Molecular Ecology Resources

WILEY

MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE OPEN ACCESS

Optimization and Evaluation of the bestRAD Sequencing Approach: Towards Ascertainment of the Invasion Routes of the Oriental Fruit Fly, *Bactrocera dorsalis*

Emeline Charbonnel[1,2,3] | Laure Benoit[1,2] | Sabine Nidelet[4] | Enrique Ortega-Abboud[5] | Bernhard Gschloessl[4,6] | Raphaël Leblois[4] | David Ouvrard[3] | Marie-Pierre Chapuis[1,2]

[1]CBGP, CIRAD, INRAE, Institut Agro, IRD, Univ Montpellier, Montpellier, France | [2]CIRAD, CBGP, Montpellier, France | [3]Plant Health Laboratory, Entomology and Botany Unit, ANSES, Montferrier-sur-Lez Cedex, France | [4]CBGP, INRAE, CIRAD, Institut Agro, IRD, Univ Montpellier, Montpellier, France | [5]CBGP, IRD, CIRAD, Institut Agro, IRD, Univ Montpellier, Montpellier, France | [6]MISTEA, Université de Montpellier, INRAE, Institut Agro, Montpellier, France

Correspondence: Marie-Pierre Chapuis (marie-pierre.chapuis@cirad.fr)

Received: 29 May 2024 | **Revised:** 31 January 2025 | **Accepted:** 1 April 2025

Handling Editor: Samridhi Chaturvedi

Funding: This work was supported by Agence Nationale de la Recherche, DISLAND (ANR-20-CE32-0012); Centre de Coopération Internationale en Recherche Agronomique pour le Développement, BACTRACK (Doctoral Fellowship); Labex CeMEB (Centre Méditerranéen de l'Environnement et de la Biodiversité), PROLAG (Exploratory Research Project) Agence Nationale de Sécurité Sanitaire de l'Alimentation, de l'Environnement et du Travail, BACTRACK (Doctoral Fellowship), ISOGEO (Project of Strategic Interest).

Keywords: genotyping error rate | high-throughput sequencing | invasion | pest | phylogeography | RAD

ABSTRACT

The bestRAD technique is a reduced genome representation approach with high-capacity sample multiplexing and physical isolation of biotin-labelled target DNA fragments using streptavidin beads, which should reduce total cost and genotyping errors. While we here formalise the relevance of this approach within the HTS landscape, our foremost aim was to improve its replicability, validity, and transparency. We first optimised the molecular laboratory protocol and shared the associated protocols (e.g., final detailed methodologies, quality control, best practices) under the FAIR principles. Using 84 worldwide individual samples of the Oriental fruit fly, *Bactrocera dorsalis*, a major invasive pest, we revealed a low rate of PCR duplicates, robustness to DNA quality and quantity, high genotype call rate, insignificant genotyping error rate, high nuclear and mitochondrial genome representativeness, and a high level of genetic information. This in-depth data quality assessment, along with total cost and handling time reduced by an estimated one-third relative to the parent RAD-Seq version, demonstrates that bestRAD is an excellent compromise between cost and quality. While we generated high-quality genomic resources for *B. dorsalis*, we also share details and recommendations for the bestRAD technique that can be readily used in any laboratory and applied to all organisms, even without published genome sequence.

1 | Introduction

Latest advances in high-throughput sequencing (HTS) have recently increased genotype calling and accuracy in ecological and evolutionary studies. Despite the continued decline in HTS cost and increased availability in high-quality reference genomes, whole-genome sequencing (WGS) may still often be unaffordable, especially in species with large genomes or in studies based on a large number of individual samples. This is often mitigated using a Pool-Seq strategy, in which genomic

DNA from a large number of individuals of the same population is pooled in equimolar proportions and sequenced together (Hivert et al. 2018). However, many studies require individual knowledge (e.g., in landscape genetics, of introgression and hybridization, etc.) or cannot rely on collecting enough individuals from each population (e.g., conservation genetics). In such contexts, a valuable alternative relies on reduced-representation sequencing (RRS), often based on the enzymatic digestion of the genome to focus the sequencing effort on small regions surrounding restriction sites rather than on all randomly sheared genomic fragments (i.e., restriction-site associated nuclear DNA sequencing or RAD-Seq; Baird et al. 2008) (Matheson and McGaughran 2022; North et al. 2021; Reid et al. 2021). The different RRS versions (Campbell et al. 2018) still allow the genotyping of tens of thousands of single-nucleotide polymorphisms (SNPs) spread along the genome, and this strategy has been increasingly used over the past decade to study a wide range of questions, including gene flow or hybridization, demographic history, and local adaptation (Andrews and Luikart 2014; North et al. 2021).

However, beyond the reduction in information content compared with WGS (e.g., fewer markers, less linkage disequilibrium (LD) information), RAD-Seq potential may be limited by specific biases that reduce genotyping calling and accuracy (Andrews et al. 2016; Bresadola et al. 2020; Cariou et al. 2013; Gautier et al. 2013). In particular, mutations at the restriction enzyme cut sites result in some individuals not being genotyped at particular loci or having a homozygous genotype called when the individual is actually heterozygous (Davey et al. 2013), which is known as allelic dropout. These false homozygous calls result in biased genetic diversity estimation and evolutionary inference (Arnold et al. 2013). This bias becomes a problem only when effective population size is very large (Andrews et al. 2016; Cariou et al. 2013; Gautier et al. 2013) or when evolutionarily distant individuals are analysed (e.g., in species-complexes). This bias is also limited when hard filters are applied to retain loci with high coverage across individuals and with allele frequencies above a cut-off (Andrews et al. 2016). Note that an insufficient number of target DNA fragments in libraries can also entail dramatic allelic dropout, or even locus dropout, i.e., when neither allele is sampled (Andrews et al. 2016).

Furthermore, library preparation almost always includes a final enrichment of targeted sequences based on polymerase chain reaction (PCR) amplification. The stochasticity with which each original DNA molecule is amplified in the first cycles of PCR leads to uneven copy numbers (Kebschull and Zador 2015). Such imbalanced representation of the template DNA molecules may create genotyping errors, such as false negative allele calls in heterozygous genotypes (Casbon et al. 2011). To offset the distortion of the sequence representation due to PCR stochasticity, it is customary to remove a posteriori all multiple copies of the same template DNA molecule, called PCR duplicates, when they are traceable. This common denoising step makes it possible to quantify the number of distinct input DNA molecules before amplification that cover each SNP site (herein referred to as non-redundant depth of coverage). Rates of PCR duplicates vary greatly across RRS studies, with high frequencies in many cases (e.g., 20%–95% in RAD-Seq studies; Andrews et al. 2014; Euclide et al. 2020; Rochette et al. 2023; Schweyen et al. 2014). This is partly expected because library complexity (i.e., the amount of the input DNA that is amplifiable) is lower in RRS (Fu et al. 2018; Rochette et al. 2023), but such high rates of PCR duplicates can increase severalfold the cost of sequencing and may even result in insufficient non-redundant depth of coverage for reliable genotype calling (Cristofari et al. 2016).

In the version of RAD-Seq developed by Ali et al. (2016), referred to as bestRAD in the literature (e.g., Rochette et al. 2019), digested DNA samples are ligated to a biotinylated tag in order to be enriched in fragments carrying the restriction site by selection on streptavidin beads. This fragment isolation step is particularly useful for increasing library complexity and is expected to considerably reduce the rate of PCR duplicates (e.g., 23%–44%; Ali et al. 2016; Rochette et al. 2023). This strategy also allows for early and efficient multiplexing of samples, which significantly reduces the workload and cost when working on a large number of individual samples. Although this technique is increasingly used to produce RRS data in population genetics and phylogeographic studies, it has mainly been used on species from the same taxonomic group (i.e., vertebrates, mostly fish species, in 85% of studies) and by authors from the same country as the seminal manuscript (i.e., 95% of studies with at least one US affiliation; see Table S1 for further details). One possible explanation is the want of precise information concerning each step (e.g., volumes, quantities, durations, materials with supplier references, expectations and recommendations), hampering any easy transfer to other laboratories and biological models. This also precludes assessing reproducibility across libraries and/or laboratories and robustness to technical variability (e.g., quality and quantity of input DNA). During the first application in our laboratory, we obtained low quality libraries, i.e., a large number of non-target DNA fragments (adapter residues) and a low overall concentration.

In this study, we first optimised a molecular laboratory protocol of the bestRAD technique meeting the FAIR criteria (Findable, Accessible, Interoperable, Reusable) and easily applicable to any organism from which sufficient DNA can be purified and by any laboratory with experience in HTS. In order to assess the quality of HTS data produced using this protocol, we then applied it to 84 individual samples of the Oriental fruit fly, *Bactrocera dorsalis* (Hendel) (Diptera, Tephritidae), representative of the distribution range of the species. Over the last two decades, *B. dorsalis* has emerged as one of the most invasive and destructive insect pests of tropical and subtropical fruits and vegetables, particularly in sub-Saharan Africa and the Indian Ocean (Drew et al. 2005; Schutze et al. 2014). In this species, an individual approach is preferable to take into account the possibility of misidentification or introgressive hybridization with closely related morphologically cryptic species (reviewed in Charbonnel et al. 2023). We took advantage of available specific nuclear and mitochondrial genomes in this species and used a rigorous bioinformatics methodology following Bresadola et al. (2020), Díaz- Arce and Rodríguez- Ezpeleta (2019), Gautier et al. (2013), Graham et al. (2020), Rivera-Colón et al. (2021) and Vaux et al. (2022). Mining RAD sequences for mitochondrial loci is an efficient strategy implying no additional sequencing effort (reviewed in

Laczkó et al. 2022), and potentially providing additional phylogeographic insights (e.g., hybridization) thanks to the distinctive characteristics of mitochondrial DNA (i.e., haploidy, absence of recombination, and maternal heredity; Hickerson et al. 2010; Wilson et al. 1985).

We described the quality of the *B. dorsalis* dataset by reporting the rate of PCR duplicates, assessing the impact of the amount and degradation of input DNA, estimating the calling and genotyping error rates through the use of independent replicates of RAD-Seq libraries for several individuals from remote localities (in order to account for systematic biases that may affect some replicates similarly, such as polymorphisms in restriction cut sites or unequal PCR amplification rates of alleles; Bresadola et al. 2020), and evaluating the performance in representing the nuclear and mitochondrial genomes and in uncovering the geographic structure and evolutionary history of the species. As regards evaluating performance, we carried out analyses of genetic differentiation (e.g., phylogenetic tree, Bayesian clustering, multidimensional scaling) and diversity (e.g., LD, private alleles) and confronted our results with those obtained in studies that have addressed the genetic structuring of the species across a large portion of its geographic range. Even though phylogeographic studies of *B. dorsalis* until recently have mainly exploited mitochondrial genetic variation (e.g., Garzón-Orduña et al. 2019; San Jose et al. 2018), a handful have been conducted on a few independent microsatellite markers (e.g., Khamis et al. 2009; Kim et al. 2021; Qin et al. 2018). Recently, two studies have focused on genome-wide SNPs (Deschepper et al. 2023; Zhang et al. 2023) using WGS, thus making it possible, by comparison with our results, to assess the trade-off between data production cost and relevance of the genetic information retrieved for population structure analyses at the global scale.

2 | Materials and Methods

2.1 | DNA Samples

A total of 84 adult *Bactrocera dorsalis* specimens were collected from 2007 to 2021, either from orchards using ME traps or from infested fruits intercepted at EU border control points. All samples had been previously analysed by Charbonnel et al. (2023), who provided sampling details in their Table S1 and confirmed for each specimen its species identification as obtained from morphological diagnosis (Drew and Romig 2013, 2016; White and Elson-Harris 1992) and molecular diagnosis based on a mitochondrial marker (COI) and two nuclear markers (ITS and EIF3L). By following Clarke et al. (2019) in considering the whole Indo-Malayan (or Oriental) biogeographic region as the native range, and thereby the Afrotropical, Australasian and Malagasy biogeographic regions as the invaded range, we can consider that samples were representative of their respective ranges (i.e., 34 specimens distributed over 13 countries and 50 specimens distributed over 22 countries, respectively) (Figure 1A). Genomic DNA from the 84 samples was extracted on the full specimen using a non-destructive protocol, i.e., without damaging the external features essential for morphological identification, as described in Charbonnel et al. (2023). DNA quantity was measured using a Qubit dsDNA HS Assay kit with a Qubit 2.0 fluorometer and DNA quality was evaluated using agarose gel electrophoresis (Table S2).

2.2 | bestRAD Optimisation

We developed an improved and reproducible laboratory protocol of the bestRAD sequencing approach of Ali et al. (2016) in order to produce high-quality libraries, based not only on a large number of target DNA fragments but also on low amounts of unexpected residues and few artefact sequences. To achieve this, we tested and validated the following critical parameters: the quality and quantity of the DNA input, the ratio of the AMpure purifications, and the number of PCR cycles of the final amplification. As required by the FAIR principles, each step of the final laboratory protocol is described on the *protocols.io* platform (Benoit et al. 2024) with special care to ensure precision (i.e., volumes, quantities, durations, materials with supplier references), along with quality controls, recommendations, and illustrations.

Each individual DNA sample was digested with the PstI restriction enzyme and ligated to a unique biotinylated adapter (the individual barcode). All barcoded DNAs were pooled in batches of 20 or 24 and mechanically fragmented at 200–500 bp by sonication using a Bioruptor Pico (Diagenode). Fragments carrying the restriction site were selected using Dynabeads M-280 streptavidin magnetic beads (Invitrogen). Finally, the pools of samples were used as inputs for the NEB Next Ultra II DNA Library Preparation Kit (New England Biolabs) to be ligated to two homemade adapters (the plate barcodes) during the final PCR amplification. Following tests to determine optimal values, for this step of library construction we used: (i) 0.65× as the ratio of the two AMpure purifications, which determine the rate of free adapter residues and adapter doublets, and (ii) nine as the number of PCR cycles of the final amplification, which determine the quantity of target DNA fragments but also the rate of chimeric fragments formed by two or more biological sequences, along with the rate of PCR duplicates.

The final libraries were normalised to 10 μM, pooled, and sequenced (150 bp paired-end) on an Illumina NovaSeq 6000 sequencing system in a single run (S1 flow-cell), except for 3 individuals (SP flow-cell) (MGX-Montpellier GenomiX platform). We expected to achieve around 25X in non-redundant depth of coverage, based on the prediction of 92,922 restriction sites (from the *B. dorsalis* genome GCA_020283865.1; Jiang et al. 2022), a low rate of PCR duplicates (i.e., around 30%), and Illumina specifications for the total number of reads.

2.3 | Bioinformatics Methodology

2.3.1 | Read Denoising

Raw sequence data were demultiplexed, first by pool of individual samples, using *cutadapt* (Martin 2011) and allowing one substitution on the plate barcode, and secondly by individual sample using *process_radtags* in Stacks2 (Rochette et al. 2019). The *process_radtags* program was also used

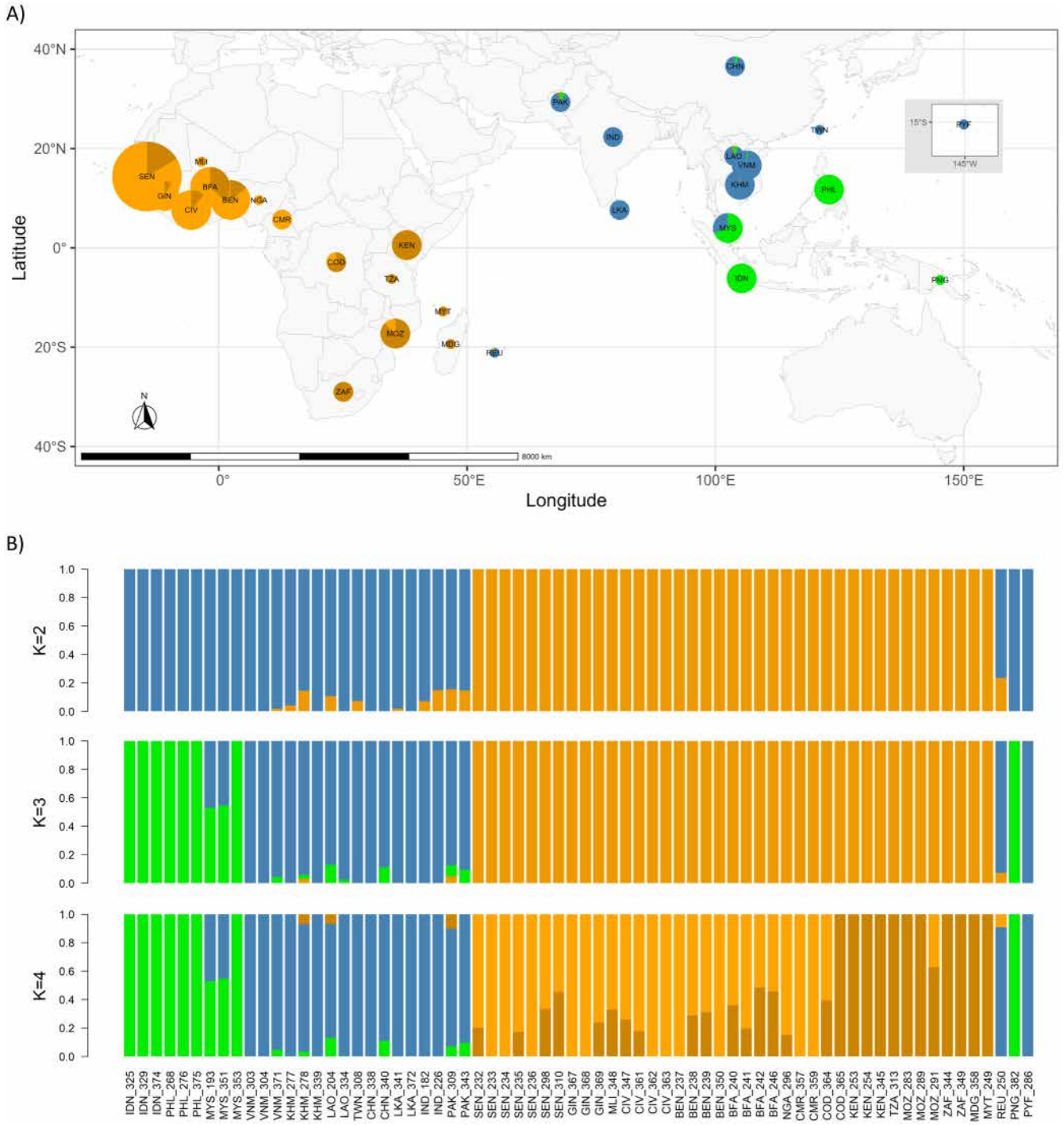

FIGURE 1 | Country-based localities of the 68 individual samples of *Bactrocera dorsalis* with their ADMIXTURE individual ancestry coefficients. We used the method of Alexander et al. (2009) on the 7168 hard-filtered and independent SNPs. (A) Each colour corresponds to one of the four genetic clusters as shown in (B). Size of dot is proportional to sampling size per country. (B) Each individual is represented by a vertical line partitioned into $K = 2$, 3 or 4 coloured segments representing the estimated fractions of its genome pertaining to each of the K clusters. Individuals are placed along the x-axis according to their sampling locality, from East to West in continental Asia and West to East in Africa.

to retain paired reads of high quality only (Phred quality score ≥ 30, and removal of any read with an uncalled base). Finally, we trimmed forward reads of the restriction site using Trimmomatic (Bolger et al. 2014), removing the first 5 nucleic bases, and used *clone_filter* in Stacks2 (Rochette et al. 2019) to identify PCR duplicates based on paired-end read sequence identity.

2.3.2 | Loci Identification

We used the *B. dorsalis* nuclear genome (Jiang et al. 2022) for aligning the passing-filtered reads with a reference using the Burrows-Wheeler Aligner (BWA-mem2; Vasimuddin et al. 2019) with default parameters, and Samtools v1.14 (Li et al. 2009) was applied to retain pairwise primary alignments

only. Repeat elements in genome sequences were masked beforehand with the software RepeatModeler2 (Flynn et al. 2020; v2.0.2a), using NCBI blastn as the search engine and subsequently providing the generated repeat library as an input for RepeatMasker (Tarailo-Graovac and Chen 2009; v4.1.1). Loci were identified using the Marukilow model in the *gstacks* program of Stacks2, using the option *–rm-pcr-duplicates* to remove remaining PCR duplicates based on mapping coordinates of the read pairs. The loci identification analysis identified 15 samples with a mean depth below 6X, all of which had more than 50% missing data (Figure S1). These 15 poor-quality samples were excluded, and we reiterated the gstacks analysis on the 69 remaining samples.

2.3.3 | SNP Calling and Hard-Filtering

Biallelic SNPs were called using the *populations* program in Stacks2 (Rochette et al. 2019) and those that failed to reach a minimum mean depth of 6X and a minimum minor allele frequency (MAF) of 0.05 (equal to a count of 7 alleles) were removed using the *vcftools* program (Danecek et al. 2011). From here, hard-filtering of SNPs (and consequently of loci that no longer contain SNPs) was carried out with R version 4.2.1 (R Core Team 2022), using the *vcfR* package (Knaus and Grünwald 2017), the *SNPfiltR* package (DeRaad 2022; Knaus and Grünwald 2017) and the *HDplot* program (McKinney et al. 2017). First, we filtered out SNPs with evidence of putative paralogs, i.e., excessively high mean depth over all samples (≥ 40; Figure S2A), excessively high observed heterozygosity (> 0.625; Figure S2B) and/or deviation from even read ratios in heterozygous genotypes (> 5; Figure S2B). The last criterion can also help to remove chimeric sequences with an incorrect combination of barcodes created from parental sequences in samples from the same pool during library PCR enrichment (Schnell et al. 2015). We then removed genotypes with quality < 30 and/or depth < 6 and subsequently, SNP sites with calling rate $< 80\%$ (and MAF < 0.05) and samples with calling rate $< 50\%$.

2.4 | Data Quality Assessment

2.4.1 | Robustness to DNA Quality

As there were 13 DNA samples with a low molecular weight (Table S2), we investigated the impact of DNA quality on data denoising and loci identification. To this end, we created a binary explanatory variable (low vs. high DNA quality, for a molecular weight predominantly below vs. above 5 kb), and tested for its effect on various sample statistics using R version 4.2.1 (R Core Team 2022): (i) read and loci counts, using a negative binomial generalised linear model, with the function *glm.nb* from the *MASS* package (Venables and Ripley 2002), (ii) mapping rate, PCR duplicates proportions, and missing rate, using a Beta regression model analysis using the *betareg* package (Cribari-Neto and Zeileis 2010), and (iii) mean depth, using an analysis of variance model, with normality verified by a Shapiro–Wilk test applied on residuals, with the *stats* package. Both former tests were followed by the application of an analysis of variance, using the *Anova* function from the *car* package (Fox and Weisberg 2018), to test for factor significance. We excluded from this analysis 5 samples that displayed both low mapping success and high prevalence of contaminant bacterial DNA, in addition to replicate samples and the three samples processed in a different sequencing run.

2.4.2 | Genotype Calling and Accuracy

In order to assess genotyping uncertainty, we replicated libraries for five samples, one of which was finally excluded from analysis for poor quality (see results and details in Table S2). To achieve this, we explored the parameter space for SNP hard-filtering, namely the minimum genotype quality (from 10 to 40), the minimum genotype depth (from 0 to 15×) and the maximum proportion of missing genotypes per SNP (from 10% to 100%), the maximum proportion of missing data per sample being set at 50%. We then used the *poppr* package (Kamvar et al. 2014) to estimate the calling error rate, i.e., the complement of the frequency with which a genotype in a sample is also called in its replicate, and the per-allele genotyping error rate, i.e., the frequency with which a genotype in a sample differs from the genotype of its replicate, weighted by two when genotypes are homozygous for alternative alleles and by one when one genotype is heterozygous and the other homozygous. Finally, we assessed the effects of the hard-filtering parameters on the proportion of SNPs and samples retained as well as on the calling rate in our dataset of 69 individual samples.

2.5 | *B. dorsalis* Worldwide Structuring Based on Nuclear SNPs

2.5.1 | Phylogenetic Tree Based on Nuclear Loci

We produced a PHYLIP file of the sequence concatenation of all the hard-filtered loci, using the option—*whitelist* in the *populations* program in *Stacks2* (Rochette et al. 2019). In compliance with the IUPAC ambiguity codes for heterozygous genotypes, variants were called only if they showed a calling rate $> 80\%$ and a MAF ≥ 0.05. The substitution model that best fitted the data was identified using ModelFinder as implemented in IQ-TREE 2.2.2.6 (Minh et al. 2020) based on BIC. We then ran ML searches with 1000 ultrafast bootstraps (Hoang et al. 2018) using the selected model. In order to explore how sensitive the RAD topology was to model specification, we also implemented a partitioned data analysis, giving as predefined partitions the start-stop positions for each sequential locus. Trees were plotted with iTOL (Letunic and Bork 2007).

2.5.2 | SNP Unlinking

For population genetics analyses, we produced a reduced set of independent hard-filtered SNPs using two standard approaches to account for the effects of linkage. First, we pruned SNPs in high LD with each other, especially since it can concern certain regions of the genome, sometimes of considerable length. For this, we considered a large chromosomal window (1 Mb) at a time, and removed a SNP from pairs whose genotypes had a correlation coefficient greater than 0.3, using the function *snpgdsLDpruning* of the *SNPRelate* package (Zheng et al. 2012).

Secondly, we thinned SNPs within a short physical distance (2 kb), using the function *distance_thin* of the *SNPfiltR* package (DeRaad 2022). We verified random genotypic associations at different SNPs above this physical distance by investigating the decay LD across all SNPs. For this, we used the package *PopLDdecay* version 3.30 (Zhang et al. 2019), the r^2 measure (Hill and Robertson 1968) and a maximal distance of 1 Mb. In addition, we used the *pcadapt* package (Luu et al. 2017), a tool for outlier detection based on principal component analysis (PCA), in order to remove the unlinked and hard-filtered SNPs that were outliers in the first two principal component dimensions. *p*-values were transformed into *q*-values using the *q*value package (Storey et al. 2023), with a specified FDR of 0.05 as recommended by Luu et al. (2017). The BED-formatted files needed for *SNPRelate, PopLDdecay*, and *pcadapt* packages were produced using PLINK 2.0 (Chang et al. 2015).

2.5.3 | Spatial Genetic Variation

Population genetics analyses were performed with R version 4.2.1 (R Core Team 2022). The VCF file of independent and hard-filtered SNPs was converted beforehand into either objects of *genind* or *genlight* class with the *vcfR* package (Knaus and Grünwald 2017), or into a *genepop* file with the *graph4lg* package (Savary et al. 2021). We started by exploring spatial genetic structuring visually with a principal component analysis (PCA), using the *glPca* function in the *adegenet* package (Jombart and Ahmed 2011). We then used the maximum likelihood model-based ADMIXTURE method v.1.3 (Alexander et al. 2009). The method first estimates the number of genetic clusters *K*, i.e., at Hardy–Weinberg and linkage equilibria, and then, for each individual, the proportions of their genome derived from each genetic cluster (ancestry coefficients). We explored successive values of *K* from 1 to 15 and performed a 5-fold cross-validation for each *K* value to determine which model had the best predictive accuracy. Because the performance of methods to resolve the correct number of genetic clusters can be poor and hierarchical population structure often leads to underestimation (Cullingham et al. 2020; Janes et al. 2017; Kalinowski 2011; Lawson et al. 2018), we also explored *K* values above the optimal value.

Within each genetic cluster (but excluding individual samples from invaded islands: Madagascar, Mayotte, La Réunion, French Polynesia), we calculated the expected heterozygosity (*He*) and inbreeding coefficient (*F*is) using *genepop* (Rousset 2008) and the allelic richness using *PopGenReport* (Adamack and Gruber 2014). In addition to these summary statistics of genetic diversity, we constructed a Venn diagram of private and shared alleles and investigated the LD decay within genetic clusters using the same methodology as described above. Finally, we tested for isolation by distance within genetic clusters, only when the subset of individual samples with sampling coordinates (i.e., orchard captures) was sufficient. To this end, we calculated between pairs of individuals the Euclidian geographic distance using the *sf* package (Pebesma 2018) and the genetic differentiation estimator a_r (Rousset 2000) using the package *genepop* (Rousset 2008). We tested for significance of the correlation between matrices of genetic differentiation and of the natural logarithm of geographic distance using a Mantel test (9999 permutations) with the *ade4* package (Dray and Dufour 2007).

2.6 | *B. dorsalis* Worldwide Structuring Based on Mitochondrial SNPs

With the aim to identify mitochondrial variants in the passing-filtered sequence reads of the 69 individual samples retained for analysing nuclear structure and diversity, we used the same procedure for read alignment as presented above except that we used the mitochondrial genome of *B. dorsalis* NC_008748.1 (15,915b; Yu et al. 2007) as reference. We produced a PHYLIP file of the sequence concatenation of the six mirrored loci located on either side of the three restriction sites of the PstI enzyme identified in the reference (Figure S3) using the option—*whitelist* in the *populations* program in *Stacks2* (Rochette et al. 2019). Variants were called only if they showed a calling rate ≥ 80%, a MAF > 0.01 (i.e., a minimum of 2 counts) and no heterozygous genotype in order to reflect the haploid nature of mitochondrial DNA and to avoid nuclear mitochondrial pseudogenes (although this procedure can also remove variants that represent mitochondrial heteroplasmy; Laczkó et al. 2022). Sequences were manually curated using Bioedit (Hall 1999) to exclude duplicated regions arising from an overlap between the forward read and the reverse read of loci originating from close restriction sites (Stobie et al. 2019; see also Figure S3), and exported in a FASTA format.

We reconstructed a phylogenetic tree following the same procedure as above, but without the partitioned data analysis, and visualised the spatial structuring by constructing a haplotype network using the TCS algorithm (Clement et al. 2002) implemented in *Popart* v.1.7 (Leigh and Bryant 2015). Within the whole mitochondrial dataset as well as each genetic cluster defined by nuclear SNPs, we estimated the number of unique haplotypes, the haplotypic and nucleotidic diversities using the *pegas* package (Paradis 2010), and the number of segregating sites using the *ape* package (Paradis and Schliep 2019).

3 | Results

3.1 | Data Quality Assessment

3.1.1 | Robustness to DNA Quality

In total, sample demultiplexing yielded 1,185,316,276 paired reads with a high-quality score, without a significant effect of the genomic DNA quality on the individual paired reads number (p-value = 0.366; Figure 2A). Conversely, the rates of PCR duplicates detected on the basis of paired-end read sequence identity were significantly inflated by poor quality of the source DNA (p-value < 0.001; Figure 2B) and averaged 18% across samples (Table S2). The averaged mapping rate of 56% was explained by high repeat content in the *B. dorsalis* nuclear genome, without a significant effect of the quality of the genomic DNA on sample values (p-value = 0.253; Figure 2C). We reached an estimate of 49% of repetitive elements (Table S3), in congruence with previous reports for *B. dorsalis* (46%; Jiang et al. 2022) or congeneric genomes (35% in *B. oleae*, Bayega et al. 2020; 31% in *B. tryoni*, Gilchrist et al. 2014). The rates of remaining PCR duplicates

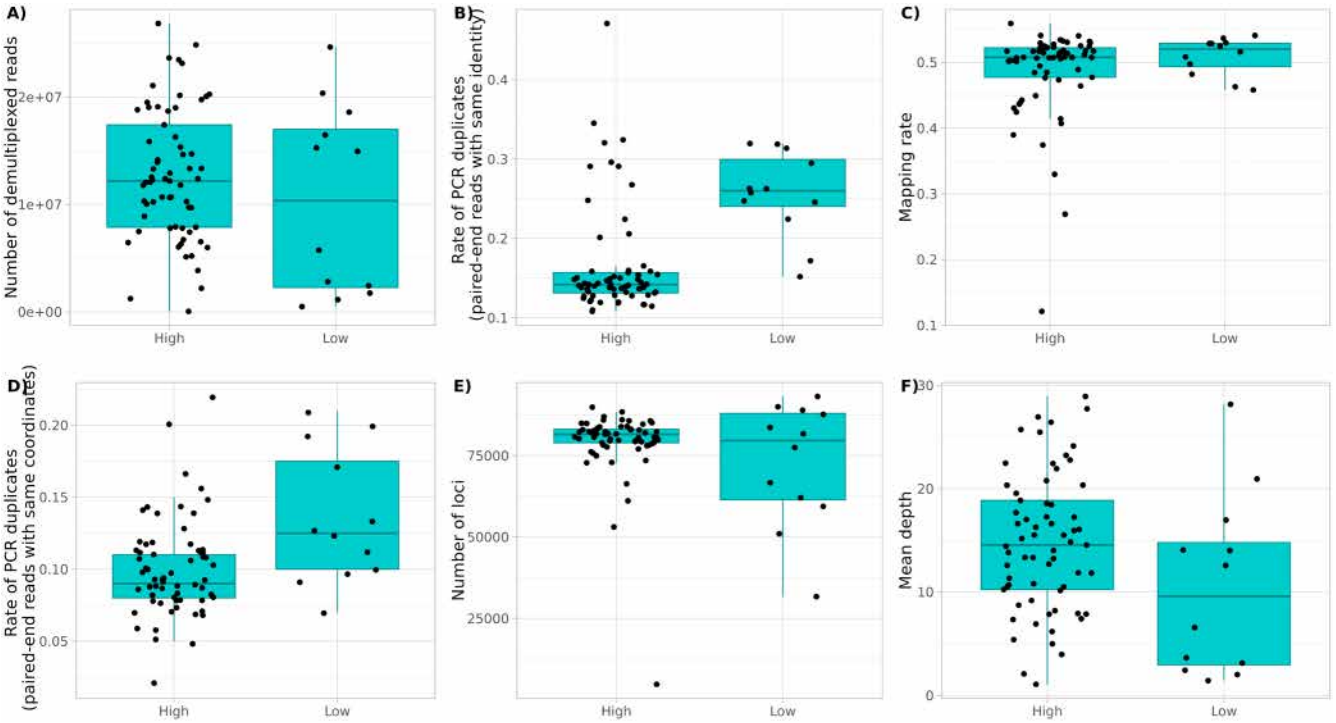

FIGURE 2 | Effects of DNA quality on read denoising and loci identification statistics. DNA quality is represented on *x*-axes under two categories: High quality (molecular weight above 5 kb), and low quality (molecular weight below 5 kb). A boxplot and dots for each DNA sample are represented on *y*-axes. Outputs came from (A) the *process_radtags* program in Stacks2, (B) the *clone_filter* program in Stacks2, (C) BWA-mem2 and (D–F) the *gstacks* program in Stacks2 using the option *–rm-pcr-duplicates*. We excluded from this analysis the 5 samples for which low mapping success was associated with high prevalence of contaminant bacterial DNA, the 5 technical replicates and the 3 samples processed in a different sequencing run (see Table S2 and main text for further details). Missingness is not represented since it was highly correlated to mean read depth.

detected, based on paired-end read sequence coordinates, were still inflated by poor quality of the source DNA (p-value = 0.003; Figure 2D) and averaged 10% across samples. DNA quality did not affect the number of loci identified on the passing-filter reads (p-value = 0.303; Figure 2E) but lowered their quality with marginal significance (i.e., non-redundant depth: p-value = 0.049; missing data: p-value = 0.019; Figure 2F). This negative effect disappeared when excluding the DNA samples that did not pass the filters (i.e., non-redundant depth < 6× and missing data > 50%), which confirms the relevance of excluding them for subsequent analyses.

3.1.2 | Genotype Calling and Accuracy

The SNP hard-filtering procedure (i.e., mean SNP depth ≥ 6×, MAF ≥ 0.05, no evidence for putative paralogs, calling rate > 80% after genotypes with a quality < 30 and/or a depth < 6 were removed) allowed us to retain 34,725 loci and excluded the ZAF_346 sample only, with a calling rate < 50%, for a final set of 68 samples. Thirty percent of SNPs were retained by the hard-filtering procedure (i.e., 265,154), with a mean non-redundant depth of 23 and a missingness of 9% (Table S4). Calling and genotyping error rates were estimated from our four technical replicates at 3.35% and 0.025% respectively. Over the parameter space for SNP hard-filtering, the calling error rate exceeded 10% only when the filter on the proportion of missing genotypes tolerated per SNP was loose (≥ 50%; Figure 3A). The calling error rate dropped to a minimal value of about 1% as soon as the filter on the maximum proportion of missing genotypes per SNP was set to 10% (Figure 3A), regardless of the values of other parameters, but at the cost of a smaller dataset due to SNPs being heavily filtered out (Figure S4). Allowing up to 20% of missing genotypes seems to be a good compromise for maximising genotype calling (i.e., error rate < 5%) while retaining a sufficient number of SNPs (i.e., several thousands, the exact number depending on the values of other parameters, Figure 3A and Figure S4). Genotyping uncertainty was low over the whole parameter space, with a maximal value of 0.25% (Figure 3B). The allele-specific genotyping error rate quickly dropped to minimal values as soon as genotype quality was above 30 or depth above 9, or for lower values when considering both genotype filters in combination (Figure 3B).

3.2 | Phylogenetic Tree Based on Nuclear Loci

The sequence concatenation of all the hard-filtered loci was 25,893,395b long and included 457,435 informative sites with a MAF > 0.05 and shared among at least 80% of the 68 samples, for a mean non-redundant depth of 15 and an overall missingness of 6%. All ML phylogenetic inferences yielded well-resolved trees, including most nodes with 100% bootstrap support, and recovered the same groups and relationships among individuals. Two major geographic clades were recognised (Figure 4A): (1) individuals of (pen)insular Asia, clearly separated according to their country of origin (Malaysia, the Philippines and Indonesia), together with the New Guinean sample grouped

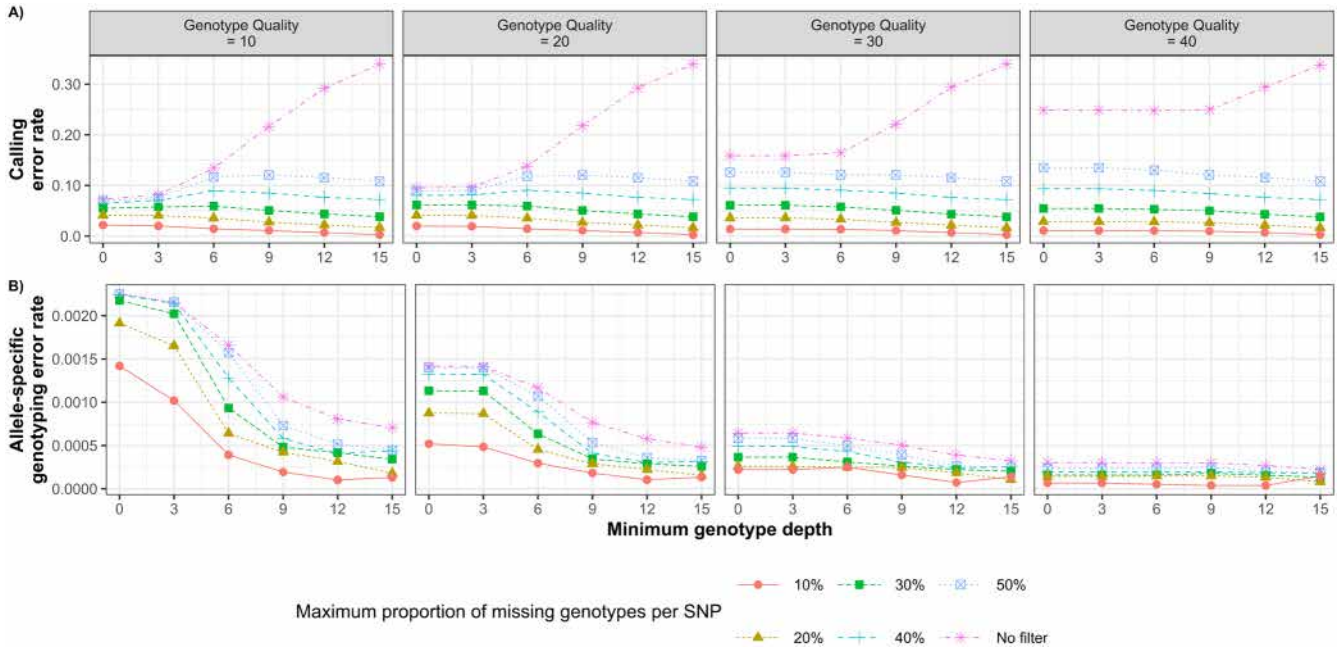

FIGURE 3 | Error rates for genotype calling (A) and allele-specific genotyping (B) as a function of minimum genotype quality (columns), minimum genotype depth (*x*-axis) and maximum proportion of missing genotypes per SNP (coloured lines). We show the average of error rates computed on the four technical replicates retained in the analysis before the hard-filtering step (the loci identification procedure identified GMB_312 and its replicate as low-quality samples, i.e., with a mean depth under 6× and more than 50% of missing data; see Figure S1).

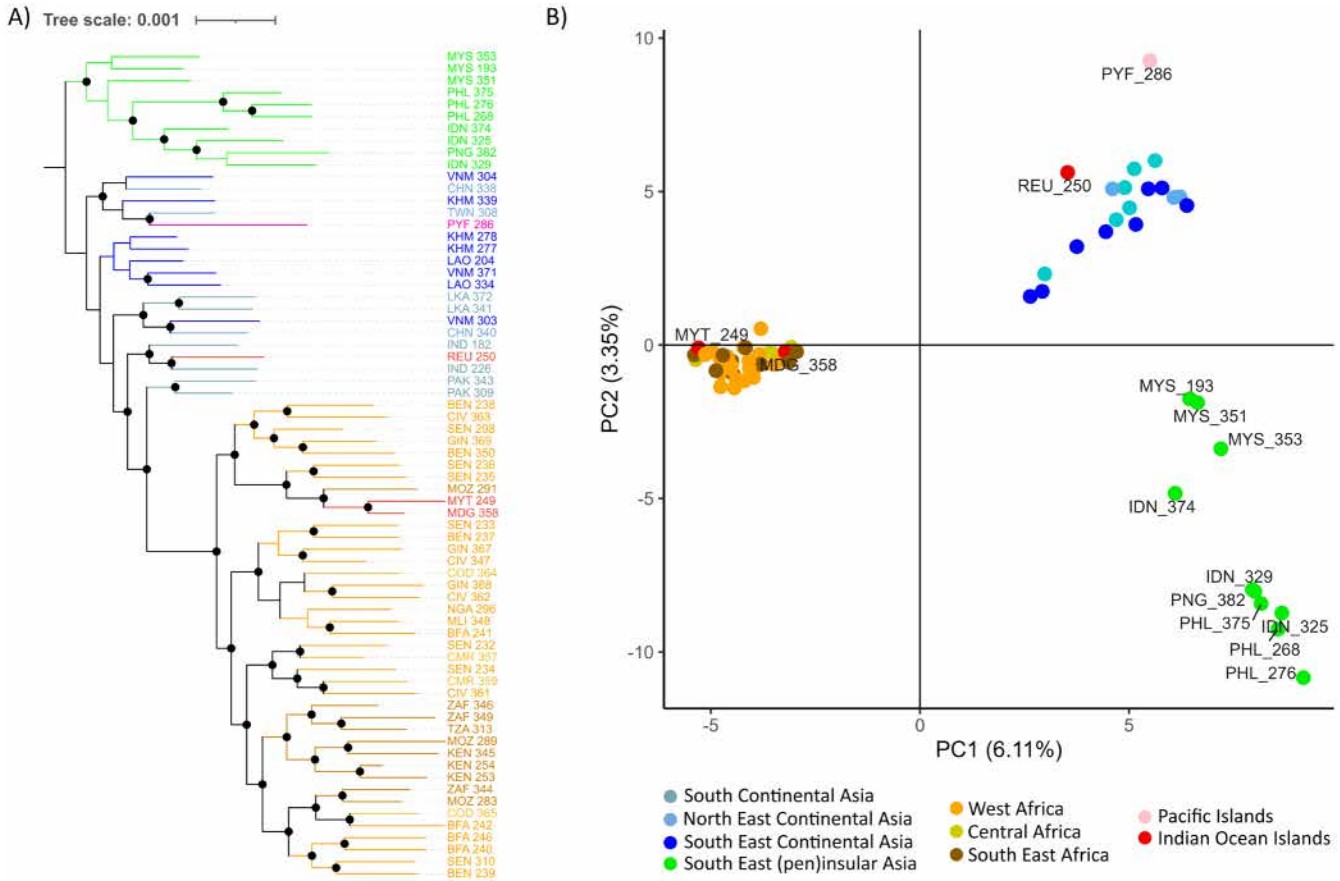

FIGURE 4 | Maximum likelihood phylogenetic tree (A) and first two principal component functions (B) inferred from the 68 individual samples of *Bactrocera dorsalis*. (A) The tree was constructed from the sequence concatenation of the 34,725 hard-filtered loci (25,893,395 bp) and rooted according to Charbonnel (2024). Bootstrap supports > 90% are shown at the nodes with a dot. (B) The PCA was obtained on the 7168 hard-filtered and independent SNPs.

TABLE 1 | Comparison of genetic diversity estimates among the four *Bactrocera dorsalis* genetic clusters assessed by the ADMIXTURE method.

	Continental Asia	Southeastern Asian islands and New Guinea	Southeastern Africa	Western Africa
Nuclear RAD-loci				
Sample size	17	10	11	26
Allelic richness	1.565	1.524	1.463	1.496
Expected heterozygosity	0.2508	0.2414	0.2055	0.2171
Inbreeding coefficient	0.1261	0.2081	0.0239	0.0278
Mitochondrial RAD-loci				
Sample size	17	10	12	26
Nb of unique haplotypes	17	9	7	14
Haplotype diversity	1.000	1.000	0.758	0.757
Nucleotide diversity	0.0034	0.0021	0.0019	0.0017
Nb of segregating sites	165	97	60	57

Note: Samples from invaded islands (i.e., PYF, REU, MYT, MDG) were excluded from this analysis. The allelic richness was computed for six alleles, i.e., the smallest number of individuals sampled across all combinations of populations and loci multiplied by two.

with the three Indonesian specimens, and (2) all invasive samples from Africa, Mayotte and Madagascar. It is noteworthy that the two samples from Pakistan are the closest relatives of this latter African clade. Individuals of continental Asia (including Hainan and Taiwan) were paraphyletic with regard to the African samples. The invasive sample from La Réunion clustered with two Indian samples, and the French Polynesia sample was connected to continental Eastern Asia. In spite of high bootstrap support, there was no clear geographic correspondence within the African continent, suggesting a lack of spatial structuring in the recently colonised continent. Yet, samples from Mayotte and Madagascar clustered together, suggesting a shared history of invasion. In general, invasive individuals displayed distinctively longer branches, suggesting high levels of drift.

3.3 | Spatial Genetic Variation Based on Nuclear SNPs

From the 265,154 hard-filtered SNPs, 7290 were retained by the stringent procedures of pruning (correlation coefficient ≤ 0.3 over a chromosomal window of 1 Mb) and thinning (above a physical distance of 2 kb). Of these, 122 were detected as outliers in the *pcadapat* analysis. The PCA applied to this SNP panel showed differentiation of the 68 individual samples in three main genetic groups along the first two principal components (Figure 4B). The first PC axis described 6.11% of the total variation and distinguished Asian and African samples, while the second PC axis described 3.35% of the total variation and primarily differentiated continental Asian countries from Southeast Asian countries (including New Guinea). Individuals from the native Asian genetic groups displayed greater differences among them than individuals from the invasive African genetic group: (i) along the PC2 axis, Southeast Asian insular samples (the Philippines and Indonesia) were clearly distinguished from peninsular Malaysian samples, and the New Guinea sample was closest to individuals from the Philippines and Indonesia, and (ii) the first two PC axes suggested a continuous geographic variation within continental Asia. The invasive samples from islands of the Pacific and Indian Oceans clearly clustered, to various degrees, with one of these main genetic groups along the first two PC axes: samples from La Réunion and French Polynesia with the continental Asian group, and samples from Madagascar and Mayotte with the invasive African group.

The ML method based on a model of distinct genetic clusters at Hardy–Weinberg equilibrium with admixture between them shed more light on the spatial genome-wide structure (Figure 1). The optimal number of genetic clusters (2; results not shown) distinguished all native Asian samples along with the invasive samples from La Réunion and French Polynesia on the one hand, from all the other samples from the invaded range (Africa, Madagascar and Mayotte) on the other hand. Exploring spatial structuring for a larger number of genetic clusters improved concordance with PCA and phylogenetic tree results. At $K = 3$, the Asian cluster separated into continental Asian countries (including the La Réunion and French Polynesian samples) on one hand, and the Southeastern Asian island countries and New Guinea on the other hand. At $K = 4$, samples from Western Africa were differentiated from Eastern and Southern African samples, including the Madagascan and Mayotte individuals. Co-ancestry estimates suggested genetic similarity between Madagascar and Mayotte islands on one hand, and between Africa and continental Asia on the other hand. This, however, should be interpreted with caution since inferences from ADMIXTURE were shown to overestimate admixture when sample sizes are small or unbalanced (Wang 2022). The two African clusters displayed the lowest level of genetic differentiation, though moderate ($F_{ST} = 7\%$). Their respective levels of genetic differentiation with respect to the continental Asian cluster were only slightly higher, more so for the Western cluster ($F_{ST} = 8\%$)

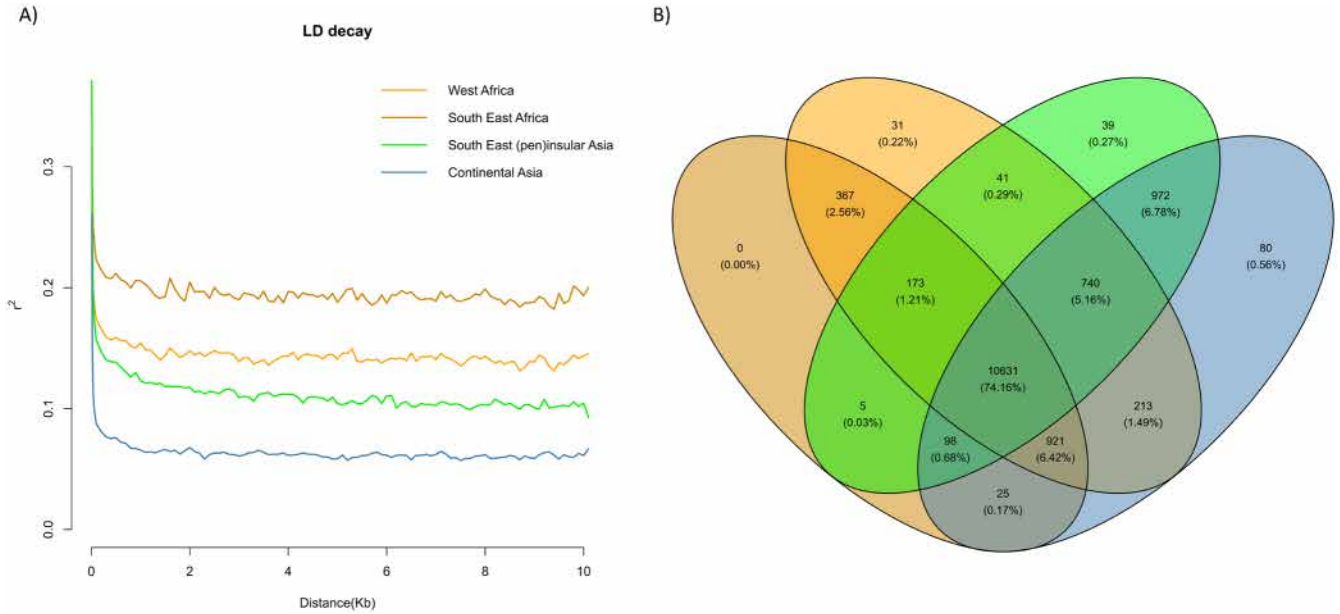

FIGURE 5 | Linkage disequilibrium across physical distance (A) and Venn Diagram of allele sharing (B), respectively within and between the four *Bactrocera dorsalis* genetic clusters assessed by the ADMIXTURE method. Individual samples from invaded islands (i.e., PYF, REU, MYT, MDG) were excluded from these analyses. We used (A) 265,154 hard-filtered SNPs, or (B) 7168 hard-filtered and independent SNPs.

than the Southeastern cluster ($F_{ST} = 10\%$). The continental and pen(insular) clusters of Asian individuals displayed a similar, moderate level of genetic differentiation ($F_{ST} = 10\%$). Higher levels of genetic differentiation were found between other pairs of clusters (results not shown). High values of the inbreeding coefficient (F_{IS}) in native Asian clusters suggested further genetic subdivisions (i.e., Walhund effect), unresolved by the ADMIXTURE genetic clustering probably due to small sampling size (Table 1).

Allelic richness and expected heterozygosity were higher in native Asian clusters than in invasive African ones while fairly even within each continent (Table 1). The great majority of alleles (74%) were shared by all four clusters, and only a limited number (1%) were restricted to a single cluster (Figure 5B). Out of the remaining quarter of alleles, 7% were exclusive to native Asia and 3% to invaded Africa, while 6% were shared between African clusters and continental Asia, and 5% were shared between Asian clusters and Western Africa (Figure 5B). We observed different signatures of LD with distance between SNPs in the four genetic clusters (Figure 5A). The continental Asian cluster displayed the fastest decay and the lowest stabilised value of LD ($r^2 \simeq 0.06$). The Southeast Asian cluster displayed a slightly slower decay and a twice higher LD with distance ($r^2 \simeq 0.12$) despite being in the native area, which may indicate smaller effective population sizes in this archipelago. The African clusters displayed the highest LD levels ($r^2 \simeq 0.15$ and 0.22 for the Western and Southeastern clusters, respectively), matching expectations regarding invasive populations having experienced founder events. The test for isolation by distance among the 33 African samples (using GPS coordinates) was significant ($R^2 = 0.205$, p-value < 0.05, Regression slope = 0.007), indicating a low but positive correlation between genetic differentiation and geographic distance on the African continent (Figure S5).

3.4 | Phylogenetic Tree and Network Based on Mitochondrial SNPs

The 1,295,238 passing-filtered reads we successfully mapped on the *B. dorsalis* mitochondrial genome allowed us to reconstruct a 5418b-long sequence covering two distinct regions (2650-6145b and 9754–11,698b, respectively). Forty-nine unique haplotypes and 211 segregating sites were identified, with a mean non-redundant depth of 170× and 2.5% of missing data (see details by sample in Table S5). Calling and genotyping error rates, estimated from our four technical replicates, were slightly lower than for nuclear data, with respective values of 2.5% and 0.0%. Overall, mitochondrial haplotype data provided a lower resolution of the spatial genetic structure than nuclear SNP data. The minimum spanning network distinguished the Asian native area from the African colonised area, with no shared haplotype at all and without strong structuring within each area (Figure S6A). Regarding the continental native range, we observed many unique or rare haplotypes, often distantly related. In contrast, the continental African samples comprised seven unique and exclusive haplotypes and three distant haplogroups, two of which were predominant (50% and 42.5%, respectively). These two major haplogroups were separated by 10 mutational steps and were distantly related to all the haplotypes from continental Asia (minimum of four and six mutations respectively). Moreover, they were found in several African countries, regardless of the SE-NW subdivision revealed by the nuclear SNP data. One of these predominant haplogroups included samples from Madagascar and Mayotte, whereas La Réunion and French Polynesia samples represented unique haplotypes, and the New Guinean sample had the same haplotype as one of the three Indonesian samples. The ML phylogenetic tree displayed low resolution, with most nodes with less than 90% bootstrap support (Figure S7). Yet, in discordance with nuclear loci and in agreement with

the mitochondrial haplotype network, African samples form the same three distinct groups. The putative monophyly of the African group suggested by the nuclear data is not supported here. Mitochondrial haplotypic and nucleotidic diversities calculated on the four genetic clusters as assessed by the ADMIXTURE method on the 7168 nuclear SNPs congruently showed much higher values in native Asia than in invaded Africa (Table 1).

4 | Discussion

RRS can be affected by inherent sources of error (e.g., high rates of PCR duplicates, allelic dropout due to polymorphic restriction sites) that may critically lower rates of genotype calling and/or accuracy. In this work, we showed that this problem can easily be addressed by choosing a RAD-Seq version designed to increase library complexity, i.e., the bestRAD technique developed by Ali et al. (2016), and a careful and meticulous approach. We first optimised the molecular laboratory protocol and reported two minor amendments to the initial protocol, which enabled us to minimise the number of unexpected residues and artefact sequences. We showed that this protocol was valid even for input DNA samples with barely detectable concentrations, i.e., at about 5 ng/μL, below recommended concentrations for most RAD sequencing protocols (but see Komoroske et al. 2019). The protocol was also robust to moderate DNA degradation as long as samples had a molecular weight ≥ 5 kb, although half of the more degraded samples could also be analysed. Our estimate of the mean rate of PCR duplicates detected on both identity and mapping coordinates of the paired-end reads (29%) lies towards the lower end of the range of PCR duplicate values reported in the literature for the standard Baird et al. (2008) method (20%–95%; Andrews et al. 2014; Euclide et al. 2020; Rochette et al. 2023; Schweyen et al. 2014) but also for the bestRAD method of Ali et al. (2016) (23%–44%; Ali et al. 2016; Rochette et al. 2023). Furthermore, our two-step approach to remove PCR duplicates is likely to be sensitive, with the risk of missing PCR duplicates lower than that of removing sequences falsely detected as such (e.g., a single-step approach would have detected either 19% (identity) or 25% (coordinates) of PCR duplicates). This result supports the addition of a restriction site-carrying fragment selection using streptavidin magnetic beads as an overall efficient means of limiting PCR duplicates. However, we showed that PCR duplicates can also depend on the quality of the input DNA, with a 50% increase in their proportion in DNA samples with a molecular weight below 5 kb, again cautioning against the inclusion of poor-quality DNA samples.

Backed by a careful bioinformatics methodology, this laboratory protocol generated tens of thousands of polymorphic loci representative of the genome. The number of retained polymorphic loci remained substantial (34,725 loci) even after filtering on genotypes, SNPs, and samples. This corresponds to the expected number of fragments of a size comprised between 150 and 1500 bp and with no bases masked for repetitive elements from an in silico analysis of the *B. dorsalis* genome, using the *rsitesearch* program from ddRADseqTools (Mora-Márquez et al. 2017). Furthermore, these loci were evenly distributed along the genome (e.g., high correlation with the total unmasked base pairs of each chromosome; Table S6) and accounted for about one-tenth of it (owing to a mean length of 745 bp). Thus, SNP number and representativeness should not limit the ability to yield an accurate description of population genetics structure and diversity and to address challenging tasks, such as tracing the origins of invasive events in the studied species. Using technical replicate samples whose library preparation and sequencing runs were different, we also obtained a low genotype calling error rate (a few percent), which would allow any user to merge different datasets produced with the same protocol for analyses (e.g., if so desired, to complete the current dataset). Finally, we revealed an insignificant per-allele genotyping error rate (a few hundredths of a percent), which was nearly one order of magnitude less than the sequencing error rate estimated using 10% PhiX sequences (0.35%).

Additionally, 0.16% of the RRS libraries sampled mitochondrial DNA, which allowed us to reconstruct one-third of the *B. dorsalis* mitochondrial genome (i.e., 5418b), in spite of a low-cut site frequency, with a great read depth (i.e., 10 times greater than for nuclear loci), few missing sites (i.e., half as that for nuclear loci) and a reliable representation of haplotypes (i.e., same haplotypes between technical replicates). The number of informative sites was increased 6-fold compared with the highly variable but 10 times shorter COI barcode, which from the same sample set yielded half as many unique haplotypes and a coarser phylogeographic reconstruction (Figure S6B). For example, the two major African haplogroups were both only two mutational steps from the same native Asian haplotype, which was common and central to many unique haplotypes. However, despite having an effective population size of one-fourth of that of the nuclear genome, leading to rapid lineage sorting, mitochondrial data yielded low resolution and mainly corroborated the higher levels of phylogeographic structure recovered by nuclear loci (but see below). It should nonetheless be possible to elicit from such additional data, at no additional sequencing effort, mitochondrial introgression due to past hybridization events between *B. dorsalis* and *B. carambolae*, *B. kandiensis* and *B. raiensis* (reviewed in Charbonnel et al. 2023).

Using population genetics approaches, our RAD sequencing of only 68 individuals from 11 countries in native Asia, 13 invaded countries in Africa, and five islands in the Indian and Pacific Oceans depicted a well-resolved genetic structure of *B. dorsalis* consistent with previously proposed patterns (Deschepper et al. 2023; Khamis et al. 2009; Kim et al. 2021; Qin et al. 2018; Zhang et al. 2023). We described the existence of four genetic clusters in *B. dorsalis*: two in the native area of the species, namely, continental Asia (from India to China) and (pen)insular Southeastern Asia (Malaysia, Indonesia and the Philippines), differentiated from each other, and two in the invaded range, namely, Southern and Eastern Africa (from the Democratic Republic of the Congo to South Africa) and Western Africa (from Senegal to Cameroon). The Western African cluster displayed a higher genetic diversity than the Southeastern African cluster. We also showed evidence that gene flow is limited by geographic distance at the African continental scale despite recent colonisation, a fact that may partly explain the observed ADMIXTURE pattern. Finally, *F*-statistics and coordinates on PC axes suggested further genetic subdivision that should be resolved with a more thorough population sampling scheme and better geographic coverage.

Our results also indicate potential sources for the five invasive samples from the Indian and Pacific Oceans, in agreement with previous NGS studies (Deschepper et al. 2023; Zhang et al. 2023) and sometimes with greater precision: (i) Southern continental Asia, probably India, for the La Réunion sample; (ii) Eastern continental Asia for the French Polynesia sample and (iii) the non-native Southeastern African group for the Madagascan and Mayotte samples. As for Africa, nuclear and mitochondrial phylogenetic placements were incongruent. In the mitochondrial data, African samples formed two main and divergent haplotypes that both branched off Asia, without further precision due to low support. In the nuclear ML tree, all invasive African samples seem to cluster as a monophyletic group that branches off from Pakistan samples. This preliminary result corroborates the WGS study of Zhang et al. (2023), who identified Southern continental Asia as the source of the African invasion. In addition, Charbonnel et al. (2023) showed that several *B. dorsalis* specimens from Africa (Cameroon and the Congo) bore a *B. kandiensis* mitochondrial COI sequence, as did 23% of *B. dorsalis* originating from the *B. kandiensis* range (i.e., Southern continental Asia). However, African phylogenetic relationships should be interpreted with caution since the existence of gene flow or genetic drift would violate the method's assumptions and we cannot exclude the possibility that the two African groups originated from different introduction events. We expect that future RRS studies on *B. dorsalis* based on high-density population sampling and high-performance model-based methods implementing admixture and bottleneck events will make useful contributions to the unresolved aspects of its invasion history.

Overall, our study shows that bestRAD is an excellent compromise between the cost of data production and the level of genetic information obtained for population genetic analyses. While WGS should, in general, provide a finer resolution, our more affordable sequencing approach showed the same ability to resolve worldwide geographic structure despite our limited sample size. We estimate that, in WGS, the cost of our study would have been roughly three times higher, considering the size of the *B. dorsalis* genome (468.7 Mb, Jiang et al. 2022), a very low rate of PCR duplicates (~5%; Ebbert et al. 2016), a tenfold increase in the cost of library construction (due to the lack of multiplexing capacity) and the same sequencing effort and platform. The bestRAD method considerably reduced the cost and time of manipulation, even compared with the parent RAD-Seq version, by increasing both non-redundant depth of coverage and sample multiplexing of the libraries. The multiplexing capacity is a multiple of the number of biotinylated tags used in the ligation process (in our laboratory, 96) and is usually greater than in the parent RAD-Seq technique of Baird et al. (2008) (in our laboratory, 32). In our laboratory, savings in total cost and handling time for producing libraries will reach about one-third as soon as several hundred individuals are analysed. We therefore expect bestRAD to long remain a valuable alternative to WGS for many evolutionary and ecological applications, as long as access to the entire genome is not required.

Author Contributions

M.-P.C., L.B., E.C. and D.O. designed the study; L.B., S.N. and E.C. optimised the bestRAD molecular biology protocol with contributions from M.-P.C.; L.B. and E.C. were in charge of DNA extractions, quality control, and sequencing libraries; B.G. undertook the genome masking; E.C., M.-P.C., R.L. and E.O.-A. carried out the bioinformatics analyses (read denoising, SNP calling and hard-filtering); E.C. and M.-P.C. performed the data analyses; M.-P.C. and E.C. wrote the manuscript with contributions from R.L., D.O. and L.B.

Acknowledgements

E.C. was supported by a doctoral fellowship funded by the French Agricultural Research Centre for International Development (CIRAD) and the French Agency for Food, Environmental and Occupational Health & Safety (ANSES) (BACTRACK). This work was supported by the DISLAND, PROLAG and ISOGEO projects and publicly funded by ANR, Labex CEMEB and ANSES. We wish to deeply thank again the phytosanitary border inspectors at Roissy airport for collecting fruit fly specimens and Emma Artige and Servane Baufumé for their assistance in complying with the Nagoya Protocol, as well as the collaborators listed in the table S1 of Charbonnel et al. (2023), where the QDAF collection should actually be called the Queensland Primary Industries Insect Collection (QDPC). We are grateful to CBGP for letting us access its Molecular Biology and Collection platforms, to Montpellier GenomiX for sequencing RAD libraries and to the Genotoul bioinformatics platform Toulouse Occitanie for providing computing and storage resources. We are also grateful for the advice and guidance offered by Sean Michael O'Rourke and Mary E. Badger on the molecular biology protocol, Charles Perrier on using the Stacks2 program and Maria Bogaerts Marquez on methods for genome masking for repeat elements. We thank Karine Berthier for her comments on the manuscript and Anya Cockle for careful English language editing.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The optimised molecular laboratory protocol for the bestRAD approach (Ali et al. 2016) was deposited on the *Protocols.io* platform (Benoit et al. 2024) and is available at dx.doi.org/10.17504/protocols.io.rm7vz3ok4gx1/v1. Demultiplexed Illumina reads for each individual sample were deposited in the Sequence Read Archive (SRA) archive and are available on the National Center for Biotechnology Information (NCBI) server under BioProject PRJNA1085726. Our study complies with the Nagoya Protocol for the Convention on Biological Diversity, with 21 legal agreements signed between CIRAD and the government agencies from the countries providing genetic samples. The contributions of all collaborators are acknowledged in Charbonnel et al. (2023) and in the Acknowledgements section, and, more broadly, our group is willing to be committed to international scientific partnerships. The research addresses a priority concern, an invasive alien species that is a threat to tropical and subtropical fruits and vegetables worldwide. Benefits from this research accrue from the sharing of our protocols and data on public databases as described above.

References

Adamack, A. T., and B. Gruber. 2014. "PopGenReport: Simplifying Basic Population Genetic Analyses in R." *Methods in Ecology and Evolution* 5, no. 4: 384–387. https://doi.org/10.1111/2041-210X.12158.

Alexander, D. H., J. Novembre, and K. Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research* 19, no. 9: 1655–1664. https://doi.org/10.1101/gr.094052.109.

Ali, O. A., C. Jeffres, and M. R. Miller. 2016. "RAD Capture (Rapture): Flexible and Efficient Sequence-Based Genotyping." *Genetics* 202, no. 2: 16. https://doi.org/10.1534/genetics.115.183665.

Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe. 2016. "Harnessing the Power of RADseq for Ecological

and Evolutionary Genomics." *Nature Reviews Genetics* 17, no. 2: 81–92. https://doi.org/10.1038/nrg.2015.28.

Andrews, K. R., P. A. Hohenlohe, M. R. Miller, B. K. Hand, J. E. Seeb, and G. Luikart. 2014. "Trade-Offs and Utility of Alternative RADseq Methods: Reply to Puritz Et al." *Molecular Ecology* 23, no. 24: 5943–5946. https://doi.org/10.1111/mec.12964.

Andrews, K. R., and G. Luikart. 2014. "Recent Novel Approaches for Population Genomics Data Analysis." *Molecular Ecology* 23, no. 7: 1661–1667. https://doi.org/10.1111/mec.12686.

Arnold, B., R. B. Corbett-Detig, D. Hartl, and K. Bomblies. 2013. "RADseq Underestimates Diversity and Introduces Genealogical Biases due to Nonrandom Haplotype Sampling." *Molecular Ecology* 22, no. 11: 3179–3190. https://doi.org/10.1111/mec.12276.

Baird, N. A., P. D. Etter, T. S. Atwood, et al. 2008. "Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers." *PLoS One* 3, no. 10: e3376. https://doi.org/10.1371/journal.pone.0003376.

Bayega, A., H. Djambazian, K. T. Tsoumani, et al. 2020. "De Novo Assembly of the Olive Fruit Fly (*Bactrocera oleae*) Genome With Linked-Reads and Long-Read Technologies Minimizes Gaps and Provides Exceptional Y Chromosome Assembly." *BMC Genomics* 21, no. 1: 259. https://doi.org/10.1186/s12864-020-6672-3.

Benoit, L., S. Nidelet, E. Charbonnel, and M.-P. Chapuis. 2024. "A FAIR Protocol of the bestRAD Sequencing Approach." Protocols.io, Version 1. https://doi.org/10.17504/protocols.io.rm7vz3ok4gx1/v1.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30, no. 15: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Bresadola, L., V. Link, C. A. Buerkle, C. Lexer, and D. Wegmann. 2020. "Estimating and Accounting for Genotyping Errors in RAD-Seq Experiments." *Molecular Ecology Resources* 20, no. 4: 856–870. https://doi.org/10.1111/1755-0998.13153.

Campbell, E. O., B. M. T. Brunet, J. R. Dupuis, and F. A. H. Sperling. 2018. "Would an RRS by any Other Name Sound as RAD?" *Methods in Ecology and Evolution* 9, no. 9: 1920–1927. https://doi.org/10.1111/2041-210X.13038.

Cariou, M., L. Duret, and S. Charlat. 2013. "Is RAD-Seq Suitable for Phylogenetic Inference? An In Silico Assessment and Optimization." *Ecology and Evolution* 3, no. 4: 846–852. https://doi.org/10.1002/ece3.512.

Casbon, J. A., R. J. Osborne, S. Brenner, and C. P. Lichtenstein. 2011. "A Method for Counting PCR Template Molecules With Application to Next-Generation Sequencing." *Nucleic Acids Research* 39, no. 12: e81. https://doi.org/10.1093/nar/gkr217.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4: 7. https://doi.org/10.1186/s13742-015-0047-8.

Charbonnel, E. 2024. "Informer sur l'Origine Géographique et le Statut Spécifique d'un Ravageur Invasif et Cryptique: Validation d'une Méthodologie Haut-débit Pour le Suivi de la Mouche Orientale des Fruits, *Bactrocera dorsalis*." Université de Montpellier. http://www.theses.fr/s259414.

Charbonnel, E., M.-P. Chapuis, A. Taddei, et al. 2023. "Evaluation of Identification Methods for Cryptic *Bactrocera dorsalis* (Diptera: Tephritidae) Specimens: Combining Morphological and Molecular Techniques." *Journal of Economic Entomology* 116, no. 6: 2193–2200. https://doi.org/10.1093/jee/toad178.

Clarke, A. R., Z. Li, Y. Qin, Z.-H. Zhao, L. Liu, and M. K. Schutze. 2019. "*Bactrocera dorsalis* (Hendel) (Diptera: Tephritidae) is Not Invasive Through Asia: It's Been There all Along." *Journal of Applied Entomology* 143, no. 8: 797–801. https://doi.org/10.1111/jen.12649.

Clement, M., Q. Snell, P. Walke, D. Posada, and K. Crandall. 2002. "TCS: Estimating Gene Genealogies." Proceedings 16th International Parallel and Distributed Processing Symposium. https://www.academia.edu/27595464/TCS_estimating_gene_genealogies.

Cribari-Neto, F., and A. Zeileis. 2010. "Beta Regression in R." *Journal of Statistical Software* 34: 1–24. https://doi.org/10.18637/jss.v034.i02.

Cristofari, R., G. Bertorelle, A. Ancel, et al. 2016. "Full Circumpolar Migration Ensures Evolutionary Unity in the Emperor Penguin." *Nature Communications* 7, no. 1: 11842. https://doi.org/10.1038/ncomms11842.

Cullingham, C. I., J. M. Miller, R. M. Peery, et al. 2020. "Confidently Identifying the Correct K Value Using the ΔK Method: When Does K = 2?" *Molecular Ecology* 29, no. 5: 862–869. https://doi.org/10.1111/mec.15374.

Danecek, P., A. Auton, G. Abecasis, et al. 2011. "The Variant Call Format and VCFtools." *Bioinformatics* 27, no. 15: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

Davey, J. W., T. Cezard, P. Fuentes-Utrilla, C. Eland, K. Gharbi, and M. L. Blaxter. 2013. "Special Features of RAD Sequencing Data: Implications for Genotyping." *Molecular Ecology* 22, no. 11: 3151–3164. https://doi.org/10.1111/mec.12084.

DeRaad, D. A. 2022. "Snpfiltr: An R Package for Interactive and Reproducible SNP Filtering." *Molecular Ecology Resources* 22, no. 6: 2443–2453. https://doi.org/10.1111/1755-0998.13618.

Deschepper, P., S. Vanbergen, Y. Zhang, et al. 2023. "*Bactrocera dorsalis* in the Indian Ocean: A Tale of Two Invasions." *Evolutionary Applications* 16, no. 1: 48–61. https://doi.org/10.1111/eva.13507.

Díaz- Arce, N., and N. Rodríguez- Ezpeleta. 2019. "Selecting RAD-Seq Data Analysis Parameters for Population Genetics: The More the Better?" *Frontiers in Genetics* 10: 533. https://doi.org/10.3389/fgene.2019.00533.

Dray, S., and A.-B. Dufour. 2007. "The ade4 Package: Implementing the Duality Diagram for Ecologists." *Journal of Statistical Software* 22: 1–20. https://doi.org/10.18637/jss.v022.i04.

Drew, R. A. I., and M. C. Romig. 2013. *Tropical Fruit Flies (Tephritidae Dacinae) of South-East Asia: Indomalaya to North-West Australasia*. CABI.

Drew, R. A. I., and M. C. Romig. 2016. *Keys to the Tropical Fruit Flies (Tephritidae: Dacinae) of South-East Asia: Indomalaya to North-West Australasia*. CABI.

Drew, R. A. I., K. Tsuruta, and I. White. 2005. "A New Species of Pest Fruit Fly (Diptera: Tephritidae: Dacinae) From Sri Lanka and Africa." *African Entomology* 13: 149–154.

Ebbert, M. T. W., M. E. Wadsworth, L. A. Staley, et al. 2016. "Evaluating the Necessity of PCR Duplicate Removal From Next-Generation Sequencing Data and a Comparison of Approaches." *BMC Bioinformatics* 17, no. 7: 239. https://doi.org/10.1186/s12859-016-1097-3.

Euclide, P. T., G. J. McKinney, M. Bootsma, C. Tarsa, M. H. Meek, and W. A. Larson. 2020. "Attack of the PCR Clones: Rates of Clonality Have Little Effect on RAD-Seq Genotype Calls." *Molecular Ecology Resources* 20, no. 1: 66–78. https://doi.org/10.1111/1755-0998.13087.

Flynn, J. M., R. Hubley, C. Goubert, et al. 2020. "RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families." *Proceedings of the National Academy of Sciences* 117, no. 17: 9451–9457. https://doi.org/10.1073/pnas.1921046117.

Fox, J., and S. Weisberg. 2018. *An R Companion to Applied Regression*. SAGE Publications.

Fu, Y., P.-H. Wu, T. Beane, P. D. Zamore, and Z. Weng. 2018. "Elimination of PCR Duplicates in RNA-Seq and Small RNA-Seq Using Unique Molecular Identifiers." *BMC Genomics* 19, no. 1: 531. https://doi.org/10.1186/s12864-018-4933-1.

Garzón-Orduña, I. J., S. M. Geib, and N. B. Barr. 2019. "The Genetic Diversity of *Bactrocera dorsalis* (Diptera: Tephritidae) in China and Neighboring Countries: A Review From Published Studies." *Journal of Economic Entomology* 112, no. 4: 2001–2006. https://doi.org/10.1093/jee/toz073.

Gautier, M., K. Gharbi, T. Cezard, et al. 2013. "The Effect of RAD Allele Dropout on the Estimation of Genetic Variation Within and Between Populations." *Molecular Ecology* 22, no. 11: 3165–3178.

Gilchrist, A. S., D. C. Shearman, M. Frommer, et al. 2014. "The Draft Genome of the Pest Tephritid Fruit Fly *Bactrocera tryoni*: Resources for the Genomic Analysis of Hybridising Species." *BMC Genomics* 15, no. 1: 1153. https://doi.org/10.1186/1471-2164-15-1153.

Graham, C. F., D. R. Boreham, R. G. Manzon, W. Stott, J. Y. Wilson, and C. M. Somers. 2020. "How "Simple" Methodological Decisions Affect Interpretation of Population Structure Based on Reduced Representation Library DNA Sequencing: A Case Study Using the Lake Whitefish." *PLoS One* 15, no. 1: e0226608. https://doi.org/10.1371/journal.pone.0226608.

Hall, T. A. 1999. "BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT." *Nucleic Acids Symposium Series* 41, no. 41: 95–98.

Hickerson, M. J., B. C. Carstens, J. Cavender-Bares, et al. 2010. "Phylogeography's Past, Present, and Future: 10 Years After Avise, 2000." *Molecular Phylogenetics and Evolution* 54, no. 1: 291–301. https://doi.org/10.1016/j.ympev.2009.09.016.

Hill, W. G., and A. Robertson. 1968. "Linkage Disequilibrium in Finite Populations." *Theoretical and Applied Genetics* 38, no. 6: 226–231. https://doi.org/10.1007/BF01245622.

Hivert, V., R. Leblois, E. J. Petit, M. Gautier, and R. Vitalis. 2018. "Measuring Genetic Differentiation From Pool-Seq Data." *Genetics* 210, no. 1: 315–330. https://doi.org/10.1534/genetics.118.300900.

Hoang, D. T., O. Chernomor, A. von Haeseler, B. Q. Minh, and L. S. Vinh. 2018. "UFBoot2: Improving the Ultrafast Bootstrap Approximation." *Molecular Biology and Evolution* 35, no. 2: 518–522. https://doi.org/10.1093/molbev/msx281.

Janes, J. K., J. M. Miller, J. R. Dupuis, et al. 2017. "The K = 2 Conundrum." *Molecular Ecology* 26, no. 14: 3594–3602. https://doi.org/10.1111/mec.14187.

Jiang, F., L. Liang, J. Wang, and S. Zhu. 2022. "Chromosome-Level Genome Assembly of *Bactrocera Dorsalis* Reveals Its Adaptation and Invasion Mechanisms." *Communications Biology* 5, no. 1: 25. https://doi.org/10.1038/s42003-021-02966-6.

Jombart, T., and I. Ahmed. 2011. "Adegenet 1.3-1: New Tools for the Analysis of Genome-Wide SNP Data." *Bioinformatics* 27: 3070–3071. https://doi.org/10.1093/bioinformatics/btr521.

Kalinowski, S. T. 2011. "The Computer Program STRUCTURE Does Not Reliably Identify the Main Genetic Clusters Within Species: Simulations and Implications for Human Population Structure." *Heredity* 106, no. 4: 625–632. https://doi.org/10.1038/hdy.2010.95.

Kamvar, Z. N., J. F. Tabima, and N. J. Grünwald. 2014. "Poppr: An R Package for Genetic Analysis of Populations With Clonal, Partially Clonal, and/or Sexual Reproduction." *PeerJ* 2: e281. https://doi.org/10.7717/peerj.281.

Kebschull, J. M., and A. M. Zador. 2015. "Sources of PCR-Induced Distortions in High-Throughput Sequencing Data Sets." *Nucleic Acids Research* 43, no. 21: e143. https://doi.org/10.1093/nar/gkv717.

Khamis, F. M., N. Karam, S. Ekesi, et al. 2009. "Uncovering the Tracks of a Recent and Rapid Invasion: The Case of the Fruit Fly Pest Bactrocera Invadens (Diptera: Tephritidae) in Africa." *Molecular Ecology* 18, no. 23: 4798–4810. https://doi.org/10.1111/j.1365-294X.2009.04391.x.

Kim, H., S. Kim, S. Kim, et al. 2021. "Population Genetics for Inferring Introduction Sources of the Oriental Fruit Fly, *Bactrocera dorsalis*: A Test for Quarantine Use in Korea." *Insects* 12, no. 10: 851. https://doi.org/10.3390/insects12100851.

Knaus, B. J., and N. J. Grünwald. 2017. "Vcfr: A Package to Manipulate and Visualize Variant Call Format Data in R." *Molecular Ecology Resources* 17, no. 1: 44–53. https://doi.org/10.1111/1755-0998.12549.

Komoroske, L. M., M. R. Miller, S. M. O'Rourke, K. R. Stewart, M. P. Jensen, and P. H. Dutton. 2019. "A Versatile Rapture (RAD-Capture) Platform for Genotyping Marine Turtles." *Molecular Ecology Resources* 19, no. 2: 497–511. https://doi.org/10.1111/1755-0998.12980.

Laczkó, L., S. Jordán, and G. Sramkó. 2022. "The RadOrgMiner Pipeline: Automated Genotyping of Organellar Loci From RADseq Data." *Methods in Ecology and Evolution* 13, no. 9: 1962–1975. https://doi.org/10.1111/2041-210X.13937.

Lawson, D. J., L. van Dorp, and D. Falush. 2018. "A Tutorial on How Not to Over-Interpret STRUCTURE and ADMIXTURE Bar Plots." *Nature Communications* 9, no. 1: 3258. https://doi.org/10.1038/s41467-018-05257-7.

Leigh, J. W., and D. Bryant. 2015. "popart: Full-Feature Software for Haplotype Network Construction." *Methods in Ecology and Evolution* 6, no. 9: 1110–1116. https://doi.org/10.1111/2041-210X.12410.

Letunic, I., and P. Bork. 2007. "Interactive Tree of Life (iTOL): An Online Tool for Phylogenetic Tree Display and Annotation." *Bioinformatics* 23, no. 1: 127–128. https://doi.org/10.1093/bioinformatics/btl529.

Li, H., B. Handsaker, A. Wysoker, et al. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25, no. 16: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Luu, K., E. Bazin, and M. G. B. Blum. 2017. "Pcadapt: An R Package to Perform Genome Scans for Selection Based on Principal Component Analysis." *Molecular Ecology Resources* 17, no. 1: 67–77. https://doi.org/10.1111/1755-0998.12592.

Martin, M. 2011. "Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads." *EMBnet.Journal* 17, no. 1: 10. https://doi.org/10.14806/ej.17.1.200.

Matheson, P., and A. McGaughran. 2022. "Genomic Data Is Missing for Many Highly Invasive Species, Restricting Our Preparedness for Escalating Incursion Rates." *Scientific Reports* 12, no. 1: 13987. https://doi.org/10.1038/s41598-022-17937-y.

McKinney, G. J., R. K. Waples, L. W. Seeb, and J. E. Seeb. 2017. "Paralogs Are Revealed by Proportion of Heterozygotes and Deviations in Read Ratios in Genotyping-By-Sequencing Data From Natural Populations." *Molecular Ecology Resources* 17, no. 4: 656–669. https://doi.org/10.1111/1755-0998.12613.

Minh, B. Q., H. A. Schmidt, O. Chernomor, et al. 2020. "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era." *Molecular Biology and Evolution* 37, no. 5: 1530–1534. https://doi.org/10.1093/molbev/msaa015.

Mora-Márquez, F., V. García-Olivares, B. C. Emerson, and U. de López Heredia. 2017. "Ddradseqtools: A Software Package for In Silico Simulation and Testing of Double-Digest RAD Seq Experiments." *Molecular Ecology Resources* 17, no. 2: 230–246. https://doi.org/10.1111/1755-0998.12550.

North, H. L., A. McGaughran, and C. D. Jiggins. 2021. "Insights Into Invasive Species From Whole-Genome Resequencing." *Molecular Ecology* 30, no. 23: 6289–6308. https://doi.org/10.1111/mec.15999.

Paradis, E. 2010. "Pegas: An R Package for Population Genetics With an Integrated–Modular Approach." *Bioinformatics* 26, no. 3: 419–420. https://doi.org/10.1093/bioinformatics/btp696.

Paradis, E., and K. Schliep. 2019. "Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R." *Bioinformatics* 35, no. 3: 526–528. https://doi.org/10.1093/bioinformatics/bty633.

Pebesma, E. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *R Journal* 10: 439–446. https://doi.org/10.32614/RJ-2018-009.

Qin, Y., M. N. Krosch, M. K. Schutze, et al. 2018. "Population Structure of a Global Agricultural Invasive Pest, *Bactrocera dorsalis* (Diptera: Tephritidae)." *Evolutionary Applications* 11, no. 10: 1990–2003. https://doi.org/10.1111/eva.12701.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing [Computer Software]*. R Foundation for Statistical Computing. https://www.R-project.org/.

Reid, B. N., R. L. Moran, C. J. Kopack, and S. W. Fitzpatrick. 2021. "Rapture-Ready Darters: Choice of Reference Genome and Genotyping Method (Whole-Genome or Sequence Capture) Influence Population Genomic Inference in Etheostoma." *Molecular Ecology Resources* 21, no. 2: 404–420. https://doi.org/10.1111/1755-0998.13275.

Rivera-Colón, A. G., N. C. Rochette, and J. M. Catchen. 2021. "Simulation With RADinitio Improves RADseq Experimental Design and Sheds Light on Sources of Missing Data." *Molecular Ecology Resources* 21, no. 2: 363–378. https://doi.org/10.1111/1755-0998.13163.

Rochette, N. C., A. G. Rivera-Colón, and J. M. Catchen. 2019. "Stacks 2: Analytical Methods for Paired-End Sequencing Improve RADseq-Based Population Genomics." *Molecular Ecology* 28, no. 21: 4737–4754. https://doi.org/10.1111/mec.15253.

Rochette, N. C., A. G. Rivera-Colón, J. Walsh, T. J. Sanger, S. C. Campbell-Staton, and J. M. Catchen. 2023. "On the Causes, Consequences, and Avoidance of PCR Duplicates: Towards a Theory of Library Complexity." *Molecular Ecology Resources* 23, no. 6: 1299–1318. https://doi.org/10.1111/1755-0998.13800.

Rousset, F. 2000. "Genetic Differentiation Between Individuals." *Journal of Evolutionary Biology* 13, no. 1: 58–62. https://doi.org/10.1046/j.1420-9101.2000.00137.x.

Rousset, F. 2008. "genepop'007: A Complete Re-Implementation of the Genepop Software for Windows and Linux." *Molecular Ecology Resources* 8, no. 1: 103–106. https://doi.org/10.1111/j.1471-8286.2007.01931.x.

San Jose, M., C. Doorenweerd, L. Leblanc, N. Barr, S. Geib, and D. Rubinoff. 2018. "Tracking the Origins of Fly Invasions; Using Mitochondrial Haplotype Diversity to Identify Potential Source Populations in Two Genetically Intertwined Fruit Fly Species (Bactrocera Carambolae and *Bactrocera dorsalis* [Diptera: Tephritidae])." *Journal of Economic Entomology* 111, no. 6: 2914–2926. https://doi.org/10.1093/jee/toy272.

Savary, P., J.-C. Foltête, H. Moal, G. Vuidel, and S. Garnier. 2021. "graph4lg: A Package for Constructing and Analysing Graphs for Landscape Genetics in R." *Methods in Ecology and Evolution* 12, no. 3: 539–547. https://doi.org/10.1111/2041-210X.13530.

Schnell, I. B., K. Bohmann, and M. T. P. Gilbert. 2015. "Tag Jumps Illuminated – Reducing Sequence-To-Sample Misidentifications in Metabarcoding Studies." *Molecular Ecology Resources* 15, no. 6: 1289–1303. https://doi.org/10.1111/1755-0998.12402.

Schutze, M. K., N. Aketarawong, W. Amornsak, et al. 2014. "Synonymization of Key Pest Species Within the *B. actrocera* Dorsalisspecies Complex (Diptera:Tephritidae): Taxonomic Changes Based on a Review of 20 Years of Integrative Morphological, Molecular, Cytogenetic, Behavioural and Chemoecological Data." *Systematic Entomology* 40, no. 2: 456–471. https://doi.org/10.1111/syen.12113.

Schweyen, H., A. Rozenberg, and F. Leese. 2014. "Detection and Removal of PCR Duplicates in Population Genomic ddRAD Studies by Addition of a Degenerate Base Region (DBR) in Sequencing Adapters." *Biological Bulletin* 227, no. 2: 146–160. https://doi.org/10.1086/BBLv227n2p146.

Stobie, C. S., M. J. Cunningham, C. J. Oosthuizen, and P. Bloomer. 2019. "Finding Stories in Noise: Mitochondrial Portraits From RAD Data." *Molecular Ecology Resources* 19, no. 1: 191–205. https://doi.org/10.1111/1755-0998.12953.

Storey, J. D., A. J. Bass, A. Dabney, D. Robinson, and G. Warnes. 2023. "qvalue: Q-Value Estimation for False Discovery Rate Control (2.32.0) [Computer Software]." Bioconductor Version: Release (3.17). https://doi.org/10.18129/B9.bioc.qvalue.

Tarailo-Graovac, M., and N. Chen. 2009. "Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences." *Current Protocols in Bioinformatics* 25: 4.10.1–4.10.14. https://doi.org/10.1002/0471250953.bi0410s25.

Vasimuddin, M., S. Misra, H. Li, and S. Aluru. 2019. "Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems." 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 314–324. https://doi.org/10.1109/IPDPS.2019.00041.

Vaux, F., L. Dutoit, C. Fraser, and J. Waters. 2022. "Genotyping-By-Sequencing for Biogeography." *Journal of Biogeography* 50: 262–281. https://doi.org/10.1111/jbi.14516.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics With S*. 4th ed. Springer.

Wang, J. 2022. "Fast and Accurate Population Admixture Inference From Genotype Data From a Few Microsatellites to Millions of SNPs." *Heredity* 129: 79–92. https://doi.org/10.1038/s41437-022-00535-z.

White, I. M., and M. M. Elson-Harris. 1992. "Fruit Flies of Economic Significance: Their Identification and Bionomics." https://www.cabdirect.org/cabdirect/abstract/19921161954.

Wilson, A. C., R. L. Cann, S. M. Carr, et al. 1985. "Mitochondrial DNA and Two Perspectives on Evolutionary Genetics." *Biological Journal of the Linnean Society* 26, no. 4: 375–400. https://doi.org/10.1111/j.1095-8312.1985.tb02048.x.

Yu, D. J., L. Xu, F. Nardi, J. G. Li, and R. J. Zhang. 2007. "The Complete Nucleotide Sequence of the Mitochondrial Genome of the Oriental Fruit Fly, *Bactrocera dorsalis* (Diptera: Tephritidae)." *Gene* 396, no. 1: 66–74. https://doi.org/10.1016/j.gene.2007.02.023.

Zhang, C., S.-S. Dong, J.-Y. Xu, W.-M. He, and T.-L. Yang. 2019. "PopLDdecay: A Fast and Effective Tool for Linkage Disequilibrium Decay Analysis Based on Variant Call Format Files." *Bioinformatics* 35, no. 10: 1786–1788. https://doi.org/10.1093/bioinformatics/bty875.

Zhang, Y., S. Liu, M. De Meyer, et al. 2023. "Genomes of the Cosmopolitan Fruit Pest *Bactrocera dorsalis* (Diptera: Tephritidae) Reveal Its Global Invasion History and Thermal Adaptation." *Journal of Advanced Research* 53: 61–74. https://doi.org/10.1016/j.jare.2022.12.012.

Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir. 2012. "A High-Performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data." *Bioinformatics* 28, no. 24: 3326–3328. https://doi.org/10.1093/bioinformatics/bts606.

Supporting Information

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