Article

Paving the Way for Gene Silencing in Lepidoptera: Integrated Sequencing Data Unveil the Rnai Core Machinery of *Leucoptera Coffeella*.

Natália F. Martins[2], Eliza F.M.B Nascimento[1], Leonardo A. Vidal[1], Vívian S. Lucena-Leandro[1], Camila I.C.V.F. Junqueira[1], Fernanda A.F. Soares[1], Marcos J.A. Viana[2,4], Pollyana N. Mendes[1], Wagner Fontes[5], Isabelle S. Luz[5], Angela Mehta[1], Eduardo Romano, Wellington R. Clarindo[3], Juliana D. Almeida[1], Roberto C. Togawa[1] and Érika V.S. Albuquerque[1*]

[1] Embrapa Genetic Resources and Biotechnology, Brasília-DF 70770-917, Brazil and [2] Embrapa Tropical Agroindustry, Rua Dra Sara Mesquita 2270, Planalto do Pici, Fortaleza, CE 60511-110, Brazil and [3] Department of General Biology,Federal University of Viçosa, Viçosa-MG, Brazil, 36570-900, Brazil and [4] Embrapa Maize and Sorghum, Rodovia MG424 Km 45, Zona Rural, Sete Lagoas-MG 35701-970, Brazil, [5]Laboratory of Protein Chemistry and Biochemistry, Department of Cell Biology, University of Brasilia, Brasilia, Brazil.

* erika.albuquerque@embrapa.br

Abstract

Background, *Leucoptera coffeella* (Guerin-Meneville, 1842) is a moth species (Lyonetiidae, Lepidoptera) pest that causes severe losses to coffee crops. Further information about its genomic data is required to allow molecular strategies for the development of sustainable pesticides and to gain in-depth knowledge on phylogenetics. However, the closest complete genome available is within the superfamily level (Yponomeutoidea). Here we report the generation of the first long-read genome, transcriptome and proteome results of *L. coffeella* and the *in silico* analysis performed in these molecular levels to investigate genes involved in the siRNA processing. **Results**, PACBio and paired-end Illumina combined DNA sequencing from pupae samples resulted in more than 436 Gb subreads and 31Mb reads with N50 read length of 15,512 nt, mean read length 13.8 Kb and max read length 420.7 Kb. Additionally, 20Gb data of short DNA sequencing was combined to produce 1,984 contigs comprising 397 Mb in total. The longest and shortest scaffold sizes are 10,809,567 nt and 15,247 nt, respectively (mean size 200,178 nt). The N50 scaffold was 275,598 nt and the GC content was 36.10%. Predicted coding DNA sequences counted 39.930 gene models. Searching of 5286 BUSCO groups revealed 91.7 percent of completeness (single and duplicated genes combined) compared to lepidoptera genomes (lepidoptera_odb10). Flow cytometry showed the 1C DNA content is approximately 295 Mb. RNA-Seq from seven development stages resulted in 28294 identified transcripts. Additionally, proteomics from immature stages resulted in 2045 proteins matching the gene models. **Conclusions**, This first nuclear genome of the *Lyonetiidae* family brings valuable molecular resources to study Lepidoptera genomes. Genome, transcriptome and proteome sequencing to raise genome annotation precision may resolve uncovered taxonomic issues. In addition, these combined approaches provide insights into plant-insect interaction players, as horizontally transferred genes (HGT) and endosymbionts. Put together, the generated data enables the development of molecular tools towards sustainable biotechnology solutions for lepidopteran pest control.

Keywords: insect; leaf miner; Coffea; pest control; biopesticide; silencing

Introduction

A Mathematician is a device for turning coffee into theorems.

—Albert Einstein

Coffee is one of the most consumed beverages and important traded crops in the world, determining economic, social and cultural values worldwide. However, several pests threaten coffee bean production, notably the coffee leaf miner (CLM) *Leucoptera coffeella*. Producing countries report losses up to 87% of the production caused by CLM attack on both *Coffea arabica* and *Coffea canephora* plantations [1]. The CLM larvae cause damage by digging mines to feed in the mesophyll, leading to necrosis and consequent diminished photosynthesis, early senescence, and defoliation, rendering the plant debilitated [2]. The CLM attack affects both the productivity and the quality of the coffee grains [3]. High temperatures and dry periods favor the CLM reproduction, which contributes to increase the pest infestations on climate change scenarios [4]. Additionally, CLM populations frequently develop resistance to certain neurotoxic insecticides from the organophosphate group [5,6].

To gain deeper knowledge into details of the biology and genetic diversity of the CLM pest, we generated the first complete genome sequencing of *Leucoptera coffeella* NCBI:txid1178041 (Fig1). We created suitable samples to long-read sequencing and performed data assembly and annotation to obtain a reference genome to *L. coffeella*. Moreover, we explored the gene models creating transcriptomic and proteomic data to use complementary "omics" approaches.

Interesting subjects are applicable to genome-scale screening of HGT in insect genomes and Lepidopterans are a particularly spotted taxa [7–9]. HGT is widespread in insects and the presence of foreign DNA sequences, mainly virus and bacteria, based on genome sequencing confers physiological traits related to immune advantages to the insect [10].

Among Lepidoptera, evidences of host-parasite interactions in moths reveal that sequences transmitted by endosymbiotic microorganisms can affect the host fitness, metabolism, reproduction, population dynamics, and genetic diversity [11]. Concerning feeding and host adaptation, digestive and metabolic benefits conferred by HGT to functional enhancement in favor of lepidopteran insects, such as b-fructofuranosidase [12] and cysteine synthase genes [13].

Genomic material allows to identify bacterial symbionts from gut microbiota [14]. Maternally inherited bacterial symbionts may be horizontally transmitted through the host plant, but also vertically via the egg stage [15]. Gut bacteria and intracellular endosymbionts exert impacting effects on their host, especially nutrition and reproductive manipulation [16]. Symbiont engineering is trending in insect endosymbionts research to solve challenges as controlling pests to protecting pollinator health [17].

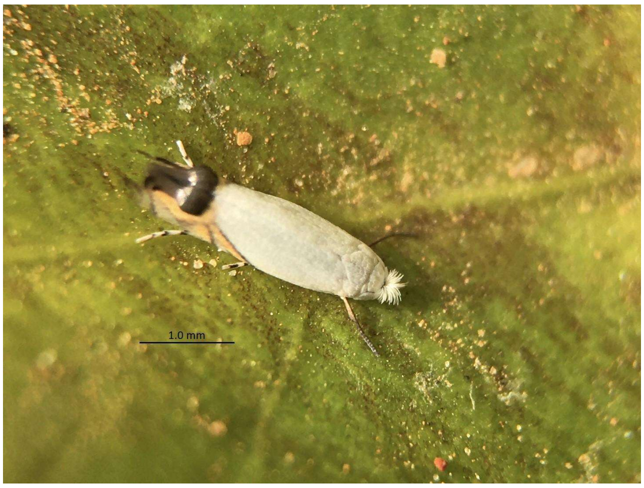

Figure 1. *Leucoptera coffeella* is a monophagous holometabolous insect that occurs naturally in Africa, along with coffee plant origins. First reported in Central America, it is currently a cosmopolitan pest present in African, Asian and Neotropical coffee producing countries.

Controlling agricultural pests through the silencing of essential genes is an increasingly sought strategy due to remarkable advantages, such as specificity to target pests, sustainability regarding human and environmental health, and applicability to commercial use by spray. The positive gain observed in relation to the methods already adopted justify the efforts to overcome some existing technical challenges [18].

To better understand and use the gene silencing technology, raising the overall knowledge about insect RNAi machinery would make a worthwhile contribution to the efficient application solutions to pest control [19]. For example, the current investigation of the main genes involved in the dsRNA processing to determine their role in the insect development, as the interference of Dicer and Argonaute affecting the molting and wing formation of *Sogatella furcifera* [20] and survival and fecundity of *Bemisia tabaci* [21]. Despite the high conservation level of the genes implicated in these processes, the comparison among agronomic pests and well characterized insects is worthy to be studied [22] because certain insects respond more efficiently to silencing due to their ability to turn on gene silencing by RNAi [23]. However, in this aspect, the order Lepidoptera lacks information from advanced studies proposed by genome initiatives of different insect pests [24].

This work presents for the first time the complete genome, the larva representative proteome and developmental stages transcriptome of an important insect pest, which permits access to silencing eligible genes. To explore the created dataset, we compared *L. coffeella* with the complete genomes of other three lepidopteran species and annotated the common genes participating in the core machinery of the dsRNA (double strain RNA) processing. Similar studies have been recently done with other insects, such as *Euschistus heros*, *Diabrotica virgifera*, *Spodoptera frugiperda* and *Nezara viridula* [23–25]. Our study focused on *L. coffeella* gene products that are involved in the uptake of dsRNA molecules and target RNA degradation, e.g. sid-like proteins, DCR-2 and AGO-2 proteins.

We generated a comprehensive data set that unveils a myriad of pathways and target genes, providing potential assets for innovative solutions to the sustainable management of insect pest control. Additionally, this information will improve the taxonomic reference to the Lyonetiidae family genomics. Furthermore, sequences may also give support to unveil feeding habits and adaptation mechanisms associated to HGT and endosymbionts. Altogether, these findings enable a better comprehension of silencing results after targeting insects and thus contribute to the biotechnology development of sustainable alternatives to control the BMC in coffee plantations.

Results and Discussion

Sequencing technologies constantly improve data acquisition. Yet, insect genome sequencing is still challenging. Up to date, there is no reference genome of the Lyonetiidae family of moths in the order Lepidoptera (butterflies and moths). In this scenario, where large read lengths require genomic DNA (gDNA) samples with high fragment size and quality [26], we found that the pupae provided better quality samples for PacBio sequencing (Precision Genomics facility). The combined long and short-read Illumina data allowed an accurate genome assembly, providing high standard sequencing results, as reflected in the N50 Read Length (15,512 nt) of the 31,637,507 subreads. The Tab1 shows a summary of the *L. coffeella* sequencing statistics.

Table 1. Genome assembly summary statistics of the leaf miner *L. coffeella*.

Sequencing technology	**PacBio SMRT, Illumina**
Number of scaffolds	1,984
Number of CDS	39,936
Assembly length (bp)	397,153,904
Longest contig (bp)	10,809,567
Shortest contig (bp)	15,247
N50 (bp)	275,598
GC %	36.09
BUSCO completeness %	91,7

There is a considerable demand for high-quality reference genome assemblies to study relevant traits of agricultural pests. According to the Genome online database (GOLD https://gold.jgi.doe.gov/), there are 1,218 genome projects from 1,039 organisms from arthropod phylum. Nevertheless, considering the superfamily Yponomeutoidea, there are 93 registered initiatives. The Taxonomy browser at NCBI shows 2 Bioprojects representative of the Lyonetiidae family, containing, respectively, only 148 protein sequences from *L. coffeella* and the mitochondrial genome of *Leucoptera malifoliella* [27]. The Leptree II initiative sequenced Yponomeutoid transcriptomes (ID 313449) and showed some phylogenetic relationships within the superfamily Yponomeutoidea [28]. A separate initiative is the 1KITE project (ID 299175), which studied the evolution and species distinction covering a thousand transcriptomic data from different insect orders, including *L. coffeella,* although sequences correspond only to adult organisms [29].

Thus, the closest reference genome to *L. coffeella* available is from *Plutella xylostella* from the Plutellidae family. Therefore, exploring the *L. coffeella* genome as a reference data substantially improves the taxonomy knowledge at the Lyonetiidae family level. Furthermore, it provides improvements in genomic data of *L. coffeella* and moth pests that unlocks unprecedented molecular tools as alternatives to chemical control.

The bioinformatics analysis showed 39,930 gene models. For functional annotation was performed by alignment against comprehensive databases (Tab2). Diamond [30] was used for genomic annotation showing 13,987 genes matches against the Ref- Seq database [31]. To analyze the completeness of the genome, we performed a BUSCO (v5.3.2 software) analysis using the annotated genome against the lepidopetra_obb10 dataset, which contains 5286 buscos. We found 4,845 complete buscos (91.7%, 3,620 single-copy and 1,225 duplicates), 55 (1%, fragmented) and 386 (7.3%, not found).

Table 2. Functional annotation tools used to analyze the 39,930 gene models

Tool	Parameters	Results
Blastp vs o_37 Lepidoptera	evalue 1e-10	31,186 hits
Blastp vs Swissprot	evalue 1e-10	17,340 hits
Diamond vs RefSeq	evalue 1e-10	13,987 hits
Diamond NR > 50 %	evalue 1e-10	23,232 hits
InterproScan v. 5.54-87.0	dp -iprlookup -goterms -pa	14,783 GOs; 20,039 PFAMs; 22,913 IPRs
Cazy	evalue <= 1e-10; - id >= 50	874 hits
InsectBase2.0	evalue <= 1e-10; - id >= 50	25,090 hits
LepBase	evalue <= 1e-10; - id >= 50	21,958 hits
Blastp vs Phibase	evalue <= 1e-10; - id >= 50	423 hits
Blastp vs BMC proteomics	evalue <= 1e-20;	2,045 hits
Blastp vs PDB	evalue <= 1e-10; - id >= 50	5,175 hits
Blastp vs BMC transcriptomics	evalue <= 1e-10; - id >= 50	28294 hits

From the 37 Lepidoptera genomes analysis, 31,186 hits gene models on identified gene families such as kinases, membrane receptors, cell communication actors and RNAi machinery (SupTab1). The gene models were supported by transcripts (28294) and by proteins (2,045) identified by BlastP alignment. We manually curated 143 gene models concerning the following specific core RNAi machinery involved in ds RNA processing: dsRNA-gated channel SID-1, endoribonuclease Dcr-1, AGO, Translin, RISC-loading complex TARBP, RNA-binding protein Staufen, DEAD/DEAH box helicase, Coatomer beta subunit appendage platform, connector enhancer of kinase suppressor of ras.

To access the genome size, we performed flow cytometry to measure the *L. coffeella* nuclear 2C value. The G0/G1 fluorescent peaks of *L. coffeella* and *D. melanogaster* showed coefficient of variation from 3.50 to 4.94. The mean nuclear 2C value were: 2C= 0.59 pg ± 0.0142 (1C = 0.295 equivalent to 288.51 Mbp) for male Viçosa, Minas Gerais, Brazil; 2C = 0.61 pg ± 0.0087 (1C = 0.305 equivalent to 293.40 Mbp) for female Viçosa, Minas Gerais, Brazil; 2C = 0.60 pg ± 0.0205 (1C = 0.300 equivalent to 585.779 Mbp) for male Barreiras, Bahia, Brazil; 2C = 0.61 pg ± 0.0176 (1C = 0.305 equiv- alent to 293.40 Mbp) for female Barreiras, Bahia, Brazil. Based on these previous results, we accomplished the flow nuclear 2C value measurements from simultaneously ganglia of the male and female of the same population, as well as only females or males of the two populations. The obtained histograms exhibited G0/G1 fluorescent peaks in the same channel. Therefore, there is no intraspecific nuclear 2C value variation, as well as we did not identify differences between the male and female nuclear 2C value. Therefore, we considered that the mean nuclear 2C value of *L. coffeella* is 2C = 0.6025 pg (1C = 0.30125 pg, 294.6225 Mbp).

In a homology search for the RNAi core genes in other insect genomes (Fig2), we compared the protein sequences deduced from *S. frugiperda*, *Plutella xylostella* and *Bombyx mori* gene models. A total of 6,029 genes were common to all of them. Furthermore, 6 common gene sets were revealed, being 3 sets involving three species. The *L. coffeella* dataset was observed in all calculated combinations. Groups of 430 genes are common between *L. coffeella*, *S. frugiperda* and *B. mori*, another group of 480 are common between *L. coffeella*, *P. xylostella* and *B. mori*. There is a third group of 618 genes among *L. coffeella, S. frugiperda* and *P. xylostella*. Three other sets involve only two species: a group of 538 genes common between *L. coffeella* and *S. frugiperda*; 725 genes between *L. coffeella* and *P. xylostella*; 575 common genes were found between *L. coffeella* and *B. mori*.

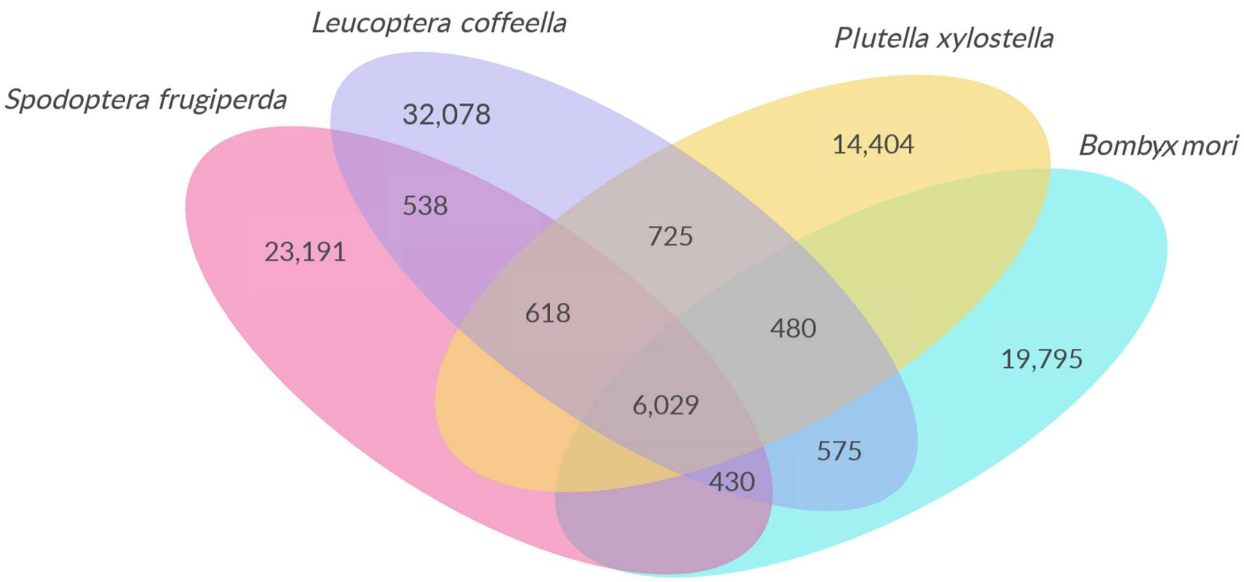

Figure 2. Venn diagram showing the orthologous groups shared among protein sequences from the genomes of *Leucoptera coffeella* and other well-studied Lepidoptera species (*S. frugiperda*, *P. xylostella* and *B. mori*).

From the common group, we found 63 common genes associated with transcripts, including Argonaute, RNA helicase, Translin and DEAD (SupTab2).

In this study, we used a molecular multilevel approach to deepen the understanding of the CLM biology. We have focused on a specific pathway of the RNAi machinery related to dsRNA processing considering the functional application to develop biopesticides by silencing. Further exploration of the proteomic data obtained can shed light on other important mechanisms for the pest survival and colonization of the host plant.

Methods

Genome

Genomic DNA (gDNA) was obtained from a pool of individuals at the pupae stage, collected in *Coffea arabica* leaves. A modified protocol of the E.Z.N.A insect DNA kit (Omega BioTek, Cat. No.: D0926-02) was used for the extractions, as described in [32]. Whole pupae tissues were macerated in the CTL buffer (provided by kit). Then, protein digestion was performed with proteinase K (Invitrogen, Cat. No.: AM2544). gDNA sample was precipitated with chloroform and isopropyl alcohol, treated with RNAse A (Invitrogen, Cat. No.: 12091021) and purified on a Mini Column HiBand DNA and eluted with MilliQ water. Quantity and quality were analyzed by NanoDrop, Qubit, agarose gel electrophoresis and Femto Pulse. The high weight gDNA was stored at 4 °C until shipping on ice to the sequencing facility (Precision Genomics).

The sequencing of the whole genome of *L. coffella* was performed with both long read from DNA Link´s PacBio RSII and short read from Illumina HISeq for the error correction. For the long read sequencing the samples were prepared using PacBio HiFi Express Prep kit and submitted to PacBio Sequel2. For the short read, a TruSeq DNA PCR-Free 350bp kit using paired end library (2x151) and insert size approximately 350 was used and sequenced by Illumina NovaSeq6000. All libraries were prepared according to the service provider DNALINK* procedures from sample standards of quantity, concentration, and quality. The loading statistics show a high productivity (81.46 percent) over the expected 60 of the SMRT Cell wells filled with templates. The PacBio sequencing generated 31,637,507 reads containing 436,466,105,362 bases. The maximum generated read length was 420,664 and N50 and mean read length was 15,512 and 13,796 respectively. For the Illumina sequencing, the results for R1 and R2 were 68,594,151 reads containing 10,357,716,801 bases.

Subsequently, the sequenced data was assembled using DNALINK pipeline customized service including the processing methods for *de novo* Assembly and Error Correction such as Zero-Mode Waveguide (ZMW), productivity, insert length, fragment contamination, filtering and removal of adapters. Falcon Genome Assembly Tool Kit was used to perform the assembly, followed by two rounds of purge haplotig to reach the final assembly. A pilon software [33] performed the improvement of the assembled genome using Illumina short reads. An additional check using the BUSCO software [34] against the lepidoptera_odb10 dataset gave a 91.7% completeness result. The *L. coffeella* was assembled into a 3.9 Gb dataset consisting of 1,984 scaffolds as summarized on Table 1. The draft assembly is available at GenBank within the BioProject ID PRJNA832598.

Automatic annotation and gene prediction used RepeatMasker to softmask the repeat regions, followed by Braker2 v 2.1.6 pipeline [35]. Homology search on the assembled gene models were performed using Diamond and Blast search against functional annotation datasets: Lepbase [36], NR and RefSeq (Genbank)[37], SwissProt [38], Interpro and IPR, Interpro Pfam [39], PDV [40], PDB [41] and PhiBase [42]. From Lepbase, we downloaded proteins of 37 species, totalizing 598,319 functionally annotated protein sequences. InsectBase 2.0, a substantially improved database for insect gene resources provided proteins from 716 organisms, resulting in 15,130,231 functionally annotated protein sequences. Diamond was used to perform a blastp on these three databases using e-value parameters 1e-10, identity of at least 50 percent and only the best hit of each sequence was recovered. To understand the genomic features and related biology, we compared *L. coffella* data with three different insect species available at NCBI/Ref-Seq: *S. frugiperda*, *P. xylostella* and *B. mori.* The comparative alignment considered at least 70% identity and an e-value of 10-8 and retrieved the best hit in a tabular output Manual verification of the common genes used alignment information from InterprotScan result.

Nuclear 2C value ganglia were dissected in commercial saline solution from adult males and females of *L. coffeella* of two populations (Viçosa MG and Barreiras BA), and of female *Drosophila melanogaster* (standard 2C = 0.36 pg for flow cytometry, https://www.genomesize.com/results.php?page=1). Nuclei were isolated from each ganglion (external standard) or simultaneously (internal standard) from crushing in 100 µL OTTO I [35] nuclear extraction buffer (0.1 M citric acid, 0.5 percent Tween 20, 2.0 mM dithiothreitol and 50 µg mL-1 RNAse, pH = 2.3), and the nuclei sus- pensions were incubated for 5 min [43,44]. One milliliter of OTTO I was added, and the suspensions were filtered through a 30 µm nylon filter into a 2.0 mL microtube and centrifuged at 100 xg for 5 min. The supernatant was poured out, and the pellets were homogenized in 100 µL OTTO I, kept at 10 min and filtered through a 20 µm filter into a cytometry tube. The nuclei suspensions were stained for 30 min in the dark with 500 µL OTTO II [35] staining buffer (0.4 µM Na2HPO4.2H2O, 75 µM propidium iodide and 50 µg mL-1 RNAse, pH = 7.8, [43,44]. We processed the nuclei suspensions in a BD Accuri™ C6 Flow Cytometer (Accuri, Belgium) equipped with a 488 nm laser source to promote emissions at FL2 (615 – 670 nm) and FL3 (> 670 nm). Flow cytometry his- tograms were analyzed using the BD Accuri™ C6 software. G0/G1 fluorescent peaks from each *L. coffeella* and *D. melanogaster* with coefficient of variation below 5 percent were considered for nuclear 2C value measurement. Mean nuclear 2C values were converted to Mbp, considering that 1C pg is equivalent to 978 Mbp [45].

Transcriptome

RNA samples were processed using illumina technology to prepare libraries, which were sequenced on Next-Generation DNA sequencing (NGS) instruments. The samples from 7 libraries of different stages of insect development, as described: Larval 1 (AVGE-1), Larval 2 (AVGE- 2) Larval 3 (AVGE-3), Larval 4 (AVGE-4), pupa (AVGE-5), male

(AVGE-6) and female (AVGE-7). Sequencing was performed in Illumina NovaSeq 6000 with a number of cycles 2x150+8+8. Base calling pipeline used is composed of NovaSeq Control Software 1.7.5, RTA v3.4.4 and bcl2fastq2.20 v2.20.0.422.

In the present study the AVGE-5 sample was selected for analysis as it was the same sample used in the DNA sequencing. The minimum quality required for this sequencing was at least 85% of the bases with Q30, which is associated with an expected error rate of 1 in 1000 (0.1%), specifically for the AVGE-5 sample, 88% were obtained from Q30 and its average quality was Q35. The number of reads in the quality standard was 54,435,117 reads.

The Star software version 2.7.10a [46] was used to map the AVGE-5 sample library against the sequenced genome of *Leuptera coffeella*, the result of this mapping was a BAM file, with the count of reads mapped to the genome, and this result was used as input by the TrinityGG version 2.14.0 software [47] for extracting transcripts. As many transcripts and many isoforms were generated, EvidentialGene tr2aacds.pl pipeline script [48] was used to group these transcripts and 21,065 were recovered. Blast software was used with default parameters and retrieved only the best hit of the transcripts file generated by TinityGG against the gene models generated by Braker2 in a previous step of DNA sequencing. 20,155 transcripts matched 13,338 gene models from DNA sequencing.

Proteome

The proteins of different stages of L2 larvae of *L. coffeella* (L2, L3, L4 and pupae) were extracted according to Mot and Vanderleyden [49]. Approximately 100 mg of larvae from each stage were added to 750 µL of extraction buffer pH 7.6 (0.7 M sucrose, 0.5 M Tris, 30 mM HCl, 50 mM EDTA, 0.1 M KCl and 40 mM DTT). The same volume (750 µL) of phenol (equilibrated at pH 8.0) was added and the samples were shaken in a vortex for 15 minutes and after that centrifuged at 10.000 x g for 3 minutes. Proteins were precipitated in 0.1 M ammonium acetate in methanol, washed with 80% acetone and solubilized in 50 mM ammonium bicarbonate (NH4HCO3 pH 8.5). RapiGest SF – Waters (0.2% v/v) was added, and samples were treated with dithiothreitol and iodoacetamide. Approximately 80 µg of total proteins were digested with trypsin (1 µg) at 37 °C for 19 hours. Proteins were quantified with a Qubit Fluorometer (Invitrogen), following the manufacturer's instructions. Samples were desalted [50], solubilized with 0.1% formic acid and injected into ESI LC/MS.

A chromatographic system (Dionex Ultimate 3000 RSLC nano UPLC, Thermo, USA), configured with a trap column (3 cm x 100 μm) containing 5 μm, 120 Å C18 particles (Reprosil-Pur, Dr. Maich GmbH) was used. Peptides (3 ug) were eluted from the trap column to the analytical column (24 cm x 75 μm), containing 3 μm C18, 120 Å (Reprosil-Pur, Dr. Maich GmbH). A linear elution gradient between solvents A (0.1% formic acid in 2% acetonitrile/water) and B (0.1% formic acid in 80% acetonitrile/water) from 2% B to 35% B for 155 min was performed. The fractions were eluted into an Orbitrap Elite mass spectrometer (Thermo Scientific™), configured in DDA (data dependent acquisition) mode. MS1 spectra were acquired in the orbitrap analyzer, with a resolution of 120000 and m/z range between 300 and 1650. The 15 most intense ions, above the intensity limit of 3000 were fragmented, generating MS2 spectra. The reanalysis of fragmented ions was inhibited by dynamic exclusion, favoring the identification of less abundant peptides. Spectra were aligned and peptides were quantified using the Progenesis® QI for proteomics v.1.0 software (Nonlinear Dynamics). For protein identification the Peaks® 7.0 software (Bioinformatics Solutions Inc.) was used. The sequences were deduced from the fragmentation information and the search performed in the database restricted to the genome sequence of the leaf miner (Taxon ID 3917), obtained in this study. The sequences were subjected to removal of redundant sequences using the FASTAtools software (http://lbqp.unb.br/LBQPtools/), resulting in 39978 sequences. de novo sequencing and PSM were performed with the following parameters: tolerance for the mass of the precursor of 10 ppm, the fragments of 0.05 Da, tolerance of up to 2 lost cleavages, carbamidomethylation of cysteines as fixed modification and methionine oxidation as variable modification. Gene Ontology was obtained by pfam2go (https://rdrr.io/github/missuse/ragp/man/pfam2go.html) software.

Availability of supporting data and materials

The data underlying this article are available at NCBI with the BioProject ID PRJNA832598 and the information management at http://lbi.cenargen.embrapa.br/CLM.

The proteomic data is available via ProteomeXchange (PXD035993).

Declarations

Competing Interests

The authors declare that they have no competing interests

Funding

This work was supported by the CONCAFE consortium (Project funding and Fellowships to P.N.M. in part by CNPq grant fellowships grants and FAPDF (publication grant).

Author's Contributions

Conceptualization—N.F.M., R.C.T., E.V.S.A.; Data curation—N.F.M., M.J.A.V., R.C.T., E.V.S.A.; Formal Analysis—R.C.T., W.F., M.J.A.V., N.F.M.; Funding acquisition—N.F.M., E.R., E.V.S.A; Investigation—N.F.M., E.F.M.B.N., L.A.V., V.S.L.L., C.I.C.V.F.J., F.A.F.S., P.N.M., I.S.Luz., W.F., A.M., W.R.C., S.R., J.D.A., E.V.S.A.; Methodology—N.F.M., M.J.A.V., R.C.T., A.M., W.F., W.R.C., E.V.S.A.; Project administration—E.V.S.A.; Resources—N.F.M., W.R.C., J.D.A., R.C.T., E.V.S.A.; Writing—original draft—N.F.M., M.J.A.V., W.R.C., S.R., J.D.A., A.M., R.C.T., E.V.S.A.; Writing—review and editing—N.F.M., M.J.A.V., R.C.T., E.V.S.A

Acknowledgements

We wish to thank Adriano Veiga from Embrapa for the insect collection; to Genomix Data, for sequencing and analysis. We also thank Andrea Queiroz Maranhão from UnB for technical advice on DNA and RNA quality analysis and Diana Fernandez for modifications on DNA extraction protocols.

References

1. Dantas, J.; Motta, I.O.; Vidal, L.A.; Nascimento, E.F.M.B.; Bilio, J.; Pupe, J.M.; Veiga, A.; Carvalho, C.; Lopes, R.B.; Rocha, T.L.; et al. A Comprehensive Review of the Coffee Leaf Miner Leucoptera Coffeella (Lepidoptera: Lyonetiidae)—A Major Pest for the Coffee Crop in Brazil and Others Neotropical Countries. *Insects* **2021**, *12*, 1130, doi:10.3390/insects12121130.
2. Souza, J.C. Bicho-mineiro-do-cafeeiro: biologia, danos e manejo integrado. *Bicho-mineiro-do-cafeeiro: biologia, danos e manejo integrado*. **1998**, 48.
3. Paulo Rebelles Reis *Reflexos Da Incidência de Pragas Na Qualidade Do Café*. 2011, pp. 104–112.

4. Leite, S.A.; Guedes, R.N.C.; Santos, M.P. dos; Costa, D.R. da; Moreira, A.A.; Matsumoto, S.N.; Lemos, O.L.; Castellani, M.A. Profile of Coffee Crops and Management of the Neotropical Coffee Leaf Miner, Leucoptera Coffeella. *Sustainability* **2020**, *12*, 8011, doi:10.3390/su12198011.

5. Daianna P. CostaTrdan, F.L.F.; Flávia M. Alves, É.M. da S.; Liliane E. Visôtto In *Resistance to Insecticides in Populations of the Coffee Leafminer*; IntechOpen, 2016; pp. 3–17 ISBN 978-953-51-2258-6.

6. Leite, S.A.; Dos Santos, M.P.; Resende-Silva, G.A.; da Costa, D.R.; Moreira, A.A.; Lemos, O.L.; Guedes, R.N.C.; Castellani, M.A. Area-Wide Survey of Chlorantraniliprole Resistance and Control Failure Likelihood of the Neotropical Coffee Leaf Miner Leucoptera Coffeella (Lepidoptera: Lyonetiidae). *J. Econ. Entomol.* **2020**, *113*, 1399–1410, doi:10.1093/jee/toaa017.

7. Li, Y.; Liu, Z.; Liu, C.; Shi, Z.; Pang, L.; Chen, C.; Chen, Y.; Pan, R.; Zhou, W.; Chen, X.; et al. HGT Is Widespread in Insects and Contributes to Male Courtship in Lepidopterans. *Cell* **2022**, *185*, 2975-2987.e10, doi:10.1016/j.cell.2022.06.014.

8. Drezen, J.-M.; Josse, T.; Bézier, A.; Gauthier, J.; Huguet, E.; Herniou, E.A. Impact of Lateral Transfers on the Genomes of Lepidoptera. *Genes* **2017**, *8*, 315, doi:10.3390/genes8110315.

9. Reiss, D.; Mialdea, G.; Miele, V.; Vienne, D.M. de; Peccoud, J.; Gilbert, C.; Duret, L.; Charlat, S. Global Survey of Mobile DNA Horizontal Transfer in Arthropods Reveals Lepidoptera as a Prime Hotspot. *PLOS Genetics* **2019**, *15*, e1007965, doi:10.1371/journal.pgen.1007965.

10. Lelio, I.D.; Illiano, A.; Astarita, F.; Gianfranceschi, L.; Horner, D.; Varricchio, P.; Amoresano, A.; Pucci, P.; Pennacchio, F.; Caccia, S. Evolution of an Insect Immune Barrier through Horizontal Gene Transfer Mediated by a Parasitic Wasp. *PLOS Genetics* **2019**, *15*, e1007998, doi:10.1371/journal.pgen.1007998.

11. Ghanavi, H.R.; Twort, V.G.; Duplouy, A. Exploring Bycatch Diversity of Organisms in Whole Genome Sequencing of Erebidae Moths (Lepidoptera). *Sci Rep* **2021**, *11*, 24499, doi:10.1038/s41598-021-03327-3.

12. Dai, X.; Kiuchi, T.; Zhou, Y.; Jia, S.; Xu, Y.; Katsuma, S.; Shimada, T.; Wang, H. Horizontal Gene Transfer and Gene Duplication of β-Fructofuranosidase Confer Lepidopteran Insects Metabolic Benefits. *Molecular Biology and Evolution* **2021**, *38*, 2897–2914, doi:10.1093/molbev/msab080.

13. Li, Y.; Zhou, Y.; Jing, W.; Xu, S.; Jin, Y.; Xu, Y.; Wang, H. Horizontally Acquired Cysteine Synthase Genes Undergo Functional Divergence in Lepidopteran Herbivores. *Heredity* **2021**, *127*, 21–34, doi:10.1038/s41437-021-00430-z.

14. One's Trash Is Someone Else's Treasure: Sequence Read Archives from Lepidoptera Genomes Provide Material for Genome Reconstruction of Their Endosymbionts Available online: https://www.researchsquare.com (accessed on 28 August 2022).

15. Paniagua Voirol, L.R.; Frago, E.; Kaltenpoth, M.; Hilker, M.; Fatouros, N.E. Bacterial Symbionts in Lepidoptera: Their Diversity, Transmission, and Impact on the Host. *Frontiers in Microbiology* **2018**, *9*.

16. Duplouy, A.; Hornett, E.A. Uncovering the Hidden Players in Lepidoptera Biology: The Heritable Microbial Endosymbionts. *PeerJ* **2018**, *6*, e4629, doi:10.7717/peerj.4629.

17. Elston, K.M.; Leonard, S.P.; Geng, P.; Bialik, S.B.; Robinson, E.; Barrick, J.E. Engineering Insects from the Endosymbiont Out. *Trends Microbiol* **2022**, *30*, 79–96, doi:10.1016/j.tim.2021.05.004.

18. Liu, S.; Jaouannet, M.; Dempsey, D.A.; Imani, J.; Coustau, C.; Kogel, K.-H. RNA-Based Technologies for Insect Control in Plant Production. *Biotechnology Advances* **2020**, *39*, 107463, doi:10.1016/j.biotechadv.2019.107463.

19. Hernández-Soto, A.; Chacón-Cerdas, R. RNAi Crop Protection Advances. *International Journal of Molecular Sciences* **2021**, *22*, 12148, doi:10.3390/ijms222212148.

20. Zeng, Q.-H.; Long, G.-Y.; Yang, X.-B.; Jia, Z.-Y.; Jin, D.-C.; Yang, H. SfDicer2 RNA Interference Inhibits Molting and Wing Expansion in Sogatella Furcifera. *Insects* **2022**, *13*, 677, doi:10.3390/insects13080677.

21. Bidari, F.; Fathipour, Y.; Asgari, S.; Mehrabadi, M. Targeting the MicroRNA Pathway Core Genes, Dicer 1 and Argonaute 1, Negatively Affects the Survival and Fecundity of Bemisia Tabaci. *Pest Management Science* **2022**, doi:10.1002/ps.7041.

22. Arraes, F.B.M.; Martins-de-Sa, D.; Noriega Vasquez, D.D.; Melo, B.P.; Faheem, M.; de Macedo, L.L.P.; Morgante, C.V.; Barbosa, J.A.R.G.; Togawa, R.C.; Moreira, V.J.V.; et al. Dissecting Protein Domain Variability in the Core RNA Interference Machinery of Five Insect Orders. *RNA Biology* **2021**, *18*, 1653–1681, doi:10.1080/15476286.2020.1861816.

23. Davis-Vogel, C.; Van Allen, B.; Van Hemert, J.L.; Sethi, A.; Nelson, M.E.; Sashital, D.G. Identification and Comparison of Key RNA Interference Machinery from Western Corn Rootworm, Fall Armyworm, and Southern Green Stink Bug. *PLoS One* **2018**, *13*, e0203160, doi:10.1371/journal.pone.0203160.

24. Childers, A.K.; Geib, S.M.; Sim, S.B.; Poelchau, M.F.; Coates, B.S.; Simmonds, T.J.; Scully, E.D.; Smith, T.P.L.; Childers, C.P.; Corpuz, R.L.; et al. The USDA-ARS Ag100Pest Initiative: High-Quality Genome Assemblies for Agricultural Pest Arthropod Research. *Insects* **2021**, *12*, 626, doi:10.3390/insects12070626.

25. Cagliari, D.; Dias, N.P.; dos Santos, E.Á.; Rickes, L.N.; Kremer, F.S.; Farias, J.R.; Lenz, G.; Galdeano, D.M.; Garcia, F.R.M.; Smagghe, G.; et al. First Transcriptome of the Neotropical Pest Euschistus Heros (Hemiptera: Pentatomidae) with Dissection of Its SiRNA Machinery. *Sci Rep* **2020**, *10*, 4856, doi:10.1038/s41598-020-60078-3.

26. Oppert, B.; Stoss, S.; Monk, A.; Smith, T. Optimized Extraction of Insect Genomic DNA for Long-Read Sequencing. *Methods and Protocols* **2019**, *2*, 89, doi:10.3390/mps2040089.

27. Wu, Y.-P.; Zhao, J.-L.; Su, T.-J.; Li, J.; Yu, F.; Chesters, D.; Fan, R.-J.; Chen, M.-C.; Wu, C.-S.; Zhu, C.-D. The Complete Mitochondrial Genome of Leucoptera Malifoliella Costa (Lepidoptera: Lyonetiidae). *DNA Cell Biol* **2012**, *31*, 1508–1522, doi:10.1089/dna.2012.1642.

28. Mitter, C.; Davis, D.R.; Cummings, M.P. Phylogeny and Evolution of Lepidoptera. *Annu Rev Entomol* **2017**, *62*, 265–283, doi:10.1146/annurev-ento-031616-035125.

29. Kawahara, A.Y.; Plotkin, D.; Espeland, M.; Meusemann, K.; Toussaint, E.F.A.; Donath, A.; Gimnich, F.; Frandsen, P.B.; Zwick, A.; Reis, M. dos; et al. Phylogenomics Reveals the Evolutionary Timing and Pattern of Butterflies and Moths. *PNAS* **2019**, *116*, 22657–22663, doi:10.1073/pnas.1907847116.

30. Buchfink, B.; Reuter, K.; Drost, H.-G. Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND. *Nat Methods* **2021**, *18*, 366–368, doi:10.1038/s41592-021-01101-x.

31. Li, W.; O'Neill, K.R.; Haft, D.H.; DiCuccio, M.; Chetvernin, V.; Badretdin, A.; Coulouris, G.; Chitsaz, F.; Derbyshire, M.K.; Durkin, A.S.; et al. RefSeq: Expanding the Prokaryotic Genome Annotation Pipeline Reach with Protein Family Model Curation. *Nucleic Acids Res* **2021**, *49*, D1020–D1028, doi:10.1093/nar/gkaa1105.

32. Nascimento, E.F. de M.B. do; Lucena-Leandro, V. dos S.; Vidal, L.A.; Junqueira, C.I.C.V.F.; Almeida, D.J.; Albuquerque, E.V.S. Optimization of Insect Genomic DNA and Total RNA Extraction Protocols for High Fidelity Gene Sequencing 2022.

33. Walker, B.J.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C.A.; Zeng, Q.; Wortman, J.; Young, S.K.; et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One* **2014**, *9*, e112963, doi:10.1371/journal.pone.0112963.

34. 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* **2015**, *31*, 3210–3212, doi:10.1093/bioinformatics/btv351.

35. Otto, F. DAPI Staining of Fixed Cells for High-Resolution Flow Cytometry of Nuclear DNA. *Methods Cell Biol* **1990**, *33*, 105–110, doi:10.1016/s0091-679x(08)60516-6.

36. Challi, R.J.; Kumar, S.; Dasmahapatra, K.K.; Jiggins, C.D.; Blaxter, M. Lepbase: The Lepidopteran Genome Database 2016, 056994.

37. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **2016**, *44*, D7–D19, doi:10.1093/nar/gkv1290.

38. Duvaud, S.; Gabella, C.; Lisacek, F.; Stockinger, H.; Ioannidis, V.; Durinx, C. Expasy, the Swiss Bioinformatics Resource Portal, as Designed by Its Users. *Nucleic Acids Research* **2021**, *49*, W216–W227, doi:10.1093/nar/gkab225.

39. Blum, M.; Chang, H.-Y.; Chuguransky, S.; Grego, T.; Kandasaamy, S.; Mitchell, A.; Nuka, G.; Paysan-Lafosse, T.; Qureshi, M.; Raj, S.; et al. The InterPro Protein Families and Domains Database: 20 Years On. *Nucleic Acids Research* **2021**, *49*, D344–D354, doi:10.1093/nar/gkaa977.

40. Li, K.; Vaudel, M.; Zhang, B.; Ren, Y.; Wen, B. PDV: An Integrative Proteomics Data Viewer. *Bioinformatics* **2019**, *35*, 1249–1251, doi:10.1093/bioinformatics/bty770.

41. wwPDB consortium Protein Data Bank: The Single Global Archive for 3D Macromolecular Structure Data. *Nucleic Acids Research* **2019**, *47*, D520–D528, doi:10.1093/nar/gky949.

42. Urban, M.; Cuzick, A.; Seager, J.; Wood, V.; Rutherford, K.; Venkatesh, S.Y.; Sahu, J.; Iyer, S.V.; Khamari, L.; De Silva, N.; et al. PHI-Base in 2022: A Multi-Species Phenotype Database for Pathogen–Host Interactions. *Nucleic Acids Research* **2022**, *50*, D837–D847, doi:10.1093/nar/gkab1037.

43. Cunha, M.S.; Soares, F. a. F.; Clarindo, W.R.; Campos, L. a. O.; Lopes, D.M. Robertsonian Rearrangements in Neotropical Meliponini Karyotype Evolution (Hymenoptera: Apidae: Meliponini). *Insect Mol Biol* **2021**, *30*, 379–389, doi:10.1111/imb.12702.

44. Lopes, D.M.; de Carvalho, C.R.; Clarindo, W.R.; Praça, M.M.; Tavares, M.G. Genome Size Estimation of Three Stingless Bee Species (Hymenoptera, Meliponinae) by Flow Cytometry. *Apidologie* **2009**, *40*, 517–523, doi:10.1051/apido/2009030.

45. Praça-Fontes, M.M.; Carvalho, C.R.; Clarindo, W.R. C-Value Reassessment of Plant Standards: An Image Cytometry Approach. *Plant Cell Rep* **2011**, *30*, 2303–2312, doi:10.1007/s00299-011-1135-6.

46. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast Universal RNA-Seq Aligner. *Bioinformatics* **2013**, *29*, 15–21, doi:10.1093/bioinformatics/bts635.

47. Trinity-v2.14.0 2022.

48. Gilbert, D. <p>Gene-Omes Built from MRNA Seq Not Genome DNA</P>. *F1000Research* **2016**, *5*, doi:10.7490/f1000research.1112594.1.

49. Mot, R.D.; Vanderleyden, J. Application of Two-Dimensional Protein Analysis for Strain Fingerprinting and Mutant Analysis of Azospirillum Species. *Can. J. Microbiol.* **1989**, *35*, 960–967, doi:10.1139/m89-158.

50. Ribeiro, D.G.; de Almeida, R.F.; Fontes, W.; de Souza Castro, M.; de Sousa, M.V.; Ricart, C.A.O.; da Cunha, R.N.V.; Lopes, R.; Scherwinski-Pereira, J.E.; Mehta, A. Stress and Cell Cycle Regulation during Somatic Embryogenesis Plays a Key Role in Oil Palm Callus Development. *J Proteomics* **2019**, *192*, 137–146, doi:10.1016/j.jprot.2018.08.015.