Received: 25 March 2021 | Revised: 26 May 2021 | Accepted: 28 May 2021

DOI: 10.1111/mec.16011

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Whole-genome sequencing reveals asymmetric introgression between two sister species of cold-resistant leaf beetles

Svitlana Lukicheva | Patrick Mardulyn

Evolutionary Biology and Ecology, Interuniversity Institute of Bioinformatics in Brussels – (IB)², Université libre de Bruxelles, Brussels, Belgium

Correspondence
Svitlana Lukicheva and Patrick Mardulyn, Evolutionary Biology and Ecology, Université libre de Bruxelles, av. FD Roosevelt 50, 1050 Brussels, Belgium.
Emails: svitlana.lukicheva@gmail.com; patrick.mardulyn@ulb.be

Funding information
Fonds De La Recherche Scientifique - FNRS, Grant/Award Number: CDR J.0075.18

Abstract

A large number of genetic variation studies have identified cases of mitochondrial genome introgression in animals, indicating that reproductive barriers among closely related species are often permeable. Because of its sheer size, the impact of hybridization on the evolution of the nuclear genome is more difficult to apprehend. Only a few studies have explored it recently thanks to recent progress in DNA sequencing and genome assembly. Here, we analysed whole-genome sequence variation among multiple individuals of two sister species of leaf beetles inside their hybrid zone, in which asymmetric mitochondrial genome introgression had previously been established. We used a machine learning approach based on computer simulations for training to identify regions of the nuclear genome that were introgressed. We inferred asymmetric introgression of ≈2% of the genome, in the same direction that was observed for the mitochondrial genome. Because a previous study based on a reduced-representation sequencing approach was not able to detect this introgression, we conclude that whole-genome sequencing is necessary when the fraction of the introgressed genome is small. We also analysed the whole-genome sequence of a hybrid individual, demonstrating that hybrids have the capacity to backcross with the species for which virtually no introgression was observed. Our data suggest that one species has recently invaded the range of the other and/or some alleles that where transferred from the invaded into the invading species could be under positive selection and may have favoured the adaptation of the invading species to the Alpine environment.

KEYWORDS
hybrid zone, machine learning, nuclear genome introgression, population genomics, reproductive barriers, speciation

1 | INTRODUCTION

Reproductive isolation between populations causes them to genetically diverge and, if maintained for a sufficient period of time, ultimately leads to speciation (Coyne & Orr, 2004). The most documented, and probably most widespread, cause of reproductive isolation is geographic separation, at the core of the process of allopatric speciation (Coyne & Orr, 2004; Mayr, 1942). Yet, geographic separation can often be interrupted before the speciation process is complete (i.e., before reproductive barriers are fully in place), creating the possibility for the diverging species to exchange genetic material (Barton & Hewitt, 1985; Hewitt, 2000). Indeed, while species barriers were once seen as impermeable to gene flow, a large number of studies have identified cases of mitochondrial (mt) genome

© 2021 John Wiley & Sons Ltd

introgression, that is the transfer of the mt genome from one species to another as a result of hybridization between them, followed by repeated backcrossing with individuals from one of the parent species. In fact, a literature survey by Mallet (2005) estimated that 25% of plant species and 10% of animal species were involved in hybridization, suggesting that introgression could potentially affect many species.

Introgression of the mt genome is usually highlighted based on the observation of a phylogenetic/phylogeographic discordance between mt and nuclear loci (Toews & Brelsford, 2012) and has been reported in a large number of animal studies (Gómez-Zurita & Vogler, 2003; Bossu & Near, 2009; Nevado et al., 2009; Mardulyn et al., 2011; Melo-Ferreira et al., 2014; Quinzin & Mardulyn, 2014). When highlighting such pattern of discordance, some studies found at the same time little or no evidence of introgression in the nuclear genome (Hedrick, 2010; Zieliński et al., 2013). To understand why introgression is more apparent in mt genomes than in nuclear genomes, Bonnet et al. (2017) have performed simulations to compare alternative hypotheses and concluded that positive selection favouring the introgressed mt genome is the most probable explanation.

It is, however, important to recognize that most studies that have reported discordant patterns of introgression between mt and nuclear genomes relied only on a small number of nuclear loci (usually a few microsatellite markers or a few DNA sequence loci). Comparing the extent of introgression between the mt and nuclear genomes based on such a small number of loci is probably biased, because (1) the nuclear genome is several orders of magnitude larger than the mt genome and (2) the mt genome is believed to never recombine in most animal species. If a single mutation confers an advantage to the introgressed mt haplotype, it will remain in its entirety in the population of the host species across generations. On the other hand, if a chromosome is transferred from one species to the other via hybridization, only a fraction of it will remain per individual after multiple generations (at least if hybridization is a rare event), even if it contained variants under positive selection, because of recombination occurring in the nuclear genome.

Before discussing hypotheses explaining why hybridization impacts the nuclear genome less than its mt counterpart, it is important to explore further whether this is indeed the case and to what extent. So far, there have been only few studies that have investigated introgression by analysing variation over the entire nuclear genome. Fortunately, with the recent progress in sequencing technologies, analysing genome variation has become increasingly accessible (Metzker, 2010). Population genetic studies are now able to explore genome-wide variation for multiple individuals. For example, Good et al., (2015) used targeted capture to sequence over 10,500 gene regions from multiple individuals of two species of chipmunks for which mitochondrial genome capture had been observed in multiple populations and found no or little evidence of nuclear DNA introgression. Kastally et al., (2019) used a RAD-seq approach to genotype >100,000 SNPs across the genome of multiple individuals of two species of leaf beetles for which mt genome introgression had occurred multiple times and reached a similar conclusion. While the number of loci was much higher in these studies compared with most earlier studies, marker density could still be too low to detect introgression, except for the most recent hybridization events, because recombination events in the nuclear genome are expected to dilute foreign sequences to a minimum after a large number of generations.

Sequencing the entire nuclear genome of multiple individuals has now become a reasonable endeavour, even for a small research group, at least for those genomes that are not too large. Whole-genome sequencing has recently been used to study introgression in several organisms such as poplars (between *Populus balsamifera* and *P. trichocarpa*; Suarez-Gonzalez et al., 2018), mosquitoes (between *Anopheles coluzzii* and *A. gambiae*; Hanemaaijer et al., 2018), hares (between *Lepus granatensis* and *L. timidus*; Seixas et al., 2018), Drosophila (between *D. simulans* and *D. sechellia*; Schrider et al., 2018) or moths (between *Helicoverpa zea* and *H. armigera*; Valencia-Montoya et al., 2020). In general, these studies have succeeded in detecting introgression within the nuclear genome, often limited to small regions that are subject to adaptive selection. It is possible that purifying selection (Valencia-Montoya et al., 2020) and recombination both contribute to decrease introgression in the nuclear genome to a very small fraction, making whole-genome sequencing the only reliable approach to detect it.

In this study, we contribute to investigating the extent of introgression in the nuclear genome by sequencing the whole genome of multiple individuals of two sister species of leaf beetle in one of their previously identified hybrid zone. More specifically, we focus on *Gonioctena quinquepunctata* and *G. intermedia*, two European cold-tolerant leaf beetles characterized by fragmented and parapatric distributions (Quinzin & Mardulyn, 2014). While they most likely differentiated first in allopatry, their current ranges overlap in different places, with their largest hybrid zone located inside the Alps. There, both species are sometimes found in the same locality and entire populations of *G. intermedia* harbour the mitochondrial genome of its sister species (Kastally et al., 2019; Quinzin & Mardulyn, 2014). Because the observed mt introgression is relatively recent (both species share identical mt COI haplotypes), it was inferred that hybridization still occurs occasionally between the two species. To investigate the extent of introgression occurring at the level of the nuclear genome, we sequenced and assembled the entire genome of 10 individuals from *G. intermedia*, 9 individuals from *G. quinquepunctata* and one individual identified as a hybrid, all collected from multiple localities in their Alpine hybrid zone. While several DNA sequence variation statistics have been proposed and used in the past to identify regions of introgression when scanning the genomes of two hybridizing species (Geneva et al., 2015; Joly et al., 2009; Rosenzweig et al., 2016), we favoured here a more powerful approach that uses a machine learning algorithm to investigate beforehand the efficiency of multiple statistics (15 in total) at detecting introgression, in the specific case of the data set analysed here (Schrider et al., 2018). For this purpose, the evolution of the two species since they started to diverge from each other was modelled using population parameters directly inferred from the genome sequence data, and this model was used to simulate multiple instances of a genomic data set sharing similar features with our observed sequence data, with and without introgression. The algorithm is then trained to detect introgressed sequences on the simulated data sets, using not one but the whole set of statistics available. Once properly trained, the algorithm estimates the extent and direction of introgression between the two species from the real data at the level of the nuclear genome, as a consequence of the occasional hybridizations that occur within their hybrid zone. We also discuss what these results tell us about the evolutionary history of the two species and how their current interaction could impact their future evolution.

2 | MATERIALS AND METHODS

2.1 | Reference genome assembly and annotation

The reference genome of *G. quinquepunctata* used for this study was assembled from a single individual (pupa) collected on 14 May 2018 in the vicinity of the 'Col d'Urbeis' (latitude 48.330 N, longitude 7.174 E). DNA extraction was performed using the Qiagen kit Genomic-tip 20/G and following the instructions from the manufacturer's protocol. Extracted genomic DNA was divided into two separate batches. A first batch of ~1.5 µg was sent to Genewiz (www.genewiz.com) for library preparation and DNA sequencing on an Illumina HiSeq 2500 platform, which resulted in generating 145.1 Gb of data (~290 million pairs of PE reads 2 × 250 bp), corresponding to a sequencing depth of about 85×. The second batch of ~1.3 µg was used for Nanopore library preparation and sequencing. Five libraries were prepared using the SQK-LSK109 Nanopore kit and sequenced on a MinION sequencer with flowcells version 9.4, generating 46.3 Gb of data (a total of 4.4 million reads associated with lengths ranging from 31 to 144,886 b, which corresponds to a sequencing depth of ~27×). The Nanopore sequences were used for the initial assembly using wtdbg2 (Ruan & Li, 2020) with the parameters -x ont -g 1.5g -t 16. The assembly was then corrected twice with the Illumina sequences following the protocol specified in the README.md file distributed with wtdbg2. The resulting assembly was 1.9 Gbp long with N50 of 326 kbp.

For helping genome annotation, we extracted RNA from 4 individuals at different developmental stages (all collected on 19 June 2018): 1 adult male and 1 fourth instar larva collected in the vicinity of 'Grand Ballon' (latitude 47.90 N, longitude 7.103 E), 1 pupa collected in the vicinity of 'Le Breitfirst' (latitude 47.95 N, longitude 7.023 E) and 1 adult collected in the vicinity of 'Col d'Urbeis' (same coordinates as before). RNA was extracted using the Qiagen RNeasy Mini Kit and following the instructions from the manufacturer's protocol. The extracted RNA was sent to Eurofins Genomics (www.eurofinsgenomics.eu) for library preparation and RNA sequencing on an Illumina HiSeq 2500 platform, which resulted in generating 51.2 Gb of data (a total of 177 million pairs of PE reads 2 × 150 bp). The resulting RNA-Seq libraries were used as evidence to annotate the genome using BRAKER2 v.2.2 with the options –softmasking and –gff3. Prior to annotation, a species-specific repeat library was built using REPEATMODELER v.1.0.11 (Smit and Hubley, 2008–2015) and was then used in combination with the Repbase library (RepeatMasker edition 20181026, Bao et al., 2015) to mask the repeated regions from the genome using REPEATMASKER v.4.0.9 (Smit and Hubley, 2013–2015) with the following options: -e ncbi -xsmall -poly -html -gff -source -frag 6000000. The genome annotation procedure predicted 57,734 coding genes, a much higher number than the range 17,000–23,000 predicted for most other Coleoptera. An even higher number of genes (75,642) were predicted for the recently published genome of another leaf beetle, *Ophraella communa* (Bouchemousse et al., 2020). It was explained by the authors as a probable overestimation resulting from the high number of transposable elements found in this genome, many of which (68%) were not included in the database of known repeat sequences. A large proportion of these predicted genes could therefore be undetected transposons. The proportion of unclassified masked sequences in the genome of *G. quinquepunctata* was also high (43%), a similar hypothesis can be proposed to explain our high number of predicted genes.

2.2 | Tissue sampling and DNA extraction

For studying introgression, individuals from both species were collected in May 2018 in multiple locations within the northwestern Alps, a region in which their ranges overlap (Kastally et al., 2019). DNA was extracted from whole individuals using the Qiagen kit Genomic-tip 20/G, following the manufacturer's protocol. Species identification (the two species are morphologically similar) was carried out by two PCRs conducted with each DNA extract using two separate primer pairs, each designed to amplify a DNA fragment of the elongation factor 1-α gene specific to one species (3′-TTTGTCAAAAGCTCCAGCGA-5′ and 3′-TCTTGGTTTATCTTAAAAATG-5′ for *G. intermedia*; 3′-ATCA GGTTATCATTTCRAAG-5′ and 3′-TCTGCGATATTTCAAAAACG-5′ for *G. quinquepunctata*; annealing temperature of 56°C; primers designed based on sequence variation highlighted in Quinzin and Mardulyn (2014)). We then selected 10 *G. intermedia* individuals, 9 *G. quinquepunctata* individuals and 1 putative hybrid (positive amplification with both primer pairs) from 7 sampling sites (Table 1) for whole-genome sequencing.

Sequencing was outsourced to Genewiz that generated ≈250 million paired-end reads (2 × 150 bp) on an Illumina HiSeq platform, which corresponds to a sequencing depth of 19 × relative to the reference genome of *G. quinquepunctata* (see Table S1).

2.3 | Variant calling

Variant calling was performed by aligning sequencing reads to the reference assembly of the *G. quinquepunctata* genome using the GATK suite v.4.1.0.0 (McKenna et al., 2010) following the best practices suggested by Van der Auwera et al., (2013) and the protocol used by Schrider et al. (2018).

TABLE 1 Sampling locations

Species	Locality	Latitude	Longitude	Sample size
G. quinquepunctata	Boudin	45.6872 N	6.5765 E	2
G. quinquepunctata	Celliers	47.4765 N	6.4116 E	1
G. quinquepunctata	Grand Naves	45.5491 N	6.5233 E	3
G. quinquepunctata	Hauteluce	45.7426 N	6.5708 E	1
G. quinquepunctata	La Thuile	45.5132 N	6.4447 E	2
G. intermedia	Boudin	45.6872 N	6.5765 E	1
G. intermedia	Hauteluce	45.7426 N	6.5708 E	2
G. intermedia	La Giettaz	45.8655 N	6.4914 E	1
G. intermedia	La Thuile	45.5132 N	6.4447 E	2
G. intermedia	Montievret	45.9785 N	6.3744 E	4
G. quinquepunctata × *G. intermedia*	La Thuile	45.5132 N	6.4447 E	1

Illumina adapters were first removed from the input sequences of all 20 individuals using BBDuk v.38.50b (https://sourceforge.net/projects/bbmap/), with the following options: minlen=100 ktrim=r k=25 mink=11 hdist=1 tpe tbo --ordered. The trimmed sequences were then aligned to the *G. quinquepunctata* reference genome using the bwa mem command of bwa v.0.7.17 (Li & Durbin, 2009), with the -R option to add read group information. The resulting SAM files were converted to BAM format and sorted with the view and sort commands of SAMTOOLS v.1.9 (Li, 2011; Li et al., 2009; Table S2). No indel-based realignment step was performed, since, according to the GATK authors, it is not useful with the recent versions of the program. Duplicated fragments were removed using GATK's MARKDUPLICATESSPARK tool.

GATK's HaplotypeCaller was run on each BAM file in discovery mode to produce a separate gVCF file for each individual. We used the –emit-ref-confidence GVCF option to compute confidence scores for each position of the genome and the --do-not-run-physical-phasing option to disable the physical phasing. Repetitive regions of the genome identified by REPEATMASKER were first converted into bed format with gff2bed of the bedops suite v.2.4.37 Neph et al., (2012) and then excluded from the HaplotypeCaller runs using the --exclude-intervals option. The individual gVCF files produced by that procedure were then used to create the VCF files needed for the downstream analyses, by first combining the relevant files into one gVCF file with GATK's CombineGVCFs and then transforming that file into a VCF file with GATK's GenotypeGVCFs with the default options, unless otherwise indicated.

2.4 | Putative hybrid individual

To determine the relationship of the putative hybrid to other individuals from the two species, we followed two different approaches: (1) we calculated pairwise distances between each pair of individuals using VCF2Dis v.1.42 (https://github.com/BGI-shenzhen/VCF2Dis), and (2) we inferred population structure from our sample a priori in a Bayesian framework using the program FASTSTRUCTURE v.1.0 (Raj et al., 2014). For both approaches, we used a VCF file that included all 20 individuals but retaining only good-quality variants. The bad-quality variants were removed from this file using GATK's hard-filtering procedure. In order to be able to apply different filters according to the variant type (SNP or indel), the VCF file was first divided into two files using GATK's SelectVariants: one containing SNPs only (with the option –select-type-to-include SNP) and one containing indels only (with the option --select-type-to-include INDEL). The two files were then filtered with GATK's VariantFiltration, annotating the field 'FILTER' of each variant with 'PASS' if it passed all the filters and with a custom annotation if one of the following conditions—retrieved from the GATK's hard-filtering tutorial available on https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants—were satisfied:

- QD < 2 for both SNPs and indels
- QUAL < 30 for both SNPs and indels
- FS > 60 for SNPs and > 200 for indels
- ReadPosRankSum < −8 for SNPs and < −20 for indels
- SOR > 3 for SNPs
- MQ40 < 40 for SNPs
- MQRankSum < −12,5

The two annotated VCF files were then merged with GATK's MergeVcfs, and the variants that passed all the filters above were selected using GATK's SelectVariants tool with the option --exclude-filtered. Following this operation, 5,012 k variants and 93 k indels were removed.

The resulting VCF file was used to compute a pairwise distance matrix using VCF2Dis with default arguments. This program computes a distance between two diploid genomes by simply adding the distances computed for each separate SNP and dividing this value by the total number of SNPs in the VCF file. A distance between the two genomes for one SNP equals 0 if both genotypes are identical, 0.5 if one allele differs between the two genotypes and 1 if both alleles are different. A phenogram was then built from the distance matrix using iTOL v.5 (Letunic & Bork, 2019).

The same VCF file was used to infer population structure with FASTSTRUCTURE, setting K = 2. Prior to this analysis, we first converted the VCF file to PLINK bed format with PLINK v1.90b6.21 Purcell et al. (2007) using the options --double-id --allow-extra-chr --make-bed.

2.5 | VCF files for FILET analyses

FILET requires two kinds of VCF files to run: one with genotype calls for each site, for data masking purposes (see next section); and one that includes only SNPs, for predicting introgression. In both cases, FILET requires one VCF file per species, leading to a total of four files. The VCF files with genotype calls for each type were produced by combining 10 *G. intermedia* and 9 *G. quinquepunctata* individuals (the hybrid individual was not included in the FILET analyses) into two files as described in Section 2.3 by adding the option --include-non-variant-sites to the GenotypeGVCFs step. These VCF files were in turn used to compute the VCF files containing only SNPs using GATK's SelectVariants with the option --select-type-to-include SNP.

2.6 | Data masking and phasing

In order to execute the FILET pipeline on good-quality data, we first masked repetitive regions and variants of poor quality in the reference genome. These steps were performed with FILET's makeMaskArmFilesFromQualAndRM.py script, by providing previously predicted repetitive regions and setting the genotype quality cut-off to 20 and the fraction of all samples that must survive this cut-off to 0.7. In its original version, the script also masks genotypes that are heterozygous for one of the species, but this behaviour was removed as in our case both species are highly heterozygous. This step resulted in a masked reference genome file in fasta format with repetitive regions and genotypes of poor quality replaced by strings of 'N's.

The masked reference, the VCF files containing genotype calls for each site and VCF files containing only SNPs were then used to mask individual genotypes (FILET's makeFilteredSnpVcfsForEachArm.py script with the quality threshold set to 20), which resulted in one VCF file per contig, with individual genotypes of poor quality masked. 10,323 VCF files containing <3 variants were removed, leaving 12,710 files for subsequent analyses. Contigs corresponding to these removed files were also removed from the reference genome. The remaining VCF files were phased using BEAGLE v.5.1 (beagle.25Nov19.28d.jar) (Browning et al., 2018; Browning & Browning, 2007) with default parameters. The phased VCF files were then transformed into phased fasta files with FILET's script phasedVcfsToFastas.py, which was modified in order to handle individuals of our species and to handle both species as diploid (while in the original script, one of the species was haploid).

2.7 | Data training

As FILET uses a machine learning approach to classify the data into three categories (no introgression, introgression from species 1 to species 2 and introgression from species 2 to species 1), it requires a data training step. This step takes as input several examples of data belonging to each of these categories and allows the classifier to learn how to classify the data. This step was performed by first modelling the shared demographic history of our species and then generating three sets of training data satisfying the model constraints through simulations.

For performing these simulations, we needed first to infer the shared demographic history of the two species from whole-genome variation data. For that purpose, we used a VCF file combining 10 *G. intermedia* and 9 *G. quinquepunctata* individuals. From that file, we removed all variants for which information for one or more individuals was missing due to insufficient data coverage using GATK's SelectVariants with the option --max-nocall-number 0. A folded SFS was then produced from that VCF file with an in-house script. The demographic history of the two species was inferred by analysing the SFS with GADMA v.1.0.1 (Noskova et al., 2020). We set an isolation–migration model with migration rates fixed at zero and estimated remaining model parameters with moments; no theta was specified, time units were set to thousands of years, and the generation time was set to 1 generation per year (which corresponds to the natural life cycle of both species). Migration was set to 0 because the two species are believed to have diverged from each other in allopatry for ≈1 million years, until recently, when they were brought again into contact in the Alps (Quinzin & Mardulyn, 2014).

The values estimated by GADMA for the size of the ancestral population (N_{anc}), the divergence time (T_{div}), and the initial and final population sizes for *G. intermedia* and *G. quinquepunctata* ($N_{\text{int\_0}}$, N_{int}, $N_{\text{qui\_0}}$, N_{qui}), as well as the value of μ estimated to $6.866*10^{-10}$, were used to simulate 3 sets of 10 kb sequence alignments (training data) with msmove (Garrigan & Geneva, 2014; Hudson, 2002): one set without introgression ('NI'), one set with introgression in direction *G. intermedia*→ *G. quinquepunctata* ('Iiq') and one set with introgression in direction *G. quinquepunctata* → *G. intermedia* ('Iqi'). For simulating these sequences, we followed Schrider et al. (2018). In order to avoid any bias produced by the demographic history inference, for each variable estimated with GADMA, we defined an interval around the estimated value v_i: $[v_i - v_i/2; v_i + v_i/2]$. In addition, for the simulations with introgression, we defined an interval of $[0;T_{div}/4]$ for the introgression time and an interval of [0.01;1] for the probability of introgression. In order to generate training sets with at least 10,000 examples per set, and since we have 7 variables to model for the set 'NI' and 9 variables for the sets 'Iiq' and 'Iqi', we generated 4 values uniformly distributed across its associated interval for the set 'NI' (resulting in 4^7 = 16,834 combinations), and 3 values for the sets 'Iiq' and 'Iqi' (resulting in 3^9 = 19,683 combinations per set); we then simulated sequence data (38 10 kb sequences, i.e., 2 sequences per individual) for all combinations of these values to create a total

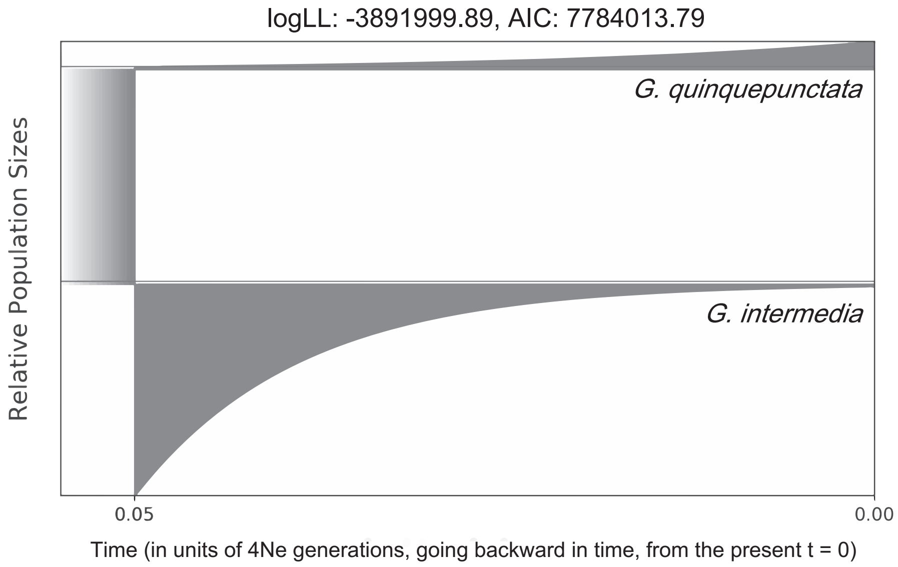

FIGURE 1 Demographic model of *G. intermedia* and *G. quinquepunctata* inferred with GADMA. Estimated ancestral population size N_{anc} = 9,808,986, divergence time T_{div} = 501,675 years, initial and final population sizes of *G. intermedia*: N_{int_0} = 9,673,164 and N_{int} = 107,758, and initial and final population sizes of *G. quinquepunctata*: N_{qui_0} =135,823 and N_{qui} =1,292,448

of 16,384 replicates of the set 'NI' and 19,683 replicates of the sets 'Iiq' and 'Iqi'.

Since repetitive regions and variants of poor quality were masked from the reference genome, a similar procedure was applied to the generated data, in order to avoid any bias related to the masking. To do so, for each 10-kb window of the reference genome, its 'masking pattern' (i.e., coordinates of the regions masked within the window) was stored in a separate file using FILET's makeMaskFilesFromMaskedRef.py script. The script was modified in order to only keep windows whose masking percentage did not exceed 75%, since this threshold will be later used in the data classification step to select the windows to classify. A total of 23,644 masking patterns were retrieved. For each of three data sets produced with msmove, we generated (1) a file containing one masking pattern per sample in the data set, in random order, and (2) a file providing the unmasked percentage (i.e., 1—the percentage of masked nucleotides) of the masked patterns used in the first file. Files with masking patterns were used to apply these patterns on the corresponding training set, by running FILET's msMaskAllRows program combined with the removeNedOutColumnsFromMsFile.py script, whereas the files containing the unmasked percentage were used during the classifier training step, described below.

Training the classifier was executed with FILET's examplePipeline.sh script available on GitHub (https://github.com/kr-colab/FILET). First, the masked simulated sequences were used to calculate, for each training set, the various single- and two-population summary statistics used by FILET for data classification. This was performed using FILET's twoPopnStats_forML combined with normalizeTwoPopnStats.py script, executed with the files containing unmasked percentage computed during the previous step. The calculated values of these statistics were then aggregated into one labelled data matrix, containing in addition one column with the appropriate class label with FILET's buildThreeClassTrainingSet.py script. In the absence of outgroup, we could not infer the ancestral state of each position in the genome; we therefore removed Fay and Wu's thetaH and H statistics from the matrix. The matrix was then used to train the classifier with FILET's trainFiletClassifier.py script.

2.8 | Data classification

For data classification, the masked reference genome was split into one file per contig. For each contig, we submitted a set of sliding windows of 10 kb to the classifier, with an overlap of 1 kb. The coordinates of these windows were generated with BEDTOOLS (Quinlan & Hall, 2010) v.2.29.0 using the options -w 10000 -s 1000, resulting in a total of 232,184 overlapping windows. The values of the various single- and two-population statistics used by FILET were computed for each window with FILET's pgStatsBedSubpop_forML with maximum fraction of missing data set to 0.75 combined with normalizePgStats.py script with windows size set to 10 kb. A significant fraction of the windows were not classified, because their masking percentage was too high, decreasing the total number of analysed contigs from 12710 to 5364. Finally, FILET's classifyChromosome.py script was used to produce classifications on the real data set with the critical posterior probability threshold for rejecting the no-migration class set to 0.05.

3 | RESULTS

3.1 | Demographic inference

The joint demographic history of 10 *G. intermedia* and 9 *G. quinquepunctata* individuals was inferred with GADMA (Noskova et al., 2020) by assuming an isolation–migration model with migration rates set to zero. The estimated best model for our data is shown in Figure 1. According to that model, the species split around 9 MYA and the population size of *G. intermedia* after the split was much larger than the one of *G. quinquepunctata* (representing 99% of the ancestral population size), whereas their sizes are more similar today, with *G. intermedia* being smaller than *G. quinquepunctata* (10^5 vs. 10^6, respectively).

3.2 | Population structure

The relationship between one sampled putative hybrid and the other sampled individuals from both species (10 *G. intermedia* and 9 *G. quinquepunctata*) was evaluated by a FASTSTRUCTURE (Raj et al., 2014) analysis in which k was set to 2. To compute the input VCF file for FASTSTRUCTURE, we aligned all 20 individuals to the reference genome of *G. quinquepunctata* and performed the variant calling procedure with GATK (McKenna et al., 2010). The resulting VCF file contained $\approx 34.7{*}10^6$ variants, including $\approx 29{*}10^6$ SNPs and $\approx 5.6{*}10^6$ indels. A hard filtering of these variants resulted in the removal of $\approx 5{*}10^6$ variants of poor quality, as shown in Table 2. The FASTSTRUCTURE barplot (Figure 2) showed a clear separation of *G. intermedia* and *G. quinquepunctata* individuals in two distinct groups, whereas the putative hybrid individual was evaluated to combine 13% of the *G. intermedia* genome with 87% of the *G. quinquepunctata* genome.

A phenogram built from a matrix of pairwise genomic distances calculated from the same VCF file (Figure 3) confirmed the inferred genetic relationships among sampled individuals. These data identify the putative hybrid as a third-generation hybrid (i.e., resulting from a two generations backcross to *G. quinquepunctata* of a F1 hybrid), thereby showing that first-generation hybrid individuals are, at least in some cases, fertile. In this example, the first-generation hybrid was able to reproduce with *G. quinquepunctata* individuals.

3.3 | Introgression detection with FILET

To study the extent of DNA exchange through hybridization between *G. intermedia* and *G. quinquepunctata*, we used FILET (Schrider et al. (2018)), a machine learning framework allowing to detect the introgression using a large set of genetic summary statistics. FILET compares intra- and inter-species variation across the genomes of two species of interest, via the computation of 18 single-population and 2 two-population statistics. It divides the genome into windows of a specific (user-defined) size and assigns each window to one of the following classes: no introgression, introgression from species 1 to species 2 and introgression from species 2 to species 1. FILET also allows users to perform a sliding-window analysis, by defining an overlap of the desired size between consecutive windows. Figure 4 shows an example of a sliding-window analysis along one contig, using 10-kb windows with 1-kb overlap, resulting in dividing the contig of interest in subwindows of 1 kb covered by one to ten 10-kb windows.

TABLE 2 Number of variants in the VCF file before and after the hard-filtering procedure

Value	Full VCF	Filtered VCF
Total variants	34,650,578	29,529,758
SNPs	29,074,338	24,046,530
Indels	5,576,240	5,483,228

We performed such a sliding-window analysis with FILET with 10-kb windows and 1-kb overlap. For this analysis, we retained 232,184 windows whose masking percentage did not exceed 75%, corresponding to 466,914 subwindows covered by 1 to 10 windows. For each subwindow, we analysed the classes assigned by FILET to the windows covering them and removed 8,757 (1.9%) subwindows that were covered by windows with different classes assigned to them. In other words, we removed all subwindows for which the classification was ambiguous. Among the remaining 458,157 subwindows, only those covered by at least 3 windows were kept, leaving 330,743 subwindows, corresponding to a total length of 330 Mb (17% of the whole genome). From this final set of subwindows, 6,550 (2%) of them were classified as introgressed with 127 (1.9%) in the direction of *G. intermedia* to *G. quinquepunctata* and 6,423 (98.1%) in the opposite direction, as summarized in Table 3 (see also Figures S1, S2).

From 6550 introgressed subwindows covered by at least 3 windows, we further retained only those having an average sequencing depth per base of 10 (considering an average per-base coverage of 20 for *G. intermedia* individuals and of 21.8 for *G. quinquepunctata* individuals), which left 6170 subwindows. These subwindows were compared with the annotated reference genome of *G. quinquepunctata*. From this comparison, we were able to establish that 1191 of them contained coding sequences, corresponding to a total of 379 coding genes.

4 | DISCUSSION

Our analyses of genetic variation across the genome of multiple individuals of both species from the Alps indicated that a fraction of ≈2% of the genome from *G. intermedia* is of foreign origin, having been transferred from the genome of *G. quinquepunctata*, presumably due to occasional hybridizations that occurred between individuals from each species, followed by successive backcrosses of the hybrids with *G. intermedia*. On the other hand, the transfer of genetic material in the reverse direction, from *G. intermedia* to *G. quinquepunctata*, was estimated to be 2 orders of magnitude lower (<0.05%) and can probably be considered to be negligible. A similar study analysing genetic variation of individuals from these two species sampled in the same area but with RAD-seq loci failed to detect any significant level of introgression (Kastally et al., 2019). Although the number of loci genotyped in that study was high (~130,000 SNPs), that number was still 2 orders of magnitude lower than in the current study, in which we have surveyed variation in the full nuclear genome. It is therefore likely that the marker density offered by a reduced-representation sequencing approach such as RAD-seq or GBS (Andrews et al., 2016)

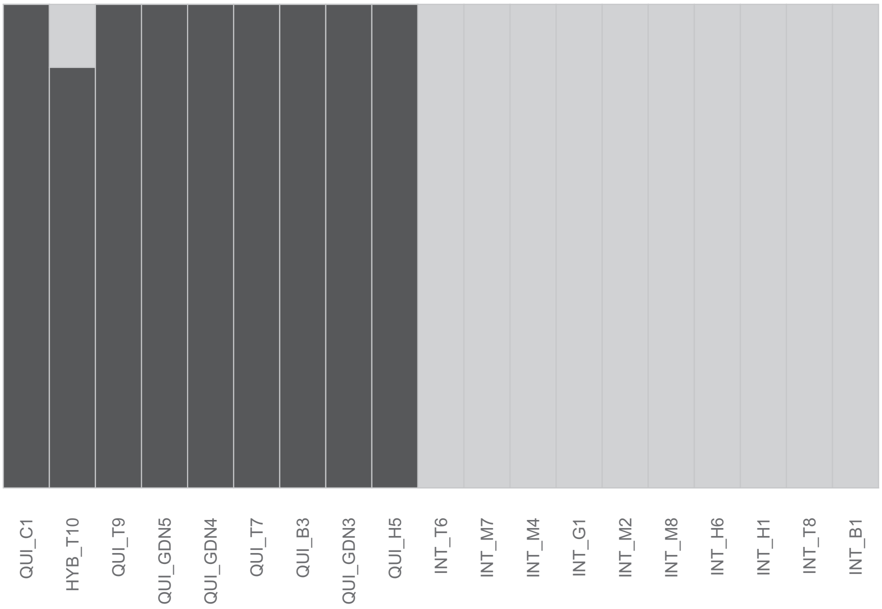

FIGURE 2 Population structure inferred with FASTSTRUCTURE. 10 *G. intermedia* and 9 *G. quinquepunctata* individuals are separated into two distinct groups (grey and black, respectively), whereas the hybrid individual is identified as 13% *G. intermedia* and 87% *G. quinquepunctata*.

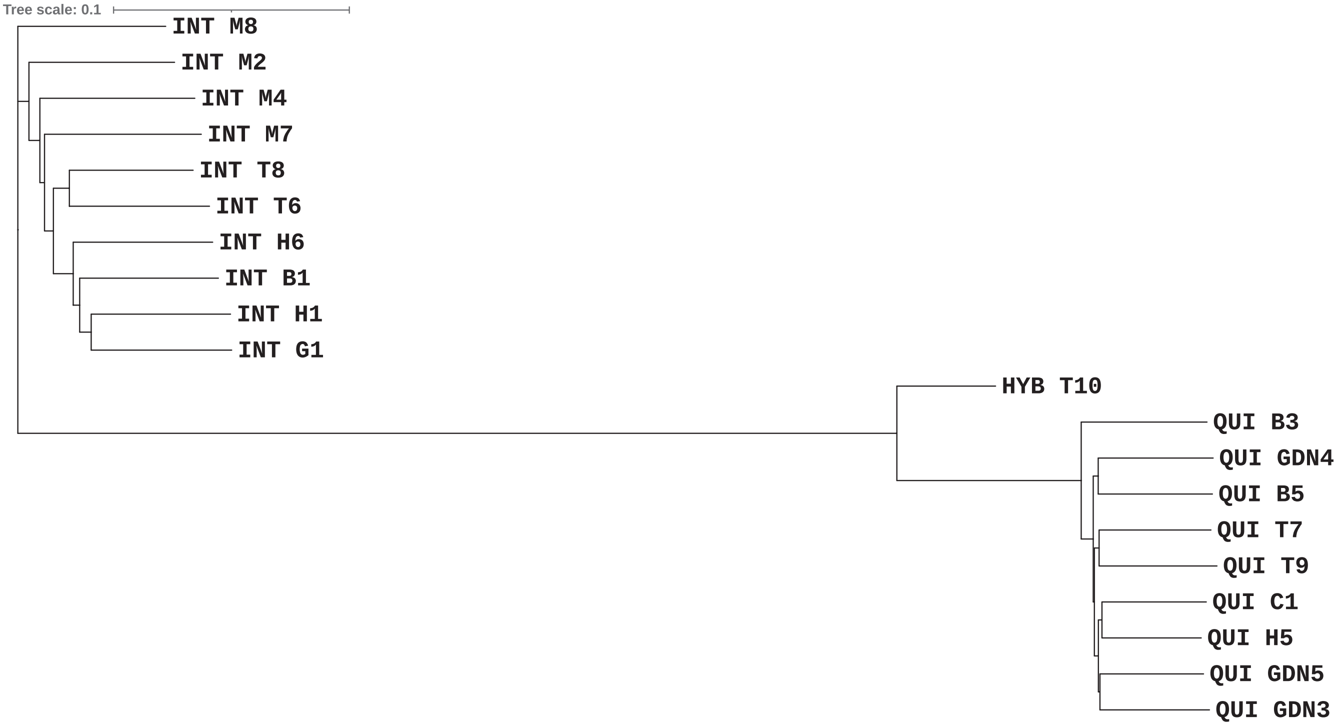

FIGURE 3 Phenogram of 10 *G. intermedia* individuals, 9 *G. quinquepunctata* individuals and one hybrid.

is not sufficient to highlight introgression if it involves only a small fraction of the genome.

The pattern of asymmetric introgression observed here at the level of the nuclear genome perfectly matches the one that was previously found for the mt genome: the mt genome of *G. quinquepunctata* was commonly found in *G. intermedia* individuals inside the Alps, but the reverse introgression pattern was never detected (Kastally et al., 2019; Quinzin & Mardulyn, 2014). Cases of asymmetric introgression have already been reported for other organisms, and different hypotheses have been proposed as explanations. One possibility is that the barriers to reproduction that have evolved during the differentiation of the two species are more effective with one type of cross than the other (Bolnick et al., 2008; Takami et al., 2007). If pre- or post-zygotic reproductive barriers prevented the crossing of *G. quinquepunctata* males with *G. intermedia* females but were less effective at preventing *G. intermedia* males × *G. quinquepunctata* females, it would account for asymmetric introgression for the mt genome. It would not account for asymmetric introgression in

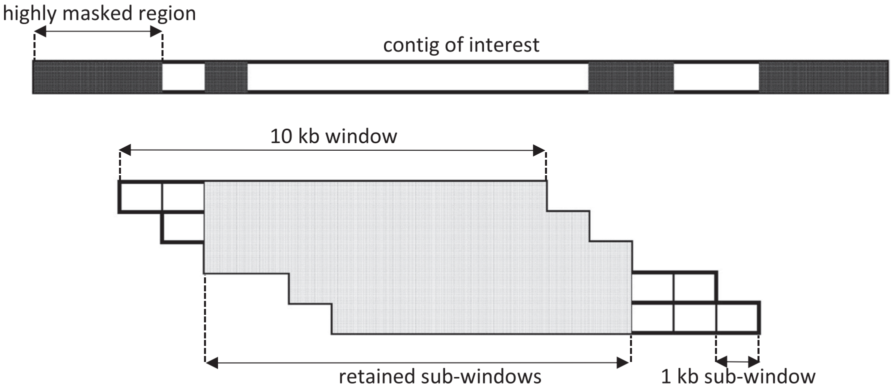

FIGURE 4 Example of defining sliding windows along a contig for FILET analysis. With size of 10 kb and overlap of 1 kb, a window is composed of 10 subwindows; each subwindow is covered by 1 to a maximum of 10 windows. Only subwindows covered by at least three windows were retained for the FILET analysis. If the maximum allowed masking percentage per window is specified, then windows exceeding that percentage will be excluded from the analysis.

TABLE 3 Subwindows assigned by FILET to each category.

Coverage[a]	Total	Introgressed	GI → GQ	GQ → GI
1	458,157	10,035	246	9,789
3	330,743	6,550	127	6,423
5	228,234	4,196	53	4,143
7	145,866	2,511	21	2,490

[a]Minimum number of windows covering each subwindow.

the nuclear genome, because hybrids would still be able to backcross with *G. quinquepunctata* individuals, importing *G. intermedia* nuclear genetic material in the *G. quinquepunctata* nuclear genome. To explain our observation of asymmetric introgression in both genomes, we would therefore be required to hypothesize further that reproductive barriers also prevented the hybrid offspring from backcrossing with *G. quinquepunctata* individuals. This prediction is directly contradicted by our analysis of the genome of one hybrid individual sampled in the field, however. This analysis revealed that 87% of this individual's nuclear genome was of *G. quinquepunctata* origin and that the remaining 13% came from *G. intermedia*, identifying this individual as a third-generation hybrid (the result of two successive backcrosses of a first-generation hybrid with *G. quinquepunctata* individuals). Reproductive barriers are therefore not capable of completely preventing hybrids from crossing with *G. quinquepunctata* individuals, and this observation suggests that reproductive isolation is not the main cause for the asymmetric introgression observed in both genomes (although the observation of additional instances of this type of hybrid would be necessary to confirm the absence of asymmetric reproductive barriers).

Two alternative hypotheses remain to explain the observed mt and nuclear asymmetric introgression, (1) a purely demographic hypothesis of invasion of the range of one species by the other species, and (2) a hypothesis involving selection. The likelihood of the first hypothesis is supported by computer simulations (Currat et al., 2008) that have shown that when one species invades the range of another, introgression occurs almost exclusively from the latter to the former. Indeed, at the wavefront of the invasion, there are only a few individuals of the invading species, surrounded by a large population from the invaded species, and these few individuals will have a higher probability of mating with individuals from the other species. Once some alleles are introgressed, their frequency in the invading population can then quickly increase, because of the important demographic growth that occurs at the wavefront of the invasion (founder effect). In that case, asymmetric introgression is expected, without the need to further invoke selection. This model prediction seems to be supported by multiple empirical studies (Currat et al., 2008) and was identified in other studies as the most likely scenario to explain current patterns of genetic variation, such as the one highlighted for two Neotropical toads (Sequeira et al., 2011) or for different temperate species of hare in the Iberian Peninsula that are believed to have replaced an arctic/boreal species after the last glaciation (Alves et al., 2008; Seixas et al., 2018). In this last example, the scenario of the invading species was coupled with male-biased dispersal, further explaining the absence of mt introgression in the other direction. If our observations are explained by a similar scenario, it would mean that *G. intermedia* has recently colonized the northern Alps and has gradually been replacing *G. quinquepunctata* in that region. We have no data at the moment allowing us to evaluate whether male-biased dispersal could occur in these beetles as well.

The other hypothesis proposes that positive selection favours at least some of the alleles transferred from another species. It appears particularly relevant in the case of the mt genome, because it is known to encode genes that play a major role in essential cell functions that are related to the respiratory chain (Dowling et al., 2008; Vafai & Mootha, 2012). The mitochondrial genome has already been shown to be under selection pressure in montane populations of another leaf beetle, *Chrysomela aeneicollis*, as its variation impacts physiology and larval development rate (Dahlhoff et al., 2019). Moreover, the functioning of the mitochondrion is ensured by genes encoded both in the mt and nuclear genomes that interact with one another and are thus likely to have co-evolved (Dowling et al., 2008). In fact, the close interaction between these mt and nuclear genes is believed to play an essential role in maintaining the integrity of species barriers, because alleles for these genes found in the same species will be strongly adapted to one another (Hill, 2015, 2019). Hybridization between species will disrupt the close association existing between these genes of mt and nuclear origin because it will lead to combining versions of them that are less compatible with each other, and thus to a decrease in individual fitness (Lima et al., 2019; Nguyen et al., 2020; Rank et al., 2020). Therefore, for mt genome introgression to be successful, it would require either (1) a strong positive selection pressure that could outweigh this expected decrease in fitness (Hill, 2019; Sloan et al., 2017) or (2) the simultaneous introgression of nuclear genes that are co-adapted to the mt genes, that is co-introgression of mt genes and nuclear genes that performs mt functions (Beck et al., 2015; Morales et al., 2018; Sloan et al., 2017). Among the 379 coding genes that we identified as introgressed in the nuclear genome of *G. intermedia*, 8 code for proteins that are located in the mitochondrion, which raises the possibility that some of them may be under positive selection when co-introgressed with the mt genome. Additional work is needed to test this hypothesis further.

Natural selection could of course also favour some of the introgressed nuclear alleles that are not co-adapted with mt alleles. Other authors have in fact suggested that most introgressed nuclear alleles are eliminated from the genome by purifying selection (Hanemaaijer et al., 2018; Valencia-Montoya et al., 2020). Under that view, if hybridization remains a rare event, the transferred DNA is diluted more and more at each generation because of recombination events and purifying selection eliminates it completely from the genome. It would then be assumed that all introgressed genes that we have identified in the nuclear genome must be either under direct positive selection or under indirect positive selection because they are located in the same region as another gene that is under positive selection (genetic hitchhiking). However, we found that only a small fraction of the introgressed regions (≈6%; Table S3) were characterized by a lower nucleotide diversity in *G. intermedia* than in *G. quinquepunctata*, an indication that these introgressed alleles may be under adaptive selection. This means that either genetic drift is the main factor governing patterns of genetic variation within these regions or that most introgression events are recent (and thus that purifying selection did not have the time to eliminate the regions that are not under positive selection). Moreover, the mere occurrence of selection in some of these regions is not enough to account for the fact that introgression at the level of the nuclear genome is asymmetric. Indeed, the probability that a significant fraction of the 379 genes transferred from *G. quinquepunctata* to *G. intermedia* are favoured by selection, but that none that were transferred from *G. intermedia* to *G. quinquepunctata* are, seems unlikely, unless *G. quinquepunctata* is better adapted to the alpine environment than *G. intermedia*. If that species has been present for a longer period of time in the Alps than its sister species, there would have been more opportunities for the appearance of new alleles allowing a better adaptation to that specific environment. On the other hand, if all introgressed genes are neutral and the fixation of introgressed alleles is mainly driven by genetic drift, we would expect to observe a symmetrical pattern of introgression (because genetic drift should lead to the fixation of a similar proportion of introgressed alleles in both sister species).

In conclusion, if we combine our inference of asymmetric introgression in both the nuclear and mt genomes (from *G. quinquepunctata* towards *G. intermedia*), with our observation of the capability of hybrids to backcross with *G. quinquepunctata* individuals, two alternative, but not necessarily exclusive, hypotheses stand out as the most probable: this pattern (1) is the result of *G. intermedia* having recently invaded the range of *G. quinquepunctata* in the Alps and/or (2) results from positive selection favouring at least a fraction of the alleles that were transferred from *G. quinquepunctata* to *G. intermedia*, allowing this last species to become better adapted to its environment. Combining both hypotheses, we suggest that *G. intermedia* is invading the range of *G. quinquepunctata*, its range shift was presumably driven by climate changes that occurred at the end of the Pleistocene (Quinzin et al., 2017). Capturing the mitochondrial genome and some nuclear genes of its sister species may have facilitated its invasion, by giving it access to genetic variation that was not available in the genome of individuals from its own species, and thereby allowing it to better adapt to the alpine habitat. Although this hypothesis needs to be investigated further by experiments that evaluate the fitness of different genotypes from both species in different environmental conditions, it could mean that *G. intermedia* will ultimately replace *G. quinquepunctata* in the Alps, at least in the northern part of the range where it is currently well established.

More generally, this study offers another example of hybridization between two diverging sister species inside a secondary contact zone that leads to introgression within the nuclear genome. Like in other studies examining whole-genome sequence variation (Hanemaaijer et al., 2018; Seixas et al., 2018; Valencia-Montoya et al., 2020), only a small fraction, representing a few per cent of the entire genome, was detected as introgressed. This may explain why introgression at the level of the nuclear genome was not as widely detected as it was for the mitochondrial genome: whole-genome sequencing may be the only option either to identify introgression of a small portion of the genome or to reject it altogether. Nevertheless, a general picture is thus emerging of more permeable species boundaries than previously thought. This has important implications for the future evolution of diverging species, as it means they still have in many cases access to the reservoir of genetic diversity from its sister species, which in some circumstances can help populations adapt to its environment.

ACKNOWLEDGEMENTS

We are grateful to two anonymous reviewers for their suggestions that helped us to improve the manuscript. This work was supported by a grant from Fonds d'Encouragement à la Recherche (FER) from the Université libre de Bruxelles and by a grant from the Belgian Fonds de la Recherche Scientifique (F.R.S.-FNRS; CDR J.0075.18) to P.M. S.L. benefitted from a 4-year PhD fellowship from

F.R.S.-FNRS and from a complementary fellowship from the David and Alice Van Buuren Fund and the Jaumotte-Demoulin Foundation. Computational resources were provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by F.R.S.-FNRS (Grant 2.5020.11) and the Walloon Region. Published with the help of the FondationUniversitaire de Belgique.

AUTHOR CONTRIBUTIONS

P.M. collected the samples and conducted the laboratory work (DNA extractions, Nanopore sequencing). S.L. performed all bioinformatic analyses. Both authors designed the study, interpreted the data and wrote the manuscript.

DATA AVAILABILITY STATEMENT

The raw sequencing reads of 20 individuals used for this study are publicly accessible at SRA NCBI database with Accession nos. SRR14696131-SRR14696150. The reference assembly, the VCF file computed for 20 individuals, the two VCF files containing SNPs per species, the two VCF files containing one entry for each site on the genome, the simulated training data sets and the classification files produced by FILET are available through the Dryad Digital Repository https://doi.org/10.5061/dryad.x3ffbg7jz. The modified FILET scripts are available on https://github.com/SvitlanaLukicheva/FILET. The program allowing to generate SFS files from a VCF file is available on https://github.com/SvitlanaLukicheva/VcfHandler. The program used to generate input values for msmove is available on https://github.com/SvitlanaLukicheva/MsmoveInputGenerator.

ORCID

Svitlana Lukicheva https://orcid.org/0000-0003-0170-6868
Patrick Mardulyn https://orcid.org/0000-0003-2154-5256

REFERENCES

Alves, P. C., Melo-Ferreira, J., Freitas, H., & Boursot, P. (2008). The ubiquitous mountain hare mitochondria: Multiple introgressive hybridization in hares, genus Lepus. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*, 2831–2839.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*, 81–92. https://doi.org/10.1038/nrg.2015.28

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1). https://doi.org/10.1186/s13100-015-0041-9

Barton, N., & Hewitt, G. (1985). Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, *16*, 113–148. https://doi.org/10.1146/annurev.es.16.110185.000553

Beck, E. A., Thompson, A. C., Sharbrough, J., Brud, E., & Llopart, A. (2015). Gene flow between *Drosophila yakuba* and *Drosophila santomea* in subunit V of cytochrome c oxidase: A potential case of cytonuclear cointrogression. *Evolution; International Journal of Organic Evolution*, *69*, 1973–1986.

Bolnick, D. I., Turelli, M., López-Fernández, H., Wainwright, P. C., & Near, T. J. (2008). Accelerated mitochondrial evolution and "Darwin's corollary": asymmetric viability of reciprocal F1 hybrids in Centrarchid fishes. *Genetics*, *178*, 1037–1048. https://doi.org/10.1534/genetics.107.081364

Bonnet, T., Leblois, R., Rousset, F., & Crochet, P.-A. (2017). A reassessment of explanations for discordant introgressions of mitochondrial and nuclear genomes. *Evolution; International Journal of Organic Evolution*, *71*, 2140–2158. https://doi.org/10.1111/evo.13296

Bossu, C. M., & Near, T. J. (2009). Gene trees reveal repeated instances of mitochondrial DNA introgression in orangethroat darters (percidae: etheostoma). *Systematic Biology*, *58*, 114–129. https://doi.org/10.1093/sysbio/syp014

Bouchemousse, S., Falquet, L., & Müller-Schärer, H. (2020). Genome assembly of the ragweed leaf beetle: a step forward to better predict rapid evolution of a weed biocontrol agent to environmental novelties. *Genome Biology and Evolution*, *12*, 1167–1173. https://doi.org/10.1093/gbe/evaa102

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *American Journal of Human Genetics*, *103*, 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Browning, S., & Browning, B. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*, 1084–1097. https://doi.org/10.1086/521987

Coyne, J. A., & Orr, H. A. (2004). *Speciation*. Sinauer Associates.

Currat, M., Ruedi, M., Petit, R. J., & Excoffier, L. (2008). The hidden side of invasions: massive introgression by local genes. *Evolution; International Journal of Organic Evolution*, *62*(1908–1920), 890. https://doi.org/10.1111/j.1558-5646.2008.00413.x

Dahlhoff, E. P., Dahlhoff, V. C., Grainger, C. A., Zavala, N. A., Otepola-Bello, D., Sargent, B. A., Roberts, K. T., Heidl, S. J., Smiley, J. T., & Rank, N. E. (2019). Getting chased up the mountain: High elevation may limit performance and fitness characters in a montane insect. *Functional Ecology*, *33*, 809–818. https://doi.org/10.1111/1365-2435.13286

Dowling, D. K., Friberg, U., & Lindell, J. (2008). Evolutionary implications of non-neutral mitochondrial genetic variation. *Trends in Ecology & Evolution*, *23*, 546–554. https://doi.org/10.1016/j.tree.2008.05.011

Garrigan, D., & Geneva, A. J. (2014). msmove. figshare. https://doi.org/10.6084/m9.figshare.1060474

Geneva, A. J., Muirhead, C. A., Kingan, S. B., & Garrigan, D. (2015). A new method to scan genomes for introgression in a secondary contact model. *PLoS One*, *10*, e0118621. https://doi.org/10.1371/journal.pone.0118621

Gómez-Zurita, J., & Vogler, A. P. (2003). Incongruent nuclear and mitochondrial phylogeographic patterns in the *Timarcha goettingensis* species complex (Coleoptera, Chrysomelidae). *Journal of Evolutionary Biology*, *16*, 833–843. https://doi.org/10.1046/j.1420-9101.2003.00599.x

Good, J. M., Vanderpool, D., Keeble, S., & Bi, K. (2015). Negligible nuclear introgression despite complete mitochondrial capture between two species of chipmunks. *Evolution*, *69*, 1961–1972. https://doi.org/10.1111/evo.12712

Hanemaaijer, M. J., Collier, T. C., Chang, A., Shott, C. C., Houston, P. D., Schmidt, H., Main, B. J., Cornel, A. J., Lee, Y., & Lanzaro, G. C. (2018). The fate of genes that cross species boundaries after a major hybridization event in a natural mosquito population. *Molecular Ecology*, *27*, 4978–4990. https://doi.org/10.1111/mec.14947

Hedrick, P. W. (2010). Cattle ancestry in bison: explanations for higher mtDNA than autosomal ancestry. *Molecular Ecology*, *19*, 3328–3335. https://doi.org/10.1111/j.1365-294X.2010.04752.x

Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. *Nature*, *405*, 907–913. https://doi.org/10.1038/35016000

Hill, G. E. (2015). Mitonuclear ecology. *Molecular Biology and Evolution*, *32*, 1917–1927. https://doi.org/10.1093/molbev/msv104

Hill, G. E. (2019). Reconciling the mitonuclear compatibility species concept with rampant mitochondrial introgression. *Integrative and Comparative Biology*, *59*, 912–924.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, *18*, 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Joly, S., McLenachan, P., & Lockhart, P. (2009). A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist*, *174*, E54–E70. https://doi.org/10.1086/600082

Kastally, C., Trasoletti, M., & Mardulyn, P. (2019). Limited gene exchange between two sister species of leaf beetles within a hybrid zone in the Alps. *Journal of Evolutionary Biology*, *32*, 1406–1417. https://doi.org/10.1111/jeb.13538

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: Recent updates and new development. *Nucleic Acids Research*, *47*, W256–W259.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lima, T. G., Burton, R. S., & Willett, C. S. (2019). Genomic scans reveal multiple mito-nuclear incompatibilities in population crosses of the copepod *Tigriopus californicus*. *Evolution; International Journal of Organic Evolution*, *73*, 609–620.

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, *20*, 229–237. https://doi.org/10.1016/j.tree.2005.02.010

Mardulyn, P., Othmezouri, N., Mikhailov, Y. E., & Pasteels, J. M. (2011). Conflicting mitochondrial and nuclear phylogeographic signals and evolution of host-plant shifts in the boreo-montane leaf beetle *Chrysomela lapponica*. *Molecular Phylogenetics and Evolution*, *61*, 686–696. https://doi.org/10.1016/j.ympev.2011.09.001

Mayr, E. (1942). *Systematics and the origin of species*. Columbia University Press.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110

Melo-Ferreira, J., Seixas, F. A., Cheng, E., Mills, L. S., & Alves, P. C. (2014). The hidden history of the snowshoe hare, *Lepus americanus*: extensive mitochondrial DNA introgression inferred from multilocus genetic variation. *Molecular Ecology*, *23*, 4617–4630.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, *11*, 31–463. https://doi.org/10.1038/nrg2626

Morales, H. E., Pavlova, A., Amos, N., Major, R., Kilian, A., Greening, C., & Sunnucks, P. (2018). Concordant divergence of mitogenomes and a mitonuclear gene cluster in bird lineages inhabiting different climates. *Nature Ecology & Evolution*, *2*, 1258–1267. https://doi.org/10.1038/s41559-018-0606-3

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., & Stamatoyannopoulos, J. A. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics*, *28*, 1919–1920. https://doi.org/10.1093/bioinformatics/bts277

Nevado, B., Koblmüller, S., Sturmbauer, C., Snoeks, J., Usano Alemany, J., & Verheyen, E. (2009). Complete mitochondrial DNA replacement in a Lake Tanganyika cichlid fish. *Molecular Ecology*, *18*, 4240–4255. https://doi.org/10.1111/j.1365-294X.2009.04348.x

Nguyen, T. H. M., Sondhi, S., Ziesel, A., Paliwal, S., & Fiumera, H. L. (2020). Mitochondrial-nuclear coadaptation revealed through mtDNA replacements in *Saccharomyces cerevisiae*. *BMC Evolutionary Biology*, *20*(128), 1000. https://doi.org/10.1186/s12862-020-01685-6

Noskova, E., Ulyantsev, V., Koepfli, K.-P., O'Brien, S. J., & Dobrynin, P. (2020). GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum dat. *GigaScience*, *9*, giaa005.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., & Sham, P. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*, 559–575.

Quinlan, A. R., & Hall, I. M. (2010). BEDTOOLS: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Quinzin, M. C., & Mardulyn, P. (2014). Multi-locus DNA sequence variation in a complex of four leaf beetle species with parapatric distributions: Mitochondrial and nuclear introgressions reveal recent hybridization. *Molecular Phylogenetics and Evolution*, *78*, 14–24. https://doi.org/10.1016/j.ympev.2014.05.003

Quinzin, M. C., Normand, S., Dellicour, S., Svenning, J. C., & Mardulyn, P. (2017). Glacial survival of trophically linked boreal species in northern Europe. *Proceedings - Royal Society. Biological Sciences*, *284*, 20162799. https://doi.org/10.1098/rspb.2016.2799

Raj, A., Stephens, M., & Pritchard, J. K. (2014). FASTSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*, 1297–1303. https://doi.org/10.1534/genetics.114.164350

Rank, N. E., Mardulyn, P., Heidl, S. J., Roberts, K. T., Zavala, N. A., Smiley, J. T., & Dahlhoff, E. P. (2020). Mitonuclear mismatch alters performance and reproductive success in naturally introgressed populations of a montane leaf beetle. *Evolution*, *74*, 1724–1740. https://doi.org/10.1111/evo.13962

Rosenzweig, B. K., Pease, J. B., Besansky, N. J., & Hahn, M. W. (2016). Powerful methods for detecting introgressed regions from population genomic data. *Molecular Ecology*, *25*, 2387–2397.

Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, *17*, 155–158. https://doi.org/10.1038/s41592-019-0669-3

Schrider, D. R., Ayroles, J., Matute, D. R., & Kern, A. D. (2018). Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLoS Genetics*, *17*, e1007341.

Seixas, F. A., Boursot, P., & Melo-Ferreira, J. (2018). The genomic impact of historical hybridization with massive mitochondrial DNA introgression. *Genome Biology*, *19*, 91. https://doi.org/10.1186/s13059-018-1471-8

Sequeira, F., Sodré, D., Ferrand, N., Bernardi, J. A., Sampaio, I., Schneider, H., & Vallinoto, M. (2011). Hybridization and massive mtDNA unidirectional introgression between the closely related Neotropical toads *Rhinella marina* and *R. schneideri* inferred from mtDNA and nuclear markers. *BMC Evolutionary Biology*, *11*, 264.

Sloan, D. B., Havird, J. C., & Sharbrough, J. (2017). The on-again, off-again relationship between mitochondrial genomes and species boundaries. *Molecular Ecology*, *26*, 2212–2236. https://doi.org/10.1111/mec.13959

Smit, A. F. A., & Hubley R. (2008–2015). REPEATMODELER open-1.0. [Cited 11 March 2021]. Retrieved from http://www.repeatmasker.org

Smit, A. F. A., & Hubley R. (2013–2015). REPEATMASKER open-4.0. [Cited 11 March 2021]. Retrieved from http://www.repeatmasker.org

Suarez-Gonzalez, A., Hefer, C. A., Lexer, C., Cronk, Q. C. B., & Douglas, C. J. (2018). Scale and direction of adaptive introgression between black cottonwood (*Populus trichocarpa*) and balsam poplar (*P. balsamifera*). *Molecular Ecology*, *27*, 1667–1680.

Takami, Y., Nagata, N., Sasabe, M., & Sota, T. (2007). Asymmetry in reproductive isolation and its effect on directional mitochondrial introgression in the parapatric ground beetles *Carabus yamato* and *C. albrechti*. *Population Ecology*, *49*, 337–346. https://doi.org/10.1007/s10144-007-0052-6

Toews, D. P. L., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, *21*, 3907–3930. https://doi.org/10.1111/j.1365-294X.2012.05664.x

Vafai, S. B., & Mootha, V. K. (2012). Mitochondrial disorders as windows into an ancient organelle. *Nature*, *491*, 374–383. https://doi.org/10.1038/nature11707

Valencia-Montoya, W. A., Elfekih, S., North, H. L., Meier, J. I., Warren, I. A., Tay, W. T., Gordon, K. H. J., Specht, A., Paula-Moraes, S. V., Rane, R., Walsh, T. K., & Jiggins, C. D. (2020). Adaptive introgression across semipermeable species boundaries between local *Helicoverpa zea* and invasive *Helicoverpa armigera* moths. *Molecular Biology and Evolution*, *37*, 2568–2583. https://doi.org/10.1093/molbev/msaa108

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*, 11.10.1–11.10.33.

Zieliński, P., Nadachowska-Brzyska, K., Wielstra, B., Szkotak, R., Covaciu-Marcov, S. D., Cogălniceanu, D., & Babik, W. (2013). No evidence for nuclear introgression despite complete mtDNA replacement in the Carpathian newt (*Lissotriton montandoni*). *Molecular Ecology*, *722*, 1884–1903.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Lukicheva, S., & Mardulyn, P. (2021). Whole-genome sequencing reveals asymmetric introgression between two sister species of cold-resistant leaf beetles. *Molecular Ecology*, 00, 1–13. https://doi.org/10.1111/mec.16011