An evaluation of pool-sequencing transcriptome-based exon capture for population genomics of 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: exome genotyping, target enrichment, non-model organism, population genomic, pool-sequencing approach, *Harmonia axyridis*, intron exon boundaries prediction

Abstract

Exon capture, coupled with high-throughput sequencing technologies, represents a cost-effective technical solution to answer specific evolutionary biology questions by focusing on areas of the genome under selection. Transcriptome-based capture, which allows exon capture for non-model species, is particularly used in phylogenomics. In the case of population genomics studies, it remains however poorly developed because the cost of sequencing a large number of indexed individuals across multiple populations of one species is prohibitively high. In this study, we evaluate the possibility of combining transcriptome-based capture and pool-seq (before extraction) as a cost-effective, generic and robust approach to estimate the population variant allelic frequencies of any species. We designed capture probes for ~5 Mb of randomly chosen *de novo* transcripts of the Asian ladybird (5,717 transcripts). From a pool of non-indexed 36 individuals, ~300,000 bi-allelic SNPs were called. We found that capture efficacy was high, and that pool-seq was as effective and accurate as individual-seq in detecting variants and estimating allele frequencies. We also propose and evaluate an approach to simplify the processing of read data, which consists of mapping reads directly to targeted transcript sequences in order to obtain coding variants. This approach is effective and does not affect the estimation of SNPs' allele frequencies, except for a small bias near some exon ends. We demonstrate that this approach can also be used to efficiently predict *a posteriori* intron-exon boundaries of targeted *de novo* transcripts, thus allowing to cancel genotyping biases around exons ends.

Introduction

A core challenge of population genomics and phylogenetic studies is to obtain a reliable set of orthologous loci respectively from a sufficient number of individuals across multiple populations and from species spanning a range of divergences. Targeted re-sequencing of a consistent subset of genomic regions represents a cost-effective technical solution compared to whole genome sequencing. Among available reduced representation methods, hybridization capture methods allow to enrich genomic DNA samples for preselected hundreds to thousands genes or DNA fragments (Hodges *et al.*, 2007). The most common application of hybridization sequence capture is exon capture in which only coding regions (and their flanking regions) of the genome are sequenced (Warr *et al.*, 2015). It provides focused information on gene function and adaptation, and it has no bias due to variation in gene expression contrary to transcriptome sequencing.

Exome capture requires knowledge of exons' sequences of the studied species to design corresponding oligonucleotide probes. As a consequence, it has been mainly developed on model species (i.e. with reference genome). Because probe specificity does not need to be absolute, and because despite millions of years of divergence, functional elements tend to be conserved, exons can be captured from a non-model species (i.e. with no published genome) using probes designed from a related model species (e.g. Cosart *et al.*, 2011; Förster *et al.*, 2018). Such approach has been particularly used to successfully resolve species phylogenies (e.g. Ilves & López-Fernández, 2014; Bossert & Danforth, 2018; Ilves *et al.*, 2018). It is also an effective method in paleontology to capture and enrich degraded and contaminated ancient DNA from extinct groups (e.g. Castellano *et al.*, 2014).

An alternative approach for capture probe design for non-model species is to use *de novo* assembled transcriptome of the studied species or related species (e.g. Bi *et al.*, 2012; Neves *et al.*, 2013). One of the limits of this approach is that the intron-exon structure is then unknown, and a probe designed from transcript sequence that span two consecutive exons will thus only partially hybridize to genomic DNA, hence reducing capture efficacy (e.g. Neves *et al.*, 2013). One way to get around with this issue is to use orthologous genes of close model species to try to predict the intron positions (Bi *et al.*, 2012; Stephens *et al.*, 2015; Bragg *et al.*, 2016). However, such approach biases the choice of targets by retaining the most evolutionary conserved genes, which can be problematic in the case of population genomics studies. Even targeting orthologous genes, the transcriptome-based exon capture studies showed a reduction of capture efficacy close to the intron-exon boundaries because no capture probe that span exons and their (unknown) non-coding flanking region could be designed (Bi *et al.*, 2012; Neves *et al.*, 2013; Portik *et al.*, 2016).

Another drawback of transcriptome-based exon capture is that it requires several tedious steps to reconstruct in a fragmented way the genomic sequence of the targeted transcripts (usually named in-target assemblies), before performing the mapping of genomic reads on them or making phylogeny. Usually the genomic raw sequences are cleaned and assembled using various combinations of k-mer and k-cov. Then the obtained genomic contigs are compared to the target sequences to be ordered and scaffolded when they belong to the same targets. Because captured fragments are usually longer than probes, the final assemblies can contain several contigs which include exons as well as non-coding flanking sequences (Bi *et al.*, 2012; Neves *et al.*, 2013; Stephens *et al.*, 2015; Bragg *et al.*, 2016; Portik *et al.*, 2016). Thus, transcriptome-based exon capture offers an opportunity to expand gene models beyond the available coding sequence and to know *a posteriori* intron exon structure of targeted genes. Such complex approach is necessary for the construction of phylogeny when the capture with the probes of a species is used to capture orthologous targets of related species, but it may appear to be non-obligatory in a context of population genomics. In particular, several studies have shown that not all targets are completely rebuilt (e.g. Bragg *et al.*, 2016; Portik *et al.*, 2016), which may possibly decrease the detection of variants. To moderate this latter issue, Bragg et al. (2016) mapped the genomic reads of the species used for probe design, directly to the transcript target sequences, allowing to cover exons that were not successfully assembled but still captured and sequenced.

Transcriptome-based exon capture method has proven its ability to generate thousands of highly informative markers. However, most previous studies have aimed at resolving phylogenetic relationships between various set of related species (Nicholls *et al.*, 2015; Stephens *et al.*, 2015; Bragg *et al.*, 2016; Heyduk *et al.*, 2016; Ilves *et al.*, 2018), and applications in the field of population genomic studies are yet rare (but see Bi *et al.*, 2013; Dauphin *et al.*, 2019). Indeed, the financial costs of analyzing large numbers of individuals across multiple populations of non-model species remain often prohibitive, even if multiplexing, by indexing individuals, can be used to optimize sequencing capacity and costs (Shearer *et al.*, 2012). Many research questions in population genomics can be addressed just by using allele frequencies computed at the population level. In this context, pooling individuals within population before DNA extraction can be an efficient low-cost and time-saving strategy. It has been shown that pool-sequencing makes it possible to estimate allelic frequencies within populations as accurately as individual genotyping, and with a much lower library construction and sequencing effort (e.g. Gautier *et al.*, 2013; Schlötterer *et al.*, 2014). To our knowledge, the estimation of allelic frequencies from the combination of targeted capture methods and pool-sequencing strategies has only be evaluated in human (Bansal *et al.*, 2011; Day-Williams *et al.*, 2011; Ramos *et al.*, 2012; Ryu *et al.*, 2018) and very recently in pine trees (Dauphin *et al.*, 2019).

Here we propose a cost-effective and efficient way to quantify the single nucleotide polymorphism (SNP) of exome at the population level in non-model species for which no genome is available. Our approach relies on the combination of transcriptome-based exon capture and pool sequencing (pool before DNA extraction). We report the development of capture probes of the Asian ladybird *Harmonia axyridis* from its *de novo* transcriptome, without considering the availability of a draft genome (HaxR v1.0; Gautier *et al.*, 2018) and thus in the absence of known gene structure as it is the case in most non-model species. Using these probes, two target enrichment experiments were performed with the same individuals and compared, the first capture with indexed multiplexed individuals, and the second with a pool of non-indexed individuals. We evaluated the efficacy of the capture and the accuracy of allelic frequency estimates from pool sequencing in this context. We also proposed and evaluated a fast approach to process read data (Bragg *et al.*, 2016; Dauphin *et al.*, 2019), without assembling them, that consists of mapping reads directly to targeted *de novo* transcript sequences in order to detect variant loci and which also allows to efficiently predict *a posteriori* the location of intron-exon boundaries of target sequences.

Methods

Design of capture probes for target enrichment

We designed capture probes for ~5 Mb of randomly chosen coding exonic regions of *H. axyridis*. From a *de novo* transcriptome obtained by RNA-Seq of various tissues and conditions/stages of the species (Vogel *et al.*, 2017), putative peptide-coding sequences were sought using FRAMEDP (Gouzy *et al.*, 2009). The corresponding partial or complete CoDing Sequences (CDS) were filtered to eliminate (i) *Wolbachia* and other putative endosymbiont sequences, (ii) CDS with more than 1% of missing data (*Ns*) or with more than four consecutive *Ns*, (iii) CDS with GC% below 25 or above 75 and (iv) CDS with length below 400 bp (to avoid proportionately capturing too much flanking region and to allow a better tiling) or above 3500 bp (to allow targeting more different genes). Among the 28,326 putative CDS found, 12,739 were retained after filters. Finally, 5,736 CDS, corresponding to 5.5 Mb, were randomly chosen. We performed a sequence similarity search against complete proteins from SwissProt database and from *Drosophila melanogaster* and *Tribolium castaneum* proteomes using BLASTX, and we found that at least 895 CDS (15.6%) were supposed to be complete. Overall, 5,120 CDS (89.3%) had a hit with *T. castaneum* proteins (e-value ≤ 10-7).

The final probe design of the 5,736 selected CDS was made by the company NimbleGen. Repetitiveness of the target sequences were determined based on the frequency of 15-mers in the genomes of the coleoptera *T. castaneum* and *Anoplophora glabripennis*. The probes with more than one close match - five or fewer single-base insertions, deletions or substitutions using SSAHA algorithm (Ning *et al.*, 2001) - in *H. axyridis de novo* transcriptome or in the draft genome of *A. glabripennis* were discarded. The probes with a match to the *H. axyridis* or *T. castaneum* mitochondrial genome were also discarded. The few residual *Ns* in target sequences were replaced by a random nucleotide to allow to tile probe across a single unknown base. The final probe set corresponded to 6,400 regions of overlapping probes (5,347,461 bp), i.e. 5,717 of our randomly selected CDS (Zenodo https://doi.org/10.5281/zenodo.2598388). This probe set was used to prepare a SeqCap EZ Developer assay by Roche NimbleGen Technologies, where probes were synthesized as biotinylated DNA oligos. One capture reaction contained 2,100,000 probes ranging from 50 to 99 bp in length with a mean of 74.71 ± 4.92.

Individual and pooled genomic DNA library preparation

Thirty-six individuals of *H. axyridis* were collected from a feral population in October 2015 in China (Beijing; 40.057128°N, 116.53989°E), and immediately stored at -20°C in RNAlater solution. Based on the use of PIFs tools proposed by Gautier et al. (2013), we theoretically expect a similar precision in allele frequency estimation when sequencing 23 individually indexed individuals at 80X coverage each, or a single pool of 36 individuals at 125X, assuming an unequal amounts of individual contribution with ε = 50%. We thus extracted DNA (i) independently from 23 individuals using half the body (without hind leg), and (ii) from a pool of 36 individuals, including the former 23 individuals, using 2 hind legs per individual. All DNA extractions were performed using Qiagen DNeasy Blood & Tissue kit following manufacturer's recommendations.

Genomic libraries were prepared following the Nimblegen SeqCap EZ Library (version 5.0). Briefly, for each of the 24 samples (1 pool and 23 individuals), DNA (2µg in 100 µl) was mechanically sheared to a mean fragment length of 200 bp using a Covaris S2 E210 device (6*30s). Fragments were end-repaired, A-tailed and indexed (one index per sample) using KAPA Library Preparation Kit Illumina platforms. After ligation of Illumina adaptors and indexes, only fragments with a length between 250 and 450 bp were retained. PCR was performed (7 cycles) using standard Illumina paired-end primers followed by purification with AMPure XP beads (Beckman). The length, quality and concentration of the prepared DNA fragments were checked on a BioAnalyzer (Agilent High Sensitivity DNA Assay) and a Qubit.

Sequence capture hybridization and sequencing

The 23 indexed individuals were pooled in equimolar concentrations prior to hybridization. For each capture (one with the pool and one with the indexed individuals), a total of 1µg of amplified DNA was used for exomes enrichments using the SeqCap EZ Developer probes described above, and strictly following the Nimblegen SeqCap EZ Library protocol, version 5.0. For each capture, two parallel post-capture PCR (14 cycles) were performed on elution solution, and the PCR products were then merged before purification with AMPure XP beads. The length, quality and concentration of the final DNA fragments were checked on a BioAnalyzer (Agilent High Sensitivity DNA Assay) and a Qubit. Each capture was sequenced on a half lane of an Illumina HiSeq3000 sequencer (on the same lane) following the manufacturer’s instructions, in paired-end mode for 150 cycles. Samples were then demultiplexed, exported as FastQ files and the 24 libraries were processed independently.

SNP calling by mapping reads to the CDS

Sequence quality was checked using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adapter removal and low-quality base pairs clipping were performed by Trimmomatic (Bolger *et al.*, 2014) with the following parameters: ILLUMINACLIP:TruSeq-file.fa:2:30:10 LEADING:30 TRAILING:30

SLIDINGWINDOW:10:30 MINLEN:75, and FASTX (http://hannonlab.cshl.edu/fastx_toolkit/) to remove the 5 first bases.

We chose to align the reads directly on the targeted CDS sequences because the target enrichment was done on the same species as the one used for the probe design. Doing so, we also avoid the complex and time-consuming steps of assembling the reads and scaffolding the obtained contigs. For each library, cleaned reads were aligned on the 5,717 targeted CDS using BOWTIE2 (v2.2.5; Langmead & Salzberg, 2012) with default parameters and *local* option to allow each genomic read not to align over its entire length on targeted CDS sequence but with a minimum alignment length of 20 bp. Reads were aligned as unpaired to allow for mapping of orphan reads. Reads with multiple alignments, and those for which the mate-pair was aligned to a different target were discarded after mapping. The aligned data in SAM format was converted in BAM format and sorted with SAMtools. Duplicate reads were estimated using picard-tools v1.113 MarkDuplicates using default settings. Coverage at each target base was calculated using mosdepth (Pedersen & Quinlan, 2018).

SNP calling was performed using reads whose alignment was larger than 40 bases and which contained less than 10 deletions (cigar format). Separately on each of the 24 libraries, SNP sites were called with the SAMtools (v0.1.19) mpileup command using options -B and varscan (v2.3) mpileup2snp command with a minimum base quality to count a read at a position of 35 and with a threshold value for the p-value of 0.01. We only considered positions with a minimum coverage of 40 reads for each individual, and 250 reads for the pool (90.3% and 93% of target bases were covered by at least 40 and 250 reads for individual and pooled exomes respectively; see Fig. S1). We required a minimum number of reads supporting a variant of 15 for each individual, and 3 for the pool. Finally, to account for the variation in read coverage across positions and samples, we required a minimum variant allele frequency threshold of 0.25 for each individual and 0.01 for the pool. An individual was considered as homozygote for the variant if the variant allele frequency f(v) was above 0.75, and heterozygote if f(v) was in-between 0.25 and 0.75.

All distinct 409,328 SNP positions independently found in the 24 libraries were then genotyped for each library using varscan2 mpileup2cns program with, for all libraries, a minimum coverage of 40, a minimum number of variant reads of 3 and a minimum variant frequency of 0.01. Several filters were then applied to all SNP positions using either in-house PERL scripts or R software (R Development Core Team, 2015). First, only biallelic positions were conserved. Second, a minimum coverage of 40X for each individual and 250X for the pool was applied. Third, only positions genotyped for at least 20 individuals were kept. Fourth, highly covered SNPs in the pool, which were likely to represent repetitive regions or paralogues of the genome, were discarded: maximum pool coverage was set to 6,960X, as computed from the formula proposed by Li et al. (2014), $c+4*\sqrt{c}$ where c is the mean coverage in the pool (c=6,634.8X). After all these filters were applied, 300,036 biallelic SNPs were retained.

At each site, and for each capture, variant allele frequency at the population level was estimated as the number of reads with the variant allele divided by the total coverage. Individual genotypes were used to compute exact p-value for Hardy-Weinberg equilibrium test with the "HardyWeinberg" package in R (Graffelman, 2015), and with a correction of significance level by the false discovery rate procedure (Benjamini & Hochberg, 1995). Finally, SNPs were annotated and categorized using the SnpEff program (Cingolani *et al.*, 2012).

SNP calling by mapping reads to the corresponding genomic regions

We used *H. axyridis* draft genome (HaxR v1.0; Gautier *et al.*, 2018) to evaluate our approach of mapping reads directly to CDS and to evaluate the approach proposed below for predicting intron-exon boundaries (IEBs, i.e. a junction of 2 exon ends). For these purposes, we selected CDS targets as proposed by Neves et al. (2013) with some adjustments. Briefly, CDS target sequences were first align against genomic scaffolds using BLASTN (e-value cutoff of 10-20 and percent identity of 90) and the best-hit were retained for each target. The target sequences were then realigned to their corresponded genomic scaffold using both GMAP (version 2015-12-31; Wu & Watanabe, 2005) and EXONERATE est2genome (v2.2.0; http://www.genome.iastate.edu/bioinfo/resources/manuals/exonerate/) softwares. We isolated targets (i) which had a genome match over their entire length, (ii) which had the same intron/exon structure predictions with both alignment softwares, (iii) which were covered by capture probes over their entire length and (iv) for which we had access to a minimum of 200 bp upstream and downstream of the targeted CDS sequences. With these filters, a total of 3,161 CDS targets were retained (3,220,718 bp, i.e. 60.23% of our targeted exome). These full or partial 3,161 CDS represented 11,754 exons with an average of 3.72 exons per target (ranging from 1 to 17 exons) and the exon average length of 274 bp (ranging from 12 to 3,417 bp). Among the 11,754 exons, 6,295 were unambiguously complete exons because they were surrounded by 2 others, and a total of 8,593 reliable IEBs were identified between consecutive exons (Zenodo https://doi.org/10.5281/zenodo.2598388).

All genomic regions corresponding to these 3,161 CDS targets were then isolated including intron sequences and 200 bp upstream and downstream of the target sequences. For each library independently, all cleaned reads were aligned on these genomic regions using BOWTIE2 (v2.2.5) with default parameters and end-to-end option (i.e. reads must be aligned to their full length). Reads with multiple alignments, and reads for which the mate-pair was aligned to another target were discarded after mapping. Coverage calculation, SNP calling and variant allele frequency estimations were performed as above. To establish the correspondence between the position of a base on the genomic region and its position on the CDS sequence, we used the vulgar format from EXONERATE output. SNP positions for which there was a deletion or an insertion between the CDS sequence and its associated genomic sequence were excluded for comparison (i.e. 559 positions).

Intron-exon boundaries detection

Our approach is based on the sole mapping of the genomic reads directly to the CDS sequences by allowing reads not to map to their full length as proposed above. This way, a genomic read with both a piece of exon (at least 20 bp) and its flanking area (intron or UTR) will start mapping on the corresponding exon on the CDS sequence, and stop mapping at the exon end (Montes *et al.*, 2013). This feature will generate a peculiar signal if we focus, on a given position, on the number of reads that either begin or end their local alignment at that position (hereafter called *nbBE*). Along a covered exon, *nbBE* should be low compared to the coverage, except at an exon end position where we should observe a large increase of *nbBE*. The strong increase in *nbBE* should occur either at the exact exon end base, or a few bases away if the beginning of the intron sequence is, by chance, the same or highly similar as the beginning of the next exon sequence on the CDS. In the latter case, this creates an abnormally inflated coverage estimation locally on few bases at the boundaries of the following exon (Fig. S2 for an example). At an IEB, depending on whether a signal offset occurs for one or both exon ends, the size of the area with abnormal local inflated coverage may be variable, but less than 10 bp in most cases. Additionally, the presence of a SNP at some bases of an exon end prevents reads with the variant from aligning up to the IEB position, which can generate a signal in addition to those of the 2 exon ends. The combination of these different scenarios is possible.

We developed and tested a method to predict the IEBs on the basis of the *nbBE* increase by focusing on the 3,161 targeted transcripts for which we have genomic sequences, and therefore the *a posteriori* known intron-exon structure. To do so, we used the mapping output BAM file obtained from the pool library. At each transcript position, *nbBE* was calculated using the start read alignment position on the transcript and the CIGAR code both present in the BAM file. The value of the *nbBE* parameter at each position is a function of the coverage (and length of the reads). To take into account the possible effect of coverage heterogeneity on *nbBE*, we calculated the median of *nbBE* on a window of 11 bases centered on the position. A signal, i.e. prediction of an IEB, was detected at a position if *nbBE* was *X* times larger than the median, *X* being a tool parameter to be defined by the user. To define *X*, the user may first have to identify a subset of transcripts that have orthologues in a close model species, and then evaluate which value of *X* allows to better predict the IEBs on this subset. Signals at the exact CDS ends were not considered. We extended the detection of a signal to the 3 bases upstream and the 3 bases downstream of a position with measured signal (default value for parameter *n*=3). A measured signal therefore predicts an IEB in a window of 7 bases in total. An IEB, composed of 2 exon ends, theoretically generates 2 close signals whose windows overlap each other. If these 2 signals are found at the exact positions of the exon ends (no deviation) then the predicted region measures 8 bp. If these 2 signals overlap (deviation from one of the signals to the other) then the region predicted for the IEB measures 7 bp. If these signals deviate up to 6 bases from each other (i.e. cumulative deviation of 8 bases, the deviation being inversely directional for both signals) then the predicted region measures 14 bp. Rarer case, if a 3rd signal is detected in the same region, then it increases the length of the region by a few bases. If the 2 signals deviate by more than 6 bases from each other, then the tool predicts for the same IEB, two signal windows separated by a few bases. Each transcript was thus divided into short regions with prediction signal (named +) alternating with regions without prediction signal (named -). Each region was then assessed as follows: true positive (TP) if region + contains an IEB, false positive (FP) if region + does not contain an IEB, true negative (TN) if region - does not contain IEB and false negative (FN) if region - contains an IEB. To evaluate the performance of the method, the percentage of true IEBs correctly predicted, i.e. sensitivity (SN), was calculated as TP/(TP+FN) and the percentage of correct predictions, i.e. specificity (SP), as TP/(TP+FP) (Burset & Guigo, 1996). For a CDS without intron (i.e. 862 of the 3,161 CDS), its SN cannot be evaluated, its SP was 100 if no IEB was predicted (good prediction) and 0 otherwise.

Results

Capture data and filtration

From the 2 target enrichment experiments, 186,998,614 and 185,457,659 *H. axyridis* raw sequence read pairs were respectively obtained for the pool library (SRA ERX3237193) and for the 23 indexed individuals (SRA ERX3237194 to ERX3237216), respectively. For each indexed individual library, an average of 8,063,376 raw read pairs were obtained, ranging from 6,912,561 to 9,176,292. After cleaning, 91.2% of the raw read sequences were retained (see detailed information for all libraries in Table S1).

Capture efficacy, coverage and duplicates

Instead of assembling the cleaned reads, we chose to map them directly on the target CDS sequences used for probe design as explained and described in the methods section. The percentage of cleaned reads that mapped properly (i.e. by meeting our mapping parameters) on the 5,717 targeted CDS, i.e. the capture specificity, was 83.2% in average (range over libraries: 80.6%-89.2%; Table S2). Aligned bases represented 74.6% (range over libraries: 72.6%-75.7%) of all cleaned read bases. The global median read alignment length was 144 pb (range over libraries: 143-144 pb, see Table S2), with 62.7% of the mapped reads (range over libraries: 60.9%-64.3%) that mapped on their entire length. Among the mapped reads, a substantial proportion (22.2%; range over libraries: 20%-25.9%) were 'orphaned' mates – i.e. only one of the two mates aligned to the target sequences. This percentage is a good indicator that our capture also allows to sequence the flanking regions of targeted regions (Yi *et al.*, 2010; Neves *et al.*, 2013), although our direct mapping approach does not allow us to quantify them.

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

Considering all targeted bases, the median and mean base coverage for individual libraries were 142X (range over libraries: 119-165X) and 283X (range over libraries: 239-323X) respectively for the 23 individual libraries, and 3361X and 6635X respectively for the pool library (see Table S3). A small number of targeted bases had very high coverage, but around 90% had a coverage <300X for individuals or <7000X for the pool (Table S3). For individuals, on average 90.3% of the targeted bases had a mean coverage greater than 40X. For the pool, 93% had a coverage greater than 250X (Fig. S1).

Base coverage varied along targeted CDS (Fig. S3A as an example). Using the known complete exons (≥ 20 bp) identified on targeted CDS that had a genomic match over their entire length on the *H. axyridis* draft genome, we confirmed that mean coverage increases steadily with exon size (Fig. S3B), and coverage within an exon decreased as you approach its ends (Fig. S3C). Note that since reads were not allowed to align with less than 20 bp with our mapping, coverage was here likely underestimated at the exon ends and by ricochet for short exons. Considering all targeted CDS, mean coverages varied between targeted CDS but were highly reproducible between libraries (Fig. S4): correlation coefficient of the target coverage between individual libraries ranged from 0.9575 to 0.9928 with a mean of 0.9867, and correlation coefficient between both captures, i.e. between pool and individuals, was 0.9988.

For the 23 individuals' libraries, on average 66.1% of reads were scored as duplicates (range over libraries: 63.6%-68.9%; Table S4). This high proportion is likely due to the markedly strong sequencing effort regarding the target region size. By pooling the reads of the 23 individual libraries, the proportion of detected duplicates increases to more than 96%, as is the case in the pool (Table S4). Consequently, reads scored as duplicates probably mostly come from different individuals in the case of the pool. Therefore, we did not discarded duplicated reads before SNP calling so as to avoid artificially biasing the estimation of allelic frequencies.

SNP calling and accuracy of allele frequency estimation

SNP calling performed independently on the individuals and on the pool from mapping directly on targeted CDS sequences enabled us to detect respectively 302,292 and 364,144 distinct putative exonic SNPs respectively, with 257,108 positions (62.8%) detected jointly in both. As expected (because there were 13 more individuals in the pool), the number of SNPs called from the pool is greater than from the 23 individuals and 26.2% of SNPs were called only from the pool. Among the 11% of SNPs called only from the individuals, the vast majority (95.6%) had minor allele frequencies lesser than 0.03, possibly explaining why these SNPs did not pass the detection thresholds for the pool.

All positions in all libraries were genotyped. After filtering we finally selected a total of 300,036 exonic bi-allelic SNPs within 4,741 targeted CDS among all libraries. Among them, 59.2% had MAFs lesser than 0.03 (Fig 1). 92.75% were polymorph in both the pool and the individuals, 7.22% were only polymorph in the pool, and 0.03% were only polymorph in the individuals. Among all SNPs, only 1,007 (i.e. 0.34%) showed significant deviation from Hardy-Weinberg equilibrium as computed with the individual genotypes. It is worth noting that this number was more than three times higher (i.e. 3,331 loci) before the removal of loci on the basis of the maximum coverage filter that was applied to the pool.

When considering all 300,036 loci, allele frequency estimations were strongly correlated between pool and individuals (r=0.9876; Fig. 2A). These results did not depend on the large fraction of SNPs with low allele frequency estimations (r=0.9833 computed on 102,494 loci with MAF ≥ 0.03).

The annotation of all SNPs showed that 74.75% of them were found on synonymous sites, while 24.92% were on missense non-synonymous sites and 0.33% on nonsense non-synonymous sites.

Comparison with SNPs found when mapping on genome

All trimmed reads were also mapped along their entire length to the genome regions corresponding to the subset of 3,161 targeted CDS that have a genomic match over their entire length. After applying the same procedure and filters for SNP calling as those used for direct mapping on CDS sequences, 290,620 filtered biallelic SNPs were genotyped including 187,591 (64.5%) on exons and 103,029 (35.5%) on the flanking regions (intron or UTR) of the exons. For this subset, we demonstrated that allele frequency estimations were strongly correlated between pool and individuals (r=0.9919; Fig. 2B).

Among exonic bi-allelic SNPs found by mapping directly to targeted CDS (Fig. 2A), 185,426 were found on the subset of 3,161 targeted CDS. Among them, 94% were also found as bi-allelic SNPs by mapping to genomic regions. For these 174,307 exonic bi-allelic SNPs found with both mapping approaches (either to genome or directly to transcripts), allele frequency estimations between mapping methods were highly correlated both for the pool (r=0.9975; Fig. 2C) and for the individuals (r=0.9977).

Among the private bi-allelic SNPs, some were variants that were found with both mapping approaches but which were discarded during filtering in one of the approaches. For example, among the private SNPs to the mapping to CDS, 1,467 bi-allelic SNPs had coverage greater than the maximum threshold and 714 were tri-allelic SNPs when mapping to genome. If we exclude the latter, 8,853 and 8,333 were respectively found as private by mapping directly to the target CDS sequences and by mapping to genomic regions. A non-exhaustive analysis of these positions allowed us to highlight 2 possible origins of private SNPs for mapping to CDS. The first is that some targeted regions (full CDS or part of CDS) have homologous copies in the genome, copies that we were unable to identify during the target selection and whose SNPs found were not discarded with the "maximum coverage" filter. By mapping the reads to the genome, for these regions, the reads mapped to several locations, were discarded and therefore no SNPs were called. For example, we have identified 43 CDS on which reads mapped to their genomic regions, but not uniquely. These CDS, with paralogs, had 3,141 SNPs, all private to CDS mapping. A second origin was identified by the following observation: 6.9% (614 SNPs) of SNPs private to CDS mapping were positioned near IEB positions (i.e. at the exact exon end or on the 3 adjacent bases). This percentage is abnormally high. For comparison, among the 174,307 SNPs found common to both mappings, only 0.6% were found close to IEBs. The detailed analysis of the read alignments on some CDS showed that the beginning of an intron can be aligned with the beginning of the next exon if it has a similarity of sequence with the latter (Fig. S2 as an example). These false alignments of only a few bases can generate false SNPs (Fig. S5 as an example).

Prediction of IEBs without assembling reads

The method we developed to detect IEBs from the mapping of genomic reads directly to CDS sequences was evaluated on the 3,161 targeted CDS for which we had genomic information, i.e. known positions of IEBs. An IEB signal was predicted at one CDS position if the number of reads that either begin or end their local alignment at that position (*nbBE*) was 20 times larger (*X*=20) than the median *nbBE* value found within the window of 11 bases around that position and a signal measured at a given position was extended to a short region of 7 bases, i.e. with a default parameter of *n*=3 (see Methods section). With parameters *X*=20 and *n*=3, the number of regions predicted as IEBs was 8,967, and their lengths ranged from 4 to 19 bp (Table 2). Regions smaller than 7 bases were regions next to regions not covered or with a measured signal just near CDS ends so that signal extension cannot be done. Among the 8,967 regions predicted as IEBs, 8,575 actually corresponded to true IEBs (i.e. method specificity = 95.63%). Among the 8,632 regions which truly contain an IEB, 8,575 were accurately predicted as IEB regions (i.e. sensitivity = 99.34%; Table 2). Among the 862 CDS with no intron, 763 were correctly predicted without IEBs. As expected, using a lower detection threshold, i.e. *X*=10, the specificity of the method decreases and its sensitivity increases (Table 2). Regions predicted with no IEB (i.e. regions -) had lengths ranging from 1 to 3,417 bp with a mean of 265.5 bp which was completely in accordance with the known distribution of exon lengths (see Methods section). 153 regions - with very small size (< 10 bp) attest to the possible cumulative deviation of signals of more than 8 bases for a small number of IEBs. Increasing the length of the signal deviation (e.g. *n*=5 bases instead of 3) allows this case to be taken into account in the vast majority of cases. Doing so slightly increases the performance but, on the other hand, increases the length, i.e. the imprecision of the position, of all predicted IEB positions (Table 2).

Discussion

Exome targeted capture associated with next-generation sequencing method has mainly been restricted to model species for which genomic resources are available (Ng *et al.*, 2009). However, transcriptome-based exome capture has proven to be an efficient alternative for dealing with non-model species (e.g. Bi *et al.*, 2012) but remain mainly confined to phylogeny studies. Here, we propose an extension of this method allowing high capture efficacy together with accurate estimation of SNPs' allelic frequencies for population genomics purposes. Our methodology is original because (i) it combines transcriptome-based exome capture with a pool-based approach performed before DNA extraction, (ii) it does not require the genome of a closely-related species, and (iii) it does not involve *de novo* assembling of genomic reads. Overall, this opens up interesting prospects for population genomics studies on non-model species because of the affordable cost and reduced labor involved.

Capture efficacy

In our study, capture efficacy of *H. axyridis*' exome was particularly high. The capture specificity (83.2%) was much higher than what can be found in more classical transcriptome-based studies, while the capture sensitivity (93.5%) approached the highest values (see Puritz & Lotterhos, 2018). Moreover, although differences in coverage are observed between targets and along a target, reproducibility among libraries was high, either in terms of specificity, sensitivity or coverage. Overall, our capture approach based on the random choice of targets on the *de novo* transcriptome of the species of interest itself, without looking for orthologous genes of related model species, has proven to be very effective. This "blind" approach had only little impact on the performance of the capture: we only observed a slight decrease in coverage for small exons and as we approached the IEBs. Such pattern is also observed in classical transcriptome-based exome capture where IEBs are taken into account by orthology, because of less efficient probes designed on two consecutive exons (e.g. Bi *et al.*, 2012; Neves *et al.*, 2013).

The random choice of targets is likely at the origin of the slightly reduced capture sensitivity that we obtained compared to some other studies (Bi *et al.*, 2012; Puritz & Lotterhos, 2018). The targets selected without looking for orthology with related species are by nature only partially known, and some could belong to other species (e.g. symbionts or parasites) or be chimeras present in the transcriptome used. This is a small price to pay given the efficiency of the method, particularly suitable for population genomics studies. Indeed, not looking for orthologous genes of related model species prevents focusing only on the most conserved genes. Moreover, our method can be applied to any species for which a closely related model species with genomic resources is not available.

SNP calling and pool-based approach

Our method allowed us to genotype a very large number of SNPs (~300,000) over the exome of *Harmonia axyridis*. The study of the subset of CDS with a complete genomic match further confirms that direct mapping on CDS has no major impact on SNP identification or on the estimation of their allelic frequencies compared to mapping on the genome. In addition to a few residual paralogue sequences inherent to an approach on non-model species, we notice at most a slightly higher number of SNPs called from mapping on CDS on areas very close to the IEBs, but the proportion of SNPs involved is minimal (0.33%).

For the same sequencing effort, we observe a high concordance between the allele frequencies resulting from individual genotyping and those resulting from the "pool-based" approach, whatever the mapping method used. This demonstrates the feasibility of this approach when individual identification of many samples is practically impossible, technically difficult, too costly and/or unnecessary. In population genomics studies, allele frequencies are sufficient to compute most summary statistics (e.g. nucleotide diversity, FST) and thus to address most questions (Schlötterer *et al.*, 2014).

Using pool of individuals, with no access to individual information, precludes the possibility to remove SNPs that are not at the Hardy-Weinberg equilibrium. Such disequilibrium can be caused by copy number variations (CNVs) or paralogous sequences that result, in addition to false-positive SNPs, in local over-coverage (Li, 2014). We show here that our "maximum coverage" filter applied to the pooled data allows us to discard almost 70% of the SNPs that were not at equilibrium as measured on individuals. The extremely small remaining proportion (0.34%) of SNP not at Hardy-Weinberg equilibrium reinforces the feasibility of our pool-based method.

Method to predict IEBs without assembling reads

We have developed a method for detecting IEBs particularly useful in the case of non-model species (or for species whose genomes are being assembled). The tool we propose does not indicate the exact location of the IEBs, but rather an area of a few bases (about 8 on average in our case) in which an IEB is supposed to be located. This method, which only requires the genomic reads and the target CDS sequences, was effective when we evaluated it on the subset of CDS for which we have a complete genomic match. Almost all of the true IEBs were identified, while false-positives represented less than 5% of the predicted IEBs. Montes et al (2013) have proposed a similar approach that they used to discard putative SNPs found in predicted IEBs area. They demonstrated better genotyping results for the selected filtered SNPs compared to other similar studies on non-model species. Here, we go further by taking into account the possible offset of the measured signal, and by evaluating the performance of the approach on a set of transcripts for which we know the true intron-exon structure. The aim here was to present evidence that this approach is promising, although it still needs to be improved and evaluated, using different mappers, on larger datasets and with various level of coverage. It also remains to develop a method to properly choose an appropriate value for parameter *X*.

General recommendations, conclusion and perspectives

The combination of approaches we propose – i.e. exome transcriptome-based capture method, random choice of targets, and direct mapping of reads on targeted CDS sequences, coupled with a pool-seq approach before DNA extraction – significantly reduces the cost, the lab time and also the complexity of sequence analyses to obtain a large number of variants on exome-scale coding regions for non-model species. However, the direct mapping on target sequences has two disadvantages. First, it does not allow the acquisition of information on flanking regions (including regulatory regions), and thus SNPs cannot be genotyped in these areas. Nevertheless, obtaining SNP polymorphism at the exonic level would be sufficient in many cases and the in-depth analysis of a large number of synonymous and non-synonymous mutations would be highly informative. It is still feasible to carry out an assembly of the genomic reads in situations in which flanking regions appear of interest. Second, the direct mapping on CDS leads to small biases of detection of SNPs in the vicinity of the IEBs. Some false-positive SNPs seem to be due to a similarity, on a few bases, between the beginning of one intron and the beginning of the next exon, resulting in a local alignment of reads from different genomic areas. To further limit the risk of detecting false SNP we hence suggest to use the IEBs detection method we propose, and then to eliminate any SNPs found in the predicted IEB regions and in short regions (≤ 20 bp) surrounded by 2 predicted IEB regions. By doing so on our data with parameters X=20 and n=3, we retained a final list of 296,736 SNPs. Among the 3,300 excluded SNPs, a large proportion (408 SNPs, i.e. 12.36%) were not at Hardy Weinberg equilibrium as measured on individuals, confirming the relevance of this filter.

We showed that the allele frequencies measured with our method are well estimated. However, it is still essential to be cautious with rare alleles (i.e. SNPs with low MAFs) as they can be confused with sequencing errors (Bansal, 2010). We recommend that future users conduct technical replicates. When possible and depending on the biological question addressed, it may also be chosen to work only on SNPs displaying MAFs above a given threshold.

In conclusion, our study highlights the possibility of acquiring at a low cost for a non-model species (i.e. without a reference genome) important and accurate information at the population level (i) on SNP polymorphism within the exome and (ii) on the general structure of this exome (intron-exon structure of targeted genes). Our tests on the Asian ladybird *Harmonia axyridis* have shown great potential, and it is now necessary to investigate other species with different genome sizes and complexities. In addition, the oversizing of our experimentation and the resulting very large coverage obtained suggest that it is possible to further reduce costs by capturing multiple indexed pools in a single reaction.

Acknowledgements

We thank Stefan Toepfer for *Harmonia axyridis* samples. The work was funded by département SPE-INRA and the INRA UMR Institut Sophia Agrobiotech.

Author contributions

ED, EL and TG conceived the study. AB, ED and EL carried out sample and library preparations. ED developed the bioinformatics pipelines. ED and EL ran the analyses. ED, EL and TG wrote the paper.

Competing interests

The authors declare that they have no competing interests.

Data accessibility

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

References

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

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

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

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

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C. & Good, J.M. 2012. Transcriptome-based 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 Predication 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.

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.

Dauphin, B., Zoller, S., Gugerli, F., Brodbeck, S. & Rellstab, C. 2019. Using transcriptome sequencing and pooled exome capture to study local adaptation in the giga-genome of Pinus cembra. *Mol. Ecol. Resour.* **19**: 536–551.

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**.

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.* 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**: ??–??

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.

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**.

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 neotropical rain forest genus Inga (Leguminosae: Mimosoideae). *Front. Plant Sci.* **6**: 1–20.

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

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

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

Puritz, J.B. & Lotterhos, K.E. 2018. Expressed exome capture sequencing: A method for cost-effective exome sequencing for all organisms. *Mol. Ecol. Resour.* 1209–1222.

R Development Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramos, E., Levinson, B.T., Chasnoff, S., Hughes, A., Young, A.L., Thornton, K., *et al.* 2012. Population-based rare variant detection via pooled exome or custom hybridization capture with or without individual indexing. *BMC Genomics* **13**: 683.

Ryu, S., Han, J., Norden-Krichmar, T.M., Schork, N.J. & Suh, Y. 2018. Effective discovery of rare variants by pooled target capture sequencing: A comparative analysis with individually indexed target capture sequencing. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **809**: 24–31. Elsevier.

Schlötterer, C., Tobler, R., Kofler, R. & Nolte, V. 2014. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15**: 749–763. Nature Publishing Group.

Shearer, A.E., Hildebrand, M.S., Ravi, H., Joshi, S., Guiffre, A.C., Novak, B., *et al.* 2012. Pre-capture multiplexing improves efficiency and cost-effectiveness of targeted genomic enrichment. *BMC Genomics* **13**.

Stephens, J.D., Rogers, W.L., Heyduk, K., Cruse-Sanders, J.M., Determann, R.O., Glenn, T.C., *et al.* 2015. Resolving phylogenetic relationships of the recently radiated carnivorous plant genus Sarracenia using target enrichment. *Mol. Phylogenet. Evol.* **85**: 76–87. Elsevier Inc.

Vogel, H., Schmidtberg, H. & Vilcinskas, A. 2017. Comparative transcriptomics in three ladybird species supports a role for immunity in invasion biology. *Dev. Comp. Immunol.* **67**: 452–456. Elsevier Ltd.

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N. & Watson, M. 2015. Exome Sequencing: Current and Future Perspectives. *Genes|Genomes|Genetics* **5**: 1543–1550.

Wu, T.D. & Watanabe, C.K. 2005. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**: 1859–1875.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., *et al.* 2010. Sequencing of Fifty Human Exomes Reveals Adaptation to High Altitude. *Science (80-.).* **329**: 75–78.

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

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

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C. & Good, J.M. 2012. Transcriptome-based 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 Predication Programs. *UNMC Mcgoogan Libr.* **655**: 12–26.

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.

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**.

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.* 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**: ??–??

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.

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**.

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 neotropical rain forest genus Inga (Leguminosae: Mimosoideae). *Front. Plant Sci.* **6**: 1–20.

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

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

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

Puritz, J.B. & Lotterhos, K.E. 2018. Expressed exome capture sequencing: A method for cost-effective exome sequencing for all organisms. *Mol. Ecol. Resour.* 1209–1222.

R Development Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramos, E., Levinson, B.T., Chasnoff, S., Hughes, A., Young, A.L., Thornton, K., *et al.* 2012. Population-based rare variant detection via pooled exome or custom hybridization capture with or without individual indexing. *BMC Genomics* **13**: 683.

Ryu, S., Han, J., Norden-Krichmar, T.M., Schork, N.J. & Suh, Y. 2018. Effective discovery of rare variants by pooled target capture sequencing: A comparative analysis with individually indexed target capture sequencing. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **809**: 24–31. Elsevier.

Schlötterer, C., Tobler, R., Kofler, R. & Nolte, V. 2014. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15**: 749–763. Nature Publishing Group.

Shearer, A.E., Hildebrand, M.S., Ravi, H., Joshi, S., Guiffre, A.C., Novak, B., *et al.* 2012. Pre-capture multiplexing improves efficiency and cost-effectiveness of targeted genomic enrichment. *BMC Genomics* **13**.

Stephens, J.D., Rogers, W.L., Heyduk, K., Cruse-Sanders, J.M., Determann, R.O., Glenn, T.C., *et al.* 2015. Resolving phylogenetic relationships of the recently radiated carnivorous plant genus Sarracenia using target enrichment. *Mol. Phylogenet. Evol.* **85**: 76–87. Elsevier Inc.

Vogel, H., Schmidtberg, H. & Vilcinskas, A. 2017. Comparative transcriptomics in three ladybird species supports a role for immunity in invasion biology. *Dev. Comp. Immunol.* **67**: 452–456. Elsevier Ltd.

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N. & Watson, M. 2015. Exome Sequencing: Current and Future Perspectives. *Genes|Genomes|Genetics* **5**: 1543–1550.

Wu, T.D. & Watanabe, C.K. 2005. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**: 1859–1875.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., *et al.* 2010. Sequencing of Fifty Human Exomes Reveals Adaptation to High Altitude. *Science (80-.).* **329**: 75–78.

Figures and Tables

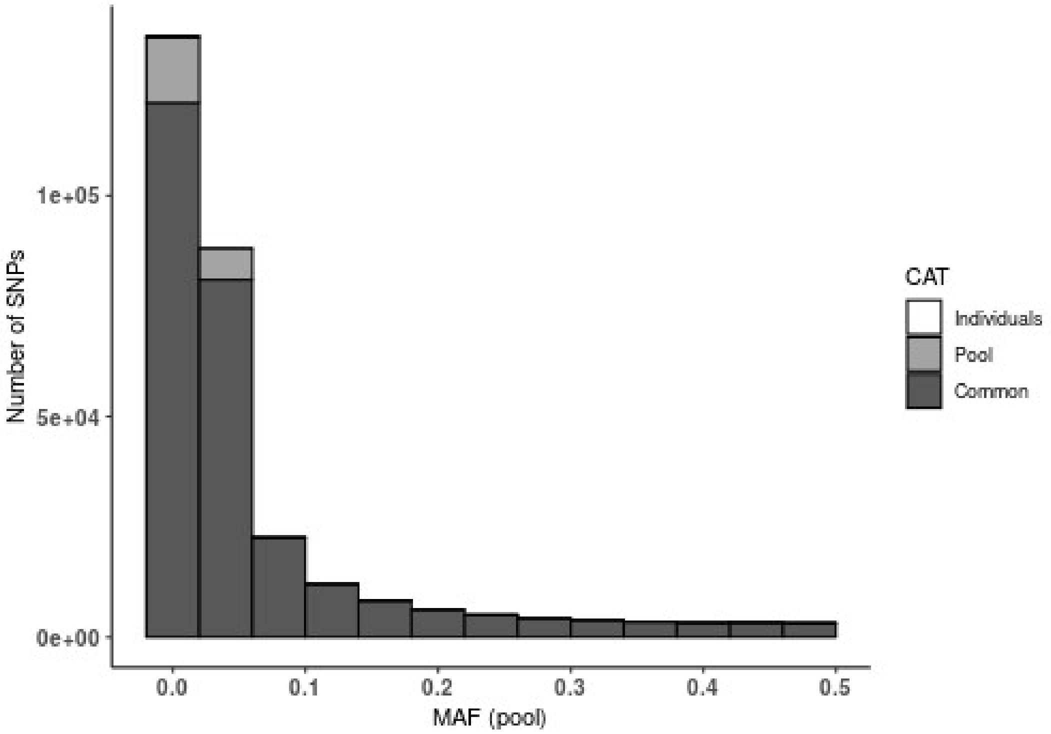
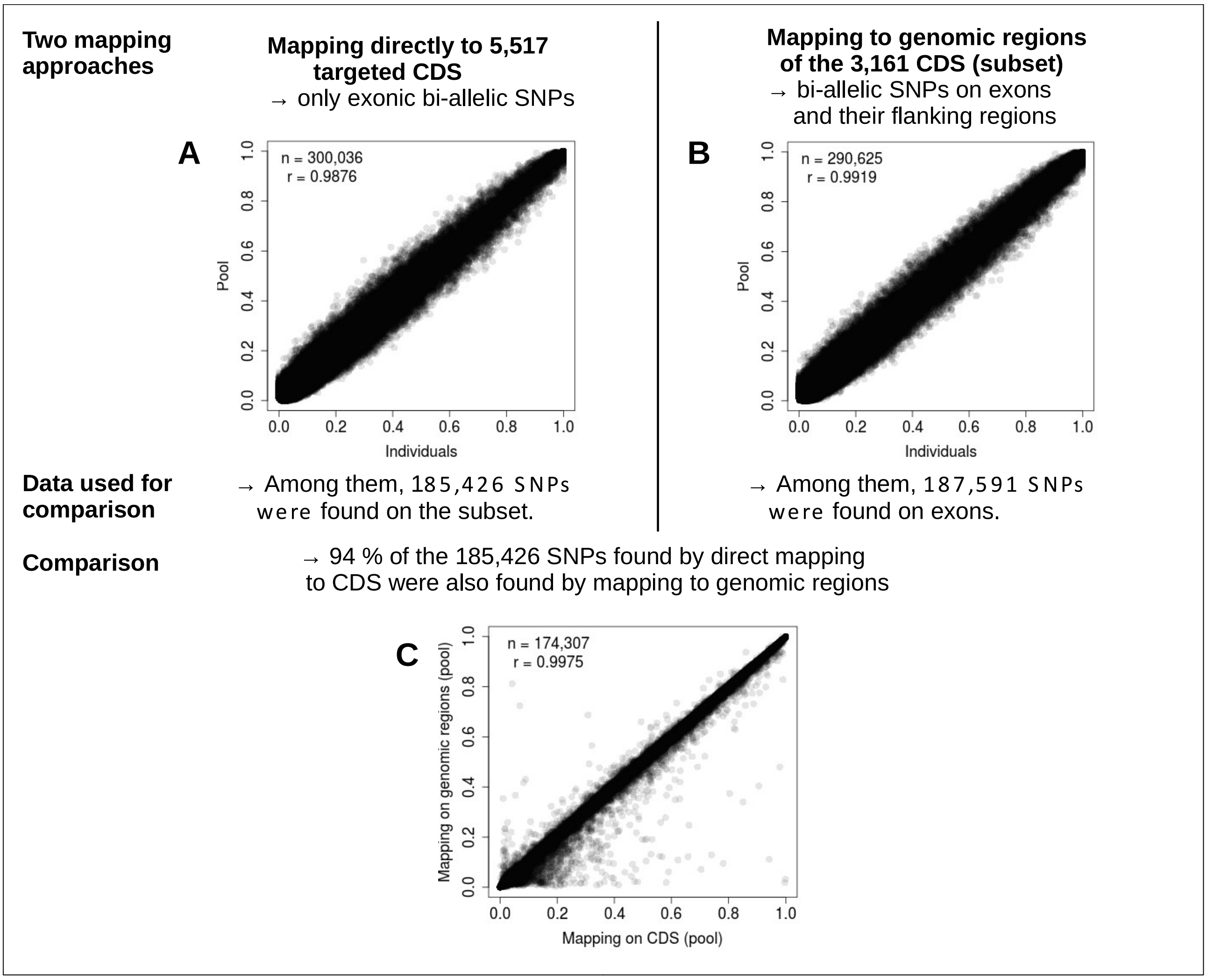


Figure 1:
Site frequency spectrum of the 300,036 filtered exonic biallelic SNPs (mapping on targeted CDS sequences) according to their polymorphism found in the pool and in the individuals. For SNPs polymorphic in both, MAF corresponds to MAF calculated on the pool.



Two mapping approaches

Mapping directly to 5,517 targeted CDS
→ only exonic bi-allelic SNPs

Mapping to genomic regions of the 3,161 CDS (subset)
→ bi-allelic SNPs on exons and their flanking regions

Data used for comparison

→ Among them, 185,426 SNPs were found on the subset.

→ Among them, 187,591 SNPs were found on exons.

Comparison

→ 94 % of the 185,426 SNPs found by direct mapping to CDS were also found by mapping to genomic regions

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

Table 1: Reproducibility of the capture sensitivity on the 24 libraries. 100 % means that all 24 samples had at least one read aligning to the target, whereas 0 % level indicate the opposite, no reads aligning to the target.

		Percentage of samples with target sequenced	
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%)

Table 2: Evaluation of the IEBs prediction method using pool data.

Parameters		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