mclellan, M.D., Lin, L., *et al.* 2012. VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**: 568–576.

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

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

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

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

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

Nicholls, J.A., Pennington, R.T., Koenen, E.J.M., Hughes, C.E., Hearn, J., Bunnefeld, L., *et al.* 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the

745 neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Front. Plant Sci.* **6**: 1–20.
746 Ning, Z., Cox, A.J. & Mullikin, J.C. 2001. SSAHA: a fast search method for large DNA databases. -
747 Abstract - UK PubMed Central. 1725–1729.
748 Pedersen, B.S. & Quinlan, A.R. 2018. Mosdepth: Quick coverage calculation for genomes and
749 exomes. *Bioinformatics* **34**: 867–868.
750 Popp, B., Ekici, A.B., Thiel, C.T., Hoyer, J., Wiesener, A., Kraus, C., *et al.* 2017. Exome Pool-Seq in
751 neurodevelopmental disorders. *Eur. J. Hum. Genet.* **25**: 1364–1376. Springer US.
752 Portik, D.M., Smith, L.L. & Bi, K. 2016. An evaluation of transcriptome-based exon capture for frog
753 phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Mol.*
754 *Ecol. Resour.* **16**: 1069–1083.
755 Puritz, J.B. & Lotterhos, K.E. 2018. Expressed exome capture sequencing: A method for cost-
756 effective exome sequencing for all organisms. *Mol. Ecol. Resour.* 1209–1222.
757 R Development Core Team. 2015. R: A language and environment for statistical computing. R
758 Foundation for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
759 Ramos, E., Levinson, B.T., Chasnoff, S., Hughes, A., Young, A.L., Thornton, K., *et al.* 2012.
760 Population-based rare variant detection via pooled exome or custom hybridization capture
761 with or without individual indexing. *BMC Genomics* **13**: 683.
762 Rellstab, C., Dauphin, B., Zoller, S., Brodbeck, S. & Gugerli, F. 2019. Using transcriptome
763 sequencing and pooled exome capture to study local adaptation in the giga-genome of *Pinus*
764 *cembra*. *Mol. Ecol. Resour.* **19**: 536–551.
765 Ryu, S., Han, J., Norden-Krichmar, T.M., Schork, N.J. & Suh, Y. 2018. Effective discovery of rare
766 variants by pooled target capture sequencing: A comparative analysis with individually
767 indexed target capture sequencing. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **809**: 24–31.
768 Elsevier.
769 Schlötterer, C., Tobler, R., Kofler, R. & Nolte, V. 2014. Sequencing pools of individuals — mining
770 genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15**: 749–763. Nature
771 Publishing Group.
772 Shearer, A.E., Hildebrand, M.S., Ravi, H., Joshi, S., Guiffre, A.C., Novak, B., *et al.* 2012. Pre-capture
773 multiplexing improves efficiency and cost-effectiveness of targeted genomic enrichment.
774 *BMC Genomics* **13**.
775 Stephens, J.D., Rogers, W.L., Heyduk, K., Cruse-Sanders, J.M., Determann, R.O., Glenn, T.C., *et al.*
776 2015. Resolving phylogenetic relationships of the recently radiated carnivorous plant genus
777 *Sarracenia* using target enrichment. *Mol. Phylogenet. Evol.* **85**: 76–87. Elsevier Inc.
778 Vilcinskas, A., Mukherjee, K. & Vogel, H. 2013. Expansion of the antimicrobial peptide repertoire in
779 the invasive ladybird *Harmonia axyridis*. *Proc. Biol. Sci.* **280**: 20122113.
780 Vogel, H., Schmidtberg, H. & Vilcinskas, A. 2017. Comparative transcriptomics in three ladybird
781 species supports a role for immunity in invasion biology. *Dev. Comp. Immunol.* **67**: 452–456.
782 Elsevier Ltd.
783 Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N. & Watson, M. 2015. Exome Sequencing:
784 Current and Future Perspectives. *Genes|Genomes|Genetics* **5**: 1543–1550.
785 Wu, T.D. & Watanabe, C.K. 2005. GMAP: A genomic mapping and alignment program for mRNA
786 and EST sequences. *Bioinformatics* **21**: 1859–1875.
787 Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., *et al.* 2010. Sequencing of Fifty
788 Human Exomes Reveals Adaptation to High Altitude. *Science (80-.).* **329**: 75–78.
789
790