






OPEN

DATA DESCRIPTOR

Integrated *de novo* transcriptome of *Culex pipiens* mosquito larvae as a resource for genetic control strategies

Valentina Mastrantonio¹, Pietro Libro², Jessica Di Martino¹, Michele Matera^{3,4}, Romeo Bellini⁵, Tiziana Castrignano², Sandra Urbanelli¹ & Daniele Porretta¹

We present a *de novo* transcriptome of the mosquito vector *Culex pipiens*, assembled by sequences of susceptible and insecticide resistant larvae. The high quality of the assembly was confirmed by TransRate and BUSCO. A mapping percentage until 94.8% was obtained by aligning contigs to Nr, SwissProt, and TrEMBL, with 27,281 sequences that simultaneously mapped on the three databases. A total of 14,966 ORFs were also functionally annotated by using the eggNOG database. Among them, we identified ORF sequences of the main gene families involved in insecticide resistance. Therefore, this resource stands as a valuable reference for further studies of differential gene expression as well as to identify genes of interest for genetic-based control tools.

Background & Summary

Insecticide resistance is a serious problem in the control of vectors and vector-borne diseases (VBDs)¹. In the last decades, many efforts have been addressed to reveal the mechanisms underlying insecticide resistant phenotypes and develop new tools alternative to chemical compounds. Genetically-based tools (e.g., RNA interference, CRISPR-Cas9) have been identified as a cutting-edge alternative to manage insecticide resistance because they enhance our ability to detect, functionally validate, and inhibit putative resistance genes and/or mutations^{2,3}. Likewise, these tools are proving effective in suppressing critical biological functions, inducing sex ratio distortion, and driving the spreading of desired genes within populations³⁻⁶. However, these technologies have a primary constraint to be effectively applied: the need for available target sequences. Increasing high-quality -omics resources is therefore mandatory, especially for non-model species.

Transcriptomes have recently been considered valuable genomic resources⁷⁻¹². Contrary to genomic data revealing what genes are present within cells, transcriptomes provide evidence about what genes are expressed under certain conditions. By comparing transcriptome data of different species, life stages, tissues, or conditions, we can thus elucidate the molecular pathways underlying specific phenotypes and identify candidate genes for traits of interest¹³⁻¹⁷. Nonetheless, this possibility to explore the multiple facets of transcriptomes mainly relies on high-quality references for comparison.

Here, we furnish a *de novo* assembled transcriptome of the mosquito *Culex pipiens* (ecotype *pipiens*), the primary vector of several pathogens such as West Nile virus, Japanese encephalitis, and lymphatic filariasis across the temperate Northern Hemisphere^{18,19}. This species belongs to the *Culex pipiens* complex, that also include *Cx. quinquefasciatus*, *Cx. pipiens pallens*, *Cx. australicus* and *Cx. globocoxitus*²⁰ and it is commonly found in urban and natural habitats. Due to its vector competence, *Cx. pipiens* has always been a priority in chemicals control applications. Recently, concern about this species has also increased because genetic resistance to the insecticide diflubenzuron (DFB), the principal larvicide used to control *Cx. pipiens*, has been found in populations across its geographic range²¹⁻²⁵. Genetic analyses showed that DFB resistance was associated with a mechanism already

¹Department of Environmental Biology, Sapienza University of Rome, 00185, Rome, Italy. ²Department of Ecological and Biological Sciences, Tuscia University, Largo dell'Università snc, 01100, Viterbo, Italy. ³Envu, 2022 ES Deutschland GmbH, Germany, Monheim, Germany. ⁴Department of Vector Biology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, United Kingdom. ⁵Centro Agricoltura Ambiente "G. Nicoli", Via Sant'Agata 835, 40014, Crevalcore, Italy. ✉e-mail: tiziana.castrignano@unitus.it

observed in other relevant arthropod pests^{2,26,27}, where point mutations in the chitin-synthase I gene (*chs1*) cause aminoacidic substitutions from wildtype Isoleucine into mutated Leucine, Methionine or Phenylalanine (i.e., I1043L; I1043M; I1043F)^{2,20,22,28,29}. A deep analysis of the homozygous I1043M-resistant strain of *Cx. pipiens* established from field populations also revealed that resistant mosquitoes had an increased cuticle thickness and constitutively up-regulate the *chs1* gene³⁰.

Since including transcripts expressed under different conditions improves the broad use of the transcriptomic references, the *de novo* transcriptome of *Cx. pipiens* presented in this paper has been assembled by including in the dataset susceptible and I1043M-resistant 4th instar larvae, which represent the target of DFB applications³¹. The bioinformatic analysis was conducted following field-best practices, employing specialized tools for quality control, sequence alignment, and transcriptomic assembly. In particular, we adopted widely recognized methodologies to ensure an accurate and comprehensive representation of the *Cx. pipiens* transcriptome. Subsequent annotation was performed using standard transcriptomic annotation procedures, ensuring a thorough understanding of the functionality of the identified transcripts. Furthermore, ORF sequences belonging to detoxifying gene families associated with insecticide resistance, such as ABC transporters, cytochrome P450, glutathione-S-transferases, UDP-glucosyltransferase, cuticular proteins and heat shock proteins were identified within the *Cx. pipiens* transcriptome.

Our resource will be a valuable reference for future studies about insecticide resistance and future comparative transcriptomic studies about mosquito biology. We cannot exclude that other mechanisms beyond target site mutation in the *chs1* gene could be associated with DFB resistance in *Cx. pipiens*. As shown by several studies, the analysis of constitutive differentially expressed genes in susceptible and resistant individuals is a main approach to identifying genes associated with insecticide resistance^{32–35}. Likewise, the analysis of differentially expressed genes under insecticide exposure could allow us to identify a larger pool of transcripts and reveal further genes associated with insecticide resistance. From a technological perspective, the transcriptome of *Cx. pipiens* will be also useful to identify new genes of interest for genetic control tools.

Methods

Mosquitoes. Mosquitoes used to generate the dataset were obtained from colonies established from field populations of *Cx. pipiens*. Immature mosquitoes were collected in the sites of Parma (Lat. 44,768382; Long. 10,319429) and Forlì (Lat. 44,21091241; Long. 12,05524495) to establish the DFB susceptible and resistant colonies, respectively. In each site, at least five breeding sites were sampled to have a representative sample of the population³⁰. The susceptible colony was homozygous for the wild-type allele I1043, while the resistant colony was homozygous for the resistant allele I1043M³⁰. Mosquitoes from 5th generation of both colonies were used for the current study.

Larval stages of both colonies were maintained in plastic trays containing 2 L of spring water and daily fed with fish food (0.85 mg/larva) (Tetra® Goldfish Granules). Adults were maintained in 45 × 45 × 45 cm Bugdorm insect-rearing cages (Watkins & Doncaster, UK) and daily fed with 10% sucrose solution and water ad libitum. Artificial blood meal was also provided to females to mature and lay eggs (Hemotek Ltd, UK). Immature and adult stages were all reared in a thermostatic chamber with constant conditions of temperature, relative humidity (RH), and photoperiod (L:D) (i.e., T = 26 ± 1 °C; RH = 70%; L:D = 16:8 hours light:dark). Two thousand L₁ larvae of each strain were put in plastic trays filled with 2 L of spring water, and the trays were set up in triplicate. In each tray, the larvae were maintained as described above, and at L₄, a pool of 10 larvae was collected and immediately stored in RNA later buffer (Thermo fisher Scientific, Ravenna, Italy) at –80 °C until RNA extraction^{36–38}.

RNA isolation, library preparation and sequencing. Pooled RNA was extracted from 6 pools of whole-body 4th-instar larvae of *Cx. pipiens* (i.e., three pools for each colony). Extraction was performed using whole-body tissues to include a larger pool of RNAs in the *de novo* transcriptome assembly of *Culex pipiens*. RNA extraction was performed using the NucleoSpin RNA Plus XS kit (Macherey-Nagel, AG), following the manufacturer's instructions. A step including a DNase treatment was also performed during RNA extraction. After RNA isolation, the quality and integrity of RNAs were checked by Qubit™ fluorimetry and the 5200 Fragment Analyzer (Agilent Technologies, USA). Libraries preparation and sequencing were performed by the Polo GGB (Polo d'Innovazione di Genomica, Genetica e Biologia), Perugia, Italy (<http://www.pologgb.com/>) using a NextSeq Illumina® platform. The libraries were prepared in accordance with the QIAseq™ Stranded mRNA Selected Kit Handbook for Illumina Paired-End Indexed Sequencing. Briefly, oligo-dT probes were covalently attached to the surface of the magnetically charged Pure mRNA Beads, and the bound RNA was then washed and eluted to provide a highly enriched pool of mRNA. Then the enriched mRNA was fragmented by heat according to input RNA quality and approximate insert size. After fragmentation, enriched mRNA was converted to cDNA using an RNase H in combination with random primers. Once the synthesis of the first strand finished, the second-strain synthesis was performed using 5' phosphorylated random primers. Samples underwent adapter ligation, and then the indexed libraries were amplified, purified, and validated. After a normalization step, libraries were pooled in equal volumes for sequencing. The Illumina NextSeq 550 sequencing system was used through the Illumina chemistry V2,2x75bp paired end run. On average 43.7 million reads for each library were obtained and they are available at the NCBI Sequence Read Archive (project ID PRJEB47420).

Pre-assembly processing stage. The bioinformatics analyses described below were conducted using high-performance computing systems through the ELIXIR-IT HPC@CINECA's call^{39–43}. The workflow of the bioinformatics pipelines, adapted from two previous studies^{44,45}, is illustrated in Fig. 1. A total of 437,233,732 pairs of reads was produced through Illumina sequencing. All of them underwent a cleaning analysis process to prepare data for assembly. The FastQC 0.11.5 tool (<http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc>) was used to assess the quality of the initial reads, which allowed to estimate the quality profiles of the RNAseq samples. Quality estimates with FastQC were performed on both raw and trimmed data. In order to eliminate adapter

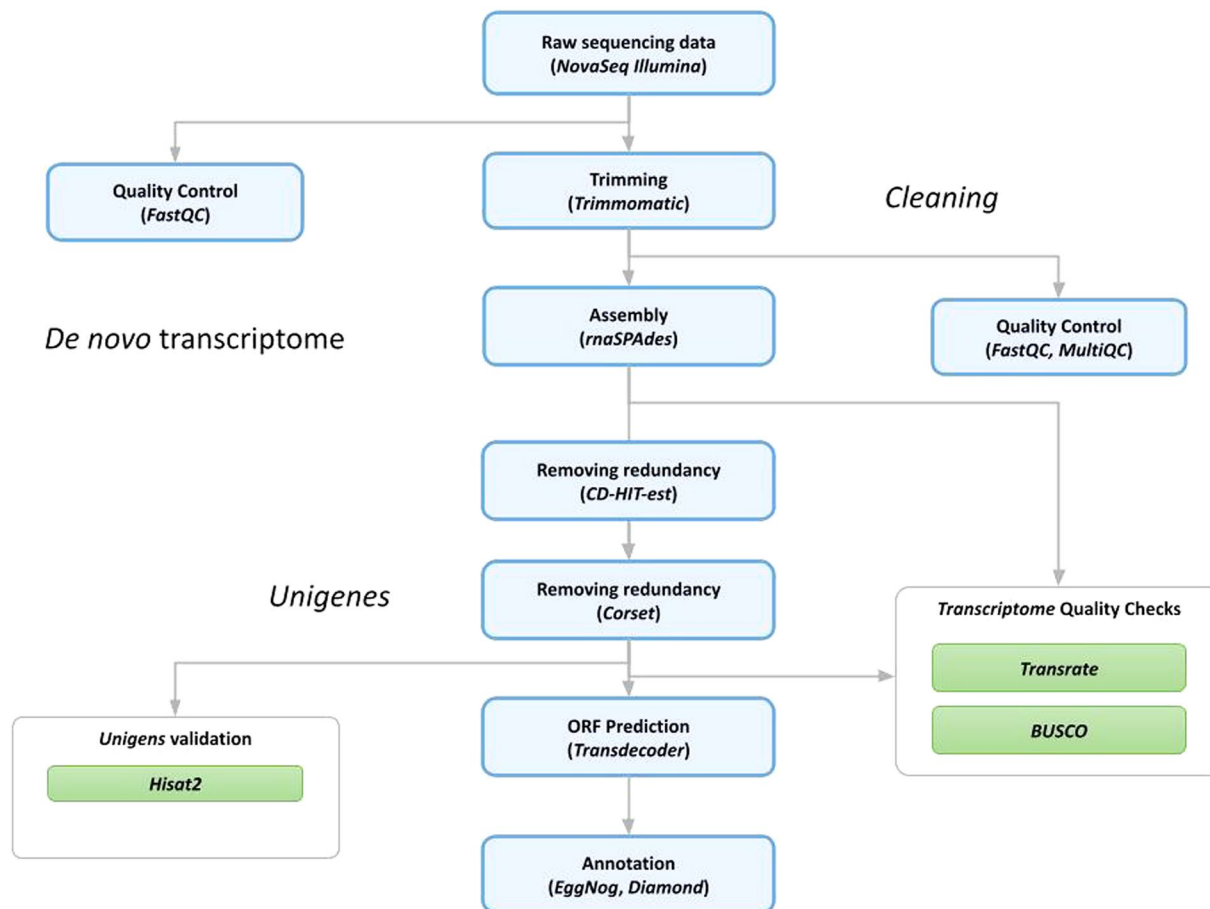


Fig. 1 Workflow of the bioinformatic pipeline for the *de novo* transcriptome assembly of *Culex pipiens*, starting from raw data and leading to annotated scripts. Each step is sequentially numbered.

Run ID	Phenotypes	RNA concentration (ng/ul)	RQN	Raw sequences	Filtered sequences (% of the total reads)
ERR10360688	DFB-Resistant	66	7,6	75,657,372	65,583,016 (87%)
ERR10360689	DFB-Resistant	60	9,6	83,210,660	76,111,468 (91%)
ERR10360692	DFB-Resistant	66	7,9	61,782,376	56,336,296 (91%)
ERR10360695	Susceptible	70	10	68,850,032	62,820,900 (91%)
ERR10360696	Susceptible	70	7,2	76,598,300	68,146,132 (89%)
ERR10360699	Susceptible	80	6,2	71,134,992	62,910,164 (88%)

Table 1. Summary of the 6 libraries deposited in the European Nucleotide Archive (ENA, BioProject: PRJEB47420)⁶⁷, in terms of number of raw and trimmed reads per sample. Information about RNA concentration and integrity (RNA Quality Number, RQN) were also shown. For RQN, higher values indicate a higher quality total RNA sample (reported on a scale of 1 to 10).

sequences and low-quality bases, an initial analysis of the raw reads was conducted using Trimmomatic, v.0.39⁴⁶ (setting the option SLIDINGWINDOW: 4: 15, MINLEN: 36, and HEADCROP: 13). All reads that were unpaired were excluded from further analysis. Executing the trimming process, a total of 391,907,976 clean reads were retained for constructing the *de novo* transcriptome assembly (which equates to approximately 90% of the original raw reads, as shown in detail in Table 1). To provide a summary quality assessment metrics view for processed data quality, results were combined across all samples and compiled into a report using the MultiQC21 v.1.9 software tool⁴⁷ (see steps 2–4 of Fig. 1). The analysis results were deposited on figshare (see Imagefile 2 on Table 2).

***De novo* transcriptome assembly and quality assessment.** Since the species under study lacks a specific annotated reference genome⁴⁸, we proceeded with the *de novo* assembly of the transcriptome. We used the assembler rnaSPAdes⁴⁹, a tool for *de novo* transcriptome assembly from RNA-Seq data, available in the SPAdes v.3.14.1 package. Using rnaSPAdes, we generated a total of 50,416 assembled transcripts, with a CG content of 47% and an N50 of 2,409 bp, as presented in Table 3.

Label	Name of data	File types	Data repository (URL)
Image file 1	Per sequence quality scores (made with MultiQC)	PDF file (.pdf)	https://doi.org/10.6084/m9.figshare.23014226
Image file 2	Mean quality scores (made with MultiQC)	PDF file (.pdf)	https://doi.org/10.6084/m9.figshare.23014184
Data file 1	rnaSPAdes RNA-seq <i>de novo</i> transcriptome assembly	Fasta file (.fasta)	https://doi.org/10.6084/m9.figshare.22512946
Data file 2	CD-HIT-est output (Unigenes)	Fasta file (.fasta)	https://doi.org/10.6084/m9.figshare.22514845
Data file 3	Corset output	Fasta file (.fasta)	https://doi.org/10.6084/m9.figshare.22515256
Data file 4	Open reading frames (ORFs) prediction	Fasta file (.cds)	https://doi.org/10.6084/m9.figshare.22515262
Data file 5	Blastx vs Nr (Functional annotation from non-redundant (Nr) NCBI)	Text file (.tsv)	https://doi.org/10.6084/m9.figshare.22561345
Data file 6	Blastx vs SwissProt (Functional annotation from Swiss-Prot)	Text file (.tsv)	https://doi.org/10.6084/m9.figshare.22561423
Data file 7	Blastx vs TrEMBL (Functional annotation from TrEMBL UniProt)	Text file (.tsv)	https://doi.org/10.6084/m9.figshare.22561450
Data file 11	Eggnog output	Text file (.tsv)	https://doi.org/10.6084/m9.figshare.23581590
Data file 12	Proteins of <i>Culex pipiens pallens</i>	Fasta file (.faa)	https://doi.org/10.6084/m9.figshare.23581704
Data file 13	Proteins of <i>Culex quinquefasciatus</i>	Fasta file (.faa)	https://doi.org/10.6084/m9.figshare.23581728
Data file 14	Proteins of <i>Aedes aegypti</i>	Fasta file (.faa)	https://doi.org/10.6084/m9.figshare.23581731
Data file 15	Proteins of <i>Anopheles gambiae</i>	Fasta file (.faa)	https://doi.org/10.6084/m9.figshare.23581743
Data file 16	Statistics comparison (OrthoFinder output)	Text file (.txt)	https://doi.org/10.6084/m9.figshare.23581752
Data file 17	Orthogroups (OrthoFinder output)	Text file (.tsv)	https://doi.org/10.6084/m9.figshare.23581770
PDF file 1	List of commands of the applied bioinformatics pipeline	PDF file (.pdf)	https://doi.org/10.6084/m9.figshare.24559084
Data file 18	Sequences of predicted ORFs associated to gene families involved in insecticide resistance	Fasta file (.fasta)	https://doi.org/10.6084/m9.figshare.25355656

Table 2. Overview of produced data files and their access on figshare⁶⁸.

Validation scores	rnaSPAdes	CD-HIT-est	Corset
Basic parameters			
Total transcripts	50416	41054	32252
N50	2409	2254	2208
GC content (%)	47	47	47
Transrate v.1.0.3			
Transrate Assembly Score	0.0828	0.1904	0.1902
Transrate Optimal Score	0.1056	0.2041	0.2070
Transrate Optimal Cutoff	0.1868	0.0682	0.0682
Good contigs	36776	38451	29969
p good contigs	0.73	0.94	0.93

Table 3. Statistics on rnaSPAdes, CD-HIT-est and Corset outputs, evaluated with the Transrate assembly validator.

In order to ensure the elimination of assembly redundancies, two filtering steps were executed. The initial step involved the utilization of CD-HIT-est (v. 4.8.1) on the output obtained from rnaSPAdes. The resulting data was then uploaded to figshare for further analysis (see Datafile 2 on Table 2). Subsequently, we employed Corset (v. 1.06)⁵⁰, a tool that has demonstrated its efficacy in a previous study⁵¹, to generate the final assembly. The Corset output showed an N50 of 2208 bp (Table 3). As a result of removing redundancies in an efficient way (two steps of removal), the final assembly contained about 63.94% of the original transcripts.

The evaluation of the assembly results involved two validation phases: one conducted after the assembly process to assess the initial assembly, and another performed after removing redundancies to evaluate the quality of the final non-redundant assembly. To accomplish this, two distinct software tools were employed, namely TransRate⁵² (v. 1.0.3) and BUSCO (Benchmarking Universal Single-Copy Orthologs)⁵³ (v. 5.4.4). These tools generate a range of metrics that serve as indicators for identifying potential errors in the assembly process and provide evidence regarding the transcriptome's quality. BUSCO offers a quantitative measure of transcriptome completeness and quality by comparing gene content against evolutionarily informed expectations derived from databases of near-universal and highly conserved protein orthologs. In this study, BUSCO analysis was performed using five orthologous gene databases: Arthropoda, Metazoa, Eukaryota, Insecta and Diptera. The degree of transcriptome completeness, as assessed by BUSCO, is presented in Table 4, while Fig. 2 illustrates the distribution of completed, fragmented, and missing genes from the four databases. Furthermore, the quality assessment of the Corset output involved the use of HISAT2 (v. 2.1.0)⁵⁴ to map trimmed reads back to the reference transcriptome (unigenes). The results of all the validation phases can be found in Table 2 and are extensively discussed in the subsequent "Technical Validation" section.

Generation of the full-length transcriptomes. Following the initial validation and evaluation phase, which involved the use of TransRate and BUSCO, the output of the assembly procedure served as the input for CD-HIT-est program⁵⁵. CD-HIT-est is a hierarchical clustering tool utilized to eliminate redundant transcripts

Busco Category	Eukaryota	Metazoa	Arthropoda	Insecta	Diptera
Complete BUSCOs (C)	243 (95.3%)	897 (94.03%)	951 (93.8%)	1249 (91,37%)	2697 (82,10%)
Complete and single-copy BUSCOs (S)	237 (92.9%)	855 (89.6%)	880 (86.9%)	1182 (86,5%)	2541 (77,5%)
Complete and duplicated BUSCOs (D)	6 (2.4%)	42 (4.4%)	71 (7.0%)	67 (4,9%)	151 (4,6%)
Fragmented BUSCOs (F)	7 (2.7%)	38 (4.0%)	30 (3.0%)	61 (4,5%)	167 (5,1%)
Missing BUSCOs (M)	5 (2.0%)	19 (2.0%)	32 (3.1%)	56 (4,1%)	420 (12,8%)
Total BUSCO groups searched	255	954	1013	1367	3285

Table 4. The BUSCO (vs. 5) validation, through the gVolante web server to five databases.

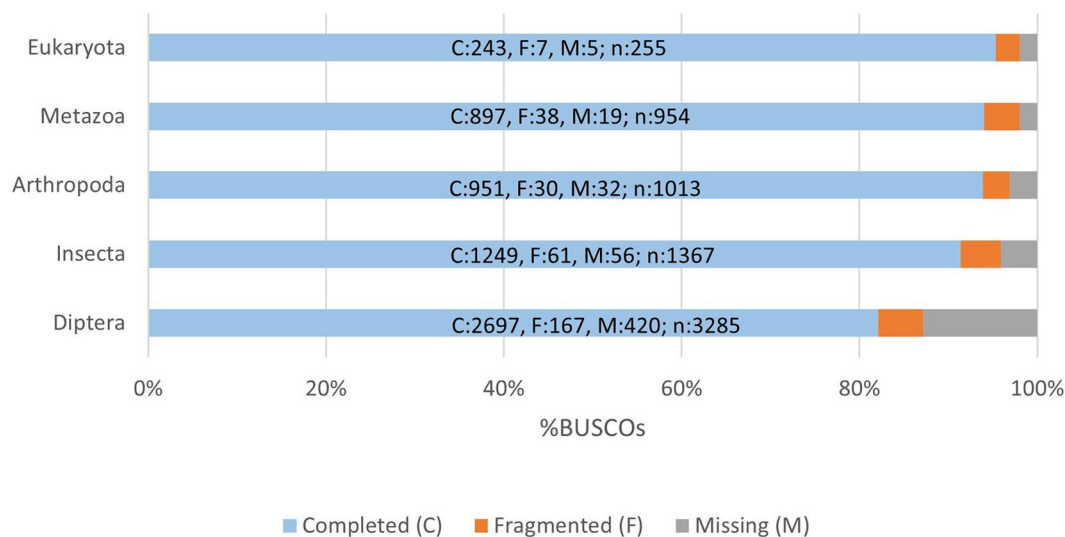


Fig. 2 BUSCO assessment results.

and fragmented assemblies, ensuring the generation of unique genes during the *de novo* assembly process. The default parameters of CD-HIT-est were employed, setting a 95% similarity threshold. To further refine the final transcriptome dataset, a subsequent hierarchical clustering phase was performed using Corset. Subsequently, the Corset output, after the validation phase with Hisat2, was subjected to analysis with TransDecoder^{56,57} (v. 5.7.0), a widely recognized tool for identifying long open reading frames (ORFs) within assembled transcripts. The default parameters of TransDecoder were used, enabling ORF prediction on both strands of assembled transcripts regardless of the sequenced library. Additionally, TransDecoder ranked the ORFs based on their completeness and assessed if the 5' end was incomplete by examining the presence of any length of amino acid codons upstream of a start codon (M) without a stop codon. The "Longest ORF" rule was adopted, selecting the translation start site with the highest 5 AUG (relative to the inframe stop codon).

Transcriptome annotation. The contigs obtained from the assembly were subjected to ORFs prediction performed with Transdecoder (vs. 5.5.0). For the predicted ORFs of the *de novo* assembly, we employed a variety of annotations. They were aligned against the Nr, SwissProt, and TrEMBL databases using the DIAMOND algorithm to extract the most relevant annotations. DIAMOND⁵⁸ is an open-source algorithm that offers a significant improvement in speed compared to BLASTX, making it highly suitable for short reads. With a comparable level of sensitivity, DIAMOND employs a double indexing approach to exhaustively determine all meaningful alignments for a given query. Traditional sequence comparison programs, such as BLASTX, typically follow a seed-and-extend paradigm. This two-phase approach involves searching for matches of seed sequences (short segments of the query sequence) in the reference database, followed by an extension phase aimed at computing a complete alignment. To configure DIAMOND, we employed the following parameter settings: DIAMOND-fast DIAMOND BLASTX -t 48 -k 250 -min-score 40 for a faster analysis, and DIAMOND-sensitive: DIAMOND BLASTX -t 48 -k 250 -sensitive -min-score 40 for a more sensitive analysis. These settings ensured efficient and accurate annotation of the assembly.

For each database, Nr, SwissProt, and TrEMBL, we selected the best annotation, resulting in the creation of an annotation matrix. The predicted ORFs were analysed using BLASTX, the tool performing nucleotide-to-protein sequence searches. The use of BLASTX against Nr, TrEMBL, and SwissProt yielded the following results: 40576 (92.7%), 41507 (94.8%), and 27289 (62.3%) contigs, respectively. These results are in line with raw reads mapping observed in the *de-novo* transcriptomes annotation of other species^{11,59,60}. Detailed information on the resulting datasets can be found in Table 5. Furthermore, an overview of the data files and datasets generated in this study, along with relevant details on the data repository and access numbers, is summarised in Table 2.

Database	Number of BLASTX results
Nr	40576 (92.7%)
TrEMBL	41507 (94.8%)
SwissProt	27289 (62.3%)

Table 5. Summary of homology annotation hits performed with BLASTX and on three different databases: Nr, SwissProt and TrEMBL.

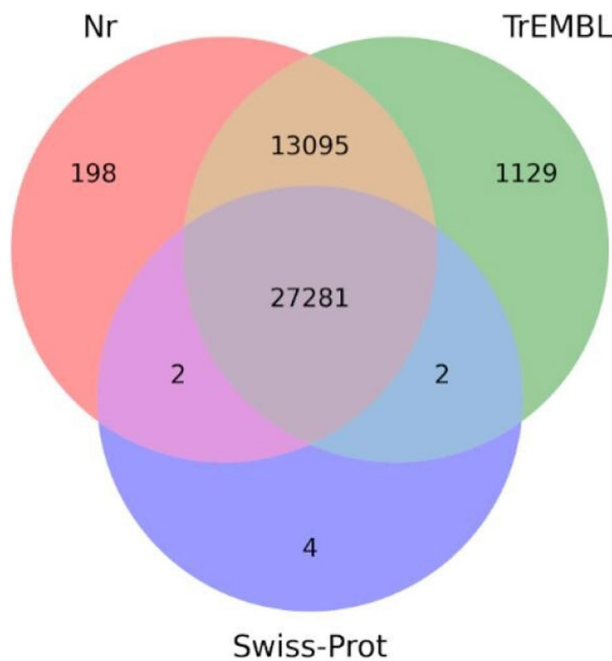


Fig. 3 Venn diagrams for the number of contigs annotated with DIAMOND (BLASTX) against the three databases: Nr, SwissProt, TrEMBL. The number of unique and shared contigs for each database is shown.

The results of the BLASTX annotation yielded a total of 27,281 sequences that were simultaneously mapped to the Nr, SwissProt, and TrEMBL databases. Venn diagrams illustrating the overlap of annotations in different databases are presented in Fig. 3, showcasing the redundancy of annotations for both DIAMOND BLASTX. Additionally, Figs. 4, 5 display the top ten most represented species and gene product hits obtained from BLASTX by aligning the transcripts with the reference database Nr. Furthermore, the ten most represented species and the ten hits of the gene product obtained with BLASTX by mapping the transcripts against the reference database Nr are shown in Fig. 4.

The total number of predicted ORFs obtained from the transcriptome assembly were also mapped onto another database of functional annotations: eggNOG (Evolutionary genealogy of genes: Non-supervised Orthologous Groups)⁶¹. The eggNOG database incorporates various taxonomic levels of orthologous groups (OGs) of proteins with functional annotations, using an algorithm that exploits previous orthologous group (COG) methodologies. Of the 43,793 total predicted ORFs obtained in our analyses, 14,966 (or 34.2%) were annotated in the eggNOG database. For details, see Datafile 11 in Table 2.

Selection of ORF sequences belonging to gene families related to insecticide resistance. Regulatory changes of genes involved in the oxidation, conjugation and extrusion of chemical compounds is a main mechanism associated with insecticide resistance⁶². By using an ad-hoc parsing script, we searched within the annotation file of predicted ORFs for gene families known to be involved in insecticide resistance. Then, we reported in Table 6 the number of found predicted ORFs in the NR annotation file by this search. Furthermore, in Table 2, we added the link to the FASTA file deposited in figshare ('Data file 18'), including the nucleotide sequences of predicted ORFs of gene families related to insecticide resistance. Further research comparing mosquitoes exposed to insecticide and under control conditions would be necessary to identify additional genes involved in DFB resistance and/or to validate the data presented here.

Comparison with other species through the orthologues. The comparison of orthologous genes with those of closely related species is a crucial step in validating the quality of a *de novo* transcriptome assembly, as outlined by several pieces of evidence in literature^{63–65}. Here, therefore we compared ORFs predicted from the *de novo* transcriptome of *Culex pipiens* with all proteins potentially transcribed based on the assembled genomes of *Aedes aegypti*, *Anopheles gambiae*, *Culex pipiens pallens* and *Culex quinquefasciatus*. The identification and orthologous grouping of all

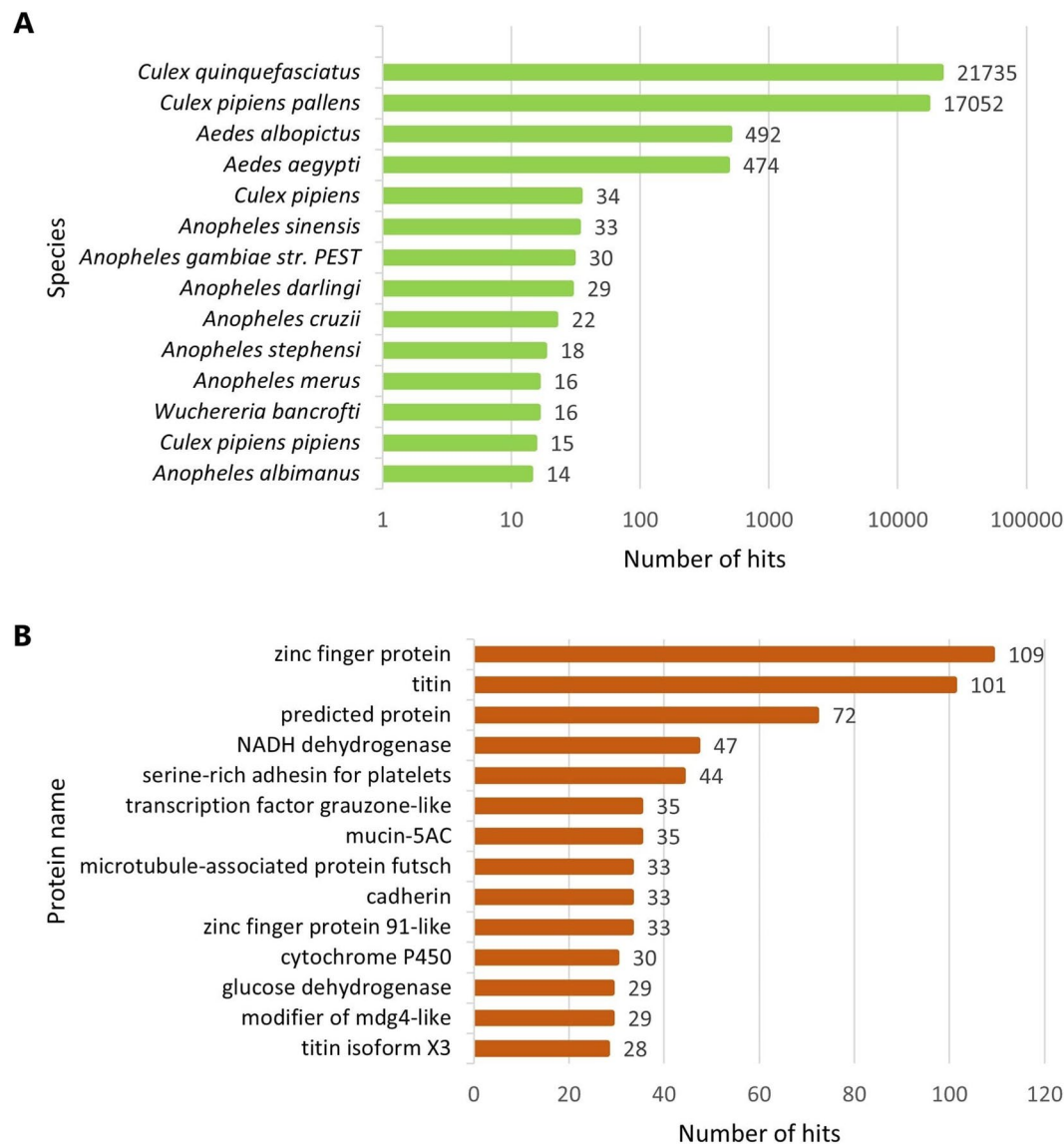


Fig. 4 Most represented species and gene product hits obtained with BLASTX by mapping the transcripts against the reference database Nr. Panel (A) shows the top 10 best species (A) and Panel (B) shows the protein hits present in the reference database.

proteins of the various species was performed with OrthoFinder⁶⁶ (v. 2.5.5). This approach also served to assess the completeness of the assembly on the basis of sequence similarity. OrthoFinder allows the identification of orthogroups, defined as a set of genes that descend from a single gene of the last common ancestor within species groups. A total of 136,663 genes were identified, grouped into 18,163 orthologues. Of these, only 12,764 genes were classified as species-specific and were grouped into 4,337 deduced orthologues. In fact, orthogroup detection showed considerable overlap in sequences in all five groups. Over 40% (7471) of the transcripts identified as putative orthologues were shared between all five species (Fig. 5). Consequently, a relatively low proportion of transcripts were identified as unique to a given assemblage (i.e., “species-specific” or “assembly-specific”); notably only 248 transcripts (1.4%) in *Aedes aegypti*, 410 transcripts (2.3%) in *Anopheles gambiae*, 243 (1.3%) in *Culex pipiens pallens*, 183 (1.0%) in *Culex quinquefasciatus* and 3253 transcripts (17.9%) in *Culex pipiens* were classified as species-specific. Therefore, the marked level of sequence overlap observed between the transcriptomes further validates the completeness and quality of the assemblage presented in this study. In addition to providing inference of the completeness of the assemblage, these results represent the first transcriptome-level comparison of four ecologically important Culicidae species. Interestingly, we found no marked difference in the number of overlapping sequences between the focal species in terms of their phylogenetic distance/proximity and each other (Fig. 5).

Data Records

All raw data generated in this project have been deposited in the European Nucleotide Archive (ENA, BioProject: PRJEB47420). The *de novo* transcriptome assembly resource is available on figshare (link: Data file 3 in Table 2). The datasets containing all files produced in this transcriptome assembly and annotation pipeline (rnaSPAdes

Detoxifying gene family	Number of ORFs
ATP-binding cassette (ABC) transporters	61
Cytochrome P450	163
Glutathione-S-transferase	16
UDP-glucosyltransferase	22
Cuticular proteins	228
Heat shock proteins	25

Table 6. Number of ORFs with Nr annotations belonging to gene families of interest known for insecticide resistance⁵⁰.

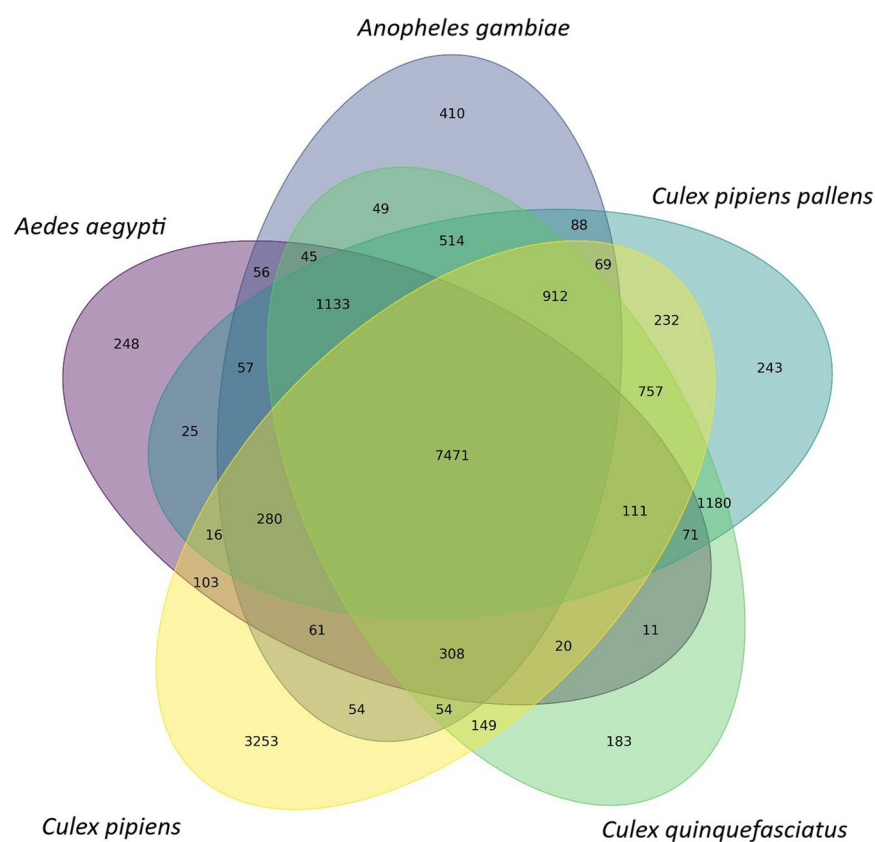


Fig. 5 Venn diagram representing the number of species-specific and overlapping protein orthogroups between the five transcriptome assemblies. The number of orthogroups were identified with OrthoFinder.

transcriptome assemblies, unigenes and functional annotation files) have also been deposited in the figshare archive (links to the pipeline results are listed in Table 2).

Technical Validation

Raw reads quality and validation of the assembly. The overall data quality was assessed using FastQC for all samples before and after cutting. Among the FastQC results, the average quality scores at each base position were above 35 (see image file 1 in Table 2). Validation of the transcriptome assembly was performed using two validation tools: BUSCO and TransRate. The results of the validation steps are shown in Table 3. BUSCO analysis was performed on five databases: Arthropoda, Metazoa, Eukaryota, Insecta and Diptera. The details of BUSCO are listed in Table 4 and some of them are represented, in the form of a histogram, in Fig. 2. A further validation evaluation was performed by mapping the clipped reads against the *de novo* assembled transcriptome of *Culex pipiens*. The HISAT2 results showed an even higher percentage of 86% (Fig. 6), confirming the high quality of the assembly. The final transcriptome (unigenes) obtained after CD-HIT-est comprised a total of 41,054 transcripts and an N50 of 2254 bp, with a completeness value of over 80% for the BUSCO evaluation in every database interrogated.

Quality control of annotation. The transcriptome was functionally annotated by running DIAMOND and eggNOG. The application of DIAMOND for annotation resulted in the identification of 43,793 predicted ORFs (for BlastX analysis) shared between the three databases. EggNOG is a comprehensive orthologous gene database

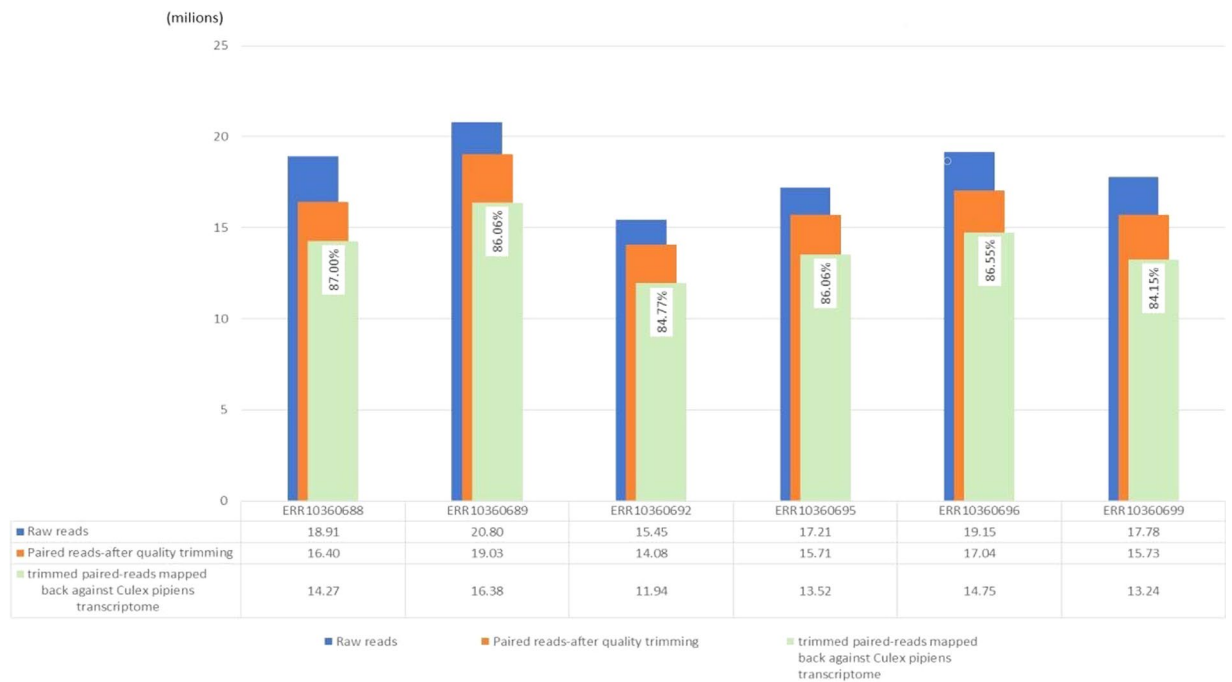


Fig. 6 For each sample, the representation of the total paired-reads is shown in blue, the total paired-reads after removal of the adapters and quality trimming is shown in orange, and the mapped trimmed paired-reads compared to the *de novo* assembled transcriptome of *Culex pipiens* is shown in green.

that provides detailed functional information for genes within each orthologous group. The EggNOG database includes a wide range of sequenced genomes from different species, providing a robust evolutionary context for our data analysis. The EggNOG analysis provided valuable insights through COG (Cluster of Orthologous Groups) assignments and KEGG (Kyoto Encyclopedia of Genes and Genomes) annotations.

Code availability

All the software programs used in this article (*de novo* transcriptome assembly, pre- and post-assembly steps and transcriptome annotation) are listed with the version in the Methods paragraph. In case of no details on parameters the programs were used with the default settings.

Received: 7 August 2023; Accepted: 19 April 2024;

Published online: 09 May 2024

References

- Nauen, R. Insecticide resistance in disease vectors of public health importance. *Pest Manag Sci.* **63**, 628e33 (2007).
- Douris, V. *et al.* Resistance mutation conserved between insects and mites unravels the benzoylurea insecticide mode of action on chitin biosynthesis. *Proc. Natl. Acad. Sci. USA* **113**, 14692–14697 (2016).
- Gantz, V. M. & Akbari, O. S. Gene editing technologies and applications for insects. *Curr. Opin. Insect Sci.* **28**, 66–72 (2018).
- Burt, A. Site-specific selfish genes as tools for the control and genetic engineering of natural populations. *Proc. Biol. Sci.* **270**, 921–928 (2003).
- Airs, P. M. & Bartholomay, L. C. RNA Interference for Mosquito and Mosquito-Borne Disease Control. *Insects* **8**, 4 (2017).
- Lester, P. J. *et al.* The potential for a CRISPR gene drive to eradicate or suppress globally invasive social wasps. *Sci. Rep.* **10**, 12398 (2020).
- Gupta A. K. & Gupta U. D. *Next Generation Sequencing and Its Applications*. In: *Animal Biotechnology*, eds. Verma, A. S. & Singh, A. pp. 345–367, (Academic Press, 2014).
- De Marco, L. *et al.* The choreography of the chemical defensome response to insecticide stress: insights into the *Anopheles stephensi* transcriptome using RNA-Seq. *Sci. Rep.* **7**, 41312 (2017).
- Morandin, C. *et al.* *De novo* transcriptome assembly and its annotation for the black ant *Formica fusca* at the larval stage. *Sci Data* **5**, 180282 (2018).
- Prado-Alvarez, M. *et al.* *De novo* transcriptome reconstruction in aquacultured early life stages of the cephalopod *Octopus vulgaris*. *Sci Data* **9**, 609 (2022).
- Palomba, M. *et al.* *De novo* transcriptome assembly and annotation of the third stage larvae of the zoonotic parasite *Anisakis pegreffii*. *BMC Res Notes* **15**, 223 (2022).
- Chowdhury, M. A. A. *et al.* Integrated transcriptome catalog of *Tenuulosa ilisha* as a resource for gene discovery and expression profiling. *Sci Data* **10**, 214 (2023).
- Pankey, M. S., Minin, V. N., Imholte, G. C., Suchard, M. A. & Oakley, T. H. Predictable transcriptome evolution in the convergent and complex bioluminescent organs of squid. *Proc Natl Acad Sci USA* **111**, E4736–E4742 (2014).
- Ingham, V. A. *et al.* Dissecting the organ specificity of insecticide resistance candidate genes in *Anopheles gambiae*: known and novel candidate genes. *BMC Genomics* **15**, 1018 (2014).
- Bharati, M. & Saha, D. Differential expression of carboxylesterases in larva and adult of *Culex quinquefasciatus* Say (Diptera: Culicidae) from sub-Himalayan West Bengal, India. *Int J Trop Insect Sci* **38**, 303–312 (2018).

16. De Marco, L. *et al.* Transcriptome of larvae representing the *Rhipicephalus sanguineus* complex. *Mol. Cell. Probes* **S0890–8508**, 30013–5 (2017).
17. Zhang, H. *et al.* Transcriptome profiling of a beach-adapted wild legume for dissecting novel mechanisms of salinity tolerance. *Sci Data* **5**, 180290 (2018).
18. Brugman, V. A. *et al.* The Role of *Culex pipiens* L. (Diptera: Culicidae) in Virus Transmission in Europe. *Int J Environ Res Public Health* **15**, 389 (2018).
19. Farajollahi, A., Fonseca, D. M., Kramer, L. D. & Marm Kilpatrick, A. “Bird biting” mosquitoes and human disease: a review of the role of *Culex pipiens* complex mosquitoes in epidemiology. *Infect Genet Evol.* **11**(7), 1577–85 (2011).
20. Harbach, R. E. *Culex pipiens*: species versus species complex taxonomic history and perspective. *J Am Mosq Control Assoc.* **28**, 10–23 (2012).
21. Grigoraki, L. *et al.* Striking diflubenzuron resistance in *Culex pipiens*, the prime vector of West Nile virus. *Sci Rep* **7**, 11699 (2017).
22. Porretta, D. *et al.* Focal distribution of diflubenzuron resistance mutations in *Culex pipiens* mosquitoes from northern Italy. *Acta Trop.* **193**, 106–112 (2019).
23. Fotakis, E. A. *et al.* Identification and detection of a novel point mutation in the Chitin Synthase gene of *Culex pipiens* associated with diflubenzuron resistance. *PLoS Negl. Trop. Dis.* **14**, e0008284 (2020).
24. Guz, N., Çağatay, N. S., Fotakis, E. A., Durmuşoğlu, E. & Vontas, J. Detection of diflubenzuron and pyrethroid resistance mutations in *Culex pipiens* from Muğla, Turkey. *Acta Trop.* **203**, 105294 (2020).
25. Vereecken, S. *et al.* Phenotypic insecticide resistance status of the *Culex pipiens* complex: a European perspective. *Parasites Vectors* **15**, 423 (2022).
26. Van Leeuwen *et al.* Population bulk segregant mapping uncovers resistance mutations and the mode of action of a chitin synthesis inhibitor in arthropods. *Proc. Natl. Acad. Sci. USA* **109**, 4407–4412 (2012).
27. Tadatsu, M. *et al.* A mutation in chitin synthase I associated with etoxazole resistance in the citrus red mite *Panonychus citri* (Acari: Tetranychidae) and its uneven geographical distribution in Japan. *Pest. Manag. Sci.* **78**, 4028–4036 (2022).
28. Mastrantonio, V. *et al.* Evolution of adaptive variation in the mosquito *Culex pipiens*: Multiple independent origins of insecticide resistance mutations. *Insects* **12**, 676 (2021).
29. Porretta, D. *et al.* Historical samples reveal a combined role of agriculture and public-health applications in vector resistance to insecticides. *Pest Manag Sci* **78**, 1567–1572 (2022).
30. Lucchesi, V. *et al.* Cuticle modifications and over-expression of the chitin-synthase gene in diflubenzuron resistant phenotype. *Insects* **13**, 1109 (2022).
31. Merzendorfer, H. Chitin synthesis inhibitors: old molecules and new developments. *Insect Sci* **20**, 121–138 (2013).
32. Li, C. X. *et al.* Identification of genes involved in pyrethroid-, propoxur-, and dichlorvos- insecticides resistance in the mosquitoes, *Culex pipiens* complex (Diptera: Culicidae). *Acta Tropica* **157**, 84–95 (2016).
33. Meng, J., Chen, X. & Zhang, C. Transcriptome-based identification and characterization of genes responding to imidacloprid in *Myzus persicae*. *Sci Rep* **9**, 13285 (2019).
34. Su, H. *et al.* Comparative transcriptome profiling reveals candidate genes related to insecticide resistance of *Glyphodes pyloalis*. *Bulletin of Entomological Research.* **110**(1), 57–67 (2020).
35. Wondji, C. S., Hearn, J., Irving, H., Wondji, M. J. & Weedall, G. RNAseq-based gene expression profiling of the *Anopheles funestus* pyrethroid-resistant strain FUM0Z highlights the predominant role of the duplicated CYP6P9a/b cytochrome P450s. *G3 (Bethesda)*. **12**(1), jkab352 (2022).
36. Cassone, B. J. *et al.* Gene expression divergence between malaria vector sibling species *Anopheles gambiae* and *An. coluzzii* from rural and urban Yaoundé Cameroon. *Mol Ecol.* **23**(9), 2242–59 (2014).
37. Wimalasiri-Yapa, B. M. C. R. *et al.* Differences in gene expression in field populations of *Wolbachia*-infected *Aedes aegypti* mosquitoes with varying release histories in northern Australia. *PLoS Negl Trop Dis* **17**(3), e0011222 (2023).
38. Main, B. J. *et al.* Genetic variation associated with increased insecticide resistance in the malaria mosquito, *Anopheles coluzzii*. *Parasites Vectors* **11**, 225 (2018).
39. Castrignano, T. *et al.* ELIXIR-IT HPC@ CINECA: high-performance computing resources for the bioinformatics community. *BMC Bioinformatics* **21**, 1–17 (2020).
40. Picardi, E., D’Antonio, M., Carrabino, D., Castrignano, T. & Pesole, G. ExpEdit: a webserver to explore human RNA editing in RNA-Seq experiments. *Bioinformatics* **27**, 1311–1312 (2011).
41. Chiara, M. *et al.* CoVaCS: a consensus variant calling system. *BMC Genom.* **19**, 1–9 (2018).
42. Castrignano, T. *et al.* ASPIC: a web resource for alternative splicing prediction and transcript isoforms characterization. *Nucleic Acids Research* **34**, W440–W443 (2006).
43. Castrignano, T. *et al.* The MEPS server for identifying protein conformational epitopes. *BMC bioinformatics* **8**, 1–5 (2007).
44. Libro, P. *et al.* First brain *de-novo* transcriptome of Tyrrhenian tree frog, *Hyla sarda*, for the study of dispersal-related behavioral variation. *Front. Ecol. Evol.* **10**, 1–6 (2022).
45. Libro, P. *et al.* *De novo* transcriptome assembly and annotation for gene discovery in *Salamandra salamandra* at the larval stage. *Sci. Data* **10**, 330 (2023).
46. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
47. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
48. Liu, W. *et al.* Chromosome-level assembly of *Culex pipiens molestus* and improved reference genome of *Culex pipiens pallens* (Culicidae, Diptera). *Mol Ecol Resour* **23**, 486–498 (2023).
49. Bushmanova, E., Antipov, D., Lapidus, A. & Prjibelski, A. D. rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. *Gigascience* **8**, giz100 (2019).
50. Davidson, N. M. & Oshlack, A. Corset: enabling differential gene expression analysis for *de novo* assembled transcriptomes. *Genome Biol.* **15**, 1–14 (2014).
51. Chiochio, A. *et al.* Brain *de novo* transcriptome assembly of a toad species showing polymorphic anti-predatory behaviour. *Sci. Data* **9**, 619 (2022).
52. Smith-Unna, R., Bournsnel, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate: Reference-free quality assessment of *de novo* transcriptome assemblies. *Genome Res.* **26**, 1134–1144 (2016).
53. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
54. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat biotechnol* **37**, 907–915 (2019).
55. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
56. Signal, B. & Kahlke, T. Borf: Improved ORF prediction in *de-novo* assembled transcriptome annotation. *BioRxiv* 2021–04 (2021).
57. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein-coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, 78 (2015).
58. Buchfink, B., Xie, C. & Huson, D. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

59. Palomba, M. *et al.* *De novo* transcriptome assembly of an Antarctic nematode for the study of thermal adaptation in marine parasites. *Scientific Data* **10**, 720 (2023).
60. Chabikwa, T. G. *et al.* *De novo* transcriptome assembly and annotation for gene discovery in avocado, macadamia and mango. *Sci Data* **7**, 9 (2020).
61. Muller, J. *et al.* eggNOG v2. 0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.* **38**, D190–D195 (2010).
62. Ranganathan, M., Narayanan, M., Kumarasamy, S. Importance of metabolic enzymes and their role in insecticide resistance. In: *New and future development in biopesticide research: Biotechnological exploration*. Springer, Singapore, pp. 243–260 (2022).
63. Mangul, S. *et al.* Transcriptome assembly and quantification from ion torrent rna-seq data. *BMC Genomics*, **15**(S5). <https://doi.org/10.1186/1471-2164-15-s5-s7> (2014).
64. O'Neil, S. T. *et al.* Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*. *BMC Genomics* **11**, 310, <https://doi.org/10.1186/1471-2164-11-310> (2010).
65. Carruthers, M. *et al.* *De novo* transcriptome assembly, annotation and comparison of four ecological and evolutionary model salmonid fish species. *BMC Genomics*, **19**(1). <https://doi.org/10.1186/s12864-017-4379-x> (2018).
66. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).
67. Libro, P. Identification of metabolic resistance mechanisms in diflubenzuron resistant *Culex pipiens* mosquitoes. *European Nucleotide Archive (ENA)* <https://www.ebi.ac.uk/ena/browser/view/PRJEB47420> (2023).
68. Libro, P. *Culex pipiens* data collection. *figshare*. <https://doi.org/10.6084/m9.figshare.c.6748110.v1> (2023).

Acknowledgements

We thank Valentina Lucchesi for technical help and Mark Eltenton for English revision. We acknowledge the CINECA and the ELIXIR-ITA HPC@CINECA initiative for providing HPC resources to our project ELIX5_porretta P.I. Daniele Porretta. This study was funded by European Union's Horizon (INFRAVEC) research and innovation programme under grant agreement No 731060 (<https://infravec2.eu/>) (Grant No. 8006) and the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. D.P. was supported by EU funding within the NextGeneration EU-MUR PNRR Extended Partnership initiative on Emerging Infectious Diseases (Project no. PE00000007, INF-ACT). V.M and S.U. received funds from the Italian Ministry for Education, University and Research (PRIN project: 2017J8JR57).

Author contributions

V.M., D.P., S.U., conceived the study; V.M., S.U. and D.P. designed the experiment; M.M. and R.B. performed sampling and sample preparation; T.C. designed and coordinated the bioinformatic analysis; P.L., J.D.M. and T.C. performed reads quality assessment, reads alignment on transcriptome, transcriptome annotation and validation; V.M., P.L. and T.C. wrote the manuscript; V.M., P.L., J.D.M., T.C., M.M., R.B., S.U. and D.P. reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024