The nuclear and mitochondrial genomes of *Frieseomelitta varia* – a highly eusocial stingless bee (Meliponini) with a permanently sterile worker caste

Flávia C. de Paula Freitas 1,2, Anete P. Lourenço 1,3, Francis M. F. Nunes 1,4, Alexandre R. Paschoal 5, Fabiano C. P. Abreu 1, Fábio O. Barbin 1, Luana Bataglia 1, Carlos A. M. Cardoso-Júnior 6, Mário S. Cervoni 6, Saura R. Silva 7, Fernanda Dalarmi 8, Marco A. Del Lama 4, Thiago S. Depintor 1, Kátia M. Ferreira 4, Paula S. Gória 4, Michael C. Jaskot 4, Denyse C. Lago 1, Danielle Luna-Lucena 1, Livia M. Moda 2, Leonardo Nascimento 8, Matheus Pedrino 4, Franciene Rabico Oliveira 1, Fernanda C. Sanches 1,4, Douglas E. Santos 6, Carolina G. Santos 6, Joseana Vieira 2, Angel R. Barchuk 2, Klaus Hartfelder 6, Zílã L. P. Simões 8, Márcia M. G. Bitondi 8, and Daniel G. Pinheiro 7

Abstract

**Background:** Most of our understanding on the social behavior and genomics of bees and other social insects is centered on the Western honey bee, *Apis mellifera*. The genus *Apis*, however, is a highly derived branch comprising less than a dozen species, four of which genomically characterized. In contrast, for the equally highly eusocial, yet taxonomically and biologically more diverse Meliponini, a full genome sequence was so far available for a single *Melipona* species only. We present here the genome sequence of *Frieseomelitta varia*, a stingless bee that has, as a peculiarity, a completely sterile worker caste.

(Continued on next page)
Background

The ecological and economic importance of bees as pollinators and their millennia-long association with man, especially of the highly eusocial honey bees (Apini) and stingless bees (Meliponini) as providers of honey, pollen, wax, and propolis, has, not surprisingly, been a key factor for including the Western honey bee *Apis mellifera* in a top priority position for genome sequencing at the beginning of this century. In fact, the honey bee nuclear genome was the third insect genome to be sequenced [1], and is now one of the best annotated ones with over 15,000 predicted protein-coding genes [2, 3]. As such, it generally serves as a major backbone for sequencing and annotation efforts of other genomes, especially so within the Hymenoptera, the phylogenetically most ancient branch within the holometabolous insects [4].

*Apis mellifera* is a model organism for understanding social organization, especially so the permanent caste systems of highly eusocial insects. Nonetheless it is actually a member of the smallest branch within the monophyletic clade of corbiculate bees [5, 6] that comprise the highly eusocial Apini and Meliponini [7], as well as the primatively eusocial bumble bees (Bombini) and the solitary to incipiently social orchid bees (Euglossini). The tribe Apini comprises a single genus, *Apis*, of less than a dozen species, and for four of these fully sequenced genomes are available (*A. mellifera* [1, 2], *A. florea* [8], *A. cerana* [9], and *A. dorsata* [10]).

This stands in strong contrast with the stingless bees (Meliponini), which comprise over 500 species classified into 48–61 genera [6, 11]. The largest number of genera and species occurs in the Neotropics, with 32 genera and 417 recognized species [12], and recent population genetics studies indicate that species numbers are likely to be even higher [13]. Nonetheless, only one of these stingless bee species, *Melipona quadrifasciata*, has a fully sequenced and annotated genome, as it was included in a comparative genomics study of bees aimed at providing insights into genomic traces of social evolution [8]. A second species, *Lepidotrigona ventralis*, a Southeast Asian species recently had raw genome sequence data deposited in GenBank [PRJNA387986], but genomic annotation is still lacking.

Stingless bees are not only a species-rich monophyletic clade, they are also phylogenetically much older than the Apini, with origins dating back to 75–80 million years ago (mya) [11], compared to the origin of Apini, which is set at 22 mya. The Gondwana origin of the Meliponini can be seen reflected in the vicariance of their biogeographical, pantropical distribution [11].

In the tropical and subtropical Americas, the stingless bees were the main pollinators until the introduction of the honey bee, *A. mellifera*, in the eighteenth century. They have higher population densities than the solitary or primatively eusocial bees, and they are generalist plant visitors [14], which makes them also ideal pollinators for economically valuable crops, including greenhouse crops. The management of stingless bees (meliponiculture) has a long history, as shown in Pre-Colombian documents, such as the Maya Codex Madrid, that records practices for *Melipona beechei* from Mesoamerica. Also, over the last

Results: The assembly of 243,974,526 high quality Illumina reads resulted in a predicted assembled genome size of 275 Mb composed of 2173 scaffolds. A BUSCO analysis for the 10,526 predicted genes showed that these represent 96.6% of the expected hymenopteran orthologs. We also predicted 169,371 repetitive genomic components, 2083 putative transposable elements, and 1946 genes for non-coding RNAs, largely long non-coding RNAs. The mitochondrial genome comprises 15,144 bp, encoding 13 proteins, 22 tRNAs and 2 rRNAs. We observed considerable rearrangement in the mitochondrial gene order compared to other bees. For an in-depth analysis of genes related to social biology, we manually checked the annotations for 533 automatically predicted gene models, including 127 genes related to reproductive processes, 104 to development, and 174 immunity-related genes. We also performed specific searches for genes containing transcription factor domains and genes related to neurogenesis and chemosensory communication.

Conclusions: The total genome size for *F. varia* is similar to the sequenced genomes of other bees. Using specific prediction methods, we identified a large number of repetitive genome components and long non-coding RNAs, which could provide the molecular basis for gene regulatory plasticity, including worker reproduction. The remarkable reshuffling in gene order in the mitochondrial genome suggests that stingless bees may be a hotspot for mtDNA evolution. Hence, while being just the second stingless bee genome sequenced, we expect that subsequent targeting of a selected set of species from this diverse clade of highly eusocial bees will reveal relevant evolutionary signals and trends related to eusociality in these important pollinators.

Keywords: Social insect, Meliponini, Illumina sequencing, Genome assembly, Synteny, Repetitive elements, Non-coding RNA, Reproductive process genes, Immunity genes
decades, meliponiculture has gained new momentum as part of subsistence agriculture [15]. Stingless bees are also highly varied in important biological aspects, including colony size, nesting sites, communication systems, and colony defense, as well as caste determination and reproductive biology. While colonies of the tiny, fruit fly-size Leurotrigona species can fit into a matchbox, colonies of the open-nesting Trigona species can be of a size comparable to that of very large honey bee colonies. In terms of nesting sites, most stingless bees are cavity nesters, mostly so in trees, but they can use pretty much any kind of cavity, including underground ones [16, 17].

With respect to caste determination, the genus Melipona has long drawn attention, as it was the first social insect species for which a genetic mechanism of queen/worker determination was proposed [18], with underlying mechanisms still under investigation [19–21]. Nonetheless, it is in their reproductive biology in general that the stingless bees differ most drastically from the honey bees, and in this respect they are actually much closer to the bumble bees, with which they have a sister group relationship [5, 6]. The queens of most stingless bees mate with a single male only, and in many species, the workers contribute to the production of males in a colony [22]. In contrast, in the genera Frieseomelitta and Leurotrigona, the workers are completely sterile, and for Frieseomelitta varia it has been shown that the ovaries of workers undergo complete programmed cell death during pupal development [23].

In the previous comparative genomics study on sociality in bees [8], M. quadrifasciata was included not only for being the first among the stingless bees to have its genome sequenced, but also because of its emblematic genetic mode of caste determination. Furthermore, M. quadrifasciata and F. varia, are the two only stingless bee species for which RNA-Seq data had previously been generated in a comparative transcriptomics study [24]. Hence, we chose here the species Frieseomelitta varia for genome sequencing of a third candidate of this largest clade of highly eusocial bees, the Meliponini.

Methods
Sampling and DNA extraction
Brood cells were removed from F. varia colonies kept in the apiary of the Department of Genetics, Ribeirão Preto Medical School, University of São Paulo, and screened for the presence of male brood. The use of haploid males, as done in previous bee genome projects [1, 8], presents a considerable advantage for genome assembly. Thus, we also opted to use whole body DNA from a single late pupal-stage male specimen with still unpigmented wings. A voucher specimen of the respective colony was deposited in the Entomological Collection RPSP (Coleção Entomológica Prof. J.M.F. Camargo, FFCLRP/USP) under the register USP_RPSP 00005682. Genomic DNA was extracted using the Wizard® Genomic DNA Purification Kit (Promega, Madison, WI) resulting in a sample of 9.7 μg total DNA.

Genomic DNA library preparation and sequencing
The DNA sample was sent to Laboratório Central de Tecnologias de Alto Desempenho em Ciências da Vida (LaCTAD, UNICAMP, Campinas, Brazil) for quality check (2100 Bioanalyzer, Agilent Technologies, Santa Clara, CA), library preparation, and sequencing. Library preparation was done using Illumina Nextera kits (Illumina, San Diego, CA), and paired-end and mate pair sequencing was done on a HiSeq 2500 platform (Illumina). The extracted DNA was used for the construction of three sequencing libraries: two paired-end (one lane each) and one mate-pair (one lane). The paired-end libraries were prepared according to the TruSeq® DNA Nano Library Preparation Protocol (Illumina) using 100 ng input DNA. After DNA shearing, 350 bp inserts were selected using a bead-based method. Inserts were amplified by 8 PCR cycles, and the sequencing reaction yielded 2 × 101 bp reads. The mate-pair library was prepared from 1 μg of input DNA, following the Nextera® Mate Pair Library Preparation Protocol (Illumina). Fragments of 3 kb were circularized and sheared followed by purification of mate-pair fragments using beads. Mate-pair fragments were amplified in 10 PCR cycles, and the sequencing reaction produced 2 × 101 bp reads.

Genome assembly
Raw reads were submitted to quality analysis using FastQC software (www.bioinformatics.babraham.ac.uk/projects/fastqc/). The paired-end reads were analyzed with Trimomatic software [25] v. 0.35, which carried out the following tasks: removal of TruSeq DNA 3′ adapters (ILLUMINACLIP:TruSeq3- PE.fa:2:30:10); removal of leading low quality or N bases (below quality 3) (LEADING:3); removal of trailing low quality or N bases (below quality 3) (TRAILING:3); scanning of the reads with a 4-base wide sliding window, and cutting when the average quality per base drops below 15 (SLIDINGWINDOW:4:15); dropping reads < 100 bases (MINLEN:100). The mate-pair reads were analyzed using NxTrim [26] v. 0.4.1 to discard low quality reads and categorize reads according to the orientation implied by the adapter location. Thus, NxTrim builds “virtual libraries” of mate pairs, paired-end reads and single-ended reads, and, also trims off adapter read-through. NxTrim was executed with an aggressive adapter search (--aggressive parameter) to retrieve only genuine mate-pair reads.
An initial assembly was obtained using SPAdes software v. 3.9.0 [27], with error correction module (BayesHammer) enabled and using multiple k-mer sizes (33–81 bp). This assembly was subsequently used as reference for read alignments using HISAT2 software v. 2.0.5 [28], considering end-to-end alignments, avoiding spliced alignments and setting the right range of insertion size for paired-end (100–600 bp) and mate-pair (1–15 Kbp) libraries. The read alignments were used by BESST software v. 2.2.4 [29] to scaffold the initial genome assembly, considering only the alignments with mapping quality greater than or equal to 30. The scaffold size distribution was calculated by using functions implemented in R Statistical Software [30]. The assembly version thus generated was named Fvar-1.2.

Gene heterozygosity, repeat content, and size were evaluated from sequencing reads by a k-mer-based statistical approach. This was done in GenomeScope software v. 1.0, using as input the histogram file from k-mer frequency counting generated in the software Jellyfish v.2.2.0 (k-mers 19–63). A genome evaluation analysis of this assembly was made using QUAST-LG v. 5.0.2. Details on softwares and scripts are available in: https://github.com/dgpinheiro/fvaria/.

**Prediction and annotation of protein coding genes**

Initial gene predictions were made using MAKER2 software [31] version 2.31.8 in conjunction with UniProt sequences for *A. mellifera*. These automated predictions were further refined by using transcriptome data from *F. varia* RNA-Seq libraries generated from abdominal and brain RNA of adult workers [24] (SRA accession number SRR098304), as well as predictions for protein coding genes of *A. mellifera* (GCF_000002195.4). Additional gene model support came from RNA-Seq libraries generated for integument RNA from preimaginal and adult stages [32]. The raw sequences of those RNA-Seq libraries (NCBI BioProject ID PRJNA490324) were assembled in Trinity software with score definitions provided by the DETONATE tool [33] and used for alignment with the Fvar-1.1 genome assembly using MAKER2.

BUSCO software v.2.0 [34] was used to evaluate the completeness of the *F. varia* gene annotation. We conducted the BUSCO analysis of single-copy orthologs using the Hymenoptera dataset (OBP9) considering the predicted transcriptome extracted from the *F. varia* genome. For the annotation of coding gene functions we used the eggNOG-mapper v. 2.0.0 software [35] and generated a fasta and GFF file for the protein sequences (Fvar-1.2-proteins.fa; Fvar-1.2.gff).

For further validation of the predicted gene models, a set of 533 genes was selected for manual curation. As a first step, the honey bee orthologs for these genes were used as queries in blastp searches against the respective protein predictions for *F. varia* [Fvar-1.2-proteins], and the prospective orthologs were mapped against the assembled *F. varia* genome using the ARTEMIS platform [36] (https://www.sanger.ac.uk/science/tools/artemis). Exon and splice site predictions, as well as position of the automatically predicted gene models were checked, and in the case of genes that were split across scaffolds this was recorded. For each manually curated gene a GFF file was created, which contained information on problems and suggestions for gene model correction.

The identification of basal and other transcription factor (TF) domains in the predicted proteins of *F. varia* followed an approach similar to that taken by Kapheim et al. [8]. The function “hmmscan” of HMMER 3.1b2 [37] was used to scan the protein sequences (option --cut_ga and E-value ≤1e⁻⁵) against the Pfam-A database [38] (Pfam-A.hmm from ftp://ftp.ebi.ac.uk/pub/databases/Pfam). The results were filtered for a curated list of TFs retrieved from the Transcription factor database v2.0 [39] (www.transcriptionfactor.org). Other TF domains and basal TF domains were retrieved from Huang et al. [40] and included in the analysis. We calculated the proportion of genes with basal and other TFs. For this, we divided the number of genes with TF domains by the total number of predicted genes, and then compared this number to the proportion of genes with TF domains reported for the 10 previously published bee species [8].

**Gene Ontology analysis**

For the predicted genes, a functional analysis was performed using Blast2GO software, version 4 (https://www.blast2go.com/), with the following steps: (1) a blastp search against the GenBank NR database, (2) an InterproScan sequence search (https://www.ebi.ac.uk/interpro/search/sequence-search), (3) mapping to retrieve Gene Ontology (GO) terms to the sequence, and (4) attributing the EC code to the respective proteins.

**Synteny analysis**

To check for synteny in the gene organization of *F. varia* with known linkage groups in *A. mellifera*, the respective orthologs were first identified by means of reciprocal blastp searches. Their genomic localization in the honey bee genome (version 4.5) was retrieved and mapped against the coordinates of the *F. varia* genome assembly version 1.1 using scripts written in Python (https://www.python.org). Synteny was plotted using Circos software package [41] for visualizing genomic data.

**Phylogenetic analysis of gene families and functional groups**

Gene families and functional groups selected for manual curation were analyzed with respect to their phylogenetic relationships among bees. For this, their orthologs were retrieved by blastp searches against the sequences...
of 13 published bee genomes (Table S1) using the Hymenopteramine tool (http://hymenopteragenome.org/hymenopteramine/begin.do) available in Hymenopterabase [2]. For each gene family or functional group of interest a FASTA file containing the respective amino acid sequences of the 14 bee species (F. varia + 13) was generated and the sequences were aligned using the MAFFT program v7.402 with the L-INS-i approach [42] (https://mafft.cbrc.jp/alignment/software/). Gene phylogenetic trees were reconstructed by means of the Randomized Axelerated Maximum Likelihood (RAxML) program version 8.2.10 [43] (https://cme.h-its.org/exelixis/web/software/raxml/), with the CAT model, the JTT substitution matrix, and 1000 bootstrap replications. All programs are implemented in the CIPRES platform [44]. The trees were edited using FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

Prediction and analysis of non-protein coding genes and repetitive genomic components

Non-coding RNA (ncRNAs) genes were predicted using sequence similarity and structural search strategies. BLAST tools were used to search against ncRNAs Ensembl Insects (Anopheles gambiae AgamP4,Apis mellifera Amel_4.5, Atta cephalotes Attacep1.0, Bombus impatiens BIMP_2.0, Bombyx mori ASM15162v1, Drosophila melanogaster BDGP6, Nasonia vitripennis NvIt_2.1 and Solenopsis invicta Si_gnG) and InsectBase databases (piRNA and lncRNA).

For the identification of microRNA genes, BLAST searches were performed against miRBase release 22.1 [45]. For structural ncRNA gene searches, we used INFERNAL version 1.1.2 [46] based on the Rfam 14.0 database [47]. The INFERNAL annotation used CMsearch software with the parameter -cut_ga. The BLAST search used filter dust, and an E-value = 0.00001 and identity/coverage of at least 95% as thresholds. All filtering and merge results steps were performed by customized Perl scripts. The insect species and respective genome versions used are listed in Table S2.

For a more detailed analysis of the microRNA gene miR-34 the corresponding sequences in arthropods were retrieved from miRBase version 21 [44] and aligned in CLUSTALW [48, 49] and analyzed in MEGA v.7 [50] using the Maximal Parsimony method and 100 bootstrap replications.

For the identification of long non-coding RNA (lncRNA) genes we first extracted intergenic regions based on coding annotation from the F. varia genome. Intergenic sequences longer than 200 nucleotides retrieved using a Perl script from the RNAplonc tool (200 nt.pl) [51] were considered for further analysis. Protein encoding potential was filtered out by using the CPAT tool [52] with ORF_size ≥100 and Coding Prob. ≥0.345, and the remaining candidate lncRNA gene sequences were analyzed against Pfam FASTA data [38] using the blastx tool with an E-value of 10^{-5} and SEGfilter for low complexity regions. All ‘no-hit’ sequences were then considered as lncRNA genes. Finally, we compared the lncRNA annotations with non-coding RNA annotations to discard lncRNA candidates that might overlap with non-coding RNA annotations, so that only bona fide lncRNA candidates would be kept in the list.

For the identification of repetitive elements we used RepeatMasker tool v4.1.0 [53] (default parameters) with RepBase version database RepBase26.10.2018, and RepeatModeler version 2.0.1. We merged both results by the overlapping of genomic regions to avoid redundancy results by using a Perl script.

Mitochondrial genome analysis

Assembly

Four different softwares, NOVOPlasty v2.59 [54], SPAdes v. 3.6.2 [27], Platanus v. 1.2.4 [55] and MitoBim v. 1.9 [56] were employed. Using all trimmed paired-end reads, as recommended, with 39 k-mers, and the COX2 sequence of Bombus hypocrita sapporensis (NC_011923) as seed, the organelle-specific software NOVOPlasty turned out to be the only procedure that generated a unique, single contig with evidence for circularization. The other softwares also generated assemblies with the same gene content, but these were still fragmented.

Validation

The first validation step was to assess the correspondence and coverage of the reads against the assembled mitochondrial genome. For this, only high-quality paired-end reads were selected and processed using PrinSeq v.2.20.3, under the following parameters: trimming was done using a sliding window approach, considering a quality score mean < 28 in a window size of 3 bp sliding 1 bp to the left, in addition to filtering sequences with at least one quality score < 30. The paired-end reads were aligned against the assembled mitochondrial genome using Bowtie v. 0.12.7 (with the parameters: end-to-end hits with up to 1 mismatch, and with a maximum insert size of 300 for paired-end). Using the same mapped reads, we used the program REAPR v.1.0.18 [57], which does not require a reference genome, for evaluation of reads coverage, counts of unique mappings, and absence of mismatches.

An additional validation approach was performed using the alignment of paired-end reads. With this we checked if there were alignments of fragments across block junctions, i.e., whether the two reads of each pair aligned in each adjacent block, thus supporting adjacency. First, with the software MAUVE v. 2.4.0 [58], we performed a multiple genome alignment for information
on genome synteny between mitochondrial genome regions of the *F. varia* assembled genome with those of *Apis mellifera* (NC001566), *Bombus hyporicta* (NC011923) and *Melipona scutellaris* (NC026198) mitochondrial genomes [59]. From this alignment, we defined six genome blocks in the assembled genome, according to rearrangements observed in the mitochondrial gene order compared to other high eusocial bee species. With these blocks we obtained the respective coordinates of the rearrangements and used them as input files in specifically developed in-house scripts (https://github.com/dgpinheiro/fvaria#assembly-of-mitochondrial-genome).

**Annotation**

This was initially done using the software MITOS2 (http://mitos2.bioinf.uni-leipzig.de/index.py). As this program does not currently detect initiation and stop codons, it was necessary to posteriorly manually adjust for these using ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/) and BLAST tools, which were also used to identify rRNA coordinates. tRNAs were identified by means of the softwares tRNAscan-SE 2.0 [60] and ARWEN [61] using standard parameters. A map of the *F. varia* mitochondrial genome was produced using OGDRAW (OrganellarGenomeDRAW00) [62].

**Phylogenetic analyses**

A multiple genome alignment was generated for the comparison of the assembled *F. varia* mitochondrial genome with complete mitochondrial genomes of the superfamily Apoidea. Three concatenated genes datasets were separated: one with all mitochondrial genes (coding sequences + tRNAs + rRNAs), with only protein coding sequences, and the third one consisting only of the tRNAs. All datasets were aligned using the MAFFT v.7 webserver [63], with standard parameters. The analyses were done using Maximum Likelihood (ML) method in the RAxML softwares using Rapid bootstrap [64] and Bayesian Inference (BI) in Mr. Bayes [65]. The evolutionary models were calculated with jModelTest v.2 [66]. All softwares were run online at the CIPRES Science Gateway. For finding the ML tree, 10,000 replicates were used and clade consistency was evaluated by 1000 bootstrap replicates. For Bayesian Inference, two runs and four chains were calculated with 5,000,000 generations until reaching an average standard deviation of split frequencies of less than 0.01. The 25% of the initial trees were discarded as burn-in. The outgroup was represented by two ant species [*Anoplolepis gracilipes* (NC_039576) and *Camponotus atroxx* (NC_029357)], and four bee species [*Megachile sculpturalis* (NC_028017), *Rediviva intermixta* (NC_030284), *Hylaeus dilatatus* (NC_026468), and *Colletes gigas* (NC_026218)]. All trees were edited with the program TreeGraph 2 [67] and iTOL (https://itol.embl.de). The evolutionary models for the trees were: GTR + G for the complete dataset and the protein coding genes, and TIM1+ G for the tRNAs. The species included in the phylogenetic analysis are all listed in Table S3.

**Results**

**Whole genome assembly**

After an initial quality check using FastQC, the 2*251, 808,069 Illumina paired-end reads (101 bp, 64.85% bases ≥ Q30) and 171,026,322 mate-pair reads (92.10% bases ≥ Q30) were trimmed using Trimmomatic and Nxtrim, respectively. This pre-processing resulted in a total of 234,357,438 high-quality paired-end and 9,671,088.161 high-quality mate pair reads. The results for the first assembly (Fvar-1.0) generated with SPAdes software was still highly fragmented, with almost 10,000 scaffolds. This assembly was considerably improved using HISAT2 and the BESST package for scaffolding. These generated the second assembly named Fvar-1.2. Details on these assemblies are given in Table 1.

The total size estimate for the assembled genome was 275 Mbp, with a GC content of 37%, and our gene annotation approach resulted in a total of 10,526 protein-coding genes. Furthermore, in a BUSCO analysis for the 4415 hymenopteran ortholog genes (OBP9) we compared *F. varia* to 10 other bees [8, 68, 69] and the parasitic wasp *Nasonia vitripennis* (Fig. 1). With this, we identified 3970 complete and 298 fragmented genes (90 and 6.7%, respectively) as hymenopteran single-copy orthologs. Only 147 genes (3.3%) were not found in the current version of the *F. varia* genome. The proportion of single-copy orthologs is widely used to assess the quality of both genome assembly and gene annotation [70]. Thus, the identification of the majority hymenopteran genes validates our genome assembly and gene prediction approaches.

**Table 1** General statistics for the first two *F. varia* genome assembly versions. The genome evaluation was made using QUAST-LG v. 5.0.2

| Statistics without reference | Fvar-1.0 | Fvar-1.2 |
|-----------------------------|----------|----------|
| # contigs (>= 0 bp)          | 9755     | 2173     |
| Largest contig              | 603,324  | 2,258,834|
| Total length (>= 100,000 bp)| 116,286,290 | 255,941,895 |
| N50                         | 85,201   | 470,005  |
| N75                         | 43,559   | 244,533  |
| L50                         | 946      | 174      |
| L75                         | 2087     | 379      |
| Total genome length         | 275,412,029 | 255,941,895 |
| GC (%)                      | 36.72    |          |

Paula Freitas et al. *BMC Genomics* (2020) 21:386
**Protein-coding genes**

The BUSCO analysis indicated that the vast majority of hymenopteran genes is represented in the *F. varia* genome sequence. One way of refining the confidence of our data is by observing their functional coherence, and aiming at this we ran a Gene Ontology analysis, comparing information from *A. mellifera* and *F. varia* protein sets. For *A. mellifera* we used RefSeq-NCBI (release 103) containing 22,451 well-annotated and non-redundant proteins. For the 10,526 predicted gene models of *F. varia* it was possible to identify isoforms for approximately 5%, and therefore, the set of data used for GO analysis corresponded to 11,115 non-redundant predicted proteins. For both bees, 56% of the sequences in each protein set (12,629 in *A. mellifera* and 6276 in *F. varia*) were associated with at least one GO term. Such incompleteness in GO term assignment is common for non-model organisms, as is the case for most insects. Thus, the data were normalized based on the proportion of GO-annotated genes prior to performing further comparisons.

Figure 2 reports the results for the top 25 Molecular Function and Biological Process categories comparing *F. varia* with *A. mellifera*. The percentages of genes with GO annotation for Biological Process and Molecular Function in the two species were similar in distribution profiles, indicating that any functional category, whether major, intermediate, or minor, is represented in approximately the same order of magnitude. This is in accordance with the view that the ontological-functional profiles are quite similar, even across large taxon borders [71].

Next, with the aim of evaluating the *F. varia* MAKER2 gene model predictions we selected 533 honey bee protein-coding genes as reference for the manual

---

### BUSCO Assessment Results

|   | Complete (C) and single-copy (S) | Complete (C) and duplicated (D) | Fragmented (F) | Missing (M) |
|---|----------------------------------|----------------------------------|----------------|-------------|
| Ador | [C:4260 S:2640 D:1620] | F:95, M:50, n:4415 |                 |             |
| Aflo | [C:4304 S:2990 D:1314] | F:93, M:50, n:4415 |                 |             |
| Amel | [C:4353 S:2482 D:1871] | F:37, M:50, n:4415 |                 |             |
| Bimp | [C:4387 S:2329 D:2058] | F:15, M:10, n:4415 |                 |             |
| Bter | [C:4368 S:2438 D:1930] | F:31, M:10, n:4415 |                 |             |
| Dnov | [C:4337 S:3659 D:978] | F:37, M:10, n:4415 |                 |             |
| Emex | [C:4346 S:3129 D:1217] | F:36, M:10, n:4415 |                 |             |
| Fvar | [C:4011 S:3960 D:51] | F:254, M:10, n:4415 |                 |             |
| Hlab | [C:4336 S:3705 D:631] | F:45, M:10, n:4415 |                 |             |
| Mqua | [C:3790 S:3784 D:8] | F:216, M:40, n:4415 |                 |             |
| Mrot | [C:4356 S:1928 D:2428] | F:33, M:20, n:4415 |                 |             |
| Nvit | [C:4195 S:2239 D:1956] | F:153, M:60, n:4415 |                 |             |

**Fig. 1** BUSCO analysis. The *F. varia* gene content was compared with that of 10 other bees and the parasitic wasp *Nasonia vitripennis* using the following databases: [Ador] *Apis dorsata* (GCF_000469605.1), [Aflo] *Apis florea* (GCF_000184785.2), [Amel] *Apis mellifera* (GCF_000002195.4), [Bimp] *Bombus impatiens* (GCF_000188095.1), [Bter] *Bombus terrestris* (GCF_00014255.1), [Dnov] *Dufourea novaangliae* (GCF_001272555.1) [Emex] *Eufrisema mexicana* (GCF_001483705.1), [Fvar] *Frieseomelitta varia* (predicted transcriptome extracted from genome annotation of Fvar-1.2), [Hlab] *Habropoda laboriosa* (GCF_001263275.1), [Mqua] *Melipona quadriangulata* (GCA_001276565.1), [Mrot] *Megachile rotundata* (GCF_000220905.1), and [Nvit] *Nasonia vitripennis* (GCF_000002325.3)
The curation of their homologs in the *F. varia* genome using the ARTEMIS platform. These 533 genes were not chosen randomly from the honey bee OGS 3.2 set, but were included because of their functional association with developmental and reproductive processes, immunity, and processes related to social communication. These can, thus, be considered as of special interest for the social biology of stingless bees.

As a result of this manual curation of 533 genes, 241 (45.03%) of the automated predictions were considered 100% correct, but for the same percentage (45.02%, 240 genes), certain problems in exon assignment were noted, primarily for the first exon (Table S4). Furthermore, for 45 genes (8.3%) we could either not identify clear orthologs in the *F. varia* genome, or they were only found after manual searches. The remaining corrections
(1.65%) were generally attributed to probable minor sequencing errors. A possible explanation for the misprediction of the first exon is a positional bias in the prediction of gene structure, in which an initial exon is less accurately identified compared to internal exons [72]. Especially, long first introns and longer introns in general [73], characteristics of higher eukaryotic genomes, impose extra challenge to accurately predict gene structures. Nonetheless, the overall quality of our genome assembly and gene annotation is confirmed by the high percentage of hymenopteran genes identified in the BUSCO analysis.

Non-coding RNA genes

For a curated annotation of ncRNA classes in the *F. varia* genome we employed a combination of similarity-based and structure-based computational approaches, and we identified a total of 1946 ncRNAs (Table 2) falling into six ncRNA classes: small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), intergenic long non-coding RNAs (lncRNAs), and microRNA precursors (miRNAs).

Among the latter we identified members belonging to 38 miRNA families. We analyzed in more detail the microRNA-34, which is highly conserved in the animal kingdom and is maternally inherited in *D. melanogaster* [74] and in *A. mellifera* [75, 76], regulating the expression of important developmental genes. Its conservation was confirmed for *F. varia* and its sequence was seen to cluster closely with the honey bee (Figure S1).

Furthermore, we performed a comparative analysis on the distribution of ncRNA families in insect genomes of different orders. The results shown in Table 3 may help in elaborating hypotheses on the evolution of these elements in insect genomes. The highest number of total ncRNAs in insect genomes was identified in the *D. melanogaster* genome, where it is close to 30% of the total number of protein coding genes. Evidently, this is due to the extensive genetics and genomics work done by the community that allowed the annotation of these loci.

| ncRNA type | number | average length (bp) | total length (bp) |
|------------|--------|---------------------|------------------|
| miRNA      | 103    | 89.81               | 9437             |
| tRNA       | 21     | 339.8               | 7136             |
| snoRNA     | 9      | 105.33              | 948              |
| snRNA      | 53     | 150.43              | 7973             |
| tRNA       | 180    | 74.08               | 13,335           |
| lncRNA     | 1580   | 687.25              | 1,087,232        |
| Total      | 1946   |                     |                  |

Table 2: Types and number of non-coding RNAs in the *F. varia* genome

Nonetheless, what is surprising is the apparent considerable variation in the number of ncRNA genes seen among hymenopteran species. The fact that the numbers are most divergent for the lncRNAs is actually not surprising, as these cannot be annotated by customary similarity-based algorithms, but there is also considerable variation in the numbers of tRNA, snRNA, and miRNA loci in these genomes. Considering this, the variation denoted in Table 3 with respect to ncRNA gene numbers is, in fact, a glimpse into a major lacuna for hypotheses on insect genome evolution.

Genome organization, synteny, and repetitive genomic components

For an overview of the general genome structure we plotted the orthologous genes predicted in the scaffolds of *F. varia* against their respective position in the linkage groups of the honey bee genome. In the Circos plot (Fig. 3a), the same-colored lines connect orthologs of the two species with regard to their respective genomic localization based on linkage groups, which are chromosomes in the case of *A. mellifera* and scaffolds for *F. varia*. For example, most of the orthologs on the FV909, FV816, and FV163 scaffolds of *F. varia* mapped all to *A. mellifera* chromosome 1 (AM1). Similarly, most of orthologs located on FV418, FV247, and FV182 were found on AM10. With this in mind, we next conducted a more in-depth analysis into gene clusters that are known to play important roles in insect and, especially so, in bee biology: the Major Royal Jelly Protein (MRJP) family, the Osiris gene cluster, and the Pbn1 genes identified in a QTL of the pollen hoarding syndrome of honey bees.

The genes encoding MRJP/MRJP-like proteins of bees are inserted within the cluster of Yellow genes, specifically between yellow *e3* and yellow *h* [77]. But while the MRJP gene family has undergone a taxon-specific expansion in the genus *Apis*, consisting of a tandem array of nine functional MRJPs [77, 78], all the other corbiculate bees, as well most ants, with the exception of *Linepithema humile*, only have a single mrjp-like gene at this genomic location [8]. This genomic architecture was also found in the *F. varia* genome, with a single copy of an MRJP gene similar to *Apis* *mrjp-9* (mrjp9-like) being flanked by the two above-mentioned Yellow genes.

In the overall synteny analysis, the Osiris gene cluster stood out because of its high degree of structural genomic conservation. Figure 3b shows the organization for the *F. varia* Osiris gene cluster in comparison to *Apis mellifera*, not only in overall cluster size (230 vs. 220 kb), but also in terms of gene number and order, as well as transcriptional direction. Osiris genes are a highly conserved cluster of ~20 genes covering a genomic region of ~160 kb [79]. They are thought to have originated...
from gene duplications in early insects about 400 mya [4, 79] and to be related to insect wing evolution and radiation.

Compared to the conserved MRJP/MRJP-like and Osiris genes, a clearly unexpected result coming out of the synteny analysis was the finding for the genes located in the pln1 QTL of the honey bee. This QTL was identified through a selection program over various generations for high vs. low levels of pollen collection and storage [80], which is a colony behavioral trait of honey bee workers. Such divergence in pollen hoarding behavior was subsequently found to have a strong association with gustatory responses and reproductive traits of A. mellifera workers [81–83]. In our genomic synteny analysis, the genes of the pln1 QTL region on chromosome 15 of the honey bee showed a high degree of conservation in both gene number and order in the F. varia genomic scaffolds, especially so in scaffolds 203 and 207 (Fig. 3c). Interestingly, for the genes mapped to the other two pollen hoarding QTLs of the honey bee, pln2 and pln3 [80], we could not find such a strong linkage in the F. varia genome.

In terms of repetitive genomic components, we identified a total of 169,371 elements in the F. varia genome, belonging in majority to unknown elements. Specific transposable elements (LTR, LINE, SINE and DNA)
represent 27.4% of the identified elements. A complete list of the identified repetitive genomic components, their respective numbers, cumulative length and % of the genome is presented in Table 4. In total, the repetitive genome represents 39.1 Mbp.

The mitochondrial genome

We assembled the mitochondrial genome of *F. varia* from the whole genome sequencing reads and validated it with REAPR, which showed 57.53% of error-free bases. Also, no errors were identified, as indicated by the reported REAPR fragment coverage distribution (FCD) error, and there was no evidence of local misassemblies. In addition, we checked whether the sites showing rearrangements were consistent with the position of each pair of paired-end reads. This showed that at least 146 paired-end reads were found between sites, therefore confirming the sites with genes rearrangement.

As a result, the *F. varia* mitochondrial genome was found to consist of 15,144 bp and as having an A + T-rich nucleotide composition (87.8%), like in other insect species [84]. We identified all the 13 protein-coding genes, the 22 tRNA and 2 rRNA genes, as well as the non-coding control region. A map of the *F. varia* mitogenome is shown in Fig. 4a. All the 13 protein-coding genes (7 NADH dehydrogenases, 3 cytochrome C oxidases, 2 ATPases and cytochrome b) are encoded in the forward strand, and all have an initiation codon typical for invertebrates. The A + T-rich control region for the initiation of replication is located between the rRNA rrnS and the tRNA trnS-TGA.

The alignment of all mitochondrial genome genes (protein-coding genes, tRNAs and rRNAs) of 84 species resulted in a concatenated matrix of 16,499 bp, 12,008 bp, and 1749 bp length, respectively. The results for all trees using the two methods (ML and BI) were similar (Fig. 4b and Figures S2, S3, S4). Only for the tRNA trees there was less phylogenetic resolution, as indicated by tree topology and branch support using both (RAxML and BI) methods (Figures S5 and S6).

The phylogenetic analysis indicated that the *F. varia* mitogenome is sister to those of the stingless bees *Melipona bicolor* and *Melipona scutellaris* (Fig. 4b), corroborating the results of a previous study based on mitochondrial 16S rDNA sequences [85].

In this context, a finding of particular interest was a major rearrangement in gene order in the *F. varia* mitogenome compared to *Apis, Bombus*, and even to the genus *Melipona* (Fig. 4c). We found a drastic shuffling in gene positions for *nad1, nad4–6, cyth*, for the two rRNAs, as well as for the A + T-rich non-coding control region. Gene translocations have previously been described in other bee and wasp genomes [86, 87], and were also reported for *M. bicolor* [88].

**Gene families and functional groups of specific interest**

**Transcription factors and epigenetic modifier genes**

By scanning the predicted protein sequences against the Pfam-A database using the hmmscan function, followed by the curation against the Transcription factor database v2.0 we identified 36 types of basal and 96 types of other transcription factor (TF) domains (Table 5). Among the basal TFs, the predominant ones were 'TFIIIC, TAFH, TBP, and TF Zn Ribbon, which is a result similar to the one reported for the silk moth, *Bombyx mori* [40]. The predominant ones among the other types of TFs were 'zf-C2H2, Homeodomain, HLH, zf-C3HC4, and PHD, which is similar to the figure shown for the 10 previously sequenced bee species [8].

Epigenetic mechanisms have emerged as important regulators of reproduction and behavior in social insects [89–91] Concerning genes involved in epigenetic modification, we identified a complete set of DNA methyltransferases (DNMTs) genes and genes associated with histone post-translational modifications (HPTM) in the *F. varia* genome. The sequences of these genes showed higher similarity scores to their orthologs in *M. quadrifasciata* and the two bumble bee species for which a complete genome sequence is available [8, 68] than to those of *A. mellifera*.

**Genes related to reproduction**

The phenomenon of complete worker sterility in the genus *Frieseomelitta* made us take a closer look at genes and gene families related to reproductive processes in insects. From literature and GO searches we compiled a list of genes with known or predicted functions in the activation of the insect ovary. This list comprises a set of 127 genes in *F. varia*, and these are represented in very similar numbers in the genomes of other bees, as well as in the parasitic wasp *Nasonia vitripennis* and the fruit fly (Table 6). Among these, 61 are members of the core sets of major signaling pathways, such as Hippo, insulin/IGF, TGF-β, Wnt, and Notch. The gene models of these 61 core genes (in KEGG) were retrieved from the 10 annotated bee genomes [8], the *Euglossa dilemma* genome [69], the *A. cerana* genome [9] and the *A. dorsata* genome, and were compared with the respective *F. varia* gene models.

A summary report representing the phylogenetic gene tree reconstructions for all these pathways is shown in Fig. 5 and details are shown for the key gene set of the Notch pathway (Fig. 6). All the respective core genes were identified in the *F. varia* genome, and the automatically predicted gene models turned out to be correct for 50 (80%) of these, while 8 genes (13%) required corrections in the automatically predicted gene models, and 4 (7%) that had not been predicted were found by manual blastp searches. As it turned out, their sequences were
| Order | Superfamily       | Total | Cumulative length (Mb) | Cumulative length (bp) | % Genome |
|-------|-------------------|-------|------------------------|------------------------|----------|
| DNA   | Unclassified      | 290   | 0.075844               | 75,844                 | 0.02758  |
|       | CMC-Chapaev-3     | 63    | 0.007718               | 7718                   | 0.00281  |
|       | Crypton           | 1     | 0.000084               | 84                     | 0.0003   |
|       | Crypton-V         | 188   | 0.045662               | 45,662                 | 0.01660  |
|       | hAT               | 114   | 0.017705               | 17,705                 | 0.00644  |
|       | hAT-Ac            | 1896  | 0.706924               | 706,924                | 0.25706  |
|       | hAT-Blackjack     | 107   | 0.019388               | 19,388                 | 0.00705  |
|       | hAT-Charlie       | 1533  | 0.422873               | 422,873                | 0.15377  |
|       | hAT-Tip100        | 99    | 0.021175               | 21,175                 | 0.00770  |
|       | Kolobok-Hydra     | 2433  | 1.172149               | 1,172,149              | 0.42624  |
|       | Kolobok-T2        | 44    | 0.012426               | 12,426                 | 0.00452  |
|       | Maverick          | 145   | 0.07668                | 76,680                 | 0.02788  |
|       | Merlin            | 423   | 0.051174               | 51,174                 | 0.01861  |
|       | MuLE-MuDR         | 57    | 0.011788               | 11,788                 | 0.00429  |
|       | PIF-Harbinger     | 262   | 0.077291               | 77,291                 | 0.02811  |
|       | PIF-Spy           | 127   | 0.017818               | 17,818                 | 0.00648  |
|       | PiggyBac          | 3472  | 0.688623               | 688,623                | 0.25041  |
|       | Sola-1            | 143   | 0.036072               | 36,072                 | 0.01312  |
|       | TcMar-Fot1        | 241   | 0.081897               | 81,897                 | 0.02978  |
|       | TcMar-ISRm11      | 479   | 0.110088               | 110,088                | 0.04003  |
|       | TcMar-Mariner     | 3520  | 0.808834               | 808,834                | 0.29412  |
|       | TcMar-Pogo        | 43    | 0.00506                | 5060                   | 0.00184  |
|       | TcMar-Stowaway    | 67    | 0.015985               | 15,985                 | 0.00581  |
|       | TcMar-Tc1         | 21,131| 5.625815               | 56,25,815              | 2.04575  |
|       | TcMar-Tc4         | 2991  | 0.61903                | 619,030                | 0.22510  |
|       | TcMar-Tigger      | 640   | 0.128285               | 128,285                | 0.04665  |
| LINE  | CR1               | 161   | 0.04403                | 44,030                 | 0.01601  |
|       | I                 | 2503  | 0.977458               | 977,458                | 0.35544  |
|       | I-Jockey          | 951   | 0.715998               | 715,998                | 0.26036  |
|       | L1                | 38    | 0.012373               | 12,373                 | 0.00450  |
|       | R1                | 262   | 0.104838               | 104,838                | 0.03812  |
|       | R1-LOA            | 40    | 0.008466               | 8446                   | 0.00307  |
|       | R2-NeSL           | 64    | 0.027806               | 27,806                 | 0.01011  |
|       | RTE-X             | 522   | 0.20592                | 20,592                 | 0.07488  |
| LTR   | Unclassified      | 40    | 0.002817               | 2817                   | 0.00102  |
|       | Copia             | 35    | 0.026156               | 26,156                 | 0.00951  |
|       | ERV1              | 285   | 0.024605               | 24,605                 | 0.00895  |
|       | ERVK              | 2     | 0.000117               | 117                    | 0.00004  |
|       | ERVL              | 21    | 0.00166                | 1660                   | 0.00060  |
|       | Gypsy             | 679   | 0.193441               | 193,441                | 0.07034  |
| RC    | Helitron          | 324   | 0.123596               | 123,596                | 0.04494  |
| Retroposon | SVA  | 1   | 0.000144               | 144                    | 0.00005  |
| Satellite | Unclassified  | 720  | 0.371618               | 371,618                | 0.13513  |
|        | acro              | 127   | 0.009609               | 9609                   | 0.00349  |
fragmented and represented on different scaffolds. Our phylogenetic tree reconstruction for the Notch pathway genes in bees (Fig. 6) confirmed that the *F. varia* orthologs were generally closely related to those of *M. quadrifasciata*.

**Immunity-related genes**

We screened the *F. varia* genome for the presence of immunity related genes. Of the 190 immune response genes predicted for *A. mellifera* [8], we found 174 homologs in the *F. varia* genome (Table 7, Figures 7a, Figure S7 and S8). Worthy of note is that we did not find orthologs of the genes encoding Cactus-2 and Cactus-3, Dorsal-2 (Relish1B), as well as of some apoptosis inhibitors (IAPs), Toll8 and CLIP. This could be due to a lack of coverage in DNA sequencing, but there was also no evidence for the respective transcript in the RNA libraries used in the mapping procedure. Furthermore, some of these genes also did not have identifiable orthologs in certain other bee genomes. An interaction network for the immunity genes of *F. varia* showing their scaffold localization in relationship to the 16 honey bee chromosomes is shown in Fig. 7b.

**Neurogenesis-related genes**

The exceptional performance of honey bee workers in learning and memory task assays has been a driver for high-throughput sequencing efforts and the establishment of a microarray platform since the early days of honey bee genomics [92, 93]. There are two central questions, one is how the behavioral sequencing efforts and the establishment of a microarray platform since the early days of honey bee genomics [92, 93]. There are two central questions, one is how the behavioral plasticity and task switching in adult workers is reflected in transcriptional brain profiles [94, 95], and the other is how gene expression in the larval brain may underlie the development of caste- and sex-specific brain anatomies [96, 97].

As genes involved in neurogenesis are candidates for the wiring and rewiring of neuronal networks, our aim here was to get an overview on neurogenesis-related genes in bees and, especially so, in the *F. varia* genome. Using the Gene Ontology Biological Process term neurogenesis (GO: 0022008) for the identification of human and *Drosophila* orthologs in 11 bee genomes, we retrieved a total of 1586 genes in the human genome and 954 genes in the fruit fly, representing 7.94 and 5.36%, respectively in terms of total gene numbers (Table 8). In the sequenced bee genomes, the number of neurogenesis-associated genes varies between 672 and 732, with respective percentages between 5.31 and 5.61. The relatively high percentage of these genes in the *F. varia* genome is intriguing, but, at present, it should not be over-interpreted, as this may change once future annotation efforts come up with a refined total gene number estimate. The Venn diagram comparing sequence similarity of neurogenesis-related genes annotated in the *F. varia* genome with those in the human genome, the fruit fly, and the honey bee (Fig. 8) shows that most of the *F. varia* neurogenesis genes have close orthologs in the honey bee (64.5%) or with the honey bee and the fruit fly combined (32.9%).

**Chemosensory repertoire and circadian rhythm genes**

To explore the chemosensory repertoire in bees, we used information from two phylogenetically close bee species, *A. mellifera* and *M. quadrifasciata* [98, 99], as input to identify the gene family sizes of odorant receptors (ORs), odorant binding proteins (OBPs), chemosensory proteins (CSPs), gustatory receptors (GRs) and ionotropic receptors (IRs) in the *F. varia* genome. In addition, we searched for protein-coding sequences of SNMPs [Sensory Neuron Membrane Proteins, Group 3 (SNMP1 and SNMP2)] in the genomes of these three bee species, and we also updated information on IR gene numbers for *A. mellifera* and *M. quadrifasciata*.

We performed a thorough manual curation that included multiple and local alignments and reciprocal BLAST searches in order to: (i) check the existence of non-automatically predicted gene models, (ii) obtain information on protein sequence lengths and structure of the coding regions, (iii) correct gene models when needed, and (iv) screen for the presence of the following canonical domains: 7tm_6 (pfam02949) for ORs, PBP_
Fig. 4 (See legend on next page.)
GOBP (pfam01395) for OBPs, OS-D (pfam03392) for CSPs, 7tm_7 (pfam08395) for GRs, PBP2_iGluR (cd13717) for IRs, and CD36 (pfam01130) for SNMPs 1 and 2.

The gene numbers for each chemosensory family that we identified in the *F. varia* genome (Table 9) are based on a more conservative approach than that of Brand and Ramírez [99], which for ORs used the earlier prediction made by Robertson and Wanner [98]. Specifically, we excluded: pseudogenes, protein-coding sequences without canonical domains, partially predicted protein sequences (incomplete ORFs or lacking an initial methionine residue), and/or sequences that presented poor alignments with those of the other bees. Notwithstanding, it is already possible to draw some conclusions. In accordance with previous studies [100, 101], we observed a variable degree of evolutionary conservation. For instance, sequence similarity for OBP and OR genes ranged from 54 to 73%, and 33 to 86%, respectively, except for the ORCO gene (also called OR2) that showed 95% similarity.

With respect to circadian rhythm genes, we manually curated the automatically predicted gene models of eight canonical clock genes in the genome of *F. varia* based on orthology to *A. mellifera* and *D. melanogaster*: clock, cycle, period, cryptochrome mammalian-like, par domain protein, vrille, timeout2, and clockwork orange. The gene models were all correct and consistent with those of other bee species.

**Discussion**

**Genome assembly and genome organization**

We report here the genome sequencing, assembly, and annotation results for the highly eusocial stingless bee *F. varia*, which, after *M. quadrifasciata*, is now the second stingless bee species with complete genomic information. For the nuclear genome we predicted an assembled genome size of 275 Mbp (and a GC content of 37%) (Table 1). This genome size estimate fits in very well with previously reported estimates for solitary and social bees [2, 8, 68, 69, 102]. Nonetheless, it differs drastically from the size estimate of 450 Mb obtained by flow cytometry [103]. Since repetitive regions represent a challenge to genome assemblies from short Illumina reads, we used a k-mer counting approach to obtain a further genome size estimate. This resulted in a haploid genome length of 345.8 Mbp. Furthermore, our analysis of repetitive genomic components indicated a repeat genome size of 39 Mbp. So, at least part of the discrepancy between the flow cytometry estimate and the assembled genome size may be due to the high number of repetitive genomic components (Table 4). The k-mer-based genome size estimate minus repeats almost represents the assembled portion of the genome and, in line with
BUSCO, the assembly, therefore, appears to be rather complete. Under- or overestimates of genome size based on flow cytometry results have previously been reported for *Bombus terrestris* [104].

The number of predicted protein-coding genes (10, 526) is close to the original prediction for *A. mellifera* gene number [1], but ca. 30% lower than the subsequently revised gene number prediction [2], and also lower than the gene numbers predicted for the other sequenced bee genomes [8]. Nonetheless, it is close to the gene number predicted for the more recently sequenced subsocial carpenter bee *Ceratina calcarata* [102], which is considered a putative model species for understanding social evolution based on matrilineal association [105].

As the assembly for the *F. varia* genome, generated in two steps from Illumina reads (Table 1), is still rather fragmented, we conducted a syntenic analysis as guidance for a future hybrid assembly, as recently completed for the honey bee [3]. This analysis showed that several genes placed on the smaller *F. varia* scaffolds had their respective orthologs placed together on one of the much larger, chromosome-size linkage groups of the honey bee (Fig. 3). Such correspondence in scaffold/chromosome localization for orthologous genes indicates a considerable degree of conservation in overall genome organization across bee genomes, despite the divergence time between the tribes Apini and Meliponini of more than 80 million years [5], and the differing ancestral haploid chromosome numbers between Apini (16) and Meliponini (range 8–18, with \(n = 9, n = 15,\) and \(n = 17\) as the main groups, with an \(n = 15\) for *F. varia* [106]). High conservation of the genetic structure was also observed between the genomes of *A. mellifera* and *B. terrestris* [103] and between *A. mellifera* and *E. dilemma* [69], suggesting slow genome evolution in Apidae. Furthermore, as a byproduct of this syntenic analysis we were able to take a closer look at some gene clusters that are known play a role in insect and specifically bee biology, such as the Major Royal Jelly Proteins (MRJPs), the Osiris gene cluster, and the *pbn1* QTL genes.

Concerning MRJP genes, already previous analyses [8, 77, 78] had shown that the tandem array of nine functional MRJP genes in the genus *Apis* represents a taxon-specific expansion expansion, and that all the other corbiculate bees, including stingless bees, have only one ortholog named *mrjp-like*. In honey bees, MRJPs are the main proteinaceous components in royal and worker jelly, but in larval food of the stingless bee *M. quadrifasciata* no MRJP-like proteins were detected by a proteomics analysis (S. Albert, G.J. Tibério, K. Hartfelder, unpublished data; MALDI/TOF/TOF analysis of tryptic peptide fragments from SDS-PAGE gel pieces of the size range 20–60 kDa). Furthermore, we also could not identify an ortholog of the *apisimin* gene in the genomes of *F. varia* and *M. quadrifasciata*. Hence, the peculiar molecular architecture combining native MRJP1, apisimin and 24-methylene cholesterol in honey bee royal jelly [107] seems to be specific to the genus *Apis*. It is likely crucial for the viscosity of royal jelly that makes it possible that honey bees can rear their queen larvae in downward pointing royal cells [108]. In contrast, stingless bees rear their brood in horizontal cells, with the larvae floating on top of a defined portion of larval food. Their larval food also has a species-specific composition of proteins, but this is quite different from that of the honey bees [109]. With this in mind, the function of the *mrjp-like* gene in stingless bees deserves a closer look, especially considering the diverse modes of caste determination in the Meliponini.

The Osiris genes of *F. varia* (Fig. 3b) form a highly conserved cluster of 18 genes. An Osiris cluster is thought to have originated already very early in the evolution of winged insects [4, 79]. Astonishingly, despite this ancient conservation in genomic organization, the Osiris genes have no Gene Ontology terms associated concerning Molecular Function, and only a putative functional protein domain is predicted. Gene expression analyses, however, showed a correlation between

### Table 6
Comparison of gene numbers with GO-BP terms associated with reproductive process in families of bees (Apidae, Megachilidae, Halictidae), the parasitic wasp *Nasonia vitripennis* and *Drosophila melanogaster*

| Species          | Genes |
|------------------|-------|
| Apidae           |       |
| *Apis mellifera* | 149   |
| *Apis dorsata*   | 132   |
| *Apis florea*    | 136   |
| *Bombus impatiens* | 140   |
| *Bombus terrestris* | 138   |
| *Eufriesea mexicana* | 117   |
| *Euglossa dilemma* | 91    |
| *Friesoeomelitta varia* | 127 |
| *Melipona quadrifasciata* | 124 |
| *Habropoda laboriosa* | 129 |
| Megachilidae     |       |
| *Megachile rotundata* | 123  |
| Halictidae       |       |
| *Dufourea novaangliae* | 126  |
| *Lasiosglossum albipes* | 127  |
| Pteromalidae     |       |
| *Nasonia vitripennis* | 130  |
| Drosophilidae    |       |
| *Drosophila melanogaster* | 106  |
neighborhood distances of the Osiris genes within the cluster to the timing of developmental events, especially so cuticle formation in the embryo and during metamorphosis, as well as differences in Osiris gene expression associated with caste development in social insects [110]. Besides their strong correlation with the timing of cuticle formation, co-expression analyses indicate that they are also involved in immune system functions and detoxification reactions in the red flower beetle, Tribolium castaneum [111].

As said, the finding of a conserved cluster of pln1 QTL genes (Fig. 3c) was unexpected, as this QTL was identified through a selection program for pollen hoarding behavior in honey bees [80]. This is a colony behavioral trait that was subsequently found to have a strong association with gustatory responses and reproductive traits of A. mellifera workers [81–83]. At present, without more in-depth analyses concerning the extent to which the pln1 cluster may be conserved across bees, we can only speculate on the possible meaning of this architectural genomic conservation. The collection of pollen, which underlies the pollen hoarding syndrome in A. mellifera, was a, if not the crucial step in the life style transition that originated the bees from a wasp ancestor. A recent phylogeny proposed the tiny Ammoplatina wasps as the sister group of the bees (Anthophila) [112]. These wasps hunt thrips on flowers, a behavior which would have facilitated the transition from a predatory to a pollen collecting life style. This evolutionary scenario of the bees could explain the conservation seen in the pln1 gene cluster, a genetic module that likely shifted and directed a wasp’s predatory food collecting behavior.
to a flower resource. Testing this hypothesis will, however, require further evidence from genomic comparisons across different families of bees and wasps.

The mitochondrial genome

With the COX2 sequence of Bombus hypocrita sappor-ensis as seed, the mitochondrial genome of F. varia was extracted from the trimmed genomic paired-end reads and assembled into a circularized gene order using the organelle-specific software NOVOPlasty (Fig. 4a). The size of the F. varia mitochondrial genome, its A + T composition, and also the number of genes was found to be consistent with the mitochondrial genomes of other insects [84]. Furthermore, the alignment of the mitochondrial genes of bees of the family Apidae and their group relationship of F. varia and Figure S2-S6) showed high support for a sister relationship and Maximim Likelihood approaches (Fig. 4b representation in a phylogenetic tree using Bayesian in-ference and maximim likelihood approaches (1000 replicates). In red are the respective orthologs for (See figure on previous page.)

Fig. 6 Gene tree for the key gene set of Notch signaling in bees. Orthologous aminoacid sequences were aligned using MAFFT, and the tree was generated using a Maximum Likelihood approach (1000 replicates). In red are the respective orthologs for F. varia. Gene names are abbreviated as: ADAM, ADAM 17-like protease; aph1, anterior pharynx defective 1; CIR, corepressor interacting with RBP1; CtxP, C-terminal binding protein; Dl, delta; DvI, segment polarity protein dishevelled homolog DVL-3; dx, deltex; frng, fringe; gro, groucho; HATs, histone acetyltransferase KAT2A; HDAC1, histone deacetylase 1; N, notch; nct, nicastriini; numb, numb; PSE2, presenilin enhancer; Psn, presenilin; Ser, serrate; SK9P, puff-specific protein Bx42; Su(H), suppressor of hairless. Species names are given as three letter acronyms: Acer, Apis cerana; Ador, Apis dorsata; Aflo, Apis florea; Amel, Apis mellifera; Bimp, Bombus impatiens; Bter, Bombus terrestris; Dnov, Dufourea novaeangliae; Edil, Euglossa dilemma; Emex, Euglossa mexicana; Fvar, Frieseomelitta varia; Hlab, Habropoda laboriosa; Llab, Lasioglossum albipes; Mrot, Megachile rotundata; Mqua, Melipona quadridasciata. Scale bars indicate the number of substitutions per site compared to members of three other tribes within the Corbiculata clade, Apis, Bombus, and Melipona (Fig. 4c). Nonetheless, while tempting, it is still early to speculate whether stingless bees may be a hotspot in mitochondrial genome evolution. In the genus Apis, dozens of mitogenomes have been sequenced for practically all the species, subspecies, and even local populations, but with respect to stingless bee, we are aware of only three fully sequenced mitogenomes, including those of M. bicolor [88], M. scutellaris [113], and now F. varia. The finding of such extensive reshuffling in gene order should, thus, clearly be an incentive for future stingless bee mitogenome sequencing projects.

Gene families and functional groups of specific interest

In the first comparative genomics study on social evolution in bees [8], transcription regulating factors, including transctiption factors (TFs) and their respective binding sites in cis-regulatory regions of potential target genes, as well as egigenetic modification and genes involved in these processes were pointed out as possible hallmarks. Hence, we performed here specific searches for TFs and epigenetic modifier genes in the F. varia

Table 7 Number of genes related to immunity pathways and functions in 11 bee species with complete genome information, including 10 previously published species [8, 67] and Frieseomelitta varia (in bold)
genome. With regard to TF families we found figures very similar to the previously reported ones [8], and this was also the case for epigenetic modifier genes.

Amongst epigenetic mechanisms, DNA methylation is clearly the best studied in the context of reproduction and behavior in social insects [89]. Nonetheless, one aspect that has recently come to light in a study on the emerging insect model organism *Tribolium castaneum*, namely the interaction between genes promoting histone post-translational modifications and juvenile hormone (JH) functions [114], could be specifically relevant with relation to reproduction and aging in highly social insects. JH is a key regulator of insect development and reproduction, as well as of caste development in social insects [115], and JH levels in the life cycle of a stingless bee, *M. scutellaris*, were recently determined [116]. Hence, we consider that epigenetic modifier genes interacting with key insect hormones and long non-coding RNAs [117] may also play a critical role in caste development and reproduction, especially so the phenomenon of complete worker sterility in *F. varia*.

**Table 8** Neurogenesis-related genes and percentage relative to total gene number in bee genomes compared to humans and *Drosophila*

| Species              | Neurogenesis-related genes | Total gene number | neurogenic/total genes (%) |
|----------------------|---------------------------|-------------------|---------------------------|
| *Homo sapiens*       | 1586                      | 19,986            | 7.94                      |
| *Frieseomelitta varia* | 674                      | 10,346            | 6.51                      |
| *Habropoda laboriosa* | 715                      | 12,279            | 5.82                      |
| *Dufourea novaangliae* | 721                      | 12,453            | 5.79                      |
| *Eufriesea mexicana*  | 696                      | 12,022            | 5.79                      |
| *Bombus impatiens*   | 732                      | 13,050            | 5.61                      |
| *Megalochile ratundata* | 717                      | 12,770            | 5.61                      |
| *Lasius niger*       | 722                      | 13,448            | 5.37                      |
| *Drosophila melanogaster* | 954                      | 17,792            | 5.36                      |
| *Bombus terrestris*  | 672                      | 12,648            | 5.31                      |
| *Apis mellifera*     | 717                      | 15,314            | 4.68                      |
| *Melipona quadrifasciata* | 703                      | 15,368            | 4.57                      |
| *Euglossa dilemma*   | 679                      | 15,904            | 4.27                      |
For a better understanding of other key processes associated with hymenopteran eusociality, such as immune system functions, reproductive biology, nervous system organization, circadian rhythm and the chemosensory system, we performed specific searches for such genes in the *F. varia* genome. We conducted a manual curation of their automatically generated predictions and a phylogenetic comparison with their orthologs in the currently sequenced bee genomes.

As for immunity-related genes, we could clearly identify 174 of the 190 genes predicted in the honey bee genome, while for the remainder we could not find orthologs in the *F. varia* genome. This could be due to the fact that the *F. varia* genome is still split into many scaffolds, as well as to the fact that some of these “missing genes” were also not found in the genomes of other social bees. Nonetheless, the phylogenetic analyses for the primary immune gene families showed that the *F. varia* genes had their closest orthologs in *M. quadrifasciata* and the two *Bombus* species (Figures S7 and S8).

Considering the phenomenon of worker sterility in *F. varia* [23], based on literature searches and GO terms we compiled a list of 127 candidate genes that are involved in reproductive processes, especially ovary function in insects. Among these, approximately 50% are related to major signaling pathways. As shown in Fig. 5, not all the genes related to these signaling pathways had been correctly predicted, including *fringe*, which is an important member of the Notch pathway, and this would have been astounding, since the Notch pathway has recently been identified as a key regulator for differential ovarian activity in honey bee queens and workers [118]. Nonetheless, by a specific search we found this gene as split between two scaffolds. Hence, besides finding “missing genes” this strategy, though laborious, can help to bridge scaffolds.

### Table 9

Comparison of the number of genes encoding for odorant receptors (ORs), odorant binding proteins (OBPs), chemosensory protein (CSPs), gustatory receptors (GRs), ionotrophic receptors (IRs), and sensory neuron membrane proteins (SNMPs) identified in the genomes of the honeybees (*Apis mellifera* and *Frieseomelitta varia*) and two stingless bees (*Melipona quadrifasciata* and *Frieseomelitta varia*)

| Family | *A. mellifera* | *M. quadrifasciata* | *F. varia* | References |
|--------|----------------|---------------------|------------|------------|
| ORs    | 177            | 196                 | –          | [98, 99]   |
|        | –              | –                   | 102        | This study |
| OBPs   | 21             | 14                  | –          | [99]       |
|        | –              | –                   | 8          | This study |
| CSPs   | 6              | 6                   | –          | [99]       |
|        | –              | –                   | 5          | This study |
| GRs    | 14             | 16                  | –          | [99]       |
|        | –              | –                   | 10         | This study |
| IRs    | 10             | 10                  | –          | [99]       |
|        | 16             | 17                  | 15         | This study |
| SNMPs, Group 3 | –       | –                   | –          | [99]       |
|        | 2              | 3                   | 3          | This study |

*Fig. 8* Neurogenesis-related genes. Venn diagram of direct orthologs of neurogenesis-related genes in the human (Hsap), *Drosophila* (Dmel), honey bee (Amel), and *F. varia* (Fvar) genomes.
Another aspect of prime importance in social insects is communication, and in this respect, bees, like most other insects, make extensive use of olfactory information in fundamental behaviors, such as food source location and recognition of conspecifics [119]. Social organization imposes an extra layer of complexity, including food source communication and brood care by workers, reproductive division of labor among queens and workers, nest defense, and in the adjustment in worker behavioral maturation with colony demography. Our results for *F. varia* are consistent with those reported for chemosensory genes in corbiculate bees and wasps, both in terms of gene numbers and the *in tandem* clustered organization in some of these [99, 119]. We also found a slightly larger number of SNMP genes in the two stingless bees (two SNMP1 and one SNMP2) compared to *A. mellifera* (one SNMP1 and one SNMP2), suggesting a possible gene loss in honey bees.

Integration of social behaviors among colony members also requires synchronization of circadian cycles. As such, the circadian clock of bees is thought to control most of the behavioral aspects of division of labor and synchronization of worker bee activities [120]. For *F. varia*, the molecular functioning and integration of clock genes in generating circadian rhythms is probably similar to *A. mellifera* [121], with a first autoregulatory feedback loop composed by Clock, Cycle, Period, and Cryptochrome mammalian-like, and a second autoregulatory feedback loop involving Clock, Par domain protein 1, and Vrille. Timeout 2, but not its paralog Timeless, is only present in Hymenoptera [122], and the *Drosophila* ortholog Clockwork orange appear to be involved in different clock functions [123, 124]. Chronobiology studies making use of this genomic information may provide important insights into the evolution of the different life styles seen in stingless bee species.

**Conclusions**

The species *F. varia* was chosen for this genome project because of its peculiar reproductive biology. Different from most other stingless bee species, *F. varia* workers are completely sterile and, thus represent an extreme end in the queen/worker and worker/worker conflict of interest over male production in a bee colony. For insights into genomic signatures related to reproduction and social behaviors, we checked and manually curated hundreds of gene models predicted for genes related to reproductive functions, including several core signaling pathways, as well as genes related to immunity, neurogenesis and chemosensory processes. Furthermore, we generated predictions for genes containing transcription factor domains, repetitive genomic components, and non-coding RNAs, including over 1500 lncRNAs.

Specific highlights of the genome analysis were the finding of a complete and highly conserved Osiris gene cluster in a single scaffold of the *F. varia* genome and the conservation of the *pln1* QTL. The Osiris gene cluster is a unique characteristic of insect genomes. Yet, despite its ancient origin and high genomic conservation, the molecular functions of these genes are still enigmatic. Clearly unexpected was the conservation in the *F. varia* genome of the *pln1* QTL found in the honey bee. This QTL has been identified and described as a key element in the reproductive ground plan underlying the social behavior of honey bees. Clearly there was no *a priori* reason to believe that this gene cluster would be so conserved also in stingless bees, considering the divergence time of over 80 mya between honey bees and stingless bees and the much more recent origin of the genus *Apis* compared to the stingless bees. This suggests that a QTL with genes related to pollen hoarding behavior may have arisen already at an early time in the phylogeny of the Apidae and since remained stabilized.

In contrast to the examples of conservation in the nuclear genome, the mitochondrial genome of *F. varia* revealed major changes in gene order in comparison to the honey bee and also to the other two stingless bee mitogenomes sequenced so far. For *F. varia* we found evidence for major reshuffling events in gene order that should be of interest for the analysis of other Meliponini mitogenomes, as this group may represent a hotspot in mitogenome structural evolution.

With several hundred species distributed in 61 genera worldwide, the stingless bees are the only bees that equal honey bees in terms of social organization and complexity. Hence, their ecological and economic importance, as well as their biological variability should more than justify further genome sequencing efforts in the tribe Meliponini to reveal genomic signatures associated with the evolution of advanced sociality in this group, in comparison and contrast to the tribe Apini.

**Supplementary information**

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-06784-8.

Additional file 1: Figure S1: Multiple alignment (A) and Maximum Parsimony phylogenetic tree (B) of miR-34.

Additional file 2: Figure S2 Phylogenetic tree based on mitochondrial genomes protein coding sequences dataset of Apoidea. The tree was generated by Bayesian Inference; the values above the branches show the posterior probability for Bayesian Inference and values below the branches represent bootstrap support values of the Maximum Likelihood analyses.

Additional file 3: Figure S3 Phylogenetic tree generated by Maximum Likelihood method, based on the complete mitochondrial genomes dataset of Apoidea.
Additional file 4: Figure S4 Phylogenetic tree generated by Maximum Likelihood method, based on mitochondrial genomes protein coding sequences dataset of Apoidea.

Additional file 5: Figure S5 Phylogenetic tree based on mitochondrial genomes rRNAs sequences dataset of Apoidea. The tree was generated by Bayesian Inference; the values above the branches show the posterior probability for Bayesian Inference and values below the branches represent bootstrap support values of the Maximum Likelihood analyses.

Additional file 6: Figure S6 Phylogenetic tree generated by Maximum Likelihood method, based on mitochondrial genomes rRNAs sequences dataset of Apoidea.

Additional file 7: Figure S7 Unrooted phylogenetic trees for core set of genes of the Toll (A) and IMD/JNK pathways. Amino acid sequences were aligned using MAFFT and the tree was generated in an ML approach (1000 replicates). In red are the orthologs of Frieseomelitta varia. Gene names are abbreviated. Species names are in three letters acronyms, Acer: Apis cerana, Ador: Apis dorsata, Aflo: Apis florea, Amel: Apis mellifera, Bter: Bombus terrestris, Birnp: Bombus impatiens, Dniv: Dufourea novaangliae, Edil: Euglossa dilemma, Emex: Euglossa mexicana, Fvar: Frieseomelitta varia, Hlb: Habropoda labiosa, Lall: Lasiosglossum albilab, Mot: Megachile rotundata, Mqua: Melipona quadridactyla.

Additional file 8: Figure S8 Unrooted phylogenetic trees for core set of genes of the JAK/STAT and RNAi pathways. Amino acid sequences were aligned using MAFFT and the tree was generated in an ML approach (1000 replicates). In red are the orthologs of Frieseomelitta varia. Gene names are abbreviated. Species names are in three letters acronyms, Acer: Apis cerana, Ador: Apis dorsata, Aflo: Apis florea, Amel: Apis mellifera, Bter: Bombus terrestris, Birnp: Bombus impatiens, Dniv: Dufourea novaangliae, Edil: Euglossa dilemma, Emex: Euglossa mexicana, Fvar: Frieseomelitta varia, Hlb: Habropoda labiosa, Lall: Lasiosglossum albilab, Mot: Megachile rotundata, Mqua: Melipona quadridactyla.

Additional file 9: Table S1 Genome databases used in blastp searches for protein-coding genes included in the gene set for manual curation of their MAKER 2 gene model predictions.

Additional file 10: Table S2 Genome databases used in the prediction of non-coding genes in the F. varia genome assembly.

Additional file 11: Table S3 Mitochondrial genomes used for the Apoidea phylogenetic tree reconstruction. * denotes the species used as outgroups.

Additional file 12: Table S4 Results of the manual curation of 533 gene models generated by automatic prediction.

Abbreviations

COX2: Cyclooxygenase 2; EC: Enzyme commission; DNMT: DNA methyltransferase; GO: Gene Ontology; HPTML: Histone post-translational modification; JH: Juvenile hormone; MRJP: Major Royal Jelly Protein; mya: million years ago; OB9P: Osrangot binding protein 9; QTL: Quantitative trait locus; RAXML: Random accelerated maximum likelihood; TF: transcription factor

Acknowledgments
We thank Jairo de Souza for maintenance of the F. varia colonies and Marcela Laure and Tiago Falcón Lopes for collecting male specimens for DNA extraction. We also thank the two anonymous reviewers for their insightful comments.

Authors’ contributions

MMGB, ZLPS, KH, FMFN and ARB conceived and supervised the project, FMFN and FCPA performed the bioinformatics analyses for assembly, transposable elements, synteny, and gene model predictions for coding and non-coding RNAs, assisted by ARB, and MFNF, DGP: SDRS, assembled and analyzed the mitochondrial genome, assisted by MFNF and LB; CACJ, CGS, DCL, DES, DLL, FCPA, FCS, FD, FOB, FOR, JH, KMF, LB, LVRM, LN, MADL, MCI, MP, MSC, PSG, and TSD performed the manual curation of gene sets; KH wrote the draft and revisions of this manuscript and all authors approved its final version.

Funding

This project was funded by the Brazilian National Science Fund (Conselho Nacional de Desenvolvimento Científico e Tecnológico, grant number 454103/2014-0) and the São Paulo Research Foundation (FAPESP, grant number 2015/06657–0). FCPF and several of the participating postgraduate students received fellowships from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES - Finance Code 001) and FAPESP. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The F. varia genome sequence assembled as Fvar_v1.2 was submitted to the NCBI genome database (GenBank accession: GCA_001392965.1), and the F. varia mitochondrial genome assembly was submitted to the NCBI nucleotide database (GenBank accession: CM022150.1). For detailed documentation on assembly and evaluation procedures, including scripts, see https://github.com/dgpinheiro/Varia.

Ethics approval and consent to participate

Biospecimens were collected in accordance with Brazilian laws. F. varia individuals were sampled under the registers: SISBIO, license number: 41883–1; authentication number: 71367685; and license number: 41883–2; authentication number: 39944978. SisGen, certificate number: A6EC418. The species does not fall under the IUCN Red List categories as a threatened species.

Consent for publication

Not applicable.

Competing interests

The authors declare no conflict of interest related to the results reported in this study.

Author details

1Departamento de Genética, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, Brazil. 2Departamento de Biologia Celular e do Desenvolvimento, Instituto de Ciências Biomédicas, Universidade Federal de Alfenas, Alfenas, MG, Brazil. 3Departamento de Ciências Biológicas, Faculdade de Ciências Biológicas e da Saúde, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina, MG, Brazil. 4Departamento de Genética e Evolução, Centro de Ciências Biológicas e da Saúde, Universidade Federal de São Carlos, São Carlos, SP, Brazil. 5Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brazil. 6Departamento de Biologia Celular e Molecular e Bioagentes Patogênicos, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Av. Bandeirantes 3900, Ribeirão Preto, SP 14049-900, Brazil. 7Departamento de Tecnologia, Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista “Júlio de Mesquita Filho”, Jaboatão, SP, Brazil. 8Departamento de Biologia, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, Brazil.

Received: 18 November 2019 Accepted: 14 May 2020

Published online: 03 June 2020

References

1. The Honey Bee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee Apis mellifera. Nature. 2006;443:931–49.
2. Elsk CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, et al. Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genomics. 2014;5:86.
3. Wallberg A, Bunkis I, Winnere-Petterson O, MÔsbêch M-B, Childers AK, Evans JD, Mikheyev AS, Roberysons HM, Robibson GE, Webster MT. A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. BMC Genomics. 2019;20:275.
4. Miyashita S, Liu S, Meisemann K, Peters RS, Donath A, Mayer C, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science. 2014;346763–7.
5. Cardinal S, Straka J, Danforth BN. Comprehensive phylogeny of apid bees reveals the evolutionary origins and antiquity of cleptoparasitism. Proc Natl Acad Sci U S A. 2010;107:16207–11.
6. Hedlie SM, Patiny S, Darforth BN. The bee tree of life: a supermatrix approach to avoid phylogeny and biogeography. BMC Evol Biol. 2013;13:138.

7. Michener CD. The Bees of the World. 2nd ed. Baltimore: Johns Hopkins University Press; 2007. p. 953.

8. Kapheim KM, Pan H, Li C, Salzberg SL, Puiu D, Magoc T, et al. Genomic signatures of evolutionary transitions from solitary to group living. Science. 2015;348:1139–43.

9. Diao Q, Sun L, Zheng H, Zeng Z, Wang S, Xu S, et al. Genomic and transcriptomic analysis of the Asian honeybee Apis cerana provides novel insights into honeybee biology. Sci Rep. 2018;8:822.

10. Oppenheim S, Cao X, Ruepell O, Rondong S, Phokasem P, Desalle R, et al. Whole genome sequencing and assembly of the Asian honey bee Apis dorsata. Genome Biol. Evol. 2020;12:2677–83.

11. Rasmussen C, Camero SA. Global stingless bee phylogeny supports ancient divergence, vicariance, and long distance dispersal. Biol J Linn Soc. 2010;99:206–32.

12. Camargo JMF, Pedro SWM. Meliponina Lepeletier, 1936. In: Moure JS, Urban D and Melo GAR, editors. Catalogue of bees (Hymenoptera, Apoidea) in the Neotropical region. Curitiba: Sociedade Brasileira de Entomologia; 2007. p. 272–578.

13. Hurtado-Burillo M, Lara J, May-Tzé W, Quezada-Euan JIG, Ruiz C, De la Rúa P. A geometric morphometric and microsatellite analyses of Scaptotrigona mexicana and S. pectoralis (Apidae: Meliponini) sheds light on the biodiversity of Mesoamerican stingless bees. J Insect Conserv. 2016;20:753–63.

14. Imperatriz-Fonseca VL, Pope N, Carvalho AT, Maia UM, Blochtein B, de Carvalho CAL, et al. Stingless bees. In: Hermann HR, editor. Social insects, vol. 3. Oxford: EUP; 2015. p. 361.

15. Jaffé R, Pope N, Carvalho AT, Maia UM, Blochtein B, de Carvalho CAL, et al. Stingless bees. In: Hermann HR, editor. Social insects, vol. 3. Oxford: EUP; 2015. p. 361.

16. Sakagami SF. Stingless bees. In: Hermann HR, editor. Social insects, vol. 3. Oxford: EUP; 2015. p. 361.

17. Rasmussen C, Cameron SA. Global stingless bee phylogeny supports ancient divergence, vicariance, and long distance dispersal. Biol J Linn Soc. 2010;99:206–32.

18. Camargo JMF, Pedro SWM. Meliponina Lepeletier, 1936. In: Moure JS, Urban D and Melo GAR, editors. Catalogue of bees (Hymenoptera, Apoidea) in the Neotropical region. Curitiba: Sociedade Brasileira de Entomologia; 2007. p. 272–578.

19. Hurtado-Burillo M, Lara J, May-Tzé W, Quezada-Euan JIG, Ruiz C, De la Rúa P. A geometric morphometric and microsatellite analyses of Scaptotrigona mexicana and S. pectoralis (Apidae: Meliponini) sheds light on the biodiversity of Mesoamerican stingless bees. J Insect Conserv. 2016;20:753–63.

20. Imperatriz-Fonseca VL, Pope N, Carvalho AT, Maia UM, Blochtein B, de Carvalho CAL, et al. Stingless bees. In: Hermann HR, editor. Social insects, vol. 3. Oxford: EUP; 2015. p. 361.

21. Jaffé R, Pope N, Carvalho AT, Maia UM, Blochtein B, de Carvalho CAL, et al. Stingless bees. In: Hermann HR, editor. Social insects, vol. 3. Oxford: EUP; 2015. p. 361.

22. Sakagami SF. Stingless bees. In: Hermann HR, editor. Social insects, vol. 3. Oxford: EUP; 2015. p. 361.

23. Rasmussen C, Camero SA. Global stingless bee phylogeny supports ancient divergence, vicariance, and long distance dispersal. Biol J Linn Soc. 2010;99:206–32.

24. Camargo JMF, Pedro SWM. Meliponina Lepeletier, 1936. In: Moure JS, Urban D and Melo GAR, editors. Catalogue of bees (Hymenoptera, Apoidea) in the Neotropical region. Curitiba: Sociedade Brasileira de Entomologia; 2007. p. 272–578.

25. Hurtado-Burillo M, Lara J, May-Tzé W, Quezada-Euan JIG, Ruiz C, De la Rúa P. A geometric morphometric and microsatellite analyses of Scaptotrigona mexicana and S. pectoralis (Apidae: Meliponini) sheds light on the biodiversity of Mesoamerican stingless bees. J Insect Conserv. 2016;20:753–63.

26. Imperatriz-Fonseca VL, Pope N, Carvalho AT, Maia UM, Blochtein B, de Carvalho CAL, et al. Stingless bees. In: Hermann HR, editor. Social insects, vol. 3. Oxford: EUP; 2015. p. 361.

27. Jaffé R, Pope N, Carvalho AT, Maia UM, Blochtein B, de Carvalho CAL, et al. Stingless bees. In: Hermann HR, editor. Social insects, vol. 3. Oxford: EUP; 2015. p. 361.

28. Sakagami SF. Stingless bees. In: Hermann HR, editor. Social insects, vol. 3. Oxford: EUP; 2015. p. 361.

29. Rasmussen C, Cameron SA. Global stingless bee phylogeny supports ancient divergence, vicariance, and long distance dispersal. Biol J Linn Soc. 2010;99:206–32.

30. Camargo JMF, Pedro SWM. Meliponina Lepeletier, 1936. In: Moure JS, Urban D and Melo GAR, editors. Catalogue of bees (Hymenoptera, Apoidea) in the Neotropical region. Curitiba: Sociedade Brasileira de Entomologia; 2007. p. 272–578.

31. Hurtado-Burillo M, Lara J, May-Tzé W, Quezada-Euan JIG, Ruiz C, De la Rúa P. A geometric morphometric and microsatellite analyses of Scaptotrigona mexicana and S. pectoralis (Apidae: Meliponini) sheds light on the biodiversity of Mesoamerican stingless bees. J Insect Conserv. 2016;20:753–63.

32. Imperatriz-Fonseca VL, Pope N, Carvalho AT, Maia UM, Blochtein B, de Carvalho CAL, et al. Stingless bees. In: Hermann HR, editor. Social insects, vol. 3. Oxford: EUP; 2015. p. 361.

33. Sakagami SF. Stingless bees. In: Hermann HR, editor. Social insects, vol. 3. Oxford: EUP; 2015. p. 361.

34. Rasmussen C, Camero SA. Global stingless bee phylogeny supports ancient divergence, vicariance, and long distance dispersal. Biol J Linn Soc. 2010;99:206–32.

35. Camargo JMF, Pedro SWM. Meliponina Lepeletier, 1936. In: Moure JS, Urban D and Melo GAR, editors. Catalogue of bees (Hymenoptera, Apoidea) in the Neotropical region. Curitiba: Sociedade Brasileira de Entomologia; 2007. p. 272–578.

36. Hurtado-Burillo M, Lara J, May-Tzé W, Quezada-Euan JIG, Ruiz C, De la Rúa P. A geometric morphometric and microsatellite analyses of Scaptotrigona mexicana and S. pectoralis (Apidae: Meliponini) sheds light on the biodiversity of Mesoamerican stingless bees. J Insect Conserv. 2016;20:753–63.

37. Imperatriz-Fonseca VL, Pope N, Carvalho AT, Maia UM, Blochtein B, de Carvalho CAL, et al. Stingless bees. In: Hermann HR, editor. Social insects, vol. 3. Oxford: EUP; 2015. p. 361.

38. Sakagami SF. Stingless bees. In: Hermann HR, editor. Social insects, vol. 3. Oxford: EUP; 2015. p. 361.

39. Rasmussen C, Cameron SA. Global stingless bee phylogeny supports ancient divergence, vicariance, and long distance dispersal. Biol J Linn Soc. 2010;99:206–32.

40. Camargo JMF, Pedro SWM. Meliponina Lepeletier, 1936. In: Moure JS, Urban D and Melo GAR, editors. Catalogue of bees (Hymenoptera, Apoidea) in the Neotropical region. Curitiba: Sociedade Brasileira de Entomologia; 2007. p. 272–578.

41. Hurtado-Burillo M, Lara J, May-Tzé W, Quezada-Euan JIG, Ruiz C, De la Rúa P. A geometric morphometric and microsatellite analyses of Scaptotrigona mexicana and S. pectoralis (Apidae: Meliponini) sheds light on the biodiversity of Mesoamerican stingless bees. J Insect Conserv. 2016;20:753–63.

42. Imperatriz-Fonseca VL, Pope N, Carvalho AT, Maia UM, Blochtein B, de Carvalho CAL, et al. Stingless bees. In: Hermann HR, editor. Social insects, vol. 3. Oxford: EUP; 2015. p. 361.

43. Sakagami SF. Stingless bees. In: Hermann HR, editor. Social insects, vol. 3. Oxford: EUP; 2015. p. 361.

44. Rasmussen C, Camero SA. Global stingless bee phylogeny supports ancient divergence, vicariance, and long distance dispersal. Biol J Linn Soc. 2010;99:206–32.
56. Hahn C, Bachmann L, Chevreux B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - a baiting and iterative mapping approach. Nucleic Acids Res. 2013;41:e129.
57. Hunt M, Kikuchi T, Sanders M, Newbold C, Bermimam M, Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 2013;14:R47.
58. Darling ACE, Mau B, Blattner FR, Perna NT. MALVE: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14:1394–403.
59. Araujo NS, Arias MC. Mitochondrial genome characterization of Melipona bicolor: insights from the control region and gene expression data. Gene. 2019;70555–9.
60. Lowe TM, Chan PP. tRNAscan-SE on-line: integrating search and context for tRNAs. Nucleic Acids Res. 2004;32:128–31.
61. Laslett D, Canbäck B. ARWEN, a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics. 2008;24:172–5.
62. Lohse M, Drechsel O, Bock R. OrganellarGenomeDRAW (OGDRAW) - a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr Genet. 2007;52:267–74.
63. Yamada KD, Tomik J, Katoh K. Application of the MAFFT sequence alignment program to large data reexamination of the usefulness of chained guide trees. Bioinformatics. 2016;32:246–51.
64. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML web servers. Syst Biol. 2008;57:758–71.
65. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogeny. Bioinformatics. 2001;17:754–5.
66. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012;9:772.
67. Stüver BC, Müller KF. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics. 2010;11:7.
68. Sadd BM, Barribeau SM, Bloch G, de Graaf DC, Dearden P, Elsik CG, et al. The dogwood tree (Cornus florida) genome reveals a complex evolutionary history. Genome Res. 2015;25:100–11.
69. Valles SM, Oi DH, Yu F, Tan XX, Buss EA. Metatranscriptomics and pyrosequencing facilitate discovery of potential viral natural enemies of the invasive Caribbean crazy ant, Melophorinae: exploring the potential of RNA-based methods. DNA Res. 2017;24:279–86.
70. Bernal A, Crammer H, Hatzigeorgiou A, Pereira P. Global discriminative gene prediction on gene loss estimates in eukaryotes by quantifying falsely inferred absences PLoS Comput Biol. 2019;15:e1007301.
71. Yamada KD, Tomik J, Katoh K. Application of the MAFFT sequence alignment program to large data reexamination of the usefulness of chained guide trees. Bioinformatics. 2016;32:246–51.
72. Pagel MR, Meyers AS. AML: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics. 2008;24:172–5.
73. Rahmsdorff O, Amad DT, Hahn C. Genetic so fr re product i on and re g u l at i on of honey bees (Apis mellifera L.) from a metagenome perspective. possibly affected by the presence of full-length mitochondrial genes. Genome Res. 2016;26:353–6.
74. Bernal A, Crammer H, Hatzigeorgiou A, Pereira P. Global discriminative gene prediction on gene loss estimates in eukaryotes by quantifying falsely inferred absences PLoS Comput Biol. 2019;15:e1007301.
75. Pires CV, Freitas FC, Cristino AS, Dearden PK, Simões ZL. Transcriptome effects of division of labor and foraging specialization in a social insect (Apis mellifera L.). Curr Topics Dev Biol. 2006;74:283–86.
76. Pires CV, Freitas FC, Cristino AS, Dearden PK, Simões ZL. miR-34 is maternally inherited in the honey bee. Mol Genet Genomics. 2008;278:995–1001.
77. Pires CV, Freitas FC, Cristino AS, Dearden PK, Simões ZL. miR-34 is maternally inherited in the honey bee. Mol Genet Genomics. 2008;278:995–1001.
78. Hahn C, Bachmann L, Chevreux B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - a baiting and iterative mapping approach. Nucleic Acids Res. 2013;41:e129.
79. Hunt M, Kikuchi T, Sanders M, Newbold C, Bermimam M, Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 2013;14:R47.
80. Darling ACE, Mau B, Blattner FR, Perna NT. MALVE: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14:1394–403.
81. Araujo NS, Arias MC. Mitochondrial genome characterization of Melipona bicolor: insights from the control region and gene expression data. Gene. 2019;70555–9.
82. Lowe TM, Chan PP. tRNAscan-SE on-line: integrating search and context for tRNAs. Nucleic Acids Res. 2004;32:128–31.
83. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML web servers. Syst Biol. 2008;57:758–71.
84. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogeny. Bioinformatics. 2001;17:754–5.
85. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012;9:772.
86. Stüver BC, Müller KF. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics. 2010;11:7.
87. Sadd BM, Barribeau SM, Bloch G, de Graaf DC, Dearden P, Eltik CG, et al. The genomes of two key bumblebee species with pristive social organization. Genome Biol. 2015;16:76.
88. Bernal A, Crammer H, Hatzigeorgiou A, Pereira P. Global discriminative learning for higher-accuracy computational gene prediction. PLoS Comput Biol. 2007;3:e64.
89. Bradnam KR, Korf I. Longer first introns are a general property of eukaryotic gene structure. PLoS One. 2008;3:e3093.
90. Soni K, Choudhary A, Patowary A, Singh AR, Bhata A, Sivasubbu S, et al. miR-34 is maternally inherited in Drosophila melanogaster and Danio rerio. Nucleic Acids Res. 2013;41:4470–80.
91. Pires CV, Freitas FC, Cristina AS, Dearden PK, Simões ZL. Transcriptome analysis of honeybee (Apis mellifera) haploid and diploid embryos reveals early zygotic transcription during cleavage. PLoS One. 2016;11:e0146447.
92. Freitas FC, Pires CV, Claudianos C, Cristina AS, Simões ZL. MicroRNA-34 directly targets pair-rule genes and cytoskeleton component in the honey bee. Sci Rep. 2017;7:40894.
93. Drapeau MD, Albert S, Kucharski R, Prusko C, Maleszka R. Evolution of the MiR-34 miRNA in a social insect (Apis mellifera L.). Curr Topics Dev Biol. 2006;74:283–86.
94. Ruppell O, Amdam GV. Genes of reproduction and regulation of honeybee (Apis mellifera L.) social behavior. Annu Rev Genet. 2012;46:97–119.
108. Buttstedt A, Mureşan CI, Lile H, Hause G, Ihling CH, Schulze S-H, et al. How honeybees defy gravity with royal jelly to raise queens. Curr Biol. 2018;28:1095–100.

109. Hartfelder K, Engels W. The composition of larval food in stingless bees: evaluating nutritional balance by chemosystematic methods. Insect Soc. 1989;36:1–14.

110. Smith CR, Morandin C, Noureddine M, Pant S. Conserved roles of Ouir genes in insect development, polymorphism and protection. J Evol Biol. 2018;31:516–29.

111. Greenwood JM, Milutinović B, Peuß R, Behrens S, Esser D, Rosenstiel P, et al. Oral immune priming with Bacillus thuringiensis induces a shift in the gene expression of Tribolium castaneum larvae. BMC Genomics. 2017;18:329.

112. Sann M, Niehuis O, Peters RS, Mayer C, Kozlov A, Podsiadlowski L, et al. Phylogenomic analysis of Apoidea sheds new light on the sister group of bees. BMC Evol Biol. 2018;18:71.

113. Pereira UD, Bonetti AM, Goulart L, dos Santos AR, de Oliveira GC, Cuadros-Orellana S, Leira-Vieira C. Complete mitochondrial genome sequence of Melipona scutellaris, a Brazilian stingless bee. Mitochondrial DNA A. 2016;27:3387–8.

114. George S, Gaddelapato SC, Palli SR. Histone deacetylase 1 supresses Krüppel homolog 1 gene expression, and influences juvenile hormone action in Tribolium castaneum. Proc. Natl. Acad Sci. U.S.A. 2019;116:17759-64.

115. Hartfelder K, Engels W. Social insect polymorphism: hormonal regulation of plasticity in development and reproduction in the honeybee. Curr Topics Dev Biol. 1998;40:45–77.

116. Cardoso CAM, Fujimura PT, Santos-Júnior CD, Borges NA, Leira-Vieira C, Hartfelder K, et al. Methyl farnesoate epoxidase (mfe) gene expression and juvenile hormone titters in the life cycle of a highly eusocial stingless bee, Melipona scutellaris. J Insect Physiol. 2017;101:185–94.

117. Humann FC, Tiberio GJ, Hartfelder K. Sequence and expression characteristics if long noncoding RNAs in honey bee caste development – potential regulators for transgressive ovary size. PLoS One. 2013;8:e78915.

118. Duncan EJ, Hyrkö O, Dearden PK. Notch signalling mediates reproductive constraint in the adult worker honeybee. Nat Commun. 2016;7:12427.

119. Robertson HM. Molecular evolution of the major arthropod chemoreceptor gene families. Annu Rev Entomol. 2019;64:227–42.

120. Bloch G. The social clock of the honeybee. J Biol Rhythm. 2010;25:307–17.

121. Rubin EB, Shemesh Y, Cohen M, Elgavish S, Robertson HM, Bloch G. Molecular and phylogenetic analyses reveal mammalian-like clockwork in the honey bee (Apis mellifera) shed new light on the molecular evolution of the circadian clock. Genome Res. 2006;16:1352–65.

122. Gu HF, Xiao JH, Niu LM, Wang B, Ma GC, Dunn DW, Huang DW. Adaptive evolution of the circadian gene timeout in insects. Sci Rep. 2014;4:4212.

123. Matsumoto A, Ueki-Tadenuma M, Yanada RG, Hou J, Uno KD, Kasukawa T, et al. A functional genomics strategy reveals clockwork orange as a transcriptional regulator in the Drosophila circadian clock. Genes Dev. 2007;21:1687–700.

124. Ingram KK, Alexander Kutovoi A, Wurm Y, Shoemaker D, Meier R, Bloch G. The molecular clockwork of the fire ant Solenopsis invicta. PLoS One. 2012;7:e45715.

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