scientific reports

OPEN

Analysis of ovarian transcriptomes reveals thousands of novel genes in the insect vector *Rhodnius prolixus*

Vitor Lima Coelho[1], Tarcísio Fontenele de Brito[1], Ingrid Alexandre de Abreu Brito[1], Maira Arruda Cardoso[1], Mateus Antonio Berni[1], Helena Maria Marcolla Araujo[1,3], Michael Sammeth[2,4] & Attilio Pane[1✉]

Rhodnius prolixus **is a Triatominae insect species and a primary vector of Chagas disease. The genome of** *R. prolixus* **has been recently sequenced and partially assembled, but few transcriptome analyses have been performed to date. In this study, we describe the stage-specific transcriptomes obtained from previtellogenic stages of oogenesis and from mature eggs. By analyzing ~ 228 million paired-end RNA-Seq reads, we significantly improved the current genome annotations for 9206 genes. We provide extended 5′ and 3′ UTRs, complete Open Reading Frames, and alternative transcript variants. Strikingly, using a combination of genome-guided and de novo transcriptome assembly we found more than two thousand novel genes, thus increasing the number of genes in** *R. prolixus* **from 15,738 to 17,864. We used the improved transcriptome to investigate stage-specific gene expression profiles during** *R. prolixus* **oogenesis. Our data reveal that 11,127 genes are expressed in the early previtellogenic stage of oogenesis and their transcripts are deposited in the developing egg including key factors regulating germline development, genome integrity, and the maternal-zygotic transition. In addition, GO term analyses show that transcripts encoding components of the steroid hormone receptor pathway, cytoskeleton, and intracellular signaling are abundant in the mature eggs, where they likely control early embryonic development upon fertilization. Our results significantly improve the** *R. prolixus* **genome and transcriptome and provide novel insight into oogenesis and early embryogenesis in this medically relevant insect.**

Triatomine bugs (Hemiptera, Reduviidae, Triatominae) include hematophagous species that are widespread in Latin America and have been firmly connected to the diffusion of Chagas disease. This life-threatening illness is caused by *Trypanosoma cruzi*, a protozoan parasite that is transferred from the insect vector to a human host due to the blood-feeding habit of the Triatomine species. Among these, *R. prolixus* represents a prominent arthropod vector of the trypanosome. Between 6 and 7 million people are affected by Chagas disease in Latin American countries[1]. However, several studies also reported the spread of Chagas disease in the United States of America, Canada, Europe, Australia, and Japan[2–5]. The sequencing and partial assembly of the *R. prolixus* genome were recently achieved and the latest RproC3 version is available in VectorBase[6,7]. With 702 Mb of genomic DNA sequenced, it was estimated that 95% of the genome, including 15,738 genes, has been covered[6]. Despite its medical relevance, few genomic resources are available for *R. prolixus*. Over the past decade, transcriptomic analyses were carried out either using expressed sequence tags or, more recently, with high-throughput sequencing techniques[8–10]. However, bioinformatic approaches coupled with molecular biology assays have recently shown that many genes are missing from the current version of the genome[11–13]. These studies highlight substantial limitations in the current *R. prolixus* genome annotations, which can hinder genetic, molecular and evolutionary studies, and underscore the importance of generating more detailed maps of the genome.

R. prolixus is a hemimetabolous insect that develops through five nymph stages before reaching the sexually mature adult stage[14,15]. Oogenesis in adult females is promoted by blood meals and each female can produce up

¹Institute of Biomedical Sciences (ICB), Federal University of Rio de Janeiro, Rio de Janeiro, Brazil. ²Institute of Biophysics Carlos Chagas Filho (IBCCF), Federal University of Rio de Janeiro, Rio de Janeiro, Brazil. ³Instituto Nacional de Ciência e Tecnologia em Entomologia Molecular (INCT-EM), Rio de Janeiro, Brazil. ⁴Department of Applied Sciences, Institute of Bioanalysis, Coburg University, Coburg, Germany. ✉email: apane@icb.ufrj.br

to 80 eggs[16]. Females harbor two ovaries, each containing 6–8 ovarioles simultaneously[17,18]. The anterior region of the ovariole is formed by a lanceolate structure known as the tropharium that harbors the trophocytes and pro-oocytes[19]. Trophocytes are polyploid cells and are arranged in a syncytium around a central cavity known as the trophic core, while oocytes arrested in meiosis I are observed at the posterior of the tropharium. Following a blood meal and the reactivation of oogenesis, the oocytes are progressively surrounded by mitotically active follicle cells to form mature egg chambers[14,15,17,20,21]. After budding off from the tropharium, egg chambers remain connected to the tropharium through cytoplasmic bridges called trophic cords[18]. Oogenesis can be divided into three phases: pre-vitellogenesis, vitellogenesis, and choriogenesis. During previtellogenesis, RNAs, proteins, and nutrients generated by the trophocytes are collected in the trophic core and then transported via the trophic cords to the growing egg chambers[18,19,22–24]. While the growth of the oocyte is slow in the previtellogenic phase, it is rapidly promoted during vitellogenesis by fat body-derived vitellogenins and nutrients that are directly absorbed by the oocyte[25]. Once the oocyte reaches 1 mm in length, the trophic cords are severed and choriogenesis begins, thus leading to the deposition of a hard chorion surrounding the oocyte. The egg displays an apparent anterior–posterior polarity with the operculum at the anterior, which is eventually displaced by the nymphs at eclosion. The dorsal–ventral axis is also readily distinguishable since the ventral region of the egg is more concave. The chorion surrounds the mature egg in *R. prolixus* and protects it against dehydration, while allowing the exchange of gases and liquids as well as fertilization[26–31].

The genetic and molecular mechanisms that control oogenesis in the model organism *Drosophila melanogaster* have been elucidated to a great extent shedding light on key factors in important developmental processes, including germline stem cell maintenance and cystoblast differentiation, DNA damage checkpoint and DNA repair enzymes, mitosis and meiosis, and dorsal–ventral and anterior–posterior axial polarization. Furthermore, in virtually all animal species, the initial stages of embryogenesis are genetically controlled by maternally-provided molecules while the zygote is still transcriptionally inactive. Some of the mRNAs and proteins deposited by the mother in the egg can control crucial developmental decisions. In *D. melanogaster*, egg chambers of the meroistic polytrophic ovary are formed by a layer of epithelial cells surrounding the germline, which in turn, comprises 15 nurse cells and the oocyte. The nurse cells provide the *gurken*, *bicoid,* and *nanos* mRNAs, which are transported and localized into the oocyte, where they control the polarization of the dorsal–ventral and anterior–posterior axes of the egg chamber and future embryo[32]. TGF-alpha-like molecules, similar to Gurken in *D. melanogaster*, also control the establishment of the dorsal–ventral axis in wasps, beetles, and crickets[33]. In *D. melanogaster*, maternally-deposited *oskar* mRNA accumulates at the posterior of the oocyte, where it determines the nucleation of the germplasm, a specialized cytoplasm that in turn contributes to the formation of the gonads in offspring[34]. In addition, maternal RNAs and proteins mediate the maternal-to-zygotic transition (MZT) during embryogenesis, when the developmental program switches from maternal to zygotic control[35]. This process entails the degradation of maternal RNAs and, to a certain extent, is controlled by active mechanisms. For instance, the Smaug and Brain Tumor enzymes together with a cluster of miRNAs are responsible for maternal mRNA decay in *D. melanogaster,* while miR-430 is a key factor in maternal RNA clearance in zebrafish[36,37]. Except for a few model organisms like *D. melanogaster* and zebrafish, the genetic basis of oogenesis and early embryogenesis is still poorly understood in other animal species including *R. prolixus.*

We have recently provided an initial analysis of the transcriptome in previtellogenic stages (PVS) of *R. prolixus* oogenesis[38]. In the current study, we combined the PVS datasets with newly generated transcriptomes from mature eggs resulting in more than 228 million paired-end reads across two biological replicates for each stage. Our analyses allowed us to dramatically improve the maps of genomic features in the *R. prolixus* genome and identify thousands of protein-coding transcripts and repetitive element variants. We used these data to characterize the genetic basis of *R. prolixus* oogenesis and the complement of maternally provided RNAs that drives early embryogenesis in *R. prolixus*. Finally, we have collected our results together with transcriptomic datasets from other laboratories in a mirror of the UCSC Genome Browser (Kent et al. 2002), that we named *Rhodnius* Integrated Omics Browser or RIO browser at the URL www.genome.rio/cgi-bin/hgGateway?db=rproC3.

Results

Improvement of the reference genome annotation with ovarian RNA-Seq datasets.
We conducted a deep interrogation of the previtellogenic stage (PVS) and mature unfertilized eggs (Egg) transcriptomes, yielding in total ~ 228 M (128 M and 100 M, respectively) sequence reads across two biological replicates for each stage. We previously published a partial analysis of the PVS libraries[38], while the Egg transcriptomes were produced de novo for the current study. Genomic mapping of the RNA-Seq libraries was performed against the genome assembly RproC3 of *R. prolixus*[6,7]. Overall library mapping rates of the PVS and Egg conditions were on average 96% and 77.5%, respectively. The mapping rate of read-pairs aligning to multiple locations in the genome (i.e. multi-mapped read-pairs) was 12.6% on average (Supplementary Table S1). The proportion of the read-pairs unambiguously or uniquely mapped to the genome was similar in PVS and Egg samples, with 85–87% of the total mapped read-pairs. Based on these observations, the uniquely mapped read-pairs were used for a genome-guided transcriptome assembly to identify novel gene models in the genome and to improve the current gene annotations of *R. prolixus*. Preliminary quantification of the genes expressed in PVS and Egg samples was performed by assigning the uniquely mapped read-pairs to the reference gene annotation available at VectorBase[7]. On average, approximately 64% and 42% of uniquely mapped read-pairs of the PVS and Egg samples were respectively assigned to the reference gene annotation (column "Assigned paired-reads to reference annotation", Supplementary Table S1). Yet, a surprisingly high percentage of uniquely mapped read-pairs, on average 36% for PVS and 58% for Egg, were not attributed to any reference genes or genomic features (Supplementary Table S1). These observations pointed to the fact that genes important for *R. prolixus* oogenesis might be missing in the current reference gene annotation. Thus, in order to expand the transcriptome landscape to
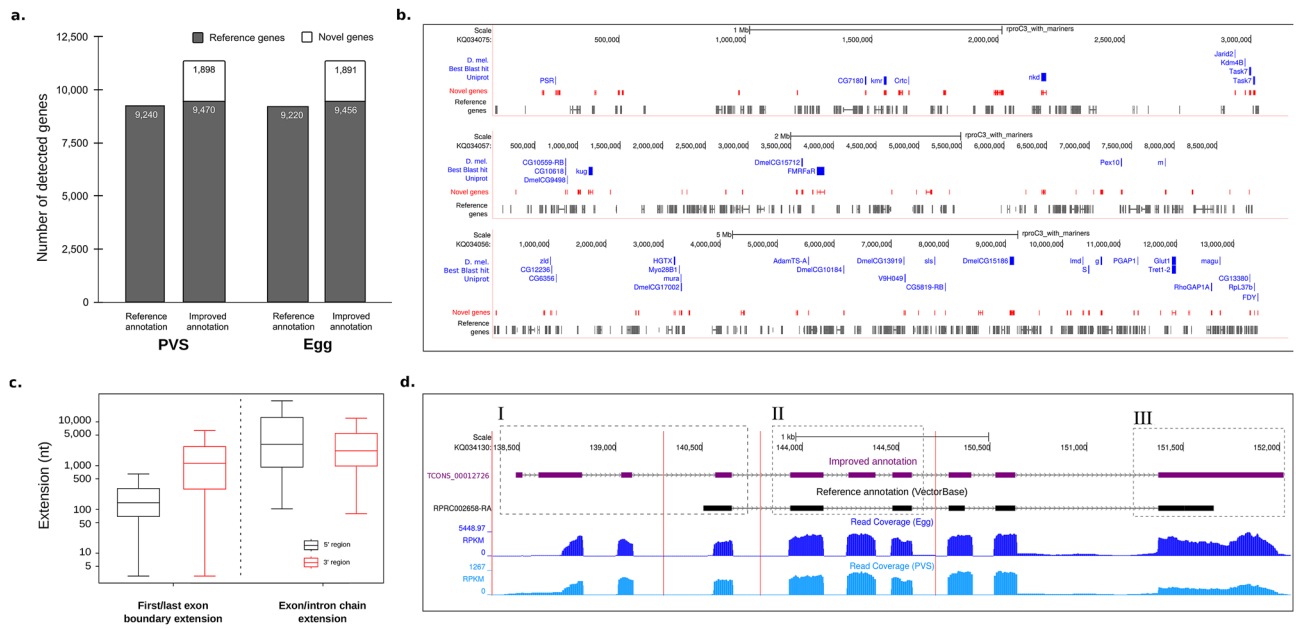
Figure 1. Improvement of genome annotations in *R. prolixus*. **(a)** Number of genes with at least 40 read-pairs overlapping the reference and improved transcriptome annotation. **(b)** Representative *R. prolixus* super-contigs with previously annotated genes (grey), novel genes (red), and their *D. melanogaster* orthologs (blue). **(c)** Distribution of the extension lengths at the 5′ (black) and 3′ (red) ends. **(d)** (I) Extension of the genomic coordinates of the gene RPRC002658 at the 5′ end due to the prediction of the alternative splice variant TCONS_00012726 (purple); (II) Prediction of a skipped internal exon compared to the reference transcript RPRC002658-RA (black); and (III) extension of the genomic coordinates at the 3′ end of the reference exon. Light and dark blue plots represent the read coverage of the samples PVS and egg, respectively.

include unassigned/unmapped reads, we used the uniquely mapped and unmapped reads as input for genome-guided transcriptome assembly and de novo transcriptome assembly approaches, respectively[39–41]. The expansion of the annotated genetic elements by the genome-guided approach provided 2126 novel putative genes. On average, the number of genes overlapped by more than 40 read-pairs increased by 23% in PVS and Egg (Fig. 1a). It is worth noting that we adopted a conservative approach; a cutoff of 10 overlapping read-pairs is generally used to identify a new gene. For 1659 novel genes (78%), we identified 1204 complete ORFs, 365 (5′ or 3′) partial ORFs, and 90 internal ORFs (Supplementary Table S3). Although the remaining 476 new transcripts display small ORFs, we cannot rule out that they represent functional long non-coding RNAs (Supplementary Table S3). Interestingly, the newly annotated genes often encode putative orthologs of *D. melanogaster* proteins that are firmly linked to important developmental processes. Some notable examples in the displayed super-contigs are the *CREB-regulated transcription coactivator* (*Crtc*), the transcription factor *zelda* (*Zld*), the type II transmembrane protein *Star* (*S*), the cadherin *kugelei* (*Kug*), the homeodomain transcription factor *HTGX*, and the BMP regulator *Magu*. For easy visualization and navigation, we uploaded our datasets in a mirrored version of the UCSC genome browser[42], that we named *Rhodnius* Integrated Omics or RIO browser (Fig. 1b). To facilitate the navigation, the RIO browser offers the possibility to use the name or acronym of the *D. melanogaster* ortholog to identify and visualize the features, cross-reference databases, functional classification and gene ontology of a *R. prolixus* gene.

In our previous study, we noticed that the annotation of certain genes (e.g. the *R. prolixus vasa* and *PIWI* genes) was imprecise as their annotated 5′ and 3′ends did not match the gene boundaries revealed by our RNA-Seq data[38]. We, therefore, sought to provide a more detailed map of the gene structure in *R. prolixus* by genome-guided assembly of our PVS and Egg datasets. We were able to extend 1960 genes at their 5′ ends, 1473 genes at their 3′ ends, and 5649 genes at both the 5′ and 3′ ends (Fig. 1c). In addition, we found approximately two new transcript variants each for 6845 genes of the reference annotation. An assembled transcript was considered a novel isoform when it displayed (i) a multi-exon structure with at least one junction match with a reference gene in the same strand, (ii) minimum abundance of 40% of the most abundant isoform (i.e. the major isoform) of the gene and (iii) Counts per Million (CPM) > 1 at least in one condition. Figure 1c shows two types of gene extensions, one occurring at the 5′/3′ genomic coordinates of the first/last exon of an annotated gene and the other that occurs due to the prediction of new chains of exons and introns for the gene (i.e. new isoforms). The 5′ and 3′ ends were extended by a median of 569 bp and 1800 bp, respectively (Supplementary Table S2). By extending the annotated coding sequence (CDS), we found 78 novel complete Open Reading Frames (ORFs) in the reference transcriptome. This also improved 1266 5′-incomplete CDSs by adding a start codon and 429 CDSs with a valid stop codon (Supplementary Table S3). A clear example of our transcriptome improvements is represented by the RPRC002658 gene encoding an uncharacterized *R. prolixus* protein (Fig. 1d). Our analyses allowed us to extend the 5′-end of this gene by approximately 2 Kb (Fig. 1d, box I) due to the prediction of the novel TCONS_00012726 transcript isoform. This alternative splice variant also revealed an extra internal exon

Samples	Genome-guided assembly		De novo assembly		Total assigned read pairs	
	Uniquely mapped read pairs	Assigned to improved gene annotation	Unmapped read pairs	Assigned to de novo transcriptome	To the reference transcriptome	To the improved transcriptome[a]
PVS_1	31,651,452	26,759,942 (83.6%)	1,565,561	871,556 (55.7%)	20,314,277 (63.4%)	27,631,498 (83.2%)
PVS_2	21,166,732	17,784,143 (83.4%)	1,028,565	532,938 (51.8%)	13,931,026 (65.3%)	18,317,081 (82.5%)
Egg_1	17,400,997	15,369,913 (84.8%)	6,360,014	5,337,815 (83.9%)	8,049,949 (44.4%)	20,707,728 (87.2%)
Egg_2	14,258,924	12,500,145 (84.8%)	4,722,559	4,009,233 (84.9%)	5,918,720 (40%)	16,509,378 (87%)

Table 1. Statistics of the read assignment to the reference and expanded transcriptomes. [a]"Expanded transcriptome" refers to the improved annotation and final putative de novo transcripts.

that was previously not described in the reference transcript RPRC002658-RA (Fig. 1d, box II). Finally, we were able to extend the last exon of the gene by approximately 362 bp (Fig. 1d, box III). Other clear examples of the improvements that we introduced in the *R. prolixus* are represented by the genes RPRC001867 (CRAL_TRIO_N domain-containing protein, ortholog gene of *D. melanogaster* gene *cralbp*), RPRC004178 (Uncharacterized protein) and RPRC014580 (ADF-H domain-containing protein, ortholog gene of *D. melanogaster* gene GMF).

Despite the substantial expansion of the RproC3 transcriptome by genome-guided assembly, increasing the percentage of the read-pairs overlapping genomic features by up to ~40% (Table 1), 28.7% of the reads in the Egg and 4% in the PVS datasets did not map to any genomic regions (Supplementary Table S1). Since the egg is not transcriptionally active, we reasoned that the stark difference between the percentage of unmapped reads in the two datasets might be due to a small number of genes whose transcripts are maternally deposited in the eggs, where they accumulate at higher levels compared to the PVS stage. To probe this hypothesis, we applied a de novo assembly approach to the unmapped reads, which initially yielded 38,856 contigs for PVS and 19,390 contigs for Egg samples. After low-expression filtering, sequence deduplication, and quality contig evaluation steps (see "Methods"), we obtained 1,993 non-redundant protein-coding transcripts (Supplementary Table S5). However, upon closer inspection, we noticed that three contigs were accountable for >80% of the unmapped reads (Supplementary Table S4 and S5; Supplementary Fig. S1). One of these contigs encodes the *R. prolixus* Rp30 eggshell protein[43]. This abundant protein has been already biochemically characterized and a partial cDNA for this protein had already been observed in transcriptomic libraries derived from the *R. prolixus* digestive tract and ESTs datasets[8,9,43]. Our approach provides the missing sequences of the transcript and the full-length ORF. The remaining two contigs encode predicted proteins that do not share similarity with known proteins in the NCBI databases, but are particularly abundant in mature eggs. Additional studies will be required to understand whether these genes are involved in the maturation of the egg chamber, like Rp30, or they are implicated in early *R. prolixus* embryogenesis.

Stage-specific gene expression during *R. prolixus* oogenesis. With our expanded version of the *R. prolixus* transcriptome, we wondered what genes are expressed during oogenesis in this species. Our quantification of the Egg datasets shows that the transcripts of 11,127 genes are present in the mature unfertilized egg (Supplementary Table S6). In *R. prolixus*, RNAs and proteins synthesized in the tropharium are transported into the oocyte via the trophic cord, while the oocyte is transcriptionally quiescent. We then performed differential expression analysis by comparing the PVS and Egg datasets in an attempt to identify stage- and tissue-specific gene expression patterns. Differentially expressed genes (DEGs) were selected according to DESeq2[44] to identify reliable fold changes between the PVS and Egg datasets, where |fold-change|≥ 2.5 and adjusted p-value ≤ 0.01[44] were set as the thresholds for significantly differentially expressed genes (Fig. 2a). Here, we refer to genes as "down-regulated" or "up-regulated" if their steady-state RNA levels are respectively lower or higher in eggs compared to PVS stages (Supplementary Table S6). Differential gene expression analysis resulted in 1480 DEGs (Fig. 2a). Of these, 851 (57.5%) genes appeared to be expressed at high levels or specifically in the PVS stage, while no expression beyond background levels was detected in Egg datasets. Surprisingly, however, 629 DEGs, which is 42.5% of the total, appeared to be highly enriched in the Egg samples. Novel putative genes represent 29% of the total DEGs, with 9.9% up-regulated and 19.1% down-regulated in PVS versus Egg stages, respectively (Fig. 2b).

Consistent with our previous study[38], among the top 25 genes that are highly represented in the PVS transcriptome, 15 encode ribosomal proteins (Supplementary Fig. S2). After removing these genes from the list, the most differentially expressed genes in this stage code for components of the transcription and translation machinery (Fig. 3a). Conversely, we find 13 genes in the top 25 expressed in eggs encoding uncharacterized proteins along with evolutionarily conserved proteins (Fig. 3a). On top of this list, we find the XLOC_000150 transcript that might represent a long non-coding RNA. Some examples are given by putative orthologs of the transforming growth factor-beta (TGF-β)-induced protein ig-h3 (RPRC001419), the activating transcription factor 4 (RPRC010216), a lysosomal-associated transmembrane protein (RPRC006528), and a Peptide Methionine Sulfoxide Reductase (PMSR) domain-containing protein (RPRC002388), among others. Enrichment analysis of the differentially expressed genes with g:Profiler[45] found 35 and 54 GO terms enriched (Benjamini–Hochberg FDR ≤ 0.05) in the PVS and Egg samples, respectively (Supplementary Table S7). Figure 3b shows ten major functionally grouped networks of these enriched GO terms produced by ClueGO[46,47]. This approach yielded four groups with 29 enriched terms in PVS (Supplementary Fig. S3a) and 12 groups with 54 enriched terms in Egg
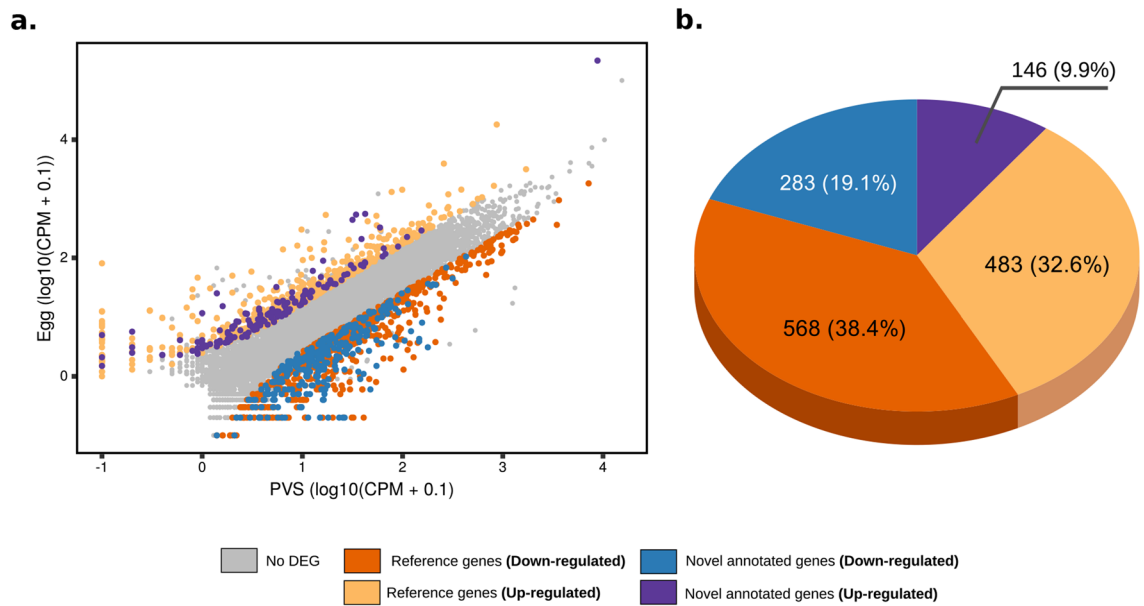
Figure 2. Differential expression analysis of previtellogenic stage and mature egg datasets of *R. prolixus*. **(a)** Scatter plot of the gene expression normalized by the average CPM (counts per million, in log10). Y axis shows Egg datasets, while X displays the PVS datasets. Upregulated reference (pink) and novel (magenta) genes and downregulated reference (orange) and novel (blue) genes are displayed. Genes with no differential expression in the two stages are shown (grey). **(b)** Composition of the total differentially expressed (DE) genes using the improved gene annotation.

samples (Supplementary Fig. S3b). Our improved transcriptome confirms previous observations that the most expressed reference genes in *R. prolixus* PVS encode putative proteins with homology to *D. melanogaster* proteins related to ribosome, translation, and splicing[38], but it also adds enriched processes related to the P-bodies, the extracellular region, and the collagen trimer (Supplementary Table S7; Supplementary Fig. S3a). Processing bodies or P-bodies are eukaryotic cytoplasmic structures that are evolutionarily conserved in organisms as distant as flies, worms and mammals[48]. In *D. melanogaster*, P-bodies are organized around the nurse cell nuclei within the egg chambers, where they form a membrane-less organelle known as "nuage". The nuage and the P-bodies share a number of enzymatic activities and have been linked to a different molecular mechanisms including RNA decay and RNA interference. Among the genes enriched in the mature unfertilized eggs, we found protein-coding genes associated with steroid hormone receptors, cellular organization, kinase activity, transmembrane transport, protein binding, and GTPase binding. These results suggest that the females provides their oocytes with a transcriptomic repertoire encoding hormone receptors (e.g. Nuclear hormone (NR) LBD domain-containing protein and ecdysone response nuclear receptor e75a), enzymes (e.g. Diacylglycerol kinases and Guanylate cyclase) and proteins with GTPase binding domains (e.g. Ran\Rab binding function and Rho GTPase exchange factor activity) critical for organismal growth, microtubule organization, cell polarity, developmental transition and regulation during the period of nutrient starvation, and transcriptional quiescence of early embryogenesis.

Key ovarian genes of *D. melanogaster* are conserved in *R. prolixus* and expressed during oogenesis. We sought to determine the evolutionary conservation and stage-specific expression levels of the *R. prolixus* orthologs of *D. melanogaster* genes that exert critical roles in oogenesis and early embryogenesis (Fig. 4). Therefore, we focussed on six cellular and developmental processes: germline stem cells, DNA damage checkpoint and repair, mitosis/meiosis, germ granules/pole cells, axial polarization, and RNA decay [37,49–51]. For each process, we selected well-characterized *D. melanogaster* genes, searched for the putative *R. prolixus* orthologs using BLAST tools, and measured their expression levels in PVS and Egg conditions using our RNA-Seq datasets (Fig. 4 and Supplementary Table S8). Interestingly, we find that factors involved in *D. melanogaster* germline stem cell maintenance and cystoblast differentiation are conserved in *R. prolixus* even though this cell type is not observed in the adult ovary. The expression of the *R. prolixus* orthologs of *pumilio* (*pum*) (RPRC000720), *decapentaplegic* (*dpp*) (RPRC000401), *mad* (RPRC008357), *benign gonial cell neoplasm* (*bgcn*) (RPRC008118), *medea* (RPRC003748), *glass bottom boat* (*gbb*) (RPRC013357), *zero-population-growth* (*zpg*) (RPRC014101), *hedgehog* (*hh*) (RPRC012384) and *Notch* (RPRC008058) is readily detected in both PVS and Egg datasets (Fig. 4a)[51–55]. The differentiation of the *D. melanogaster* germline stem cells requires the activity of *bag-of-marbles* (*bam*), a partner of BGCN in cystoblast differentiation. However, we did not find a *bam* ortholog in the *R. prolixus* genome, or in the improved transcriptome. Interestingly, *dpp* transcripts are detected only in PVS and do not accumulate in the egg.

We also examined the expression profile of genes involved in meiosis and mitosis, including *cdk1* (RPRC009342), *cyclin B* (*cycB*) (RPRC012735), *twine* (RPRC008472), *greatwall* (RPRC005266), *polo* (RPRC002498), *orb* (RPRC006637), *orb2* (RPRC001021), *calcineurin B* (RPRC008393), *sarah* (RPRC006424),
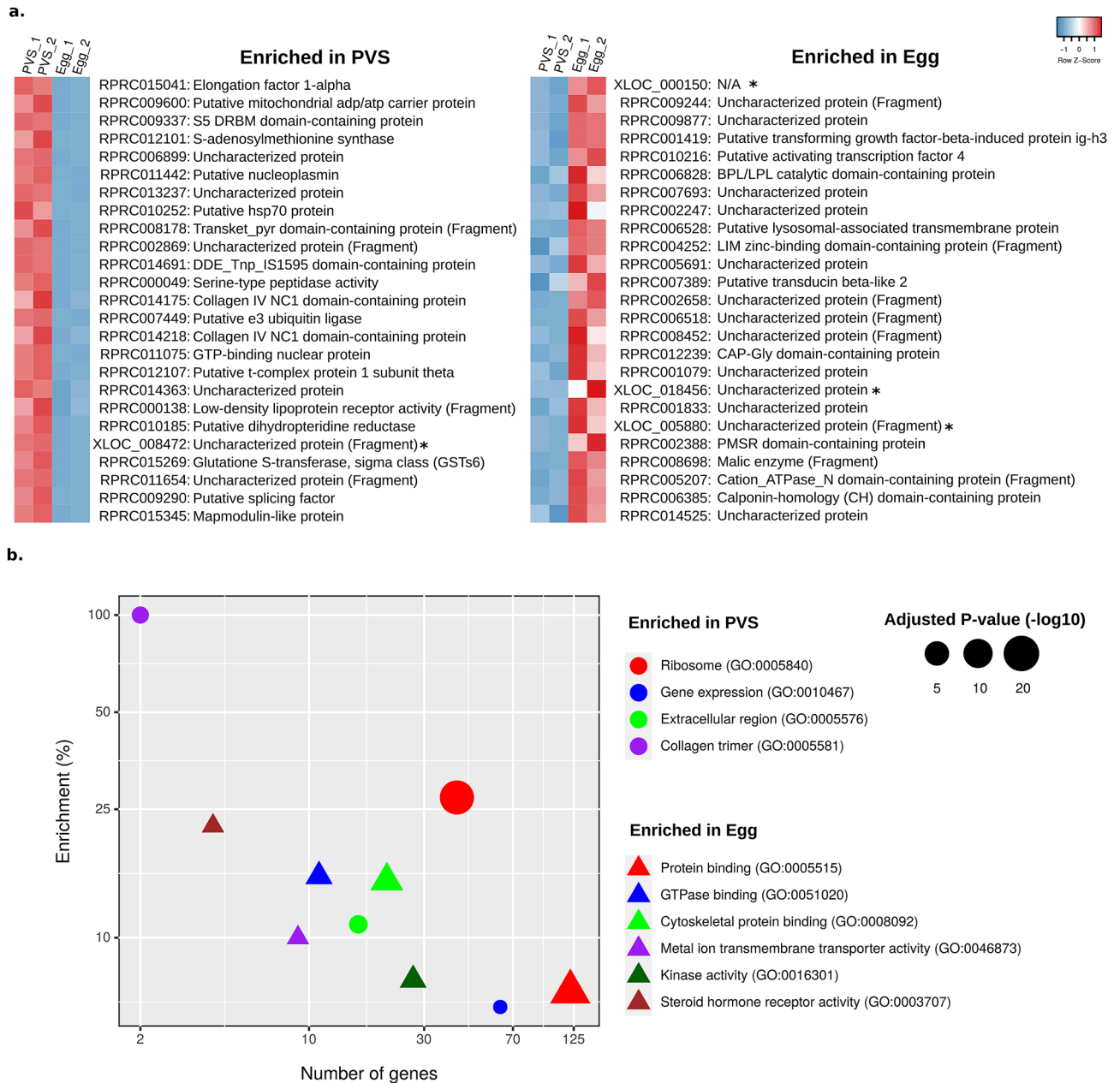
Figure 3. Functional categories enriched in previtellogenic stage and mature eggs. **(a)** Left: Top 25 differentially expressed genes (non-ribosomal-related protein) ordered by average normalized expression (CPM) for PVS, and the encoded putative protein according to the Uniprot database. Right: Top 25 most differentially expressed genes at the Egg stage. Novel genes (identifiers with suffix "XLOC" and asterisks, *) were assigned to protein products by BLAST sequence similarity. "N/A" indicates that a significant hit was not obtained (e-value ≤ 0.05). **(b)** Enriched GO terms representing 10 functionally related groups identified by ClueGO in the PVS (circle) and Egg (triangle) conditions. Enrichment (y-axis) is given by the number of genes in the input related to a GO term (x-axis) divided by the total number of genes associated with this GO term in %.

verthandi (RPRC005603), *stromalin* (RPRC005579), *SMC1* (RPRC009822) and *SMC3* (RPRC009208)[56–60]. All of the genes appear to be expressed in PVS and deposited in the mature eggs with one notable exception of *cycB*. We find two putative orthologs of *cycB* in the *R. prolixus* genome, RPRC012735 and RPRC001188, whose putative protein products share 43% and 37% amino acid sequence identity, respectively, with the *D. melanogaster cycB* gene product (Supplementary Table S8). Yet, neither is expressed at detectable levels in *R. prolixus* ovaries, nor do we find additional putative orthologs in our improved transcriptome. Given that *cycB* expression is downregulated in endocycling cells like *D. melanogaster* nurse cells, and because most of the PVS tissues in *R. prolixus* are occupied by giant polyploid nuclei, it is possible that the expression of *cycB* genes in *R. prolixus* ovaries is below detectable levels.
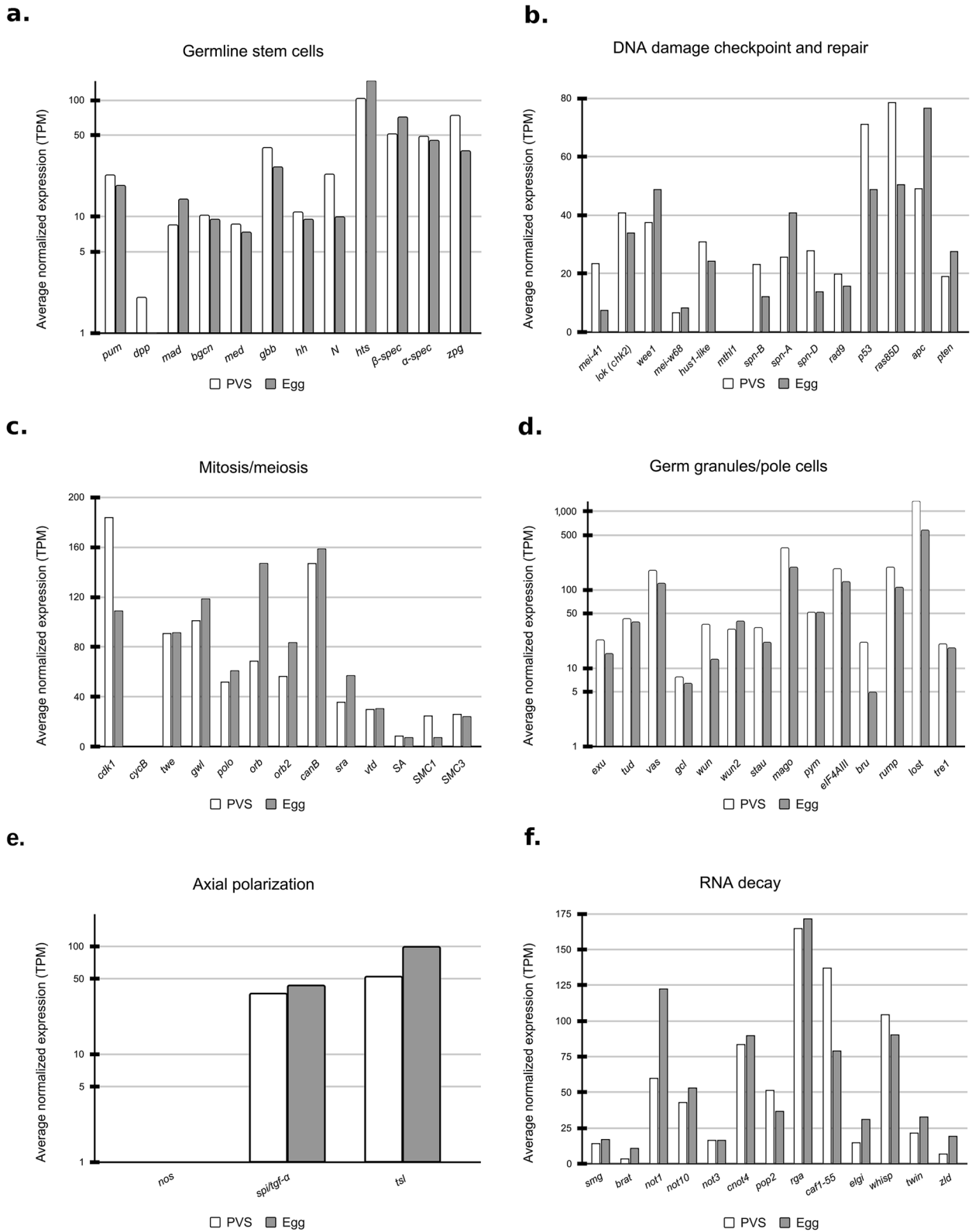
Figure 4. Gene expression levels of the *R. prolixus* orthologs of *D. melanogaster* genes involved in oogenesis and early embryogenesis. (**a**) Genes involved in maintaining germline stem cells. (**b**) Genes related to DNA damage checkpoint and repair processes. (**c**) Genes associated with mitosis and meiosis processes. (**d**) Essential genes for the germ granules and pole cells (**e**) Genes related to the axial polarization process. (**f**) Genes associated with RNA decay process. Y-axis displays expressions levels are normalized by transcripts per million (TPM).

A set of evolutionarily conserved genes is dedicated to the DNA damage checkpoint and repair in organisms ranging from yeast to mammals. DNA damage can be introduced by UV light and mutagenic agents, but DNA double-strand breaks (DSBs) introduced by *mei-W68*, the *D. melanogaster* ortholog of yeast *spo11*, are also required for meiotic recombination[50,61–68]. DNA damage and DSBs are sensed and repaired by a set of genes including *mei-41* (RPRC004298), *chk-2* (RPRC005512), *wee1* (RPRC013702), *mei-w68* (RPRC004962), *hus1-like* (RPRC004826), *methuselah1* (*mthl1*) (RPRC001933), *spn-A* (RPRC006488), *spn-B* (RPRC010106), *spn-D* (RPRC000377), *Rad9* (RPRC001146), *p53* (RPRC003641), *ras85D* (RPRC008553), *apc* (RPRC002106), and *pten* (RPRC014866). The putative orthologs of these genes are present in the *R. prolixus* genome and the corresponding transcripts can be readily detected both in early oogenesis as well as in mature eggs (Fig. 4b).

In *D. melanogaster*, the germ plasm is a specialized cytoplasm that accumulates at the posterior of the oocyte. At the beginning of cellularization during embryogenesis, the germplasm is incorporated in the first cells appearing at the posterior of the embryo. These cells, known as pole cells, will later migrate during gastrulation and join gonadal somatic precursor cells to initiate the formation of the future gonad. Although there is no evidence of a germplasm and germ granules in *R. prolixus,* we do observe the expression of the putative orthologs of the following *D. melanogaster* genes that have been firmly connected to germplasm assembly and pole cell specification and migration: *exuperantia* (*exu*) (RPRC001415), *tudor* (RPRC012896), *valois* (RPRC005681), *vasa* (RPRC009661), *germ cell-less* (*gcl*), (RPRC010622), *wunen* (RPRC004485), *wunen2* (RPRC009835), *staufen* (RPRC010019), *mago nashi* (RPRC010985), *pym* (RPRC009783), *eIF4AIII* (RPRC002208), *bruno* (RPRC007498), *rump* (RPRC014343), *lost* (RPRC014809), and *tre1* (RPRC011268)[69–75]. All of the corresponding *R. prolixus* genes seem to be expressed early in oogenesis and accumulate in the mature eggs. The steady-state RNA levels of the putative *lost* ortholog in *R. prolixus* (Transcripts per Million or TPM > 1000) suggest that this gene exerts a prominent role during *R. prolixus* oogenesis. As expected, we did not find a *R. prolixus* ortholog of the *oskar* gene, that is known to be restricted to Diptera, while, consistent with our previous findings, the *R. prolixus vasa* gene is expressed in PVS and the transcripts are deposited in mature chorionated eggs[38].

Axial polarization in *D. melanogaster* is dictated by maternal transcripts encoding the Bicoid and Gurken proteins, whose localized translation sets the anterior–posterior and dorsal–ventral axes, respectively, of the egg and future embryo[76,77]. Neither of these two genes appears to be conserved in *R. prolixus*, although we found a *spitz/tgf-a* gene (RPRC009386) that is expressed in both the PVS and Egg datasets. Members of the TGF-alpha family were shown to control the formation of the dorsal–ventral axis in crickets, wasps, and beetles[78]. We also find putative orthologs of *nanos* (RPRC002927) and *torso-like* (RPRC006513), but not of *torso*, which are involved in the specification of the anterior–posterior axis in fruit flies, aphids and honeybees[79].

Finally, we investigate the expression profiles of genes associated with the RNA decay pathway. The early phases of embryonic development in *D. melanogaster* and other organisms are driven by maternally contributed RNAs and proteins. However, the maternal program must be erased at the onset of MZT so that the embryo can follow a zygotically-driven transcription program. The degradation of maternal RNAs in *D. melanogaster* is mostly guided by the protein products of a group of genes comprising *smaug* (RPRC007649), *brain tumor* (RPRC007254), *not1* (RPRC005691), *not3* (RPRC003870), *not10* (RPRC001957), *cnot4* (RPRC001525), *pop2* (RPRC004491), *regena* (RPRC003111), *caf1* (RPRC015316), *early girl* (RPRC004083), *wispy* (RPRC007755), *twin/CCR4* (RPRC006991), and *string* (RPRC008472)[80–85]. In addition, *zelda* (RPRC000020) encodes a transcription factor that is pivotal for the onset of zygotic transcription [86]. We find that all of these genes are expressed in PVS and their transcripts are detected in mature eggs. Unexpectedly, the levels of *smaug*, *wispy*, and *zelda* are low and might suggest that other mechanisms are in place during early *R. prolixus* embryogenesis to degrade maternal RNAs and prompt zygotic transcription. We did not find *R. prolixus* orthologs for the following key *D. melanogaster* genes in the genome or in our updated transcriptome: *fs(1)Yb*, *swallow*, *trunk*, *barentsz*, *atr/tefu*, *cup*, and *matrimony*[50,87–89]. Despite the extensive conservation of crucial genes between *D. melanogaster* and *R. prolixus*, we also detect important differences, which might explain, at least in part, the striking differences between oogenesis and embryogenesis in these two insect species.

Expression levels of transposable and repetitive sequences during *R. prolixus* oogenesis. Transposable elements (TE) are mobile sequences that often constitute a significant proportion of animal and plant genomes. While the biological function of these selfish sequences is still debated, it is well established that transposon mobilization is often connected to DNA damage, insertional mutagenesis, and genome instability[90–92]. RepeatMasker annotation of RproC3, available at VectorBase, shows that more than half of *R. prolixus* mobilome is dominated by repetitive elements (longer than 0.5 Kb) of the type Tc1-*mariner* (28%), *Helitron* (14%), *Tc1* (10%) and a class of repetitive elements labeled as *Unknown* (9.8%) (Supplementary Table S9). The prevalence of Tc1-*mariner* elements has been reinforced using de novo and library-based approaches for transposon detection, revealing that approximately 75% of the transposons dispersed in the *R. prolixus* genome belong to the Tc1-*Mariner* family[93]. Given the incomplete assembly of the genome and the challenges imposed by the repetitive nature of these elements, it is likely that the mobilome of *R. prolixus* is not completely characterized. Furthermore, the expression levels of these sequences have not been investigated yet in *R. prolixus*. We therefore employed the available data together with our stage-specific RNA-Seq datasets to determine the expression levels of all known transposon families in *R. prolixus* oogenesis. As for the protein-coding genes, we also compared the early (PVS) and late (Egg) stages of oogenesis to identify transposons potentially displaying differential activity in these two stages (Fig. 5). First, we calculated the RNA-Seq data coverage of the transposable elements by assigning all of the mapped read-pairs to the RproC3 RepeatMasker annotation and our improved gene annotation. Read-pairs assigned to more than one feature or genomic element class (e.g., protein-coding gene, transposable element, simple repeats) were considered ambiguous and, therefore, not included in the pie charts (less than 7% of the total assigned read-pairs) (Fig. 5a). In agreement with their relative abundance
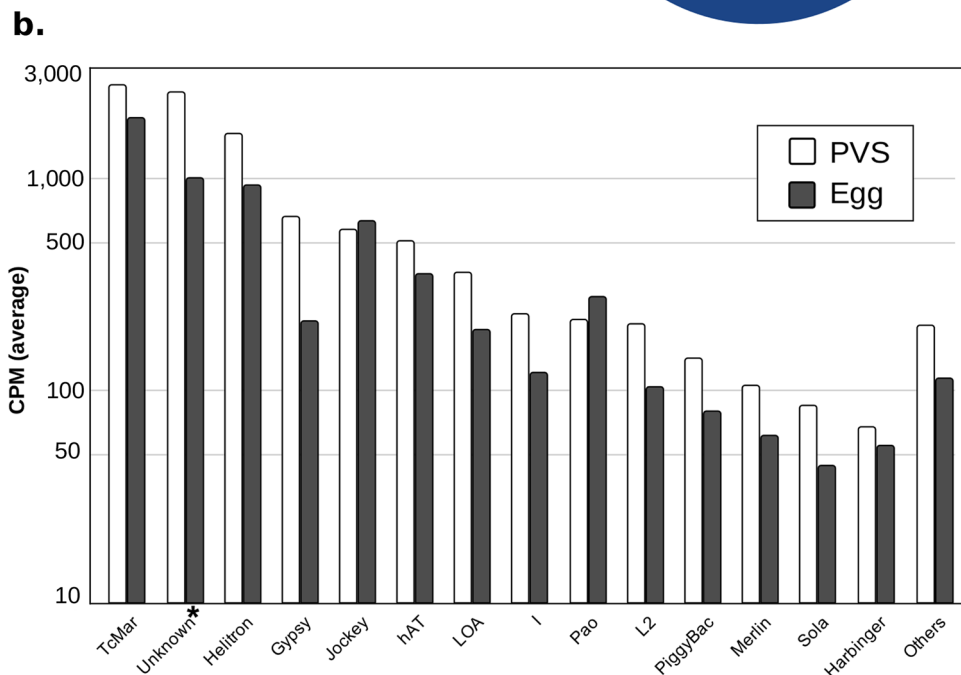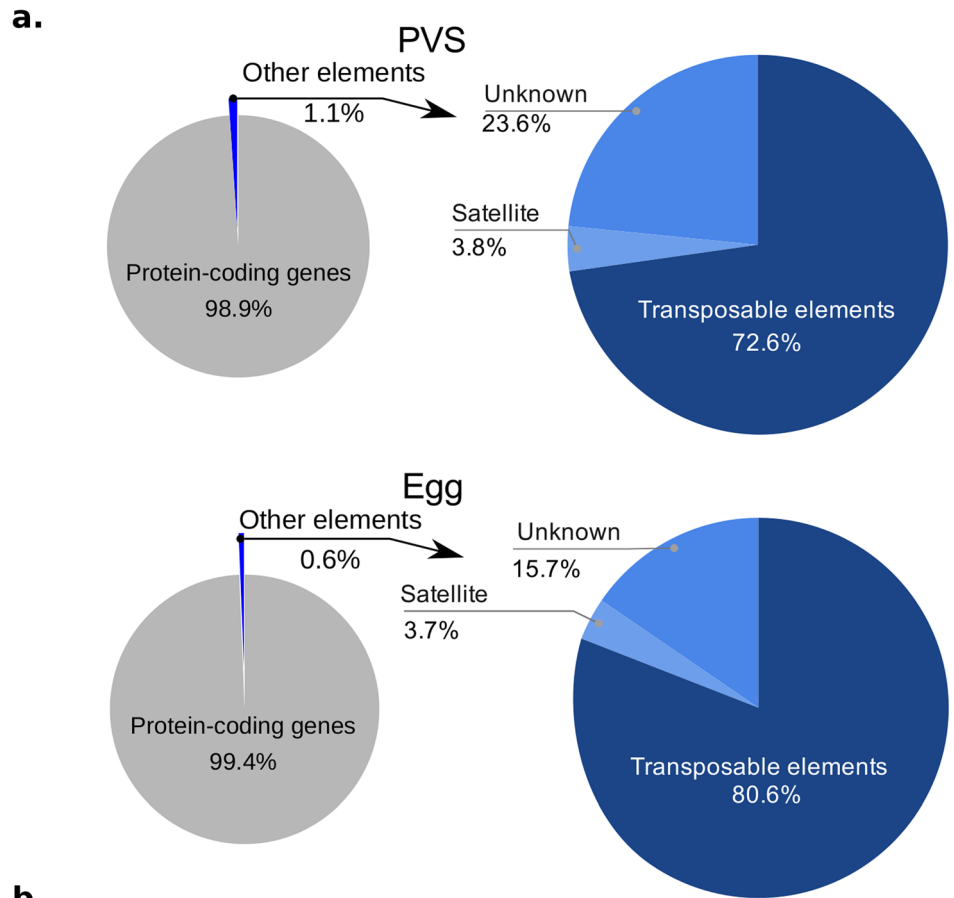
Figure 5. RNA levels of transposable and repetitive elements. **(a)** Left: pie chart shows the assignment of the multi- and uniquely-mapped paired-end reads to the protein-coding gene annotation (gray) and repetitive element regions (blue). Right: pie chart shows a more detailed distribution of mappings for satellite simple repeats, transposable elements, and unknown classes of repetitive elements. **(b)** Cumulative quantification of expression for each class of transposable elements longer than 0.5 Kb (kilobases) for the conditions PVS (white) and egg (gray).

Identifier	Stage	Dfam classification	E-value	Score	Rep. 1 (CPM)	Rep. 2 (CPM)
DN6993_c2_g2_i2	PVS	R1 (DF0004710)	3.6e−08	32.2	119.3	1714.6
DN6822_c0_g1_i1	PVS	Mariner (DF0006767)	4.3e−38	129.6	705.3	611.9
DN6491_c0_g1_i1	PVS	Satellite (DF0001947)	4.3e−17	61.4	345.8	462.5
DN6980_c0_g1_i1	PVS	R1 (DF0004887)	2.40e−09	35.6	114.9	544.3
DN4130_c0_g1_i1	PVS	Mariner (DF0004831)	1.8e−14	52.5	202.9	257.9
DN6969_c2_g5_i1	PVS	L2 (DF0006592)	9.9e−16	55.1	209.1	216.3
DN6874_c0_g1_i3	PVS	R1 (DF0004775)	4.3e−33	113.5	57.1	307.8
DN6975_c2_g1_i1	PVS	RTE-BovB (DF0004797)	6.2e−09	34.8	96.7	129.9
DN4053_c0_g1_i1	PVS	Unknown (DF0005939)	4.4e−14	51	187.7	1.8
DN6127_c0_g1_i2	PVS	DNA transposon TIR (DF0005060)	5.8e−10	38.2	173.2	0
DN5699_c0_g1_i1	PVS	RTE-BovB (DF0005461)	1.5e−08	31.7	73.9	81.8
DN2864_c3_g2_i1	Egg	Unknown (DF0004525)	9.6e−44	148.4	32,006	28,084

Table 2. Putative repetitive elements and transposons non-redundant sequence variants.

in the genome, the Tc1-*mariner* family of transposons appeared to have the highest expression levels both in PVS and in mature eggs (Fig. 5b). High expression levels are also observed for a class of "unknown" transposons and the *Helitron* family. When comparing the steady-state levels in PVS versus Egg datasets, we observed that some classes of DNA transposons, LTR and Non-LTR retrotransposons, have greater coverage in the previtellogenic stage (like the "unknown" TE class and *gypsy*), while others have greater coverage in the mature unfertilized egg (like Tc1-*mariner* and *helitron*). Despite these differences, however, it seems that all of the transposon transcripts that are generated in the tropharium are efficiently transported to the growing oocyte. We were also interested in identifying novel repetitive element (RE) variants, more specifically transposons. To do so, we performed homology searches based on hidden Markov model (HMM) profiles of transposable elements[94,95] against our de novo assembled contigs. In total, homology searches resulted in 167 contigs with significant hits of non-rRNA repetitive elements. To reduce the sequence redundancy between stages, we aligned the new repetitive elements with each other, resulting in 20 non-redundant putative REs (Supplementary Table S10). We also aimed to identify whether our transposons were already characterized in the library of RproC3 repetitive elements available at VectorBase or found in published Tc1-*mariner* consensus sequences[93]. Interestingly, we found 12 non-redundant transposon consensus sequences with no significant similarity or with low alignment coverage with RproC3 repetitive elements[6,7] or Tc1-*mariner* consensus sequences[93] (Table 2). Abundances of these sequences are represented by the number of estimated reads normalized by the total of estimated reads (estimated CPM) for each replicate ("Rep.") of the PVS and Egg de novo contigs. Our study clearly shows that a variety of transposons belonging to known or uncharacterized families are expressed in early oogenesis and their transcripts accumulate in the mature eggs.

Discussion

R. prolixus is a primary vector of *Trypanosoma cruzi*, the etiologic agent of Chagas disease. Sequencing of the *R. prolixus* genome in recent years together with studies on specific pathways or gene sets provided an important resource to understand the biology of this insect[6,96] . In this study, we sought to improve genome annotations and use the improved transcriptome to gain insight into *R. prolixus* oogenesis and the maternal contribution to early embryonic development. Using a combination of genome-guided and de novo transcriptome assembly, we were able to increase the number of putative protein-coding genes in *R. prolixus* from the 15,738 that are currently annotated in the genome (RproC3 version) to 17,864. Many of these genes likely code for critical factors in *R. prolixus* development as they include histone modification enzymes and chromatin remodeling factors, homeotic genes, signaling molecules and kinases, and metabolic enzymes, many of which are evolutionarily conserved. Furthermore, we significantly improved the 5′ and 3′ boundaries of previously annotated genes, thus shedding light on their 5′ and 3′ UTRs. Importantly, for 2,709 of these genes, we revealed the complete ORF and the encoded full-length putative protein. Finally, the paired-end nature of our libraries allowed us to determine transcript variants generated by alternative splicing events for 6845 genes.

Based on the updated transcriptome, we investigated the complement of genes expressed in early stages of *R. prolixus* oogenesis that drive egg chamber formation and oocyte maturation, as well as the atlas of transcripts that are maternally deposited in mature unfertilized eggs. The latter genes are essential during the first stages of embryogenesis, when zygotic transcription is not yet active. The morphology and architecture of the ovarioles differ profoundly in *R. prolixus* and *D. melanogaster*. While in *D. melanogaster* the oocytes are produced from germline stem cells located at the anterior tip of the ovariole, this cell type seems to be restricted to nymph stages in *R. prolixus*[97]. Instead, the anterior region of the ovariole harbors actively dividing trophocytes in this species[17–19]. Despite this difference, we find orthologs of *Drosophila* genes involved in germline stem cell specification and maintenance, like *hts, α-spec, bgcn* and *dpp*, in early stages of *R. prolixus* oogenesis. Particularly surprising is the expression of *hts* and *a-spec*, given that the products of these genes in *D. melanogaster* mark the spectrosome/fusome, a cytoplasmic structure that initially labels the germline stem cells, but once the cystoblast differentiates and start diving, it forms a bridge connecting the germ cells within the mitotic cyst[98]. The spectrosome/fusome exerts an important role in oocyte specification in the fruit fly. Thus, it will be of great interest to characterize

its localization and function during *R. prolixus* oogenesis, where the pro-oocytes are already specified in earlier nymph stages and germline stem cells are not detected in the adult ovary. Our analysis also shows that many *D. melanogaster* genes controlling critical processes like meiosis/mitosis, DNA damage checkpoint and repair, germ plasm assembly and pole cell formation, maintenance and migration and RNA decay are conserved in *R. prolixus*. Remarkably, the *R. prolixus* orthologs of several genes, whose expression in *D. melanogaster* is restricted to early oogenesis like *hts*, *α-spec*, *bgcn*, and *mei-W68*, are instead expressed in PVS and their transcripts deposited in mature eggs. This might represent a consequence of the particular structure of *R. prolixus* egg chambers and ovarioles. The egg chamber is exclusively formed by somatic follicle cells surrounding the oocyte and is constantly communicating with the tropharium through the trophic cords until choriogenesis begins. Differently, the egg chamber in the fruit fly is formed by a layer of follicle cells surrounding the germline, where the germline comprises the oocyte and fifteen nurse cells. Once the egg chamber buds off the germarium in *D. melanogaster*, it remains connected to the germarium through stalk cells, but the supply of RNAs and proteins for the growth and patterning of the oocyte is provided by the adjacent nurse cells through ring canals. Therefore, RNAs and factors expressed in early stages of oogenesis (i.e. tropharium/germarium) can be found in the mature eggs of *R. prolixus*, but not in those of *D. melanogaster*. The presence of transcripts encoding components of the RNA decay pathway, including *smaug*, *wispy*, and the CCR4/Not complex, suggest that maternal RNAs are actively degraded in *R. prolixus* to promote the maternal-to-zygotic transition. As expected, some patterning genes like *gurken* and *bicoid*, which are responsible for axial polarization in *D. melanogaster* are not detected in the *R. prolixus* genome given that they are restricted to Diptera[33]. However, we identified a putative ortholog of *D. melanogaster spitz*, which encodes a TGF-α protein, expressed in *R. prolixus* oogenesis. Members of this family, which include Gurken, have been connected to the establishment of the dorsal–ventral polarity in crickets, wasps, and beetles[33]. Also, we did not find an obvious ortholog of the germ plasm factor *oskar*, although a *oskar* gene was identified in cricket, a hemimetabolous insect like *R. prolixus*[99]. Another patterning gene expressed in *R. prolixus* oogenesis is *nanos*. Nanos is responsible for the specification of the posterior segments in a range of distantly related insect species[78,100,101]. Although these observations suggest that the mechanisms controlling axial polarization and germplasm formation are partially conserved in *R. prolixus*, functional studies are required to gain further insights into these processes.

Our data show that the transcripts encoding a variety of ribosomal proteins rank among the top 25 genes that are expressed early during oogenesis, but are under-represented in mature unfertilized eggs. It is possible that in *R. prolixus* ribosomal proteins rather than the corresponding transcripts are deposited in the egg and used in early stages of embryogenesis. This strategy might guarantee that ribosomal proteins are readily available to rapidly assemble ribosomes. Surprisingly, we found 629 genes to be highly represented in the Egg samples, but not in the PVS. It seems reasonable to assume that these genes are activated in the tropharium in response to the blood meal, but their expression is not sustained over time. The transport of the corresponding transcripts to developing oocytes would, therefore, progressively deplete the tropharium of these RNAs while they accumulate in the egg. Among the top 20 of DEGs with high expression in mature eggs, we find a putative ortholog of *Yellow-g2* (RPRC009244). This gene has been associated with the resistance of mosquito eggs to desiccation as well as to the production of cuticle pigments in *R. prolixus*[102]. At least two genes might be activated in oogenesis to control the oxidative stress caused by the blood feeding habits of *R. prolixus*. In fact, RPRC010216 encodes a putative ortholog of the cyclic AMP-dependent transcription factor ATF-4, a key regulator in metabolic and redox processes as well as a crucial factor in the integrated stress response, and RPRC002388 encoding a predicted PMSR-containing protein. Peptide Methionine Sulfoxide Reductase (PMSR) was associated with the repair of proteins and peptides that have been damaged by oxidizing environments[103]. It is well-established that the blood-feeding habit of *R. prolixus* generates oxidative species. Hence, RPRC002388 might exert a crucial role in the repair of proteins during *R. prolixus* oogenesis and early embryonic development. Another gene that might be linked to the heme metabolism is RPRC001419 encoding a putative ortholog of the venom periostin-like protein 1 from *Pristhesancus plagipennis,* also an assassin bug like *R. prolixus*, but preying on other insects and not on mammals. *Pristhesancus* injects the venom into the prey to paralyse it and liquify the tissues to facilitate the feeding. Finally, some promising genes need to be validated by other molecular biology techniques. For instance, RRPC009877 coding for a putative ortholog of a proline-proline-glutamic acid (PPE) protein from *Mycobacterium*. The PPE protein family has been connected to cell wall remodeling and virulence in bacteria. It is noteworthy however, that the neighboring gene in the contig, namely RPRC009884, also encodes for a putative bacterial protein. It will be important to determine whether these genes are truly present in the *R. prolixus* genome or they are bacterial genes accidentally assembled in the insect genome.

Our data also reveal that a variety of transposable elements are expressed in the early stages of oogenesis and their transcripts accumulate in the developing oocyte. It was previously shown that approximately 6% of the *R. prolixus* genome is made up of transposable sequences, with the Tc1-*mariner* elements comprising 75% of the mobilome. In accordance, we find that the Tc1-*mariner* transposons display the highest steady-state RNA levels in both PVS and Egg samples compared to other transposon families. However, the *helitron* elements also appear to display comparable expression levels to Tc1-*mariner* together with a class of "unknown" repetitive sequences, even though they are less represented in the genome. Overall, transposable and repetitive sequences account for only 1.1% and 0.6% of the unambiguous reads in PVS and Egg stages, respectively. We previously showed that key components of the piRNA pathway are conserved in *R. prolixus* and are important for female adult fertility[38]. Together with the observation that *R. prolixus* ovaries express piRNAs[38], these data suggest that the piRNA pathway efficiently downregulates transposable and repetitive sequences in *R. prolixus* oogenesis. For all of the transposons, the RNA levels in PVS and Egg stages are generally comparable, thus pointing to a mechanism whereby transposon transcripts are expressed in the trophocytes and transported from the tropharium onto the oocyte through the trophic cords. Although our analysis does not allow us to draw conclusions on the transcriptional activity of the transposons in the follicle cells versus the germline in *R. prolixus*, the fact that transcripts

corresponding to all known transposon families are detected both in PVS and in Egg stages demonstrates that none of these families is restricted to the follicle cells. It remains to be elucidated whether transposons can be transcribed in the follicle cells and their transcripts transferred to the oocyte during oogenesis or rather the transcription of these repetitive sequences occurs both in somatic and germ cells.

Chagas disease is a main threat to human health worldwide, but cures or vaccines for this illness are not yet available. The most promising strategies to reduce its diffusion rely on targeting triatomine vectors. Our study sheds light on oogenesis, early embryogenesis, and adult female fertility in *R. prolixus,* a primary vector of the Chagas disease in Central and South America. Together with substantial improvements to genome annotation in this species and novel bioinformatic resources, we provide a framework for future genetic, genomic and functional studies that might pave the way to the development of novel *R. prolixus* population control or replacement strategies.

Methods

R. prolixus handling, total RNA extraction, and library preparation.

The *R. prolixus* colony was raised at 28 °C and 75% relative humidity and regularly fed white rabbit blood at 3-week intervals in the laboratory of Insect Biochemistry at the Institute of Medical Biochemistry at the Federal University of Rio de Janeiro, Brazil. All animal care and experimental protocols were conducted in accordance with the guidelines of the Committee for Evaluation of Animal Use for Research (Universidade Federal do Rio de Janeiro, CAUAP-UFRJ) and the NIH Guide for the Care and Use of Laboratory Animals (ISBN 0-309-05377-3). Protocols were approved by CAUAP-UFRJ under registry #IBQM155/13. Dedicated technicians in the animal facility localized at the Instituto de Bioquímica Médica Leopoldo de Meis (UFRJ) carried out all protocols related to rabbit husbandry under strict guidelines, with supervision of veterinarians to ensure appropriate animal handling. Approximately, 100 mature chorionated eggs were dissected in ice-cold Phosphate Buffered Saline (PBS) from 10/15 adult females two weeks after a blood meal, and RNA-Seq libraries were produced as previously described[38]. Briefly, total RNA was extracted with TRIzol Reagent (Life Technologies) according to manufacturer instructions, treated with TURBO DNA-free kit (Ambion), and subjected to paired-end RNA sequencing (RNA-Seq) library production (Illumina). The libraries were generated and sequenced on HiSeq Illumina platforms at Lactad Facility (University of Campinas, Brazil) as previously described[38]. PVS libraries were generated previously and are available from NCBI Sequence Read Archive (SRA)[38].

Pre-processing and genomic mapping of the RNA-Seq reads.

Different tools were used in order to ensure the qualities of the reads we worked on as well as map them to *Rhodnius'* genomic sequence. FastQC[104] was adopted to check the quality of the libraries in regard to per-base sequence quality, sequence duplication level, overrepresented sequences, GC content, and presence of Illumina adapters. Raw read sequences were pre-processed by filtering out read sequences with less than 20 bp in length (-length 20); removing Ns from the read ends (-trim-n); trimming Illumina adapters (-illumina) and low-quality bases (-q 20) by Trim Galore! version 0.6.1[105,106]. A genome index was built with the genome assembly of *R. prolixus* RproC3[7] and *R. prolixus* Tc1-*mariner* consensus sequences[93] for the first-pass genomic mapping. Read alignment was performed by STAR (Spliced Transcripts Alignment to a Reference) version 2.7.2b[107] reporting read-pairs with at most 30 multiple alignments (-outFilterMultimapNmax) and a ratio between mismatches and mapped bases less or equal to 0.1 (-outFilterMismatchNoverLmax). To rescue unmapped reads in this first-pass genomic mapping, we trimmed these sequences for poly-A/T tails (tail length > 4) and we filtered by a minimum length (> 25nt) and GC-content (> 20%) using prinseq-lite v0.20.4[108] and re-mapped again (Supplementary Table S1).

Transcriptome improvements in the reference annotation.

Novel genes and genomic coordinate updates of reference genes were identified using the genomic mappings by Cufflinks. Uniquely mapped read-pairs were used to improve the transcriptome landscape by genome-guided transcriptome assembly using Cufflinks pipeline version 2.2.1[41] with the following parameters: at least 20 mapped read-pairs to support an assembled transcript (-min-frags-per-transfrag 20), novel isoform abundance ≥ 40% of the major isoform abundance (-min-isoform-fraction 0.4), reference assembly RproC3 (*Rhodnius-prolixus*-CDC_SCAFFOLDS_RproC3.fa) was used for bias detection (-b) and to exclude artifacts (e.g. repetitive elements) (-s); and the reference gene annotation of RproC3 (*Rhodnius-prolixus*-CDC_BASEFEATURES_RproC3.3.gtf) was used to guide the assembly, to compare with reference genes and to classify the assembled elements. We considered the assembled elements with a complete match of intron chain (class code "="), potentially novel isoform (class code "j"), and unknown/intergenic transcript (class code "u"). Novel isoform abundances were calculated by Salmon v1.2.1[109] and assembled isoforms with estimated CPM < 1 (Counts per Million) on both previtellogenic and egg stage were removed from the transcriptome assembly. Transdecoder v5.5[110] was applied to predict ORF candidates in the reference transcripts and to compare with the predicted ORFs in the improved transcriptome (Supplementary Tables S2 and S3).

Discovery of novel transcripts and repetitive elements by de novo assembly.

RNA-Seq sequences that failed to map to the reference genome were used by the tool Trinity to assemble transcripts and repetitive elements. Unmapped read-pairs were used for de novo transcriptome assembly with the Trinity method v2.5.1[40] with default parameters to identify novel transcripts that were not previously detected due to incomplete sequencing and assembly of the *R. prolixus* genome for each development stage, PVS and egg. De novo assembled transcripts were post-processed to filter out potential artifacts based on abundance and quality. First, expression levels of the de novo assembled transcripts were quantified by Salmon v1.2.1[109] and isoforms with TPM < 10 were removed. Next, deduplication of redundant contig sequences was performed by CD-HIT-

EST v4.6[111,112] at a nucleotide identity of 95%. TransRate software version 1.0.3[113] used the unmapped reads and the remaining contigs from the previous filtering steps as input to evaluate common assembly errors (chimeras, structural errors, incomplete assembly, and base errors) and to produce a diagnostic quality score for each contig, thereby removing possible artifacts. Assembled transcripts remaining from this last step were considered to be good quality transcripts (Supplementary Table S4). To identify potential protein-coding transcripts, we compared the assembled contigs against the protein databases of *R. prolixus, D. melanogaster, Caenorhabditis elegans* and *Homo sapiens* from Uniprot release 2020_01[114] using BLASTX v2.9 (e-value < 1e−10). Uniprot protein coverage was calculated using the scripts blast_outfmt6_group_segments.pl and blast_outfmt6_group_segments. tophit_coverage.pl provided by Trinity[40]. In addition, we performed predictions of candidate coding regions (at least 100 amino acids in length) using Transdecoder[110] v5.5 (https://github.com/TransDecoder). ORF candidates can be classified by their completeness into complete, partial (5′ or 3′) and internal. If more than one type of ORF is detected in a contig, we assign the most complete type (Supplementary Table S5). De novo assembled transcriptomes of PVS and eggs were deduplicated using CD-HIT-EST (identity > 95%) to produce a final non-redundant set of transcripts. Putative protein-coding transcripts with an estimated number of reads lower than 20 were removed. To identify putative transposable elements, we performed homology searches using HMMER v3.1b2 (-cut_ga) with profile hidden Markov models (HMMs) of repetitive DNA elements from Dfam database release 3.1[94,95]. We verified the uniqueness of the putative transposon sequences by comparing (using BLASTN v2.9) with the library of repetitive elements (*Rhodnius-prolixus*-CDC_REPEATS.lib) of the assembly RproC3 available in VectorBase[7] and consensus sequences from Tc1-*mariner*[93] (Supplementary Table S10).

Read coverage of the improved gene annotation and repetitive elements.

Read coverage of protein-coding genes was assayed by counting the number of genes of the reference and improved gene annotation overlapped by uniquely mapped read-pairs (at least 40 mapped read-pairs) using featureCounts version 2. We set featureCounts to only consider read-pairs that have both ends aligned and to not include read-pairs that have their ends mapping to different chromosomes or mapping to the same chromosome but on different strands (-p -B -C -largestOverlap -O). To quantify the read coverage of the repetitive element classes, we included the multi-mapped read-pairs in the read assignment (-fraction -M) to the RepeatMasker and improved gene annotation of *R. prolixus*. Read-pairs assigned to more than one class of genomic elements (protein-coding gene, transposable element, simple repeat, multi-copy genes, etc.) were considered ambiguous. Reference gene annotation, Repeat-Masker annotation, and genome assembly of the RproC3 are available at VectorBase[7]. To reduce the ambiguity in the read-pair assignment, we merged the overlapped repetitive elements from the same class and strand (Supplementary Table S9). Total coverage for each repetitive element class was calculated by summing the partial coverage of individual elements of the same class. We considered TE merged regions longer than 0.5 Kb[7].

Quantification and differential gene expression analysis.

Quantification of the improved gene annotation was performed by Salmon 1.2.1[109], which was set to produce aggregated gene-level abundance estimates (-g). The transcriptome index was built using k-mer size equal to 19 (-k 19) and the DNA sequences of the improved transcriptome. These sequences were extracted using the script gtf_genome_to_cdna_fasta.pl (https://github.com/TransDecoder). The estimated number of reads was used for the differential gene expression analysis. Genes with low expression levels were removed; we retained genes with counts scaled by total mapped reads per million (CPM) greater than 1 in at least two samples for differential gene expression analysis. We applied Trimmed Mean of M values[115] to normalize gene counts by sequencing depth and RNA composition. DESeq2[44] was applied to evaluate the differential gene expression during the transition of the two developmental stages using DEBrowser v1.16.1[116]. Genes were considered differentially expressed when they had an adjusted p-value ≤ 0.01 (FDR); fold-change ≤ − 2.5 (most abundant in PVS) or fold-change ≥ 2.5 (most abundant in egg data sets). Differentially expressed genes with negative or positive fold-change are labeled as "down-regulated" and "up-regulated" genes, if their expression is respectively lower or higher in eggs versus PVS stages. Heatmaps of the top 25 most differentially expressed genes were generated by Heatmapper[117] using the normalized expression (CPM) for each sample. Gene product annotation was obtained from Uniprot[114].

Enrichment analysis.

We compared the enriched functional categories annotated for differentially expressed genes of each stage. We used Uniprot protein's best hits (based on score and e-value) of *R. prolixus* assigned to the DNA sequences of the *R. prolixus* transcripts by BLASTX v2.9 [118] (-evalue 0.05) as the input for gene ontology enrichment analysis using g:Profiler[45]. All known *R. prolixus* genes were used as statistical domain scope or background. GO terms were considered enriched when Benjamini–Hochberg FDR ≤ 0.05. ClueGO[46,47] was used to fuse the GO parent-terms based on similar associated genes and produce functionally grouped networks (Supplementary Fig. S3).

Data availability
RNA-Seq datasets can be accessed at SRA NIH, accession number …. ; all other data can be found in the manuscript and supporting information files. A visual representation of the improved transcriptome that we elaborated in our work is available at http://www.genome.rio/cgi-bin/hgGateway?db=rproC3.

References
1. Lidani, K. C. F. *et al.* Chagas disease: From discovery to a worldwide health problem. *Front. Public Health* **7**, 166 (2019).

2. Yadon, Z. E. & Schmunis, G. A. Congenital Chagas disease: Estimating the potential risk in the United States. *Am. J. Trop. Med. Hyg.* **81**, 927–933 (2009).
3. Antinori, S. *et al.* Chagas disease in Europe: A review for the internist in the globalized world. *Eur. J. Intern. Med.* **43**, 6–15 (2017).
4. Pinto, A., Pett, S. & Jackson, Y. Identifying Chagas disease in Australia: An emerging challenge for general practitioners. *Aust. Fam. Phys.* **43**, 440–442 (2014).
5. Imai, K. *et al.* Chagas disease: A report of 17 suspected cases in Japan, 2012–2017. *Trop. Med. Health* **47**, 38 (2019).
6. Mesquita, R. D. *et al.* Genome of *Rhodnius prolixus*, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 14936–14941 (2015).
7. Giraldo-Calderón, G. I. *et al.* VectorBase: An updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.* **43**, D707–D713 (2015).
8. Medeiros, M. N. *et al.* Transcriptome and gene expression profile of ovarian follicle tissue of the triatomine bug *Rhodnius prolixus*. *Insect Biochem. Mol. Biol.* **41**, 823–831 (2011).
9. Ribeiro, J. M. C. *et al.* An insight into the transcriptome of the digestive tract of the bloodsucking bug, *Rhodnius prolixus*. *PLoS Negl. Trop. Dis.* **8**, e2594 (2014).
10. Leyria, J., Orchard, I. & Lange, A. B. Transcriptomic analysis of regulatory pathways involved in female reproductive physiology of *Rhodnius prolixus* under different nutritional states. *Sci. Rep.* **10**, 1–16 (2020).
11. Salcedo-Porras, N. & Lowenberger, C. The innate immune system of kissing bugs, vectors of chagas disease. *Dev. Comp. Immunol.* **98**, 119–128 (2019).
12. César, J. *et al.* Analysis of the testis transcriptome of the Chagas' disease vector *Rhodnius prolixus*. bioRxiv 616193, https://doi.org/10.1101/616193 (2019).
13. Cardoso, J. C. *et al.* Analysis of the testicle's transcriptome of the Chagas disease vector R*hodnius prolixus*. *Genomics* 33 (2019).
14. Nunes-da-Fonseca, R., Berni, M., Tobias-Santos, V., Pane, A. & Araujo, H. M. *Rhodnius prolixus*: From classical physiology to modern developmental biology. *Genesis* **55**, 22995 (2017).
15. Buxton, P. A. The biology of a blood-sucking bug, *Rhodnius prolixus*. *Trans. R. Entomol. Soc. Lond.* **78**, 227–256 (2009).
16. Valle, D., Lima, J. E., Goldenberg, S. & Garcia, E. S. *Rhodnius prolixus* vitellogenesis: Dependence upon the blood source. *J. Insect Physiol.* **33**, 249–254 (1987).
17. Lutz, D. A. & Huebner, E. Development and cellular differentiation of an insect telotrophic ovary (*Rhodnius prolixus*). *Tissue Cell* **12**, 773–794 (1980).
18. Huebner, E. Nurse cell-oocyte interaction in the telotrophic ovarioles of an insect, *Rhodnius prolixus*. *Tissue Cell* **13**, 105–125 (1981).
19. Büning, J. *The Insect Ovary: Ultrastructure, Previtellogenic Growth and Evolution*. (Springer, 1994).
20. Huebner, E. Oocyte—Follicle cell interaction during normal oogenesis and atresia in an insect. *J. Ultrastruct. Res.* **74**, 95–104 (1981).
21. Lutz, D. A. & Huebner, E. Development of nurse cell-oocyte interactions in the insect telotrophic ovary (*Rhodnius prolixus*). *Tissue Cell* **13**, 321–335 (1981).
22. Huebner, E. & Anderson, E. A cytological study of the ovary of *Rhodnius p*rolixus II. Oocyte differentiation. *J Morphol.* **137**, 385–415 (1972).
23. Valdimarsson, G. & Huebner, E. The development of microtubular arrays in the germ tissue of an insect telotrophic ovary. *Tissue Cell* **21**, 123–138 (1989).
24. Harrison, R. E. & Huebner, E. Unipolar microtubule array is directly involved in nurse cell-oocyte transport. *Cell Motil. Cytoskelet.* **36**, 355–362 (1997).
25. Valle, D. Vitellogenesis in insects and other groups: A review. *Mem. Inst. Oswaldo Cruz* **88**, 1–26 (1993).
26. Beament, J. W. L. The formation and structure of the chorion of the egg in an hemipteran, *Rhodnius prolixus*. *Q. J. Microsc. Sci.* **87**, 393–439 (1946).
27. Tuft, P. H. The structure of the insect egg-shell in relation to the respiration of the embryo. *J. Exp. Biol.* **26**, 327–334 (1950).
28. Fakhouri, M. *et al.* Minor proteins and enzymes of the *Drosophila* eggshell matrix. *Dev. Biol.* **293**, 127–141 (2006).
29. Telfer, W. H. & Anderson, L. M. Functional transformations accompanying the initiation of a terminal growth phase in the cecropia moth oocyte. *Dev. Biol.* **17**, 512–535 (1968).
30. Waring, G. L. & Mahowald, A. P. Identification and time of synthesis of chorion proteins in *Drosophila melanogaster*. *Cell* **16**, 599–607 (1979).
31. Mazur, G. D., Regier, J. C. & Kafatos, F. C. Order and defects in the silkmoth chorion, a biological analogue of a cholesteric liquid crystal. *Insect Ultrastruct.* 150–185 (1982). https://doi.org/10.1007/978-1-4615-7266-4_6.
32. Kugler, J.-M. & Lasko, P. Localization, anchoring and translational control of oskar, gurken, bicoid and nanos mRNA during *Drosophila* oogenesis. *Fly* **3**, 15–28 (2009).
33. Lynch, J. A., Peel, A. D., Drechsler, A., Averof, M. & Roth, S. EGF signaling and the origin of axial polarity among the insects. *Curr. Biol.* **20**, 1042–1047 (2010).
34. Ephrussi, A. & Lehmann, R. Induction of germ cell formation by oskar. *Nature* **358**, 387–392 (1992).
35. Hamm, D. C. & Harrison, M. M. Regulatory principles governing the maternal-to-zygotic transition: Insights from *Drosophila* melanogaster. *Open Biol.* **8**, 180183 (2018).
36. Giraldez, A. J. *et al.* Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science* **312**, 75–79 (2006).
37. Laver, J. D., Marsolais, A. J., Smibert, C. A. & Lipshitz, H. D. Regulation and function of maternal gene products during the maternal-to-zygotic transition in *Drosophila*. *Curr. Top. Dev. Biol.* **113**, 43–84 (2015).
38. Brito, T. *et al.* Transcriptomic and functional analyses of the piRNA pathway in the Chagas disease vector *Rhodnius prolixus*. *PLoS Negl. Trop. Dis.* **12**, e0006760 (2018).
39. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
40. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
41. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
42. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
43. Bouts, D. M. D. *et al.* Biochemical properties of the major proteins from *Rhodnius prolixus* eggshell. *Insect Biochem. Mol. Biol.* **37**, 1207–1221 (2007).
44. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
45. Reimand, J. *et al.* g:Profiler—A web server for functional interpretation of gene lists (2016). *Nucleic Acids Res.* **44**, W83–W89 (2016).
46. Bindea, G. *et al.* ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).
47. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

48. Luo, Y., Na, Z. & Slavoff, S. A. P-Bodies: Composition, properties, and functions. *Biochemistry* **57**, 2424–2431 (2018).
49. Greenspan, L. J., de Cuevas, M. & Matunis, E. Genetics of gonadal stem cell renewal. *Annu. Rev. Cell Dev. Biol.* **31**, 291–315 (2015).
50. Hughes, S. E., Miller, D. E., Miller, A. L. & Hawley, R. S. Female meiosis: Synapsis, recombination, and segregation in. *Genetics* **208**, 875–908 (2018).
51. Bastock, R. & Johnston, D. S. *Drosophila* oogenesis. *Curr. Biol.* **18**, R1082–R1087 (2008).
52. Chen, S., Wang, S. & Xie, T. Restricting self-renewal signals within the stem cell niche: multiple levels of control. *Curr. Opin. Genet. Dev.* **21**, 684–689 (2011).
53. Chen, D. & McKearin, D. M. A discrete transcriptional silencer in the bam gene determines asymmetric division of the *Drosophila* germline stem cell. *Development* **130**, 1159–1170 (2003).
54. Klusza, S. & Deng, W.-M. At the crossroads of differentiation and proliferation: Precise control of cell-cycle changes by multiple signaling pathways in *Drosophila* follicle cells. *BioEssays* **33**, 124–134 (2011).
55. Huang, J., Reilein, A. & Kalderon, D. Yorkie and Hedgehog independently restrict BMP production in escort cells to permit germline differentiation in the ovary. *Development* **144**, 2584–2594 (2017).
56. Ohkura, H. Meiosis: An overview of key differences from mitosis. *Cold Spring Harb. Perspect. Biol.* **7**, (2015).
57. Cahoon, C. K. & Hawley, R. S. Regulating the construction and demolition of the synaptonemal complex. *Nat. Struct. Mol. Biol.* **23**, 369–377 (2016).
58. Barr, J., Gilmutdinov, R., Wang, L., Shidlovskii, Y. & Schedl, P. The *Drosophila* CPEB protein orb specifies oocyte fate by a 3′UTR-dependent autoregulatory loop. *Genetics* **213**, 1431–1446 (2019).
59. Garrido, D. *et al.* Cyclin B3 activates the anaphase-promoting complex/cyclosome in meiosis and mitosis. *PLoS Genet.* **16**, e1009184 (2020).
60. Pek, J. W., Ng, B. F. & Kai, T. Polo-mediated phosphorylation of Maelstrom regulates oocyte determination during oogenesis in *Drosophila*. *Development* **139**, 4505–4513 (2012).
61. Abdu, U., Brodsky, M. & Schüpbach, T. Activation of a meiotic checkpoint during *Drosophila* oogenesis regulates the translation of Gurken through Chk2/Mnk. *Curr. Biol.* **12**, 1645–1651 (2002).
62. Bakhrat, A., Pritchett, T., Peretz, G., McCall, K. & Abdu, U. *Drosophila* Chk2 and p53 proteins induce stage-specific cell death independently during oogenesis. *Apoptosis* **15**, 1425–1434 (2010).
63. Brodsky, M. H. *et al. Drosophila melanogaster* MNK/Chk2 and p53 regulate multiple DNA repair and apoptotic pathways following DNA damage. *Mol. Cell. Biol.* **24**, 1219–1231 (2004).
64. Morris, J. & Lehmann, R. *Drosophila* oogenesis: Versatile spn doctors. *Curr. Biol.* **9**, R55–R58 (1999).
65. LaRocque, J. R., Jaklevic, B., Su, T. T. & Sekelsky, J. *Drosophila* ATR in double-strand break repair. *Genetics* **175**, 1023–1033 (2007).
66. Ghabrial, A. & Schüpbach, T. Activation of a meiotic checkpoint regulates translation of Gurken during *Drosophila* oogenesis. *Nat. Cell Biol.* **1**, 354–357 (1999).
67. Ghabrial, A., Ray, R. P. & Schüpbach, T. okra and spindle-B encode components of the RAD52 DNA repair pathway and affect meiosis and patterning in *Drosophila* oogenesis. *Genes Dev.* **12**, 2711–2723 (1998).
68. González-Reyes, A. DNA repair and pattern formation come together. *Nat. Cell Biol.* **1**, E150–E152 (1999).
69. Eichler, C. E., Hakes, A. C., Hull, B. & Gavis, E. R. Compartmentalized oskar degradation in the germ plasm safeguards germline development. *eLife* **9** (2020).
70. Dehghani, M. & Lasko, P. Multiple functions of the DEAD-Box helicase vasa in *Drosophila* oogenesis. *Results Probl. Cell Differ.* **63**, 127–147 (2017).
71. Lehmann, R. Germ plasm biogenesis–An Oskar-centric perspective. *Curr. Top. Dev. Biol.* **116**, 679–707 (2016).
72. Sinsimer, K. S., Jain, R. A., Chatterjee, S. & Gavis, E. R. A late phase of germ plasm accumulation during *Drosophila* oogenesis requires lost and rumpelstiltskin. *Development* **138**, 3431–3440 (2011).
73. Mahowald, A. P. Assembly of the *Drosophila* germ plasm. *Int. Rev. Cytol.* 187–213. https://doi.org/10.1016/s0074-7696(01)03007-8 (2001).
74. Trcek, T. & Lehmann, R. Germ granules in *Drosophila*. *Traffic* **20**, 650–660 (2019).
75. Breitwieser, W., Markussen, F. H., Horstmann, H. & Ephrussi, A. Oskar protein interaction with Vasa represents an essential step in polar granule assembly. *Genes Dev.* **10**, 2179–2188 (1996).
76. Neuman-Silberberg, F. S. & Schüpbach, T. The *Drosophila* dorsoventral patterning gene gurken produces a dorsally localized RNA and encodes a TGF alpha-like protein. *Cell* **75**, 165–174 (1993).
77. Ephrussi, A. & St Johnston, D. Seeing is believing: The bicoid morphogen gradient matures. *Cell* **116**, 143–152 (2004).
78. Lynch, J. A. & Desplan, C. Novel modes of localization and function of nanos in the wasp *Nasonia*. *Development* **137**, 3813–3821 (2010).
79. Duncan, E. J., Benton, M. A. & Dearden, P. K. Canonical terminal patterning is an evolutionary novelty. *Dev. Biol.* **377**, 245–261 (2013).
80. Gatfield, D., Unterholzner, L., Ciccarelli, F. D., Bork, P. & Izaurralde, E. Nonsense-mediated mRNA decay in *Drosophila*: At the intersection of the yeast and mammalian pathways. *EMBO J.* **22**, 3960–3970 (2003).
81. Sgromo, A. *et al. Drosophila* Bag-of-marbles directly interacts with the CAF40 subunit of the CCR4–NOT complex to elicit repression of mRNA targets. *RNA* **24**, 381–395 (2018).
82. Niinuma, S., Fukaya, T. & Tomari, Y. CCR4 and CAF1 deadenylases have an intrinsic activity to remove the post-poly(A) sequence. *RNA* **22**, 1550–1559 (2016).
83. Towler, B. P. & Newbury, S. F. Regulation of cytoplasmic RNA stability: Lessons from *Drosophila*. *Wiley Interdiscip. Rev. RNA* **9**, e1499 (2018).
84. Conti, E. & Izaurralde, E. Nonsense-mediated mRNA decay: Molecular insights and mechanistic variations across species. *Curr. Opin. Cell Biol.* **17**, 316–325 (2005).
85. Rebbapragada, I. & Lykke-Andersen, J. Execution of nonsense-mediated mRNA decay: What defines a substrate?. *Curr. Opin. Cell Biol.* **21**, 394–402 (2009).
86. Ribeiro, L. *et al.* Evolution and multiple roles of the Pancrustacea specific transcription factor zelda in insects. *PLoS Genet.* **13**, e1006868 (2017).
87. van Eeden, F. J., Palacios, I. M., Petronczki, M., Weston, M. J. & St Johnston, D. Barentsz is essential for the posterior localization of Oskar mRNA and colocalizes with it to the posterior pole. *J. Cell Biol.* **154**, 511–523 (2001).
88. Menon, K. P., Carrillo, R. A. & Zinn, K. The translational regulator Cup controls NMJ presynaptic terminal morphology. *Mol. Cell. Neurosci.* **67**, 126–136 (2015).
89. Schnorrer, F., Bohmann, K. & Nüsslein-Volhard, C. The molecular motor dynein is involved in targeting swallow and bicoid RNA to the anterior pole of *Drosophila* oocytes. *Nat. Cell Biol.* **2**, 185–190 (2000).
90. Chénais, B., Caruso, A., Hiard, S. & Casse, N. The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene* **509**, 7–15 (2012).
91. McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U. S. A.* **36**, 344–355 (1950).
92. Miller, W. J. & Capy, P. Mobile genetic elements as natural tools for genome evolution. *Methods Mol. Biol.* **260**, 1–20 (2004).

93. Filée, J., Rouault, J.-D., Harry, M. & Hua-Van, A. Mariner transposons are sailing in the genome of the blood-sucking bug *Rhodnius prolixus*. *BMC Genomics* **16**, 1061 (2015).
94. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
95. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).
96. Henriques, B. S. *et al.* Genome wide mapping of peptidases. Identification of protease gene duplications, horizontally transferred proteases and analysis of peptidase A1 structures, with considerations on their role in the evolution of hematophagy in triatominae. *Front. Physiol.* **8**, 1051 (2017).
97. Telfer, W. H. Development and physiology of the oöcyte-nurse cell syncytium. *Adv. Insect Physiol.* 223–319, https://doi.org/10.1016/s0065-2806(08)60164-2 (1975).
98. de Cuevas, M., Lilly, M. A. & Spradling, A. C. Germline cyst formation in *Drosophila*. *Annu. Rev. Genet.* **31**, 405–428 (1997).
99. Ewen-Campen, B., Srouji, J. R., Schwager, E. E. & Extavour, C. G. Oskar predates the evolution of germ plasm in insects. *Curr. Biol.* **22**, 2278–2283 (2012).
100. Chang, C.-C., Lee, W.-C., Cook, C. E., Lin, G.-W. & Chang, T. Germ-plasm specification and germline development in the parthenogenetic pea aphid *Acyrthosiphon pisum*: Vasa and nanos as markers. *Int. J. Dev. Biol.* **50**, 413–421 (2006).
101. De Keuckelaere, E., Hulpiau, P., Saeys, Y., Berx, G. & van Roy, F. Nanos genes and their role in development and beyond. *Cell. Mol. Life Sci.* **75**, 1929–1946 (2018).
102. Berni, M. *et al.* Pigmentation loci as markers for genome editing in the Chagas disease vector *Rhodnius prolixus*. bioRxiv 2020.04.29.067934, https://doi.org/10.1101/2020.04.29.067934 (2020).
103. Weissbach, H. *et al.* Peptide methionine sulfoxide reductase: Structure, mechanism of action, and biological function. *Arch. Biochem. Biophys.* **397**, 172–178 (2002).
104. Andrews, S. FastQC. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed Jul 2020 (2017).
105. Krueger, F. Trim Galore! https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. Accessed Jul 2020 (2019).
106. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.j* **17**, 10 (2011).
107. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
108. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
109. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
110. Brian, H. & Papanicolaou, A. *TransDecoder (Find Coding Regions Within Transcripts)*. https://github.com/TransDecoder/TransDecoder (2018).
111. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283 (2001).
112. Li, W., Jaroszewski, L. & Godzik, A. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* **18**, 77–82 (2002).
113. Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate: Reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.* **26**, 1134–1144 (2016).
114. UniProt Consortium. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
115. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
116. Kucukural, A., Yukselen, O., Ozata, D. M., Moore, M. J. & Garber, M. DEBrowser: Interactive differential expression analysis and visualization tool for count data. *BMC Genomics* **20**, 6 (2019).
117. Babicki, S. *et al.* Heatmapper: Web-enabled heat mapping for all. *Nucleic Acids Res.* **44**, W147–W153 (2016).
118. Altschul, S. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

Acknowledgements

Author contributions

A.P. conceived the study. V.C., T.B. and A.P. performed the bioinformatic analyses. M.S. composed the RIO browser with inputs from A.P. I.B., M.C. and M.B. carried on the wet biology experiments. M.C. and M.B. supervised and oriented younger students involved in the project. V.C., T.B., M.C., M.B., H.A. and A.P. contributed to the interpretation of the data. H.A. and A.P. provided financial support. V.C. and A.P. wrote the manuscript with input and help from all the authors.

Funding

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-81387-1.

Correspondence and requests for materials should be addressed to A.P.

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

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