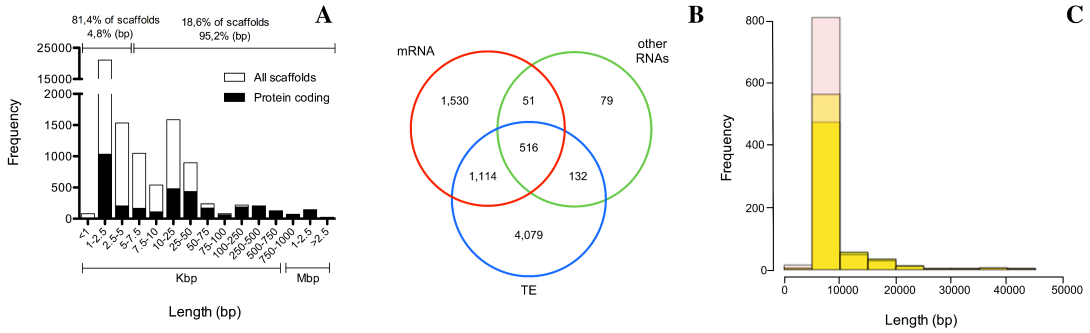

Figure A1: Genome basics: (A) Histogram showing length distribution of scaffolds. The length distribution of all scaffolds is shown as white bars (placed behind) and to scaffolds having protein-coding gene as black bars (placed in front). Relations between number of scaffolds and summed length (both as percentage) are shown above the pannel. (B) Venn diagram showing features distribution in the scaffolds. Features present in the consensus gene prediction (RNAs) and transposable element (TE) prediction populated 7,501 scaffolds (27%) from a universe of 27,870. (C) Duplicated regions length distribution histogram. There are three superimposed distributions: yellow (base) representing 99% identity regions, orange (middle) representing 95% and pink (top) representing 90% identity.

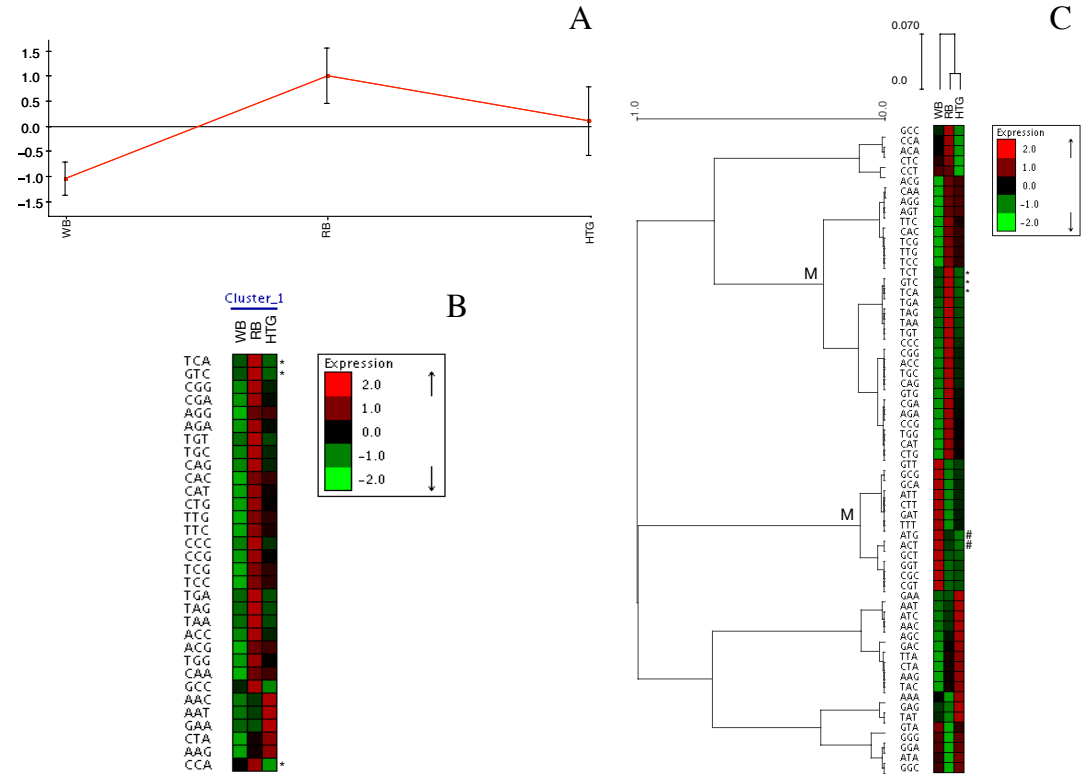

Figure A2: Codon usage frequency clustering: Coding backgrounds from *Rhodnius* (RB), *Wolbachia* (WB) and the horizontally transferred genes from *Wolbachia* (HTG) had their codon usage frequencies calculated. Automatic clustering used Click algorithm (A and B) inside Expander v6 and clustered codons with intermediate usage frequency (named *middle*), comparing to WB and RB. (A) Shows clustered average frequency profile, (B) The individual codon expression profiles. Hierarchical clustering (C) used Expander v6. Nodes marked as "M" contains codons with intermediate usage frequency, comparing to WB and RB (named *middle)*."*" Codons presenting HTG frequency difference comparing to WB smaller than 5%. *"#"*Codons presenting HTG frequency lower than WB and RB, despite shifted to RB. Codons marked with "*" or "#"were not considered as *middle* in Table D1.4 in Dataset.

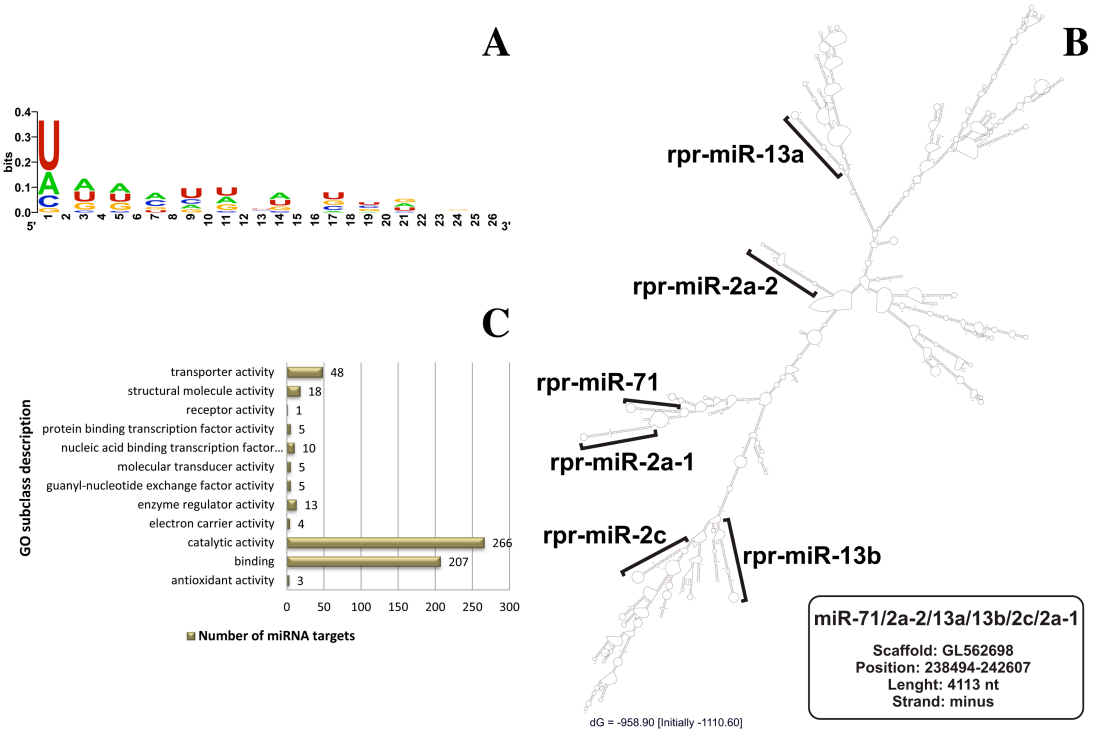

Figure A3: miRNA: *R. prolixus* mature miRNA sequences logo (**A**). RNA secondary structure of the cluster rpr-miR-71/2a-2/13a/13b/2c/2a-1 (**B**). GO subclass description of *R. prolixus* mature miRNA target genes (**C**).

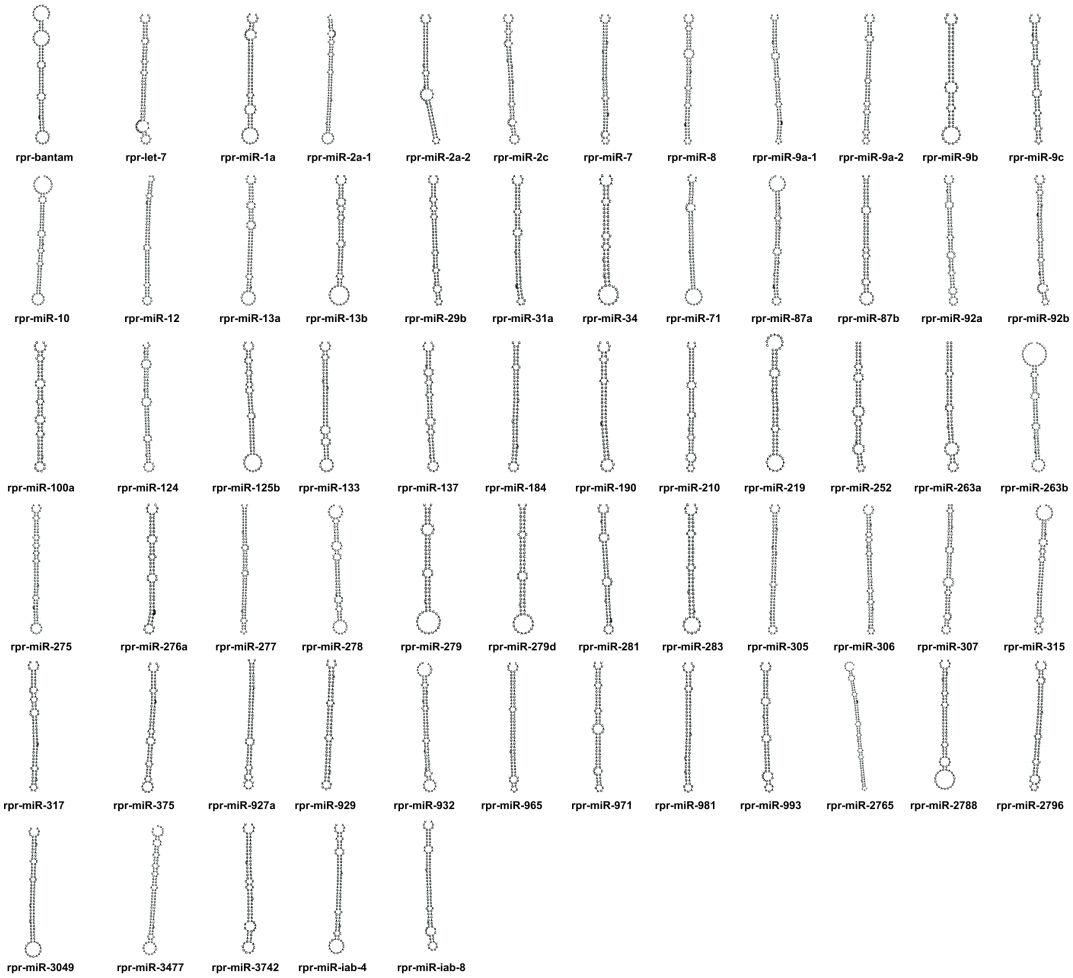

Figure A4: Secondary structures of the *R. prolixus* pre-miRNA sequences using RNAfold.

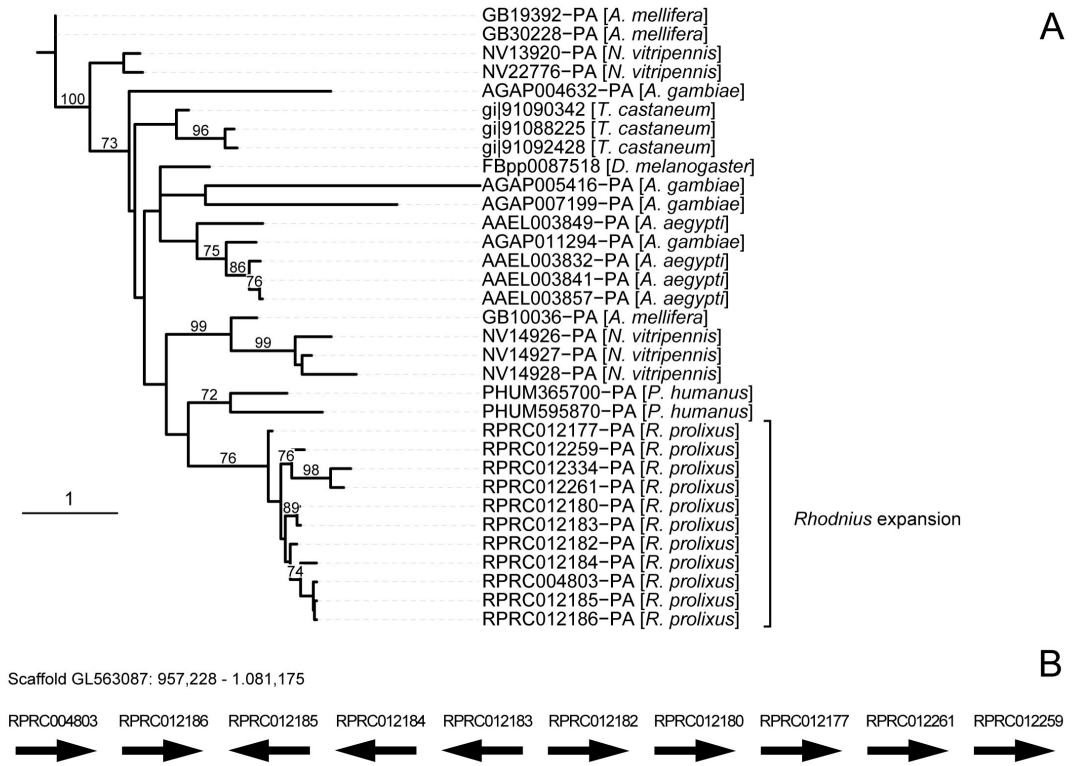

Figure A5: **Defensins phylogeny and tandem arrangement:** Sequences clustered by OrthoMCL in group 560 were joined with unclustered sequences presenting conserved domain PF01097. (**A**) Defensins phylogeny. Numbers in the tree represent bootstrap support of 500 replicates. (**B**) Tandem arrangement of defensins in scaffold GL563087.

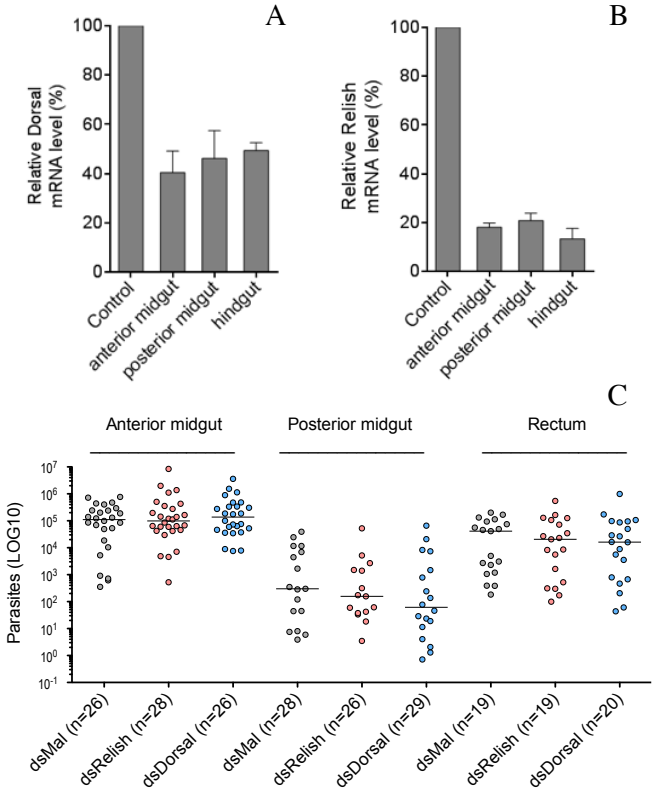

Figure A6: **Immunity:** To Knock-down rpRelish and rpDorsal, the insects were injected intrathoracically with 1µg of dsRNA for rpDorsal (**A**) or rpRelish (**B**). Three days after injection the different tissues were dissected, RNA was extracted and rpRelish and rpDorsal RNA levels were estimated by q PCR. The IMD pathway do not control of T. cruzi replication (**C**). Upon knock down of either rpRelish or rpDorsal *T cruzi* levels are not altered in any section of the digestive tract 7 days after infection.

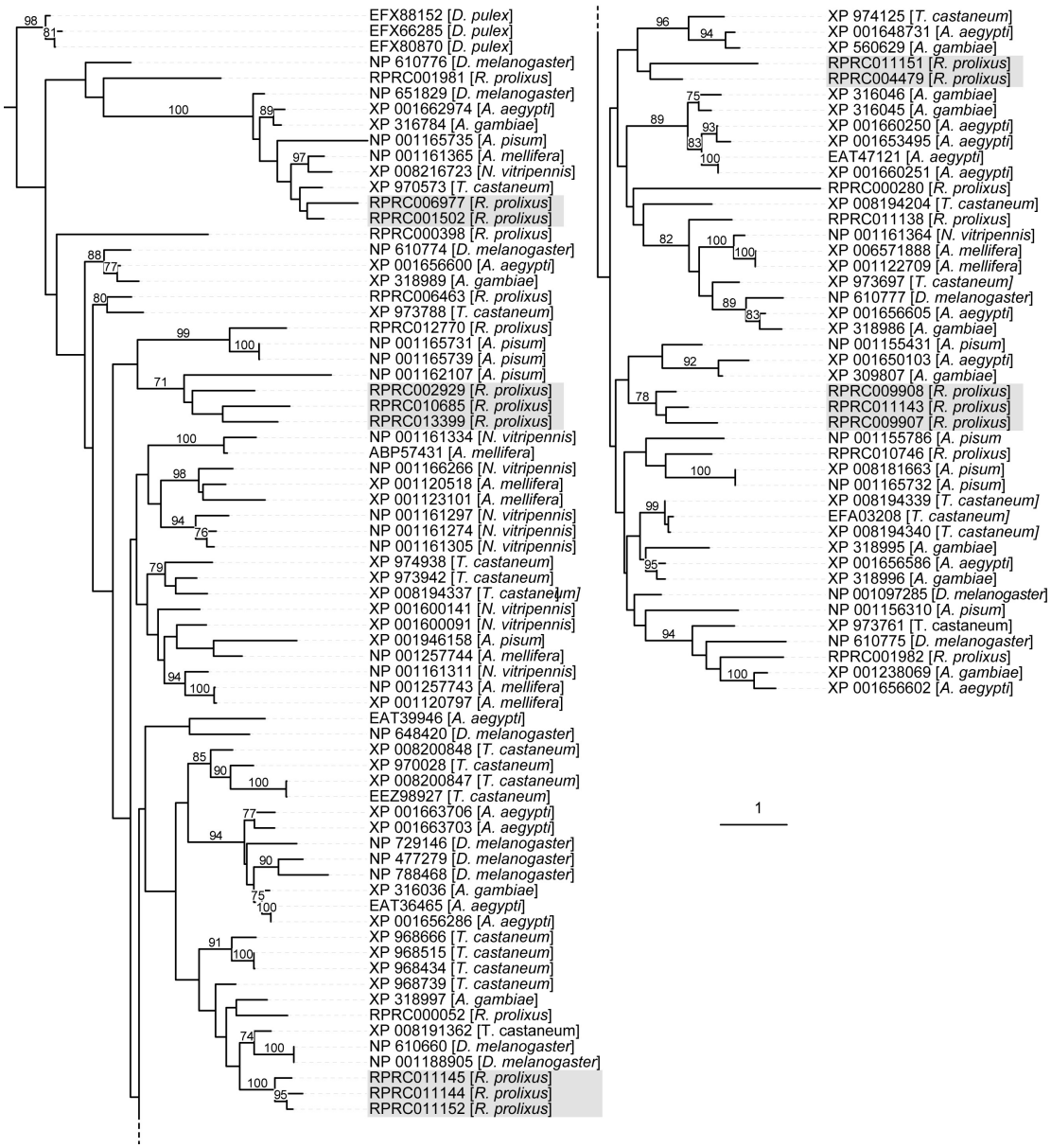

Figure A7: Phylogenetic tree of RR-1 domain CPR cuticular protein family: Grey background highlight *R. prolixus* expansions.

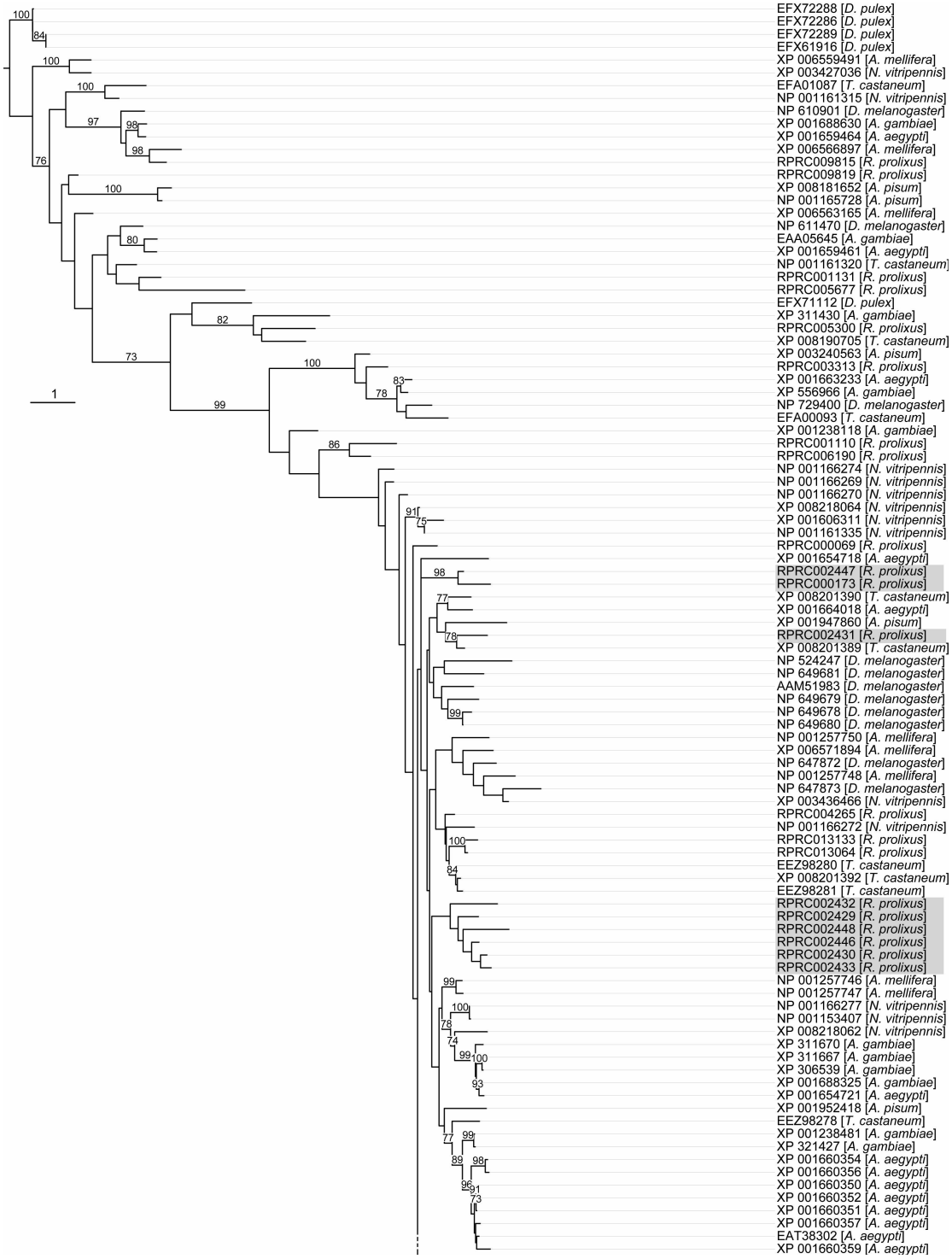

Figure A8: Phylogenetic tree of RR-2 domain CPR cuticular protein family. Part 1/2 : Grey background highlight *R. prolixus* expansions identified in OrthoMCL clustering.

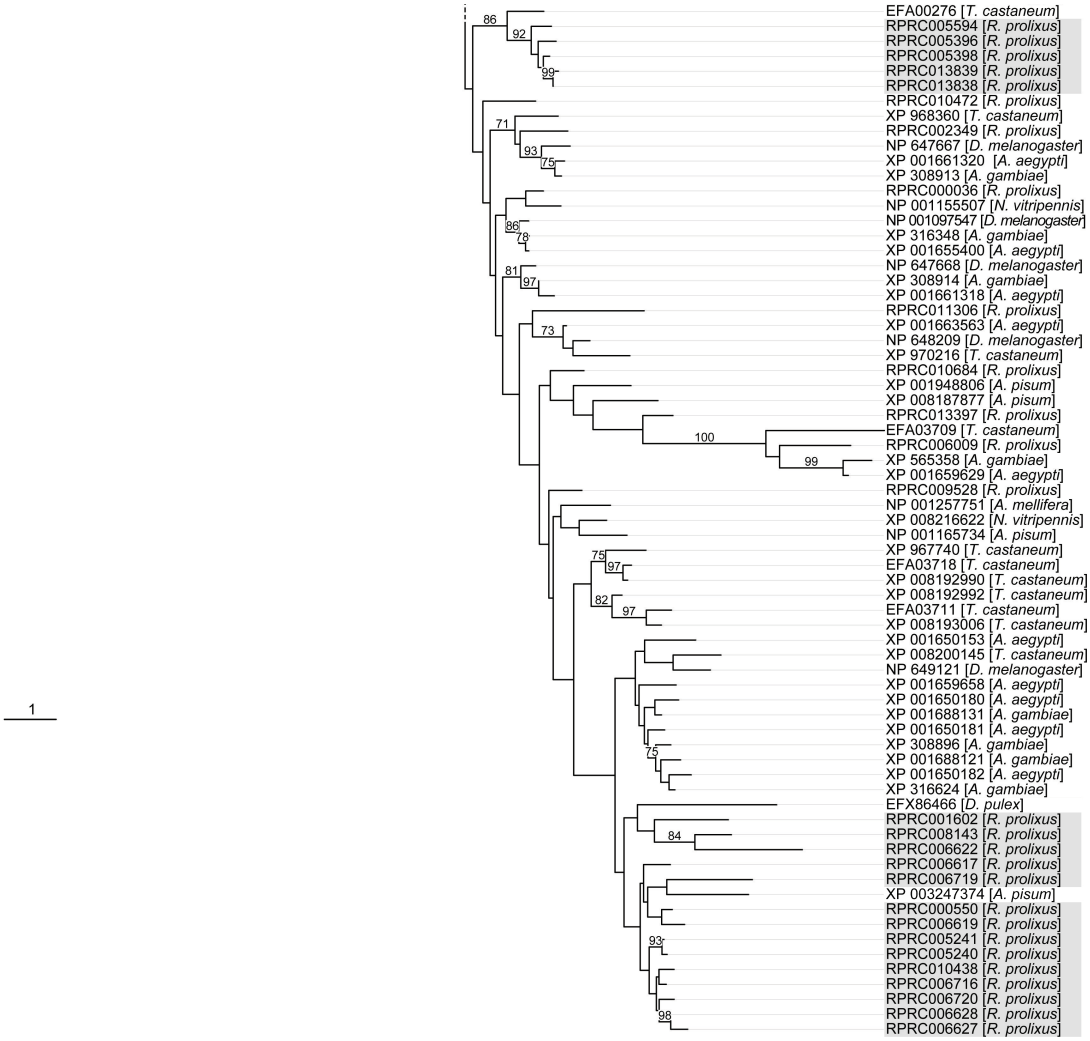

Figure A8: Phylogenetic tree of RR-2 domain CPR cuticular protein family. Part 2/2. Grey background highlight *R. prolixus* expansions identified in OrthoMCL clustering.

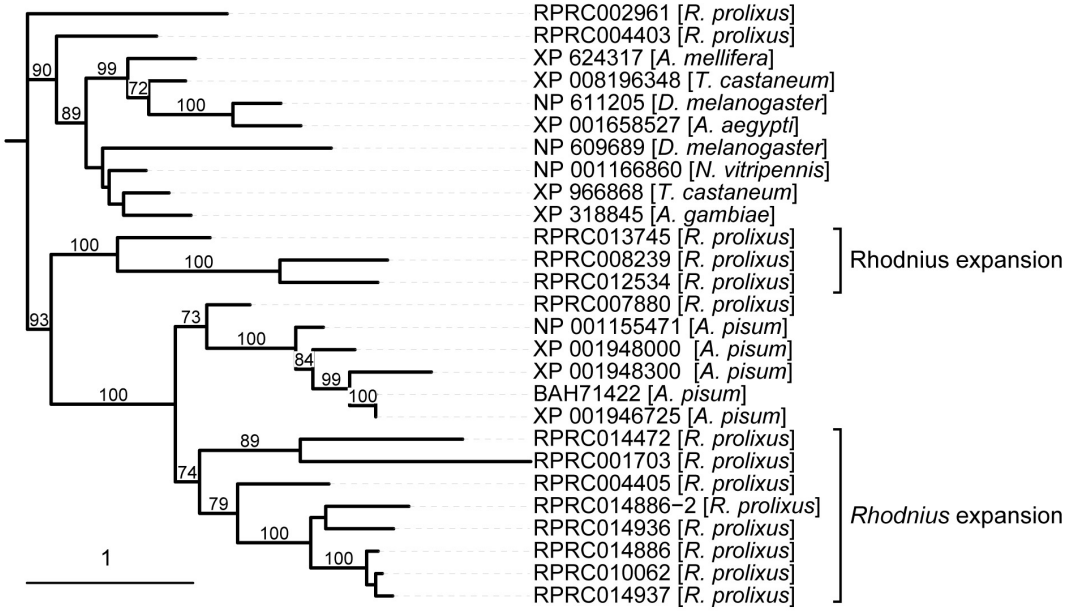

Figure A9: Phylogenetic tree of CPLCP cuticular protein family.

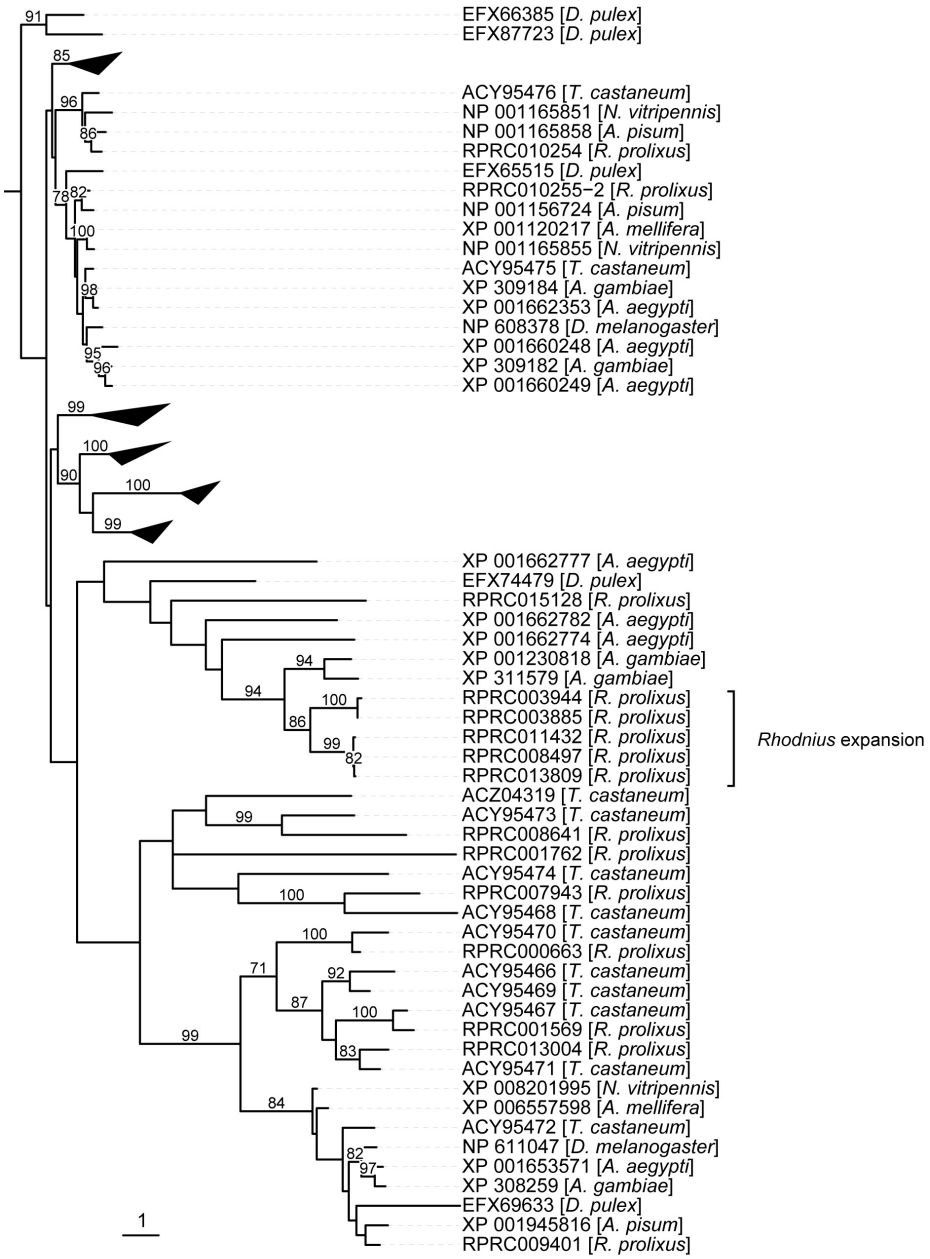

Figure A10: Phylogenetic tree of CPAP3 cuticular protein family: Monophyletic clades were collapsed and are represented as black triangles.

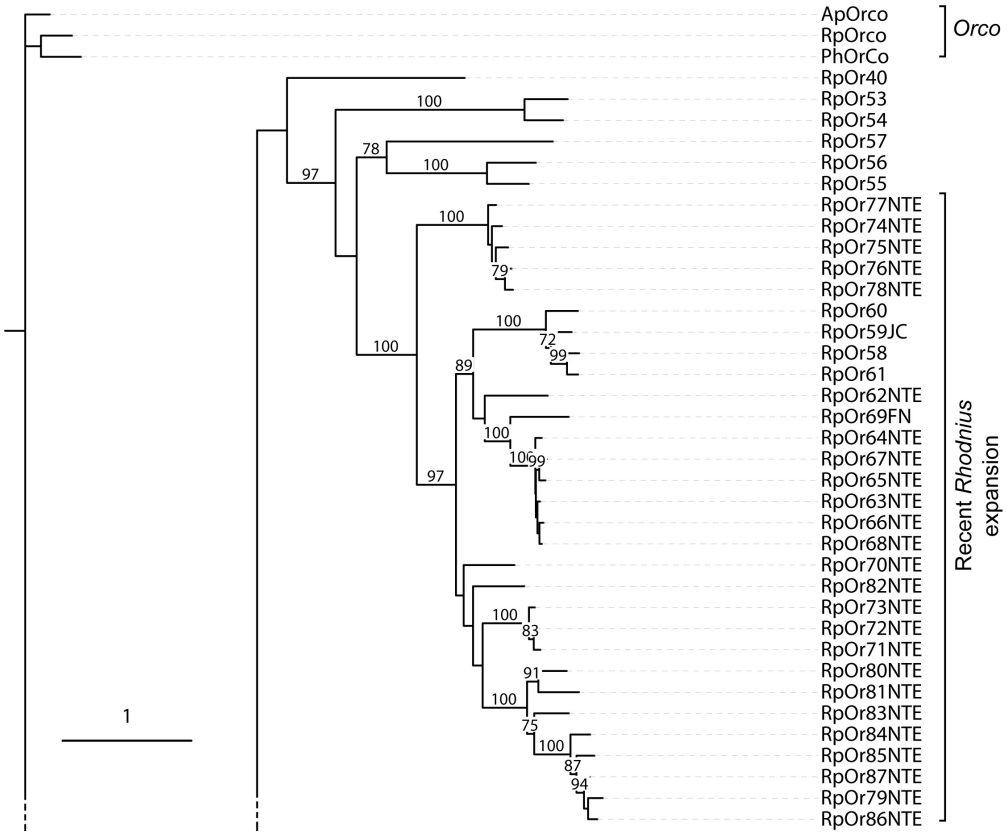

Figure A11: Phylogenetic tree of the ORs – Part 1/4. Comments on each gene lineage are on the right. Suffixes after the gene/protein names are: PSE – pseudogene; NTE – N-terminus missing; CTE – C-terminus missing; INT – internal sequence missing; FIX – sequence fixed with raw reads; JOI – gene model joined across scaffolds; multiple suffixes are abbreviated to single letters.

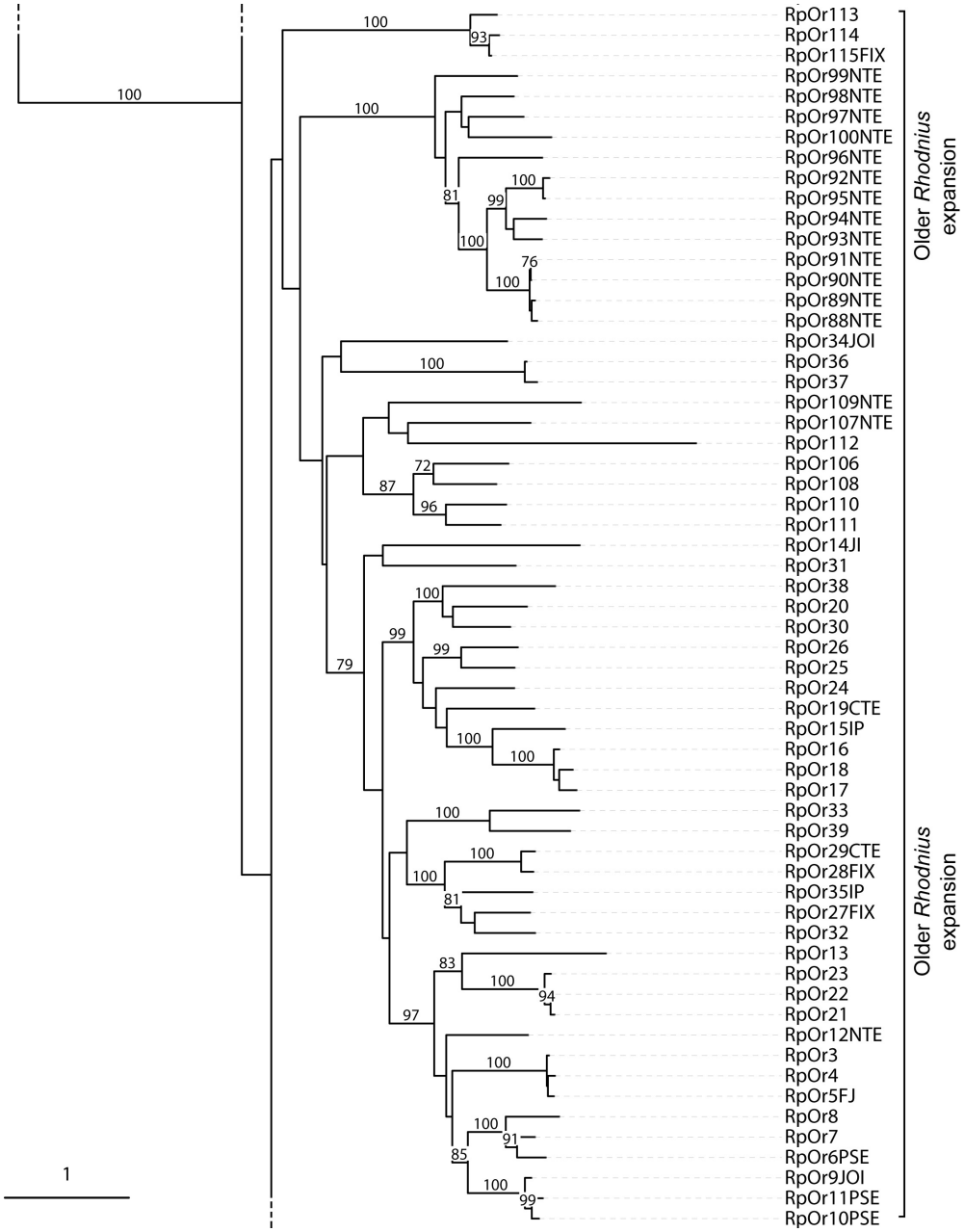

Figure A11: Phylogenetic tree of the ORs – Part 2/4: Comments on each gene lineage are on the right. Suffixes after the gene/protein names are: PSE – pseudogene; NTE – N-terminus missing; CTE – C-terminus missing; INT – internal sequence missing; FIX – sequence fixed with raw reads; JOI – gene model joined across scaffolds; multiple suffixes are abbreviated to single letters.

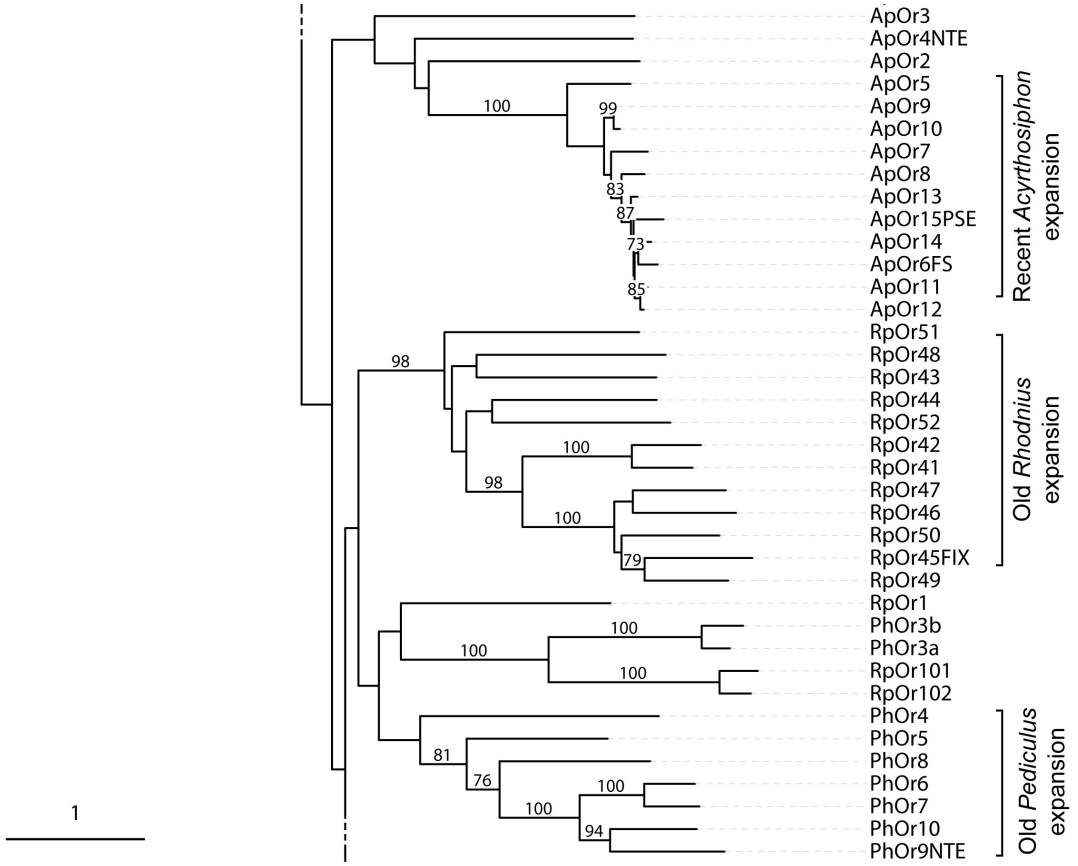

Figure A11: Phylogenetic tree of the ORs – Part 3/4: Comments on each gene lineage are on the right. Suffixes after the gene/protein names are: PSE – pseudogene; NTE – N-terminus missing; CTE – C-terminus missing; INT – internal sequence missing; FIX – sequence fixed with raw reads; JOI – gene model joined across scaffolds; multiple suffixes are abbreviated to single letters.

Figure A11: Phylogenetic tree of the ORs – Part 4/4: Comments on each gene lineage are on the right. Suffixes after the gene/protein names are: PSE – pseudogene; NTE – N-terminus missing; CTE – C-terminus missing; INT – internal sequence missing; FIX – sequence fixed with raw reads; JOI – gene model joined across scaffolds; multiple suffixes are abbreviated to single letters.

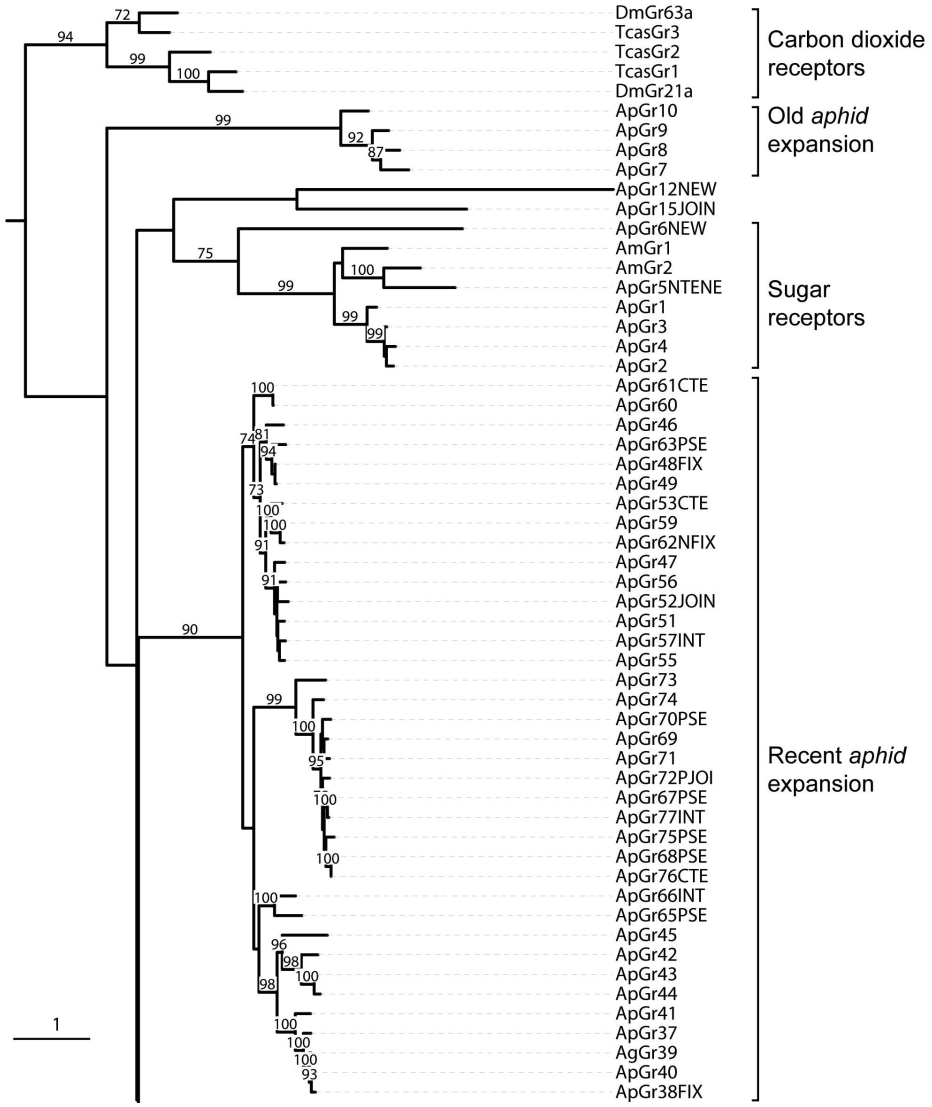

Figure A12: Phylogenetic tree of the GRs – Part 1/2: Organisms included: *Rhodnius*, *Acyrthosiphon*, *Pediculus*, *Anopheles*, *Apis*, *Bombyx*, *Nasonia*, *Tribolium*, and *Drosophila*. The tree was rooted with the carbon dioxide receptors. All other details as in Figure A11.

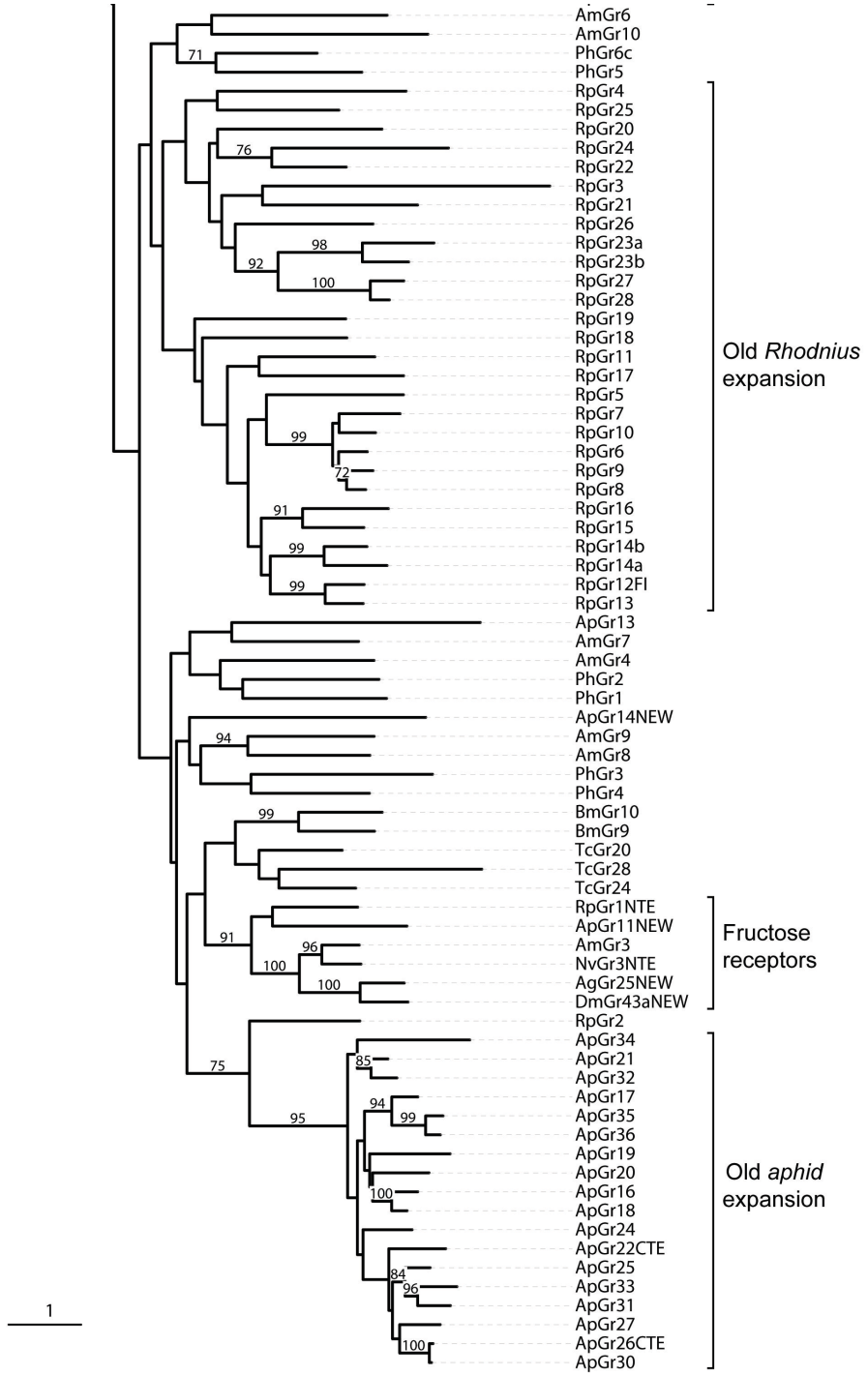

Figure A12: Phylogenetic tree of the GRs – Part 2/2: Organisms included: *Rhodnius, Acyrthosiphon, Pediculus, Anopheles*, *Apis, Bombyx*, *Nasonia*, *Tribolium*, and *Drosophila*. The tree was rooted with the carbon dioxide receptors. All other details as in Figure A11.

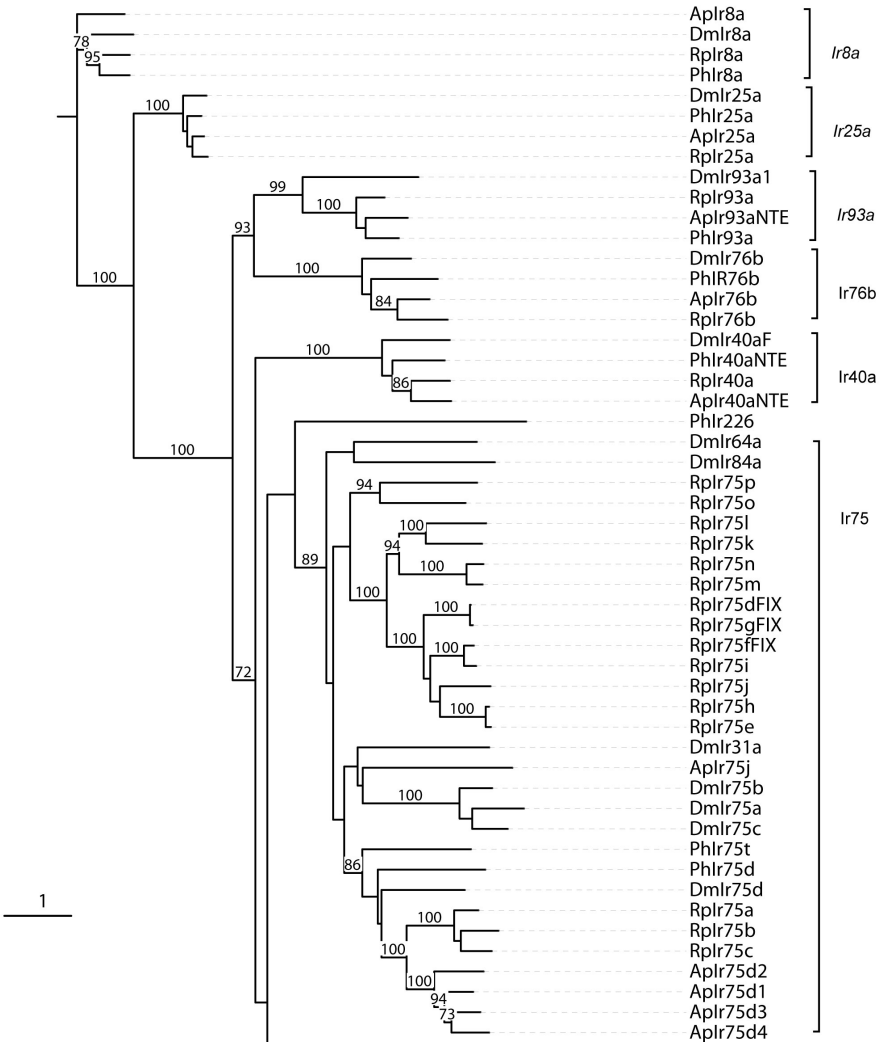

Figure A13: Phylogenetic tree of the IRs – Part 1/2 : Organisms included: *Rhodnius*, *Acyrthosiphon*, *Pediculus*, and *Drosophila*. The tree was rooted with the IR25a and 8a proteins. All other details as in Figure A11 .

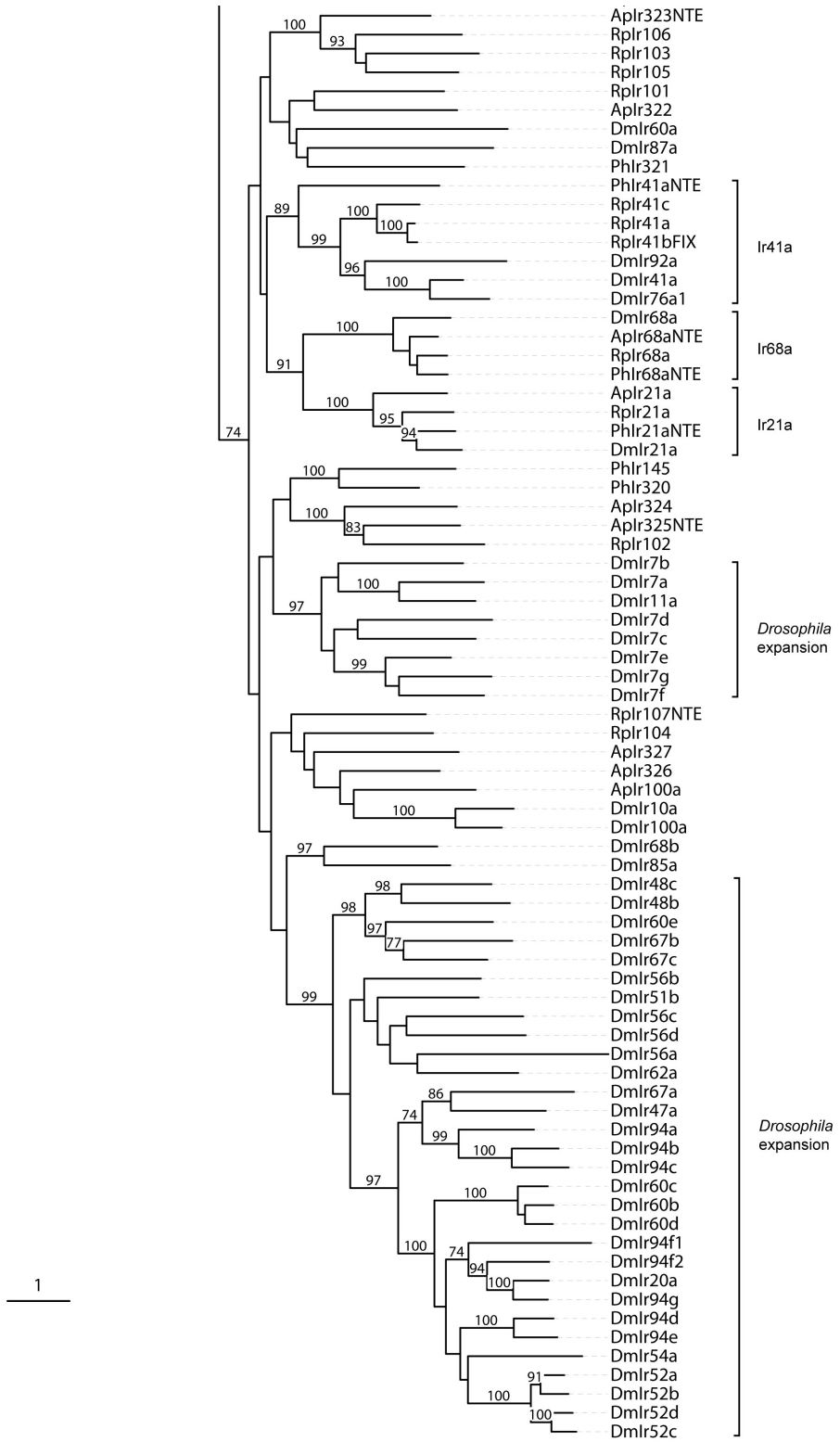

Figure A13: Phylogenetic tree of the IRs – Part 2/2 : Organisms included: *Rhodnius*, *Acyrthosiphon*, *Pediculus*, and *Drosophila*. The tree was rooted with the IR25a and 8a proteins. All other details as in Figure A11 .

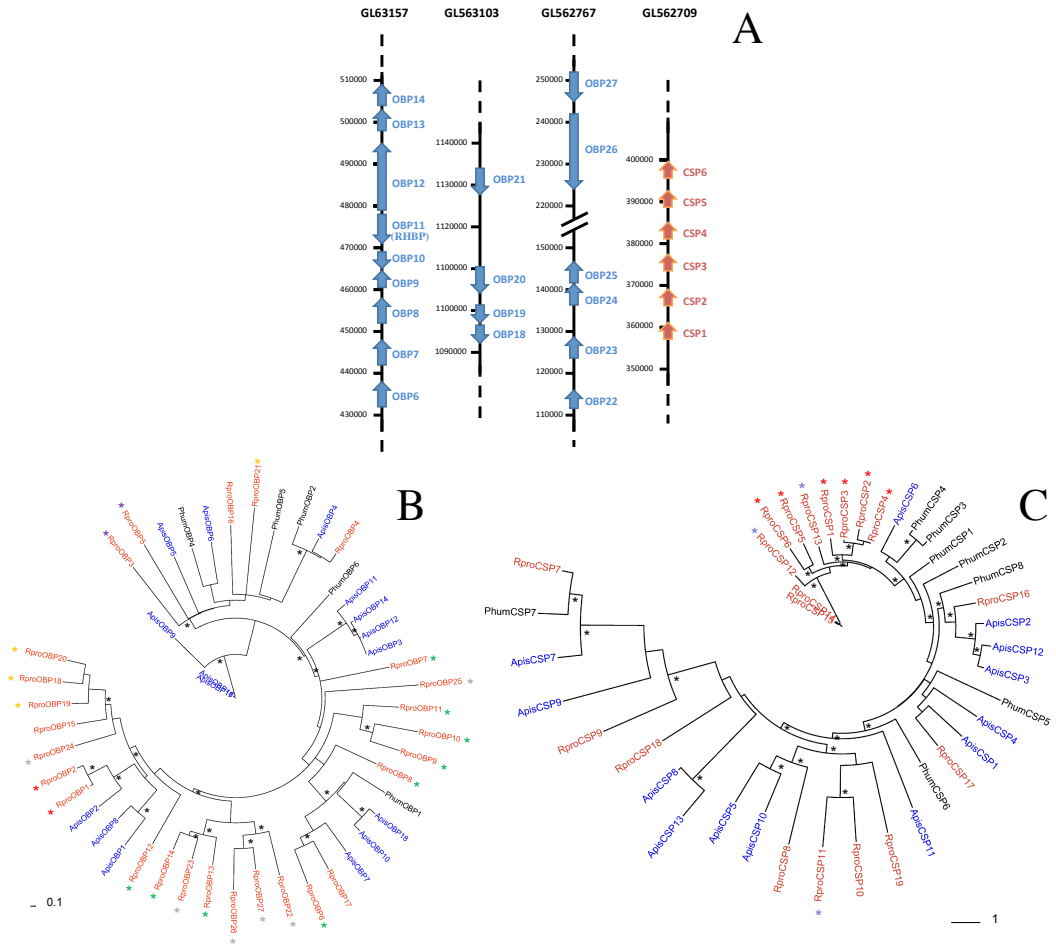

Figure A14: Clusters of OBPs and CSPs in *R. prolixus*: Tandem arrangementof OBPs and CSPs (**A**). Scaffolds ID are presented on the top of each chromosome. Blue and red arrows represent each gene and its position on the scaffold region. Chemosensory proteins tree (**B**) and Odorant-binding proteins tree (**C**) contain the organisms Rpro, *R. prolixus*; Apis, *Ac. pisum*; and Phum, *P. Humanus*. All included organisms belong to the infraclass Paraneoptera. Genes from *R. prolixus*, *Ac. pisum* and *P. humanus* are shown in red, blue and black, respectively. Black stars indicate bootstrap values >70%. Stars of the same colour indicate genes located on the same scaffold.

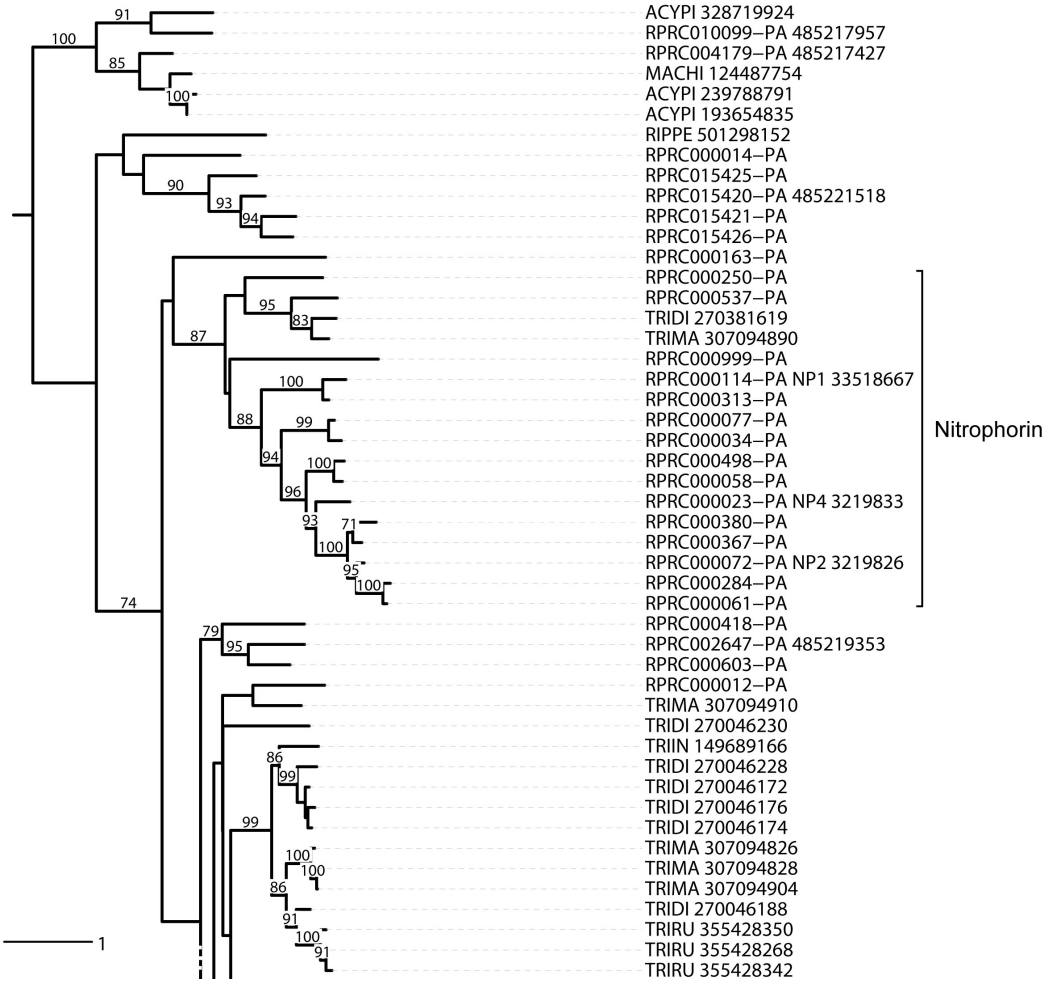

Figure A15: Lipocalins phylogenetic tree – Part 1/3 : *R. prolixus* lipocalins and related *Triatoma* proteins were included. Previously known nitrophorins NP1, NP2 and NP4 are identified by the RPRC number followed by their NCBI accession numbers. Sequences are represented by the first three letters of the genus name, followed by the first three letters of the species name, followed by their Genbank accession number. ACYPI: *Acyrthosiphon pisum;* MACHI: *Maconellicoccus hirsutus;* RIPPE: *Riptortus pedestris;* TRIDI: *Triatoma dimidiata;* TRIMA: *T. matogrossensis;* TRIRU: *T. rubida;* TRIIN: *T. Infestans;*

Figure A15: Lipocalins phylogenetic tree – Part 2/3: *R. prolixus* lipocalins and related *Triatoma* proteins were included. Previously known nitrophorins NP1, NP2 and NP4 are identified by the RPRC number followed by their NCBI accession numbers. Sequences are represented by the first three letters of the genus name, followed by the first three letters of the species name, followed by their Genbank accession number. ACYPI: *Acyrthosiphon pisum;* MACHI: *Maconellicoccus hirsutus;* RIPPE: *Riptortus pedestris;* TRIDI: *Triatoma dimidiata;* TRIMA: *T. matogrossensis;* TRIRU: *T. rubida;* TRIIN: *T. Infestans;* TRIBR: *T. brasiliensis;* TRIPA*: T. pallidipennis;* DIPMA: *Dipetalogaster maximus*.

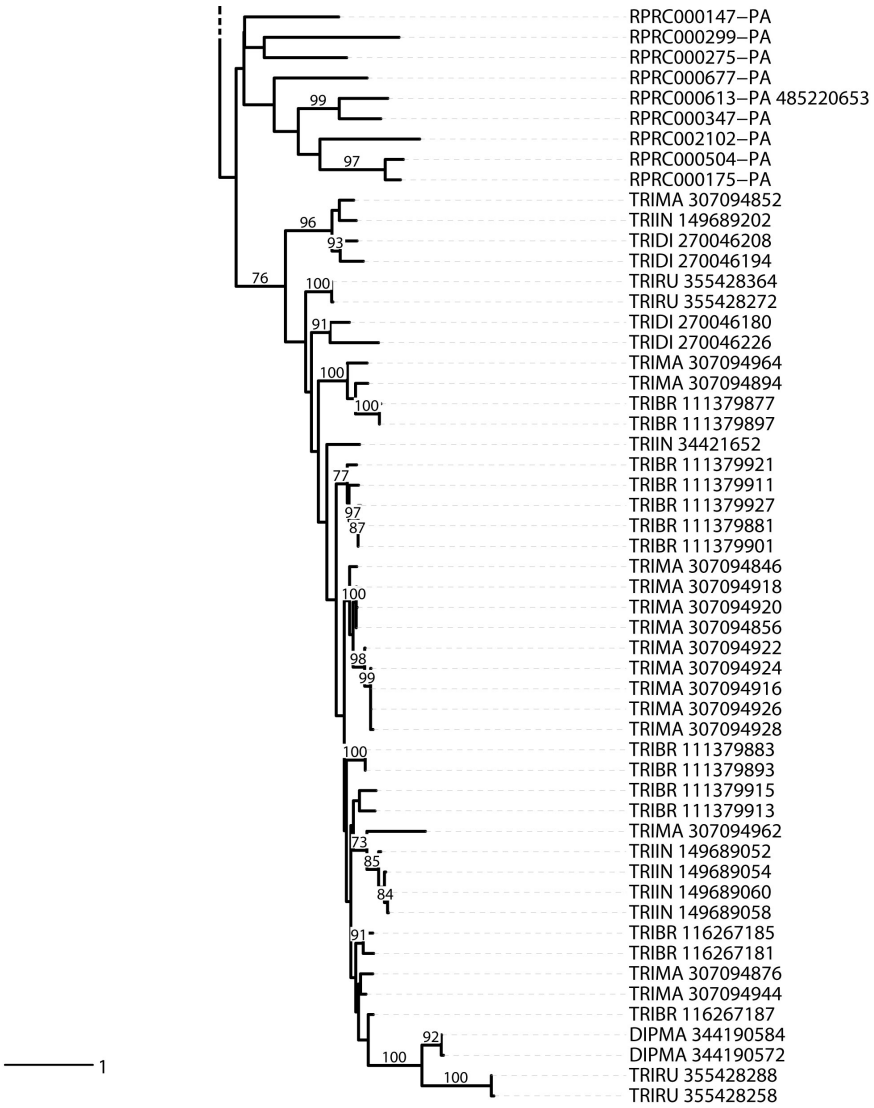

Figure A15: Lipocalins phylogenetic tree – Part 3/3: *R. prolixus* lipocalins and related *Triatoma* proteins were included. Previously known nitrophorins NP1, NP2 and NP4 are identified by the RPRC number followed by their NCBI accession numbers. Sequences are represented by the first three letters of the genus name, followed by the first three letters of the species name, followed by their Genbank accession number. ACYPI: *Acyrthosiphon pisum;* MACHI: *Maconellicoccus hirsutus;* RIPPE: *Riptortus pedestris;* TRIDI: *Triatoma dimidiata;* TRIMA: *T. matogrossensis;* TRIRU: *T. rubida;* TRIIN: *T. Infestans;* TRIBR: *T. brasiliensis;* TRIPA*: T. pallidipennis;* DIPMA: *Dipetalogaster maximus*.

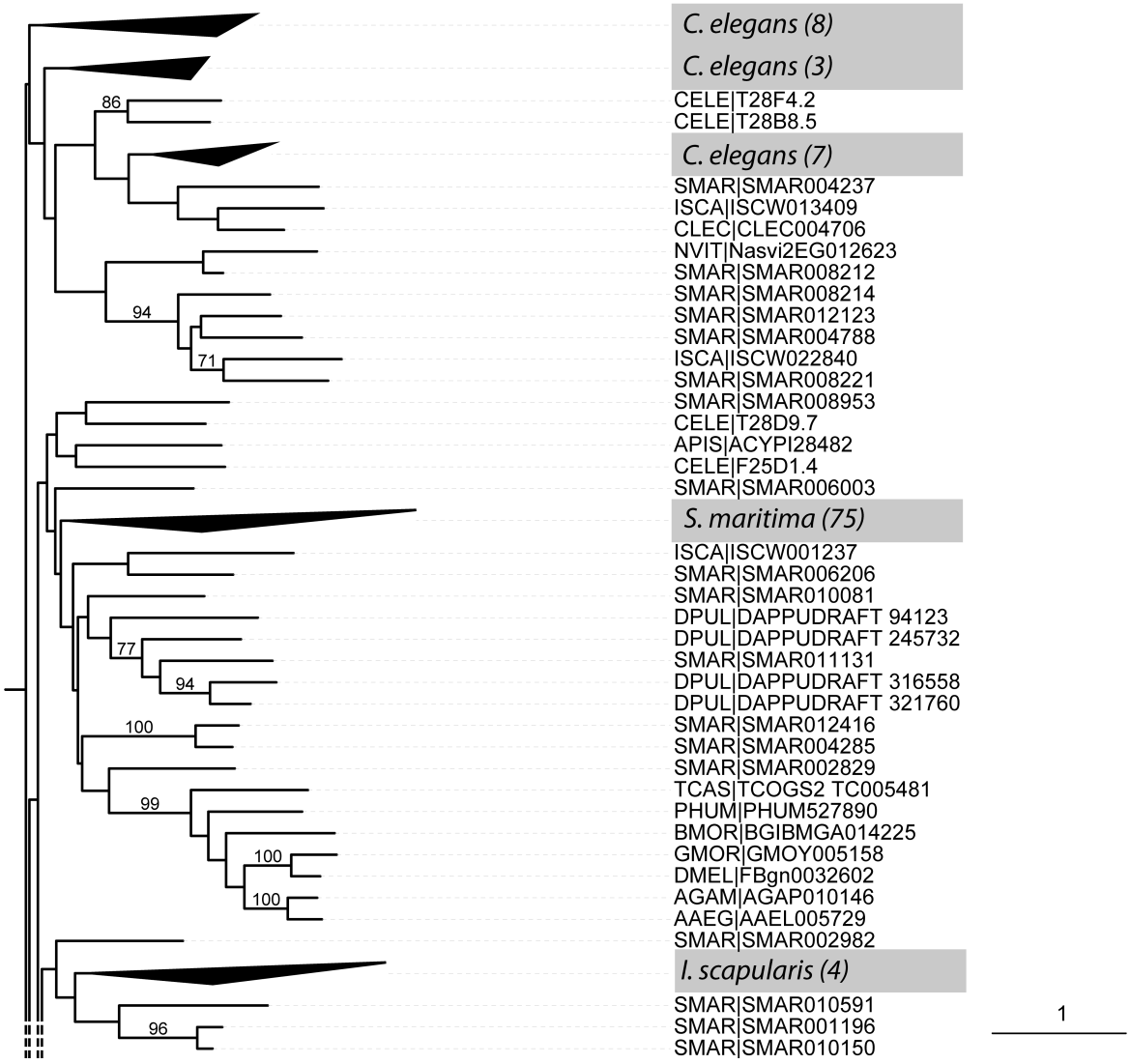

Figure A16: Amiloride-sensitive sodium channels tree – Part 1/4: All proteins containing the amiloride-sensitive sodium channel conserved domain (IPR001873) were included. Sequences are identified by the species code followed by their gene code. LSEs were collapsed and highlighted in grey and are represented by the species name followed by the number of sequences. Vertical bars show the monophyletic clades. Numbers inside the tree represent the bootstrap support. RPRO: *R. prolixus* (black dots)*;* APIS: *Acyrthosiphon pisum;* CLEC: *Cimex lectularius;* PHUM: *Pediculus humanus;* NVIT: *Nasonia vitripennis;* AMEL: *Apis mellifera;* BMOR: *Bombix mori;* TCAS: *Tribolium. castaneum;* AAEG: *Aedes aegypti;* AGAM: *Anopheles gambiae;* GMOR: *Glossina morsitans;* DMEL: *Drosophila melanogaster;* ISCA: *Ixodes scapularis;* DPUL: *Daphnia pulex;* SMAR: *Strigamina maritime;* CELE: *Caenorhabditis elegans*.

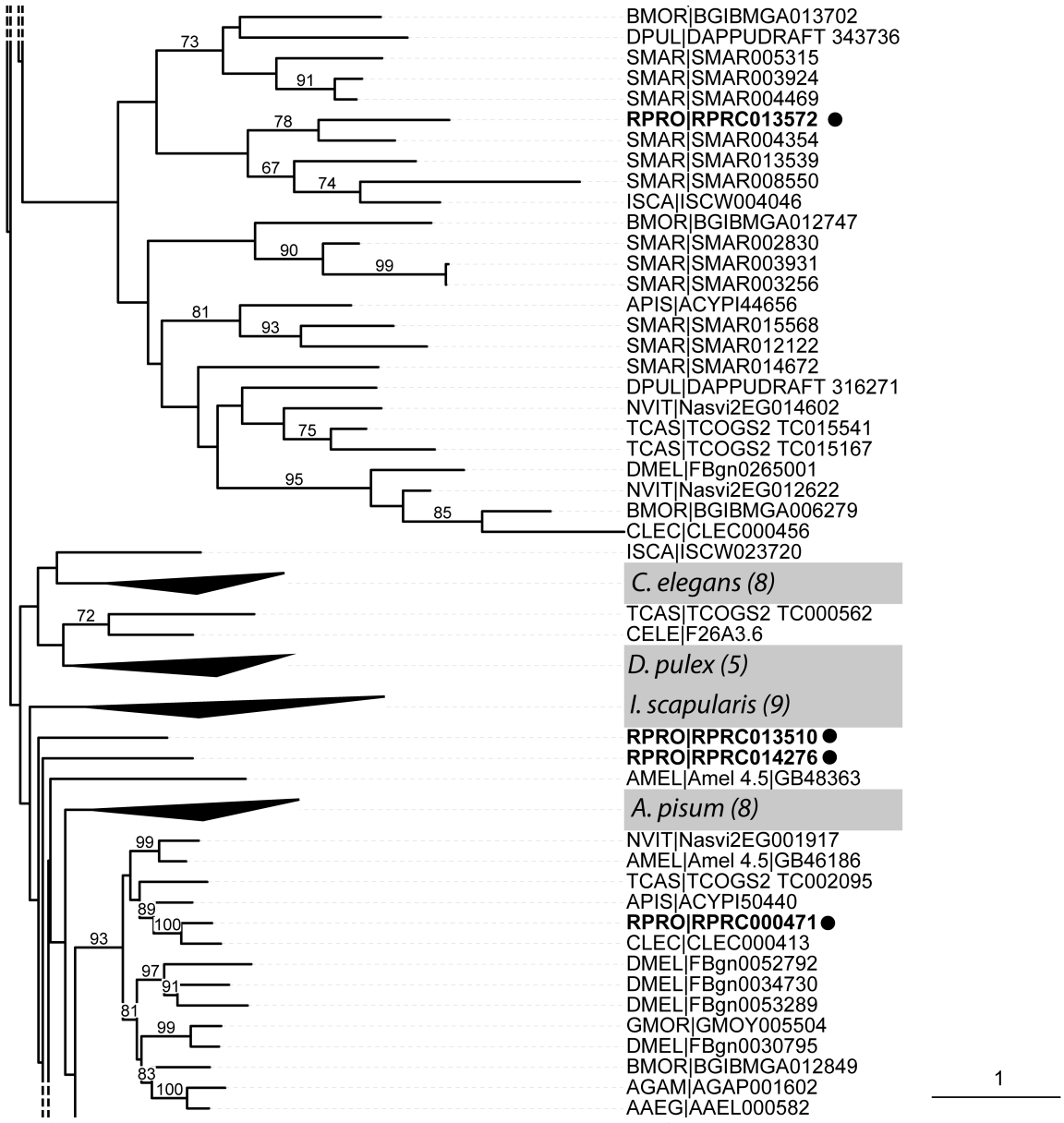

Figure A16: Amiloride-sensitive sodium channels tree – Part 2/4: All proteins containing the amiloride-sensitive sodium channel conserved domain (IPR001873) were included. Sequences are identified by the species code followed by their gene code. LSEs were collapsed and highlighted in grey and are represented by the species name followed by the number of sequences. Vertical bars show the monophyletic clades. Numbers inside the tree represent the bootstrap support. RPRO: *R. prolixus* (black dots)*;* APIS: *Acyrthosiphon pisum;* CLEC: *Cimex lectularius;* PHUM: *Pediculus humanus;* NVIT: *Nasonia vitripennis;* AMEL: *Apis mellifera;* BMOR: *Bombix mori;* TCAS: *Tribolium. castaneum;* AAEG: *Aedes aegypti;* AGAM: *Anopheles gambiae;* GMOR: *Glossina morsitans;* DMEL: *Drosophila melanogaster;* ISCA: *Ixodes scapularis;* DPUL: *Daphnia pulex;* SMAR: *Strigamina maritime;* CELE: *Caenorhabditis elegans*.

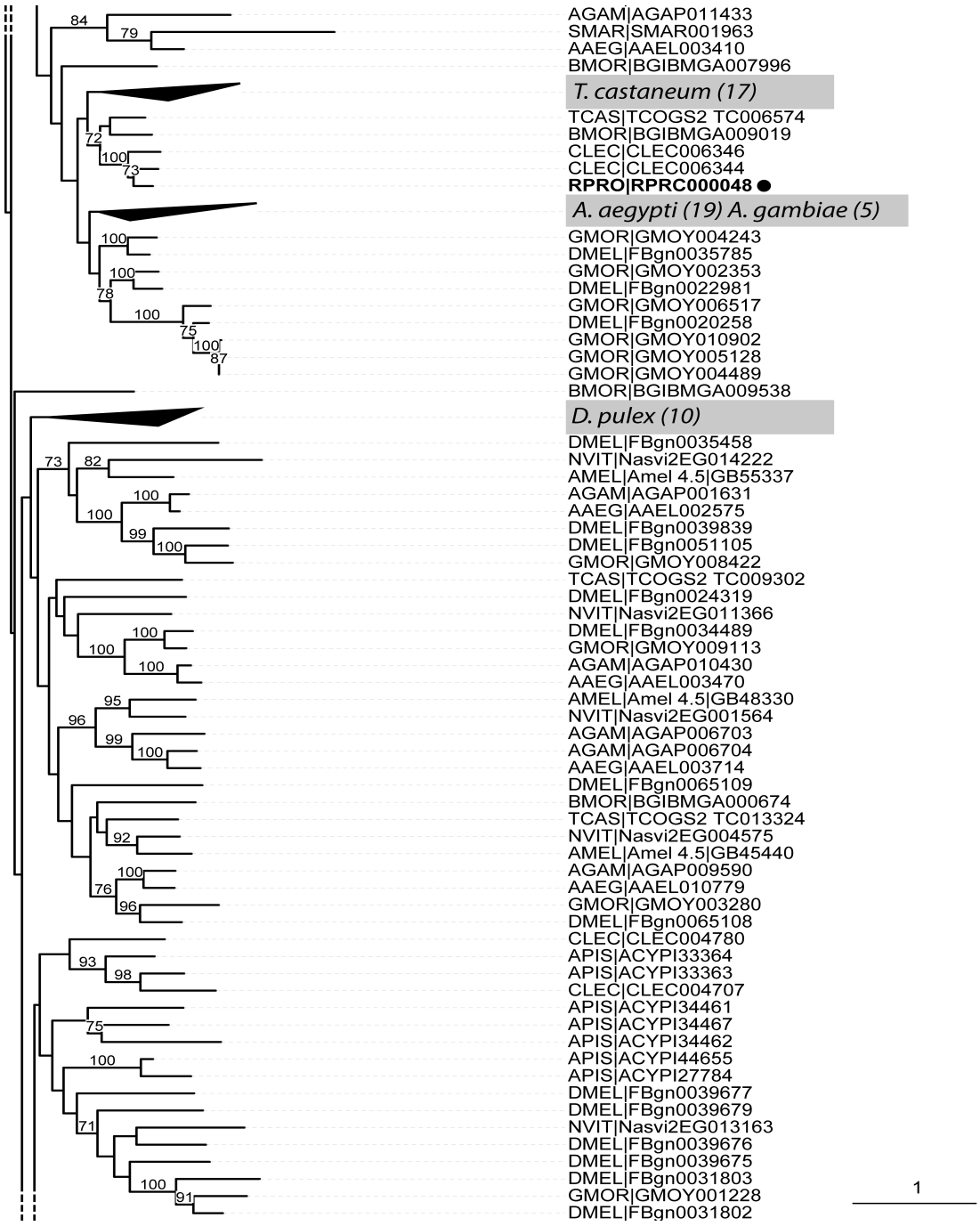

Figure A16: Amiloride-sensitive sodium channels tree – Part 3/4: All proteins containing the amiloride-sensitive sodium channel conserved domain (IPR001873) were included. Sequences are identified by the species code followed by their gene code. LSEs were collapsed and highlighted in grey and are represented by the species name followed by the number of sequences. Vertical bars show the monophyletic clades. Numbers inside the tree represent the bootstrap support. RPRO: *R. prolixus* (black dots)*;* APIS: *Acyrthosiphon pisum;* CLEC: *Cimex lectularius;* PHUM: *Pediculus humanus;* NVIT: *Nasonia vitripennis;* AMEL: *Apis mellifera;* BMOR: *Bombix mori;* TCAS: *Tribolium. castaneum;* AAEG: *Aedes aegypti;* AGAM: *Anopheles gambiae;* GMOR: *Glossina morsitans;* DMEL: *Drosophila melanogaster;* ISCA: *Ixodes scapularis;* DPUL: *Daphnia pulex;* SMAR: *Strigamina maritime;* CELE: *Caenorhabditis elegans*.

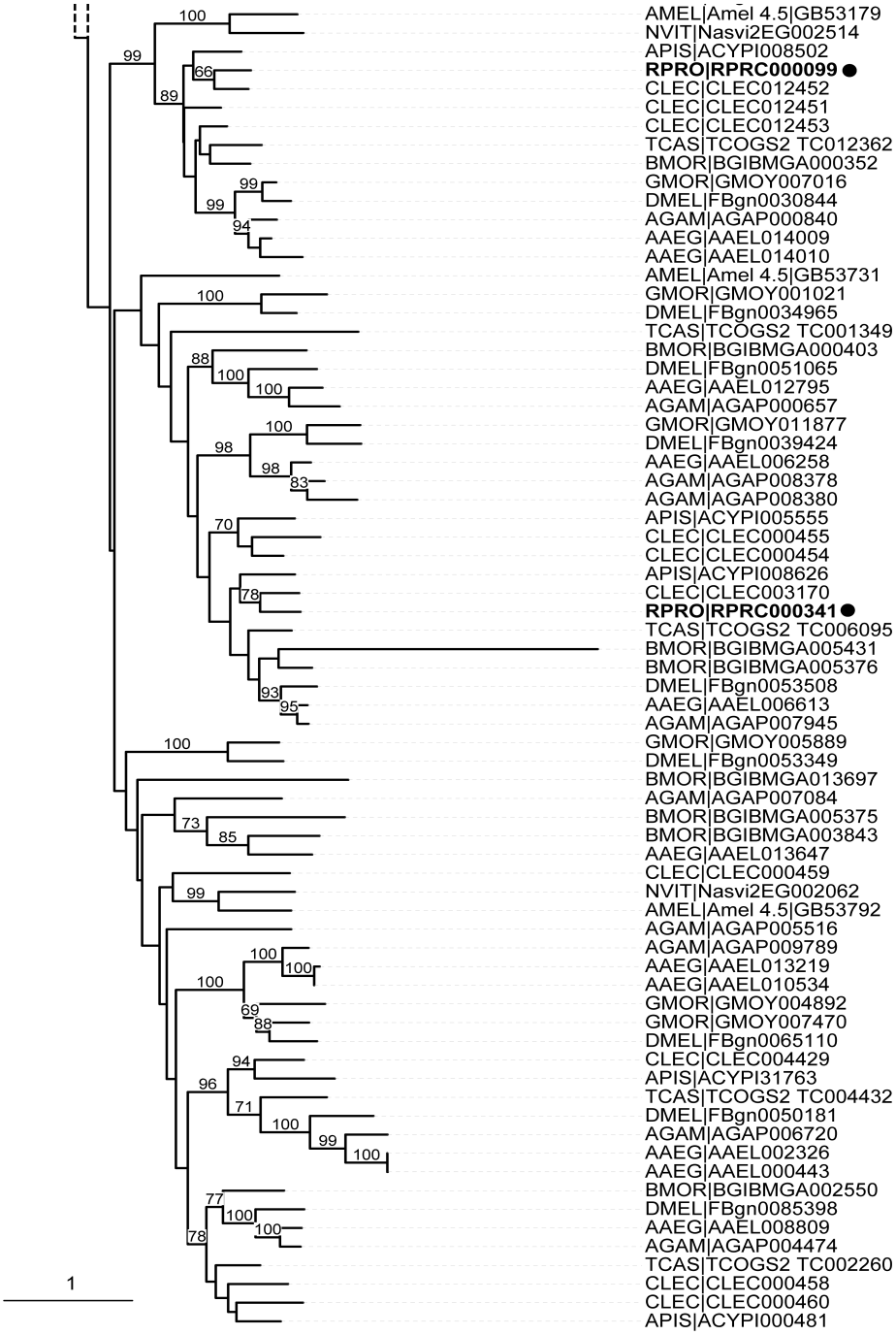

Figure A16: Amiloride-sensitive sodium channels tree – Part 4/4 : All proteins containing the amiloride-sensitive sodium channel conserved domain (IPR001873) were included. Sequences are identified by the species code followed by their gene code. LSEs were collapsed and highlighted in grey and are represented by the species name followed by the number of sequences. Vertical bars show the monophyletic clades. Numbers inside the tree represent the bootstrap support. RPRO: *R. prolixus* (black dots)*;* APIS: *Acyrthosiphon pisum;* CLEC: *Cimex lectularius;* PHUM: *Pediculus humanus;* NVIT: *Nasonia vitripennis;* AMEL: *Apis mellifera;* BMOR: *Bombix mori;* TCAS: *Tribolium. castaneum;* AAEG: *Aedes aegypti;* AGAM: *Anopheles gambiae;* GMOR: *Glossina morsitans;* DMEL: *Drosophila melanogaster;* ISCA: *Ixodes scapularis;* DPUL: *Daphnia pulex;* SMAR: *Strigamina maritime;* CELE: *Caenorhabditis elegans*.

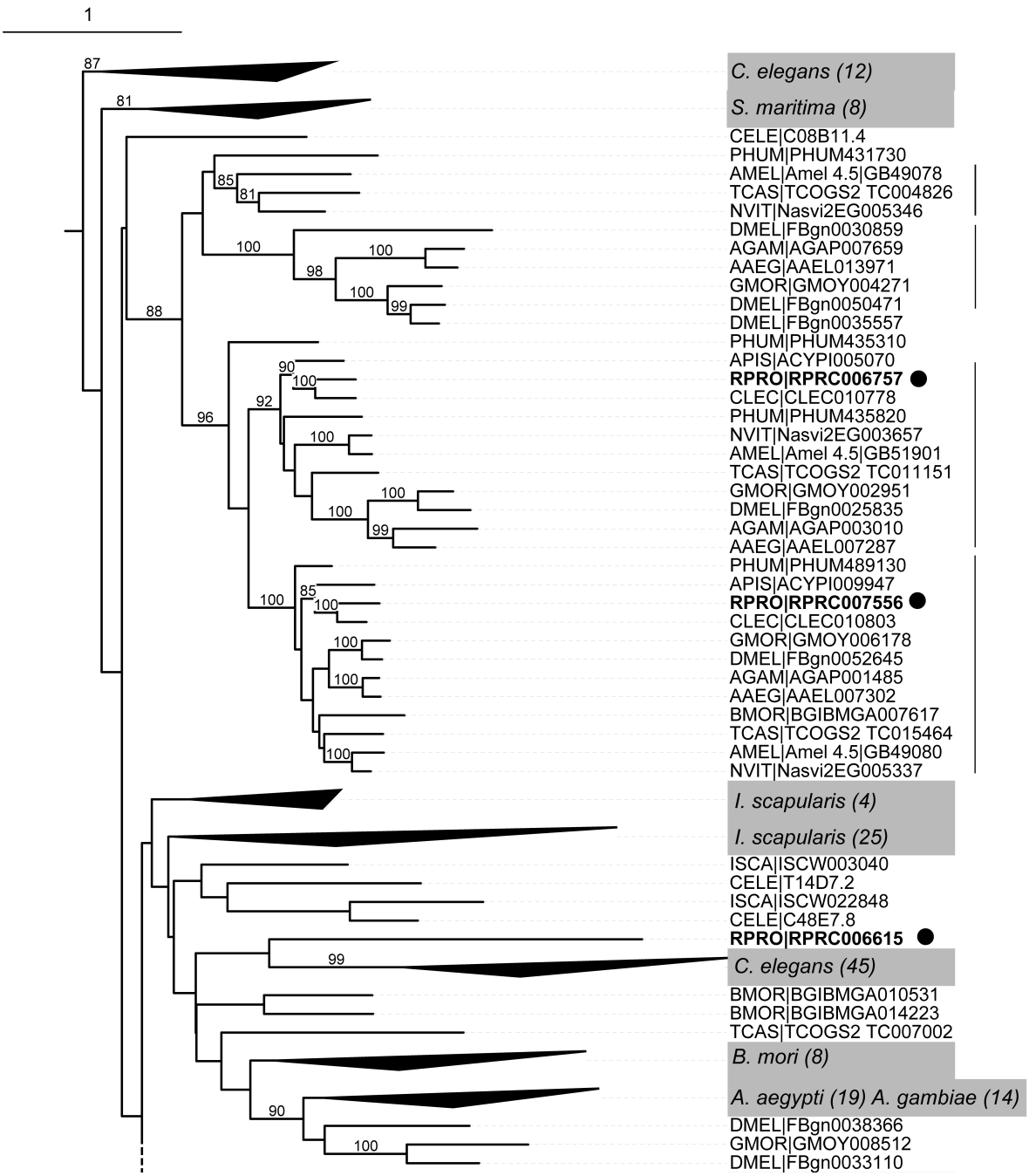

Figure A17: Acyltransferase 3 tree – Part 1/2 : All proteins containing the acyltransferase-3 domain (IPR002656) were included. Sequences are identified by the species code followed by their gene code. LSEs were collapsed and highlighted in grey and are represented by the species name followed by the number of sequences. Vertical bars show the monophyletic clades. Numbers inside the tree represent the bootstrap support. RPRO: *R. prolixus* (black dots)*;* APIS: *Acyrthosiphon pisum;* CLEC: *Cimex lectularius;* PHUM: *Pediculus humanus;* NVIT: *Nasonia vitripennis;* AMEL: *Apis mellifera;* BMOR: *Bombix mori;* TCAS: *Tribolium. castaneum;* AAEG: *Aedes aegypti;* AGAM: *Anopheles gambiae;* GMOR: *Glossina morsitans;* DMEL: *Drosophila melanogaster;* ISCA: *Ixodes scapularis;* DPUL: *Daphnia pulex;* SMAR: *Strigamina maritime;* CELE: *Caenorhabditis elegans*.

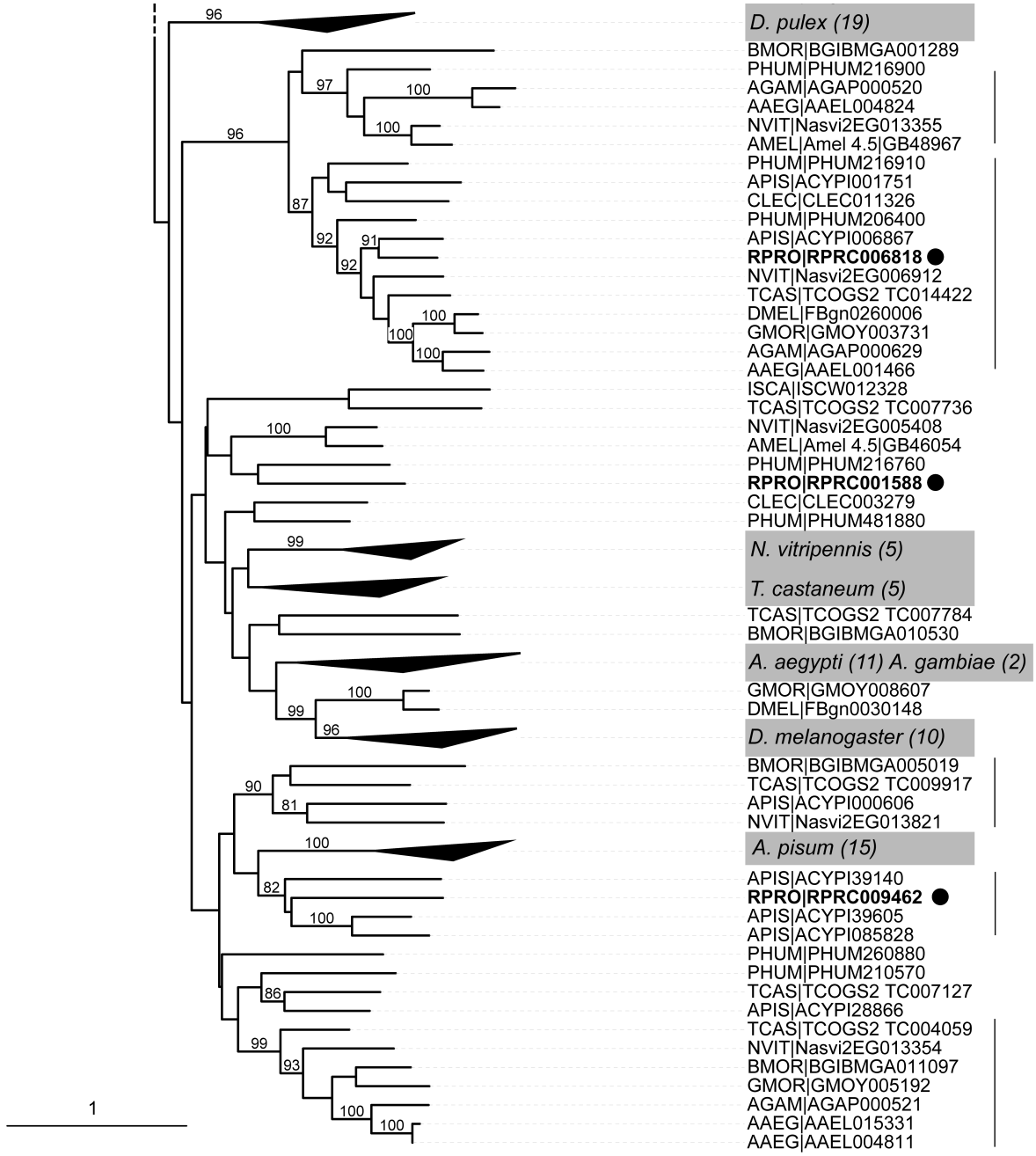

Figure A17: Acyltransferase 3 tree – Part 2/2 : All proteins containing the acyltransferase-3 domain (IPR002656) were included. Sequences are identified by the species code followed by their gene code. LSEs were collapsed and highlighted in grey and are represented by the species name followed by the number of sequences. Vertical bars show the monophyletic clades. Numbers inside the tree represent the bootstrap support. RPRO: *R. prolixus* (black dots)*;* APIS: *Acyrthosiphon pisum;* CLEC: *Cimex lectularius;* PHUM: *Pediculus humanus;* NVIT: *Nasonia vitripennis;* AMEL: *Apis mellifera;* BMOR: *Bombix mori;* TCAS: *Tribolium. castaneum;* AAEG: *Aedes aegypti;* AGAM: *Anopheles gambiae;* GMOR: *Glossina morsitans;* DMEL: *Drosophila melanogaster;* ISCA: *Ixodes scapularis;* DPUL: *Daphnia pulex;* SMAR: *Strigamina maritime;* CELE: *Caenorhabditis elegans*.

Table A1: Number of annotated sequences belonging to different TE superfamilies : The TE copy number and the total of genomic bases occupied by TEs were calculated considering sequences larger than 500 bp and 1000 bp. Total bases: Sum of the length of the elements; #: Copy Number; %G: Percentage of the genome occupied by elements.

Element Class	Prototype sequences	Total Bases > 500 bp	#	%G	TE %	Total Bases > 1000 bp	#	%G	% TE
Class I									
LTR									
Pao	7	4.419.767	5.230	0,6290	11,1971	113.014	47	0,0161	0,7085
Gypsy	16	297.673	206	0,0424	0,7541	213.867	80	0,0304	1,3407
Copia	6	86.827	53	0,0124	0,2200	69.695	28	0,0099	0,4369
NLTR									
Jockey	70	4.303.525	6.862	0,6125	10,9026	632.965	458	0,0901	3,9680
Rose	154	1.098.992	1.081	0,1564	2,7842	626.665	391	0,0892	3,9285
LoA	6	1.092.791	1.436	0,1555	2,7685	283.281	201	0,0403	1,7759
Felyant	139	985.493	931	0,1403	2,4967	595.650	355	0,0848	3,7341
Totopi	28	827.181	980	0,1177	2,0956	323.917	219	0,0461	2,0306
CR1-like	15	731.870	921	0,1042	1,8541	208.981	156	0,0297	1,3101
RTE	36	379.346	444	0,0540	0,9610	149.816	99	0,0213	0,9392
I	17	165.491	154	0,0236	0,4193	99.123	55	0,0141	0,6214
R1	14	140.130	141	0,0199	0,3550	75.061	47	0,0107	0,4705
Ingi	4	33.361	30	0,0047	0,0845	19.341	10	0,0028	0,1212
CLASS II									
Mariner	696	18.705.545	19.403	2,6622	47,3888	11.600.387	9.098	1,6510	72,7216
Tc-1	45	1.144.473	1.311	0,1629	2,8994	483.743	372	0,0688	3,0325
hAT	27	987.866	1.271	0,1406	2,5027	220.561	158	0,0314	1,3827
Maverick	66	140.077	158	0,0199	0,3549	66.545	48	0,0095	0,4172
Helitron	11	118.395	121	0,0168	0,2999	61.336	38	0,0087	0,3845
Pogo	7	99.651	108	0,0142	0,2525	56.543	36	0,0080	0,3545
PiggyBac	22	65.979	53	0,0094	0,1672	47.241	25	0,0067	0,2961
P	1	2.085	1	0,0003	0,0053	2.085	1	0,0003	0,0131
Sola	1	1.965	1	0,0003	0,0050	1.965	1	0,0003	0,0123
MITEs	12	3.644.040	12.000	0,5186	9,2318	nd	nd	nd	nd
TOTAL	**1.400**	**39.472.523**	**52.896**	**5,6177**		**15.951.782**	**11.923**	**2,2702**	

Table A2: Copy number and percentage of the genome occupy by TEs with respect to length.: Total bases: Sum of the length of the elements; #: Copy Number; %G: Percentage of the genome occupied by elements.

Clade	Total Bases > 500 bp	#	%G	Total Bases >1000 bp	#	%G	Total Bases >2000 bp	#	%G	Total Bases >3000 bp	#	%G	Total Bases >4000 bp	#	%G	Total Bases >5000 bp	#	%G
LTR																		
Pao	4.419.767	5230	0,629	113.014	47	0,016	77569	21	0,011	55.696	12	0,008	42.461	8	0,006	20.076	3	0,003
Gypsy	297.673	206	0,042	213.867	80	0,030	159185	41	0,023	132.287	30	0,019	83.405	16	0,012	34.341	5	0,005
Copia	86.827	53	0,012	69.695	28	0,010	58.093	19	0,008	26.405	7	0,004	9.027	2	0,001	0	0	0,000
NLTR																		
Jockey	4.303.525	6862	0,612	632.965	458	0,090	106098	42	0,015	22.667	7	0,003	0	0	0,000	0	0	0,000
Rose	1.098.992	1081	0,156	626.665	391	0,089	199086	78	0,028	36.092	11	0,005	0	0	0,000	0	0	0,000
LoA	1.092.791	1436	0,156	283.281	201	0,040	38374	17	0,005	3.687	1	0,001	0	0	0,000	0	0	0,000
Felyant	985.493	931	0,140	595.650	355	0,085	219596	87	0,031	41.276	12	0,006	8.324	2	0,001	0	0	0,000
Totopi	827.181	980	0,118	323.917	219	0,046	85992	34	0,012	16.542	5	0,002	0	0	0,000	0	0	0,000
CR1-like	731.870	921	0,104	208.981	156	0,030	16007	7	0,002	0	0	0,000	0	0	0,000	0	0	0,000
RTE	379.346	444	0,054	149.816	99	0,021	48172	20	0,007	6.892	2	0,001	0	0	0,000	0	0	0,000
I	165.491	154	0,024	99.123	55	0,014	53050	19	0,008	24.406	7	0,003	0	0	0,000	0	0	0,000
R1	140.130	141	0,020	75.061	47	0,011	24386	10	0,003	3.003	1	0,000	0	0	0,000	0	0	0,000
Ingi	33.361	30	0,005	19.341	10	0,003	10754	4	0,002	3.702	1	0,001	0	0	0,000	0	0	0,000

Supplementary Material and Methods:

The genome of *Rhodnius prolixus*, a vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection

Colony selection

Since *Rhodnius prolixus* is morphologically indistinguishable from *Rhodnius robustus* (1), we relied on DNA sequence analyses to select a legitimate *R. prolixus* colony with accurate taxonomic identification (1-3). To do this, we genotyped twenty specimens from each of fifteen candidate colonies using two markers of proven efficacy in the discrimination of the species that comprise the *R. prolixus*/*R. robustus* complex: 1) a mitochondrial cytochrome b gene fragment (*mtCytb*) and 2) the nuclear ribosomal second internal transcribed spacer (*ITS-2*) (4). We determined species identity using a phylogenetic comparison of DNA sequences of the candidate specimens with the following reference samples: *R. prolixus* from Barinas, Venezuela (N=2); *R. robustus* I from Trujillo, Venezuela (N=2); *R. robustus* II from Porto Velho, Rondônia, Brazil (N=4); *R. robustus* III from Novo Repartimento, Pará, Brazil (N=4); and *R. robustus* IV from Cayenne, French Guyana (N=6). Of the fifteen presumed *R. prolixus* colonies, only four were indeed *R. prolixus*, since one colony presented both mitochondrial and nuclear DNA sequences of *R. robustus* II, while the other ten colonies showed evidence of introgression with *R. robustus* IV mitochondrial DNA. From the four true *R. prolixus* colonies, we chose specimens from the Atlanta/CDC and the Medical Entomology Research Training Unit, Guatemala (MERTU) colony for further sequencing analyses. The *R. prolixus* colony from UFRJ/IBqM was used in the functional experiments described on this manuscript because it was one of the largest and more stable colonies available, in spite of having some degree of *R. robustus* mitochondrial introgression. Most of the transcriptomic information available now for *Rhodnius prolixus* was generated from this specific colony - See paper from Ribeiro*, et al.* (5) and comparisons with CDC confirmed genetic identity of this colony as true *Rhodnius prolixus*.

Genome Survey Sequencing (GSS), which provides a global view of the genome, was used to estimate the quantity and variability of repetitive regions of the genome. Since high heterozygosity levels complicate genome assembly, we applied GSS to obtain estimates of repetitive sequences and genetic variability values and also for assessing genome coverage, library construction quality, and scaffold formation (6). GSS was also used to select a non-introgressed *R. prolixus* colony for sequencing. To

perform GSS, we extracted DNA from ovaries or testis, as previously described (2). From a WGS plasmid library, 12,626 random sequences were generated on an ABI 3730 automated sequencer, as described previously (7). GSS reads were screened to extract sequences with unmasked regions of 400-500bp. These were then assessed with a custom primer-calling pipeline to design 96 primer pairs, as previously described (2). We surveyed genome sequences measuring 7-22kb in length and revealed that *R. prolixus* colonies from Atlanta/CDC and MERTU displayed heterozygosity levels an order of magnitude lower than wild *R. prolixus* from Venezuela [1 in 2.0kb (0.05%), and 1 in 3.0kb (0.03%), respectively, in comparison to 1 in 0.3kb (0.5%)]. Since Atlanta/CDC and MERTU colonies gave comparable results with respect to heterozygosity levels, the former was chosen because of its stability.

Genome sequencing and assembly

To reduce the amount of symbiotic bacterial contamination, testes and ovaries were dissected from virgin adults and extracted using the Gentra high pure system. For BAC library preparation and construction, two grams of freshly hatched N1s were frozen and submitted to AmpliconExpress (Pullman, Washington, USA). All sequencing occurred at the Genome Institute, Washington University School of Medicine. Using an estimated genome size of 679Mb, we targeted coverage to a depth of 15X. All reads were assembled using the CABOG v6.0 assembler. This version of the genome (v3.0.1) assembled to 702,645,054bp including gap lengths, and had a final output coverage of 8.1X. Approximately 46% of reads were not utilized in the final genome assembly, presumably due to the presence of high AT-rich regions across an estimated 66% of the genome. The 8,752,534 input sequences consisted of fragments and 3kb inserts generated with 454 (Roche), as well as 5kb plasmids and BAC end sequences from an Amplicon Express BAC library that were generated with Sanger technology on a 3730 instrument (ABI). We assembled 58,559 contigs into 27,872 scaffolds, and the respective contig and scaffold N50 lengths were 27,483bp and 847,873bp (Figure A1 in Appendix). We excluded from the final genome assembly any contig <200bp, as well as contigs identified as *E. coli* or cloning vectors (N=290 contigs). The *R. prolixus* whole genome shotgun (WGS) project (Accession ACPB00000000.2) consists of sequences ACPB02000001-ACPB02058559. Bulk downloads of the sequence and annotation data are available via GenBank, Ensembl, DDBJ, and the UCSC Genome Browser. The complete set of sequence reads is available at the NCBI Trace archive. http://www.ncbi.nlm.nih.gov/sra?LinkName=nuccore_sra_wgs&from_uid=313569759

Duplicated regions

We identified duplicated regions through a similarity comparison of all scaffolds against each other using BLASTN (8). We parsed the tabular result using an internally developed PERL script, which selected regions >5kb at different identity levels (90%, 95%, and 99%) and then generated input files that were formatted for visualization with histograms. We selected a minimum length cutoff to avoid transposable elements detection in this analysis, and we identified a small number of regions with little variation (934-596) among the range of identity space (934 regions with 90% identity and 596 with 99%). The largest region identified by this analysis measured 41kb, and the majority of variation occurred in regions measuring between 5kb and 10kb, possibly due to the longest transposable elements or areas that are rich in these structures (Figure A1 in Appendix).

Y-chromosome scaffolding and gene identification

R. prolixus has a XY sex-chromosome system, however, as with other repeat-rich regions, current methods for sequencing, assembly, and gene identification along the Y-chromosome are problematic (9-11). Recent research on *Drosophila* and mosquitoes suggests that separate sequencing of male and female DNA provides an appropriate approach for identifying Y chromosome sequences (12-14). Accordingly, we employed sex-specific DNA libraries and designated as candidates for Y-linkage any scaffold that assembled exclusively from male traces. Computational analysis of the reference assembly indicated that 766 scaffolds were Y-linked; a PCR test with ten randomly chosen scaffolds confirmed Y-linkage for all of them. BLAST analyses against different databases suggested that 568 scaffolds contain genes or part of putative genes. We selected the 36 most promising single-copy genes (those with hits in the transcriptome) to test for Y-linkage, and found nine new Y-linked genes in the *R. prolixus* genome. To solve gaps, we re-sequenced transcripts and corrected any sequencing errors in these Y-linked genes. We were able to identify putative functions for four Y-linked genes: a Metaloprotease (*Met*-Y), an Aconitase (*Aco*-Y) and two Zinc Finger proteins (*Zfn-Y1* and *Zfn*-Y2). The remaining five genes have no similarity to proteins with known functions or motifs, and were named as (*Rpr-Y1*, *Rpr-Y2*, *Rpr-Y3*, *Rpr-Y4* and *Rpr-Y5*). As with most Y-linked genes in other animals (12-16) most Y-linked have a testis specific expression (only *Aco-Y* is expressed specifically in the gut). Our discovery and description of Y-linked chromosome sequences and genes suggests that this approach (14) is more efficient and robust than traditional genome projects in which male and female DNA are mixed before sequencing. The Y-linked genes described here provide unique markers for population and phylogenetic analyses using R. prolixus. Further analyses of these Y-linked genes and their

functions will likely expand our knowledge of sexual dimorphism, fertility, and the origins of sex chromosomes in *R. prolixus* and across other insects

Gene prediction

Protein-coding gene prediction was undertaken using both *ab initio* and similarity-based methods with subsequent aggregation using the MAKER software (17). The annotation process can be broken down into the following phases: Identification of *de novo* repeat sequences using RepeatScout (18) and RECON (19). These were supplemented with publicly available repeat sequences from GenBank and mapped to the genome assembly using RepeatMasker (20). Repetitive regions where excluded from further analyses with regard to the prediction of protein-coding loci. Repeat regions were checked for protein similarities to avoid over-prediction, which would potentially mask valid protein-coding genes in other, downstream steps of the annotation process. We used the first instance transcript consensus sequences for training *ab initio* gene prediction programs SNAP (21) and Augustus (22). Subsequent rounds of re-training were based on the output of the prediction programs themselves. Similarity-based gene predictions were generated using exonerate (23), which incorporated alignments of EST sequences and protein-based predictions using taxonomically constrained subsets of the non-redundant (nr) protein database and UniProt (24). Gene predictions from both the *ab initio* and similarity approaches were aggregated into a final set using a layering approach within the Ensembl gene prediction system (25). The first two rounds were designed to iteratively improve the training of the *ab initio* gene predictions, and the final round incorporated protein similarities to all metazoan sequences in the nr protein database to guide the final predictions. The candidate gene sets were assessed for completeness with filtering based on comparative analyses and finally screened for potential transposable elements. The resulting data sets formed the basis for community-led prediction appraisal and improvement.

Alternative prediction for protein-coding genes used GeneID v1.3 (26, 27). GeneID was trained and evaluated with real genes from *R. prolixus* prior to whole genome gene prediction. First, full-length and 3'-truncated transcripts (FLT and 3T) from the *R. prolixus* transcriptome (5) were mapped to the genome using a custom PERL script named Sim4_wizzard. This script utilized Megablast (28) and chose the best-hit genomic scaffold to each transcript and mapped it with SIM4 (29). The mapped transcripts were filtered to select the ones having an exon maximum error rate of 1% in all exons and a complete donor/acceptor splicing sites (GT/AG or GC/AG) in all intron boundaries. After these filters,

the transcripts were labeled as *complete* (single-location complete mapping), *3'-truncated* (single-location but without stop codon mapping), *repetitive* (multiple-location mapping), and *fragmented* (all others). The training step used GeneID scripts to extract start codon (from *complete* and *3'-truncated* mapping) and splicing site regions (only from *complete* mapping), named signals, and calculated a Hidden-Markov Model (HMM) comparing the real signals with background (unspecific) signals. The coding/noncoding HMM was created using all transcripts as the coding region and all introns from *complete* transcripts as the non-coding region. Additional parameters were estimated based on mapped transcripts, as minimum and maximum intron length and minimum inter-genic distance. To evaluate the training, two hundred completely mapped transcripts were saved only to training. Their genomic region was extracted, along with 500 extra nucleotides on each side, and these were concatenated into an artificial chromosome using a custom Perl script. This set was used to evaluate the prediction performance of the calculated *R. prolixus* parameters regarding sensitivity and specificity at the nucleotide level, sensitivity and specificity at the exon level, the ratio of missing and wrong exons, and the ratio of missing and wrong genes. Gene prediction software used the parameters file and the genome to predict all genes despite the *complete* transcripts. The mapping information used from the *3'-truncated* and *fragmented* transcripts was limited to the completely mapped exons. For these genes, GeneID used this information to predict only the missing exons. All post-prediction filters were run as custom Perl scripts to clean false positive gene models and transposable elements. We included the following steps: 1) elimination of gene models that had a BLASTN result with e-value $<10^{-10}$ against any sequence in a TE database (provided by JMR); 2) selection of all gene models presenting any PFAM-A or PFAM-B conserved domain; 3) selection of all gene models that had a BLASTN result with e-value $<10^{-10}$ against any sequence in the transcript database; 4) selection of all gene models having a predicted signal peptide using SignalP (30); 5) elimination of gene models <40 amino acids and unannotated in the NCBI protein database; and 6) elimination of premature stop codons that shortened gene models. We considered all gene models remaining in the principal group as protein coding genes, and we considered all gene models identified with the first filter as TEs. All others were eliminated. RNA genes were annotated with the pipeline used by Ensembl Genomes, which combined results from Rfam, tRNAScan, and RNAmmer. The pipeline is summarized at http://ensemblgenomes.org/info/data/ncrna.

After performing two independent gene predictions (described above), the standard gene models provided by VectorBase were complemented with 1474 selected protein-coding gene models found

exclusively in our alternative predictions (named Lagerblad 3.1). The current consensus gene prediction (VectorBase 1.3) has 15,456 protein-coding gene models and 738 RNA-coding gene models. Consensus prediction is available at vectorbase.org as *fasta* and *gff* files; the site also offers browse and BLAST capabilities. Results from our alternative prediction are available under request as *fasta*, *gff* and anoXcel (31) annotated *xls* files. Protein-coding regions showed GC content very similar to *Acyrthosiphon pisum* (32) and *Apis mellifera* (33), averaging 40.4±6.8%. Only 12.3% of scaffolds have at least one feature predicted (Figure A1 in Appendix) and those having protein-coding gene models showed a distribution similar to the whole group (Figure A1 in Appendix) Therefore, small scaffolds also have genes, although at a lower density.

Gene annotation and gene family clusters

We performed automatic annotation and detected paralogous genes using AnoXcel and Cluster5 (31). Proteins and their coding sequences were categorized together, along with their matches by BLASTP, BLASTN, or RPSBLAST to several databases, as previously described for AnoXcel (31). Annotation was automated by parsing keywords from the significant matches from BLAST. We also incorporated information regarding the genomic location of the genes, as well as MW, pI, presence of signal peptides (34), transmembrane domains (35), N-linked glycosylation (36) and furin (37) cleavage sites of the predicted polypeptides. We also clustered the protein sequences based on their similarities at different stringencies. The cluster stringency of 80% identity in at least 50% of length was chosen based on the correct grouping of a manually curated cluster of carboxylestherases.

We clustered orthologous genes by running OrthoMCL (38) with default parameters on the predicted proteins from sixteen genomes (four Hemimetabola - *R. prolixus*, *Ac. pisum*, *Cimex lectularius, Pediculus humanus;* four Holometabola - *N. vitripennis, Ap. mellifera, Bomyix mori, T. castaneum;* four Diptera - *Ae. aegypti, An. gambiae, Glossina morsitans, D. melanogaster* and four outgroups - *Ixodes scapularis, Daphnia pulex, Strigamina maritime* and *Caenorhabditis elegans*). We parsed all orthologous clusters to identify those exclusive to taxonomic groups, and we used a script to identify genes with widespread ortology among the groups analyzed.

Comparing gene family sizes based on protein domain annotations revealed several putative lineage-specific reductions (LSR) where *R. prolixus* appears to have substantially fewer genes. LSRs were searched based on the conserved domain counting of all proteins using interproscan conserved domain

identification (39). Conserved domains with counts of a half or less in *Rhodnius* comparing to at least 2 other organisms from each holometabola, hemimetabola, diptera and outgroup taxonomic divisions were analyzed (Table D1.30 in Dataset). Reductions with absolute difference equal or smaller than 4 copies, also considering at least 2 other organisms from each holometabola, hemimetabola, diptera and outgroup taxonomic divisions were considered as unreliable and discarded.

Transposable elements

Transposable elements constitute a small fraction of the *R. prolixus* genome, and we identified most of the superfamilies' belonging to both classes (I and II) and orders (LTR and Non-LTR) that are present in other insect genomes. We also identified several canonical sequences, including complete copies with full-length ORFs, which represented putatively active elements belonging to the LTR order and Class II superfamily.

Three different approaches were used for the discovery and identification of transposable elements in the *R. prolixus* genome. To identify Class II and non-long terminal repeats (NLTR), we created reverse position specific matrices using PSI-BLAST (40) from the deducted coding sequences of all transposable elements found in the TEFAM (http://tefam.biochem.vt.edu/tefam/index.php) and REPBASE (41) databases. We generated 50kb windows of the *R. prolixus* genome using a step-size of 40kb (for an overlap of 10kb), and we queried our database with the tool RPS-BLAST (40) using an e-value cutoff of 10^{-15}. If matches exceeded 800bp, then coordinates were extended by 500bp to include flanking regions. We fused overlapping coordinates before retrieving putative transposable elements. All sequences were catalogued along with their respective coding sequence (CDS) and terminal inverted repeats (TIR). The sequences were trimmed at the TIR, or alternatively, at the CDS when repeats were not found. Finally, sequences were clustered by 90% identity when their lengths were within 90% of the larger sequence pair being compared. We classified different superfamilies of NLTR through phylogenetic analyses of reverse transcriptase domain sequences (>1000bp) plus complete open reading frames.

We identified long terminal repeats (LTR) using a homology-based approach, as previously described (42) and refined (43). Briefly, the canonical sequences of LTR retrotransposons from several genomes of insects were recruited from REPBASE (41) and TEFAM. We used TBLASTN (40) to search for sequences homologous to the pol regions of representative LTR retrotransposons in the *Rhodnius*

genome. Matches showing >30% amino acid identity over >80% of the length of the query sequence were subjected to further analyses in order to identify both LTRs of each element using the program Blast 2 sequences (8). This strategy allowed us to identify canonical sequences, which correspond to those constructed after alignment of at least three complete copies of each LTR retrotransposon element in the genome. We then performed BLASTN searches (40), using as queries each one of the consensus/canonical sequences of each LTR retrotransposon element. We provided a list of coordinates for each putative element in the genome, although if the identity of two copies was >90% at the nucleotide level, then we categorized these copies as belonging to the same LTR retrotransposon element.

Miniature inverted-repeat transposable elements (MITEs) are non-autonomous elements composed of terminal inverted repeats (TIR) flanking non-coding regions. These elements need the presence of active transposase for amplification and are usually found in very high numbers in some eukaryotic genomes. To identify MITEs, a repeat library was produced using the *R. prolixus* genome and RepeatScout (18) at a kmer size of 15. The repeat library was used to run Findmite (44) with the following parameters: no requirement of direct repeat; terminal inverted repeat at 12bp allowing one mismatch, and MITE length at 60-700bp. The resulting candidates were used as query to run TEalign, which is a pipeline that runs BLAST against the *R. prolixus* genome, retrieves matching copies plus flanking sequences, and performs Clustal alignments (45). The results from TEalign were used to manually confirm and classify each MITE on the basis of clear boundaries shared by multiple copies, terminal inverted repeats, and target site duplications. After obtaining this initial list of MITEs, self-to-self BLAST analyses were performed to remove redundancy using a cut-off of >80% identity. The non-redundant MITEs were finally used as a library to perform RepeatMasker with options "-div 20" (http://www.repeatmasker.org/), and this output was used to count MITE copy number and the percentage of genome occupancy.

To determine the overall representation of transposable elements within the genome, we ran the program BLAT (46) using all annotated sequences corresponding to LTRs, Non-LTRs, and Class II elements, while considering different lengths of the matches. Genomic coordinates were retrieved, as above, redundancy was eliminated, and results were sorted, fused and their lengths computed based on different TE lengths.

LTRs (Class I – Retrotransposons), representing the smallest fraction of transposable elements, occupied <1% of the *R. prolixus* genome. Still, we found 29 canonical elements belonging to the three main groups of LTRs: the Ty3/gypsy, the Pao/Bel, and the Ty1/copia groups (Table A1 in Appendix). We analyzed the occupancy of these elements based on different lengths (Table A2 in Appendix). For instance, when we considered short fragments (<500bp), the Pao/Bel elements represented the majority of all LTRs, however, for larger elements (>500bp), the Ty3/gypsy elements were the most abundant. We identified nine full-length elements (>5kb; Pao/Bel [N=3]; Ty3/gypsy [N=5]; Ty1/copia [N=1]). The Ty3/gypsy group was the most abundant and the most diverse, containing elements belonging to five known lineages (gypsy, Mag, CsRn1, Osvaldo-like and Mdg3), and we identified a new lineage of Ty3/gypsy based on three full-length sequences that were phylogenetically clustered with 100% bootstrap support. The Mdg1 lineage is not present in *R. prolixus*.

We identified 784 different sequences representing non-LTR elements. Together, all the Non-LTRs corresponded to 1.4% of the genome. We annotated 483 of these since the corresponding RT coding domain >1000bp and since complete ORFs were present. These sequences represented seven different clades: Jockey, LoA, CR-1, RTE, I, R1, and Ingi. The last clade represents a recently identified element (47) of the seventeen known Non-LTR retrotransposons. In addition, we identified three novel clades of Non-LTRs (e.g., Felyant, Rose and Totopi). We identified these clades based on phylogenetic analyses of the RT domain along with reference sequences representing all known Non-LTR clades, and our classifications received high bootstrap support. Although Jockey was the most abundant Non-LTR based on copy number and occupied bases, when considering fragments >500bp, the only clade with sequences >4000bp was the novel clade, Felyant. We identified twelve copies >3000bp (Table A2 in Appendix), and further analysis will provide more detailed descriptions of these novel Non-LTR families.

The great majority of the transposable elements found in the *R. prolixus* genome belonged to the Class II superfamily, which represented nearly 3% of the genome. This overrepresentation was due to the amplification of only one family of mariners. We describe seven new families: DTTRP1, DTTRP2, DTTRP3, DTTRP4, DTTRP5, DTTRP6 and DTTRP7; many of these sequences have full-length transposases and ORFs, indicating that they were recently amplified in a spate of transposition. We annotated other class II transposons in the *Rhodnius* genome, including Tc1/mariner, hAt, P, Maverick, Pogo, PiggyBac, Sola, and Helitrons.

MITEs occupied 0.5% of the genome and represented the second most abundant element in *R. prolixus*. We identified twelve different families, most of which were within the expected size range (<700bp); however, some were larger and contained partial coding sequences of DNA transposons, which suggested that they were undergoing processes of deterioration.

Lateral gene transfers from *Wolbachia*

Wolbachia is an intracellular bacterial symbiont of many invertebrates that induces a variety of reproductive phenotypes (48-52), and many genomes have been reported to contain chromosomal insertions originating from *Wolbachia,* (53-57). We therefore obtained sixteen *Wolbachia* genomes from NCBI (GIs: 481068109, 481066914, 225591853, 58418577, 42410857, 575882256, 545683404, 482891467, 482888170, 402496444, 225629872, 58584261, 190570478, 42519920, 398649717, and 190356750). Upon comparing each with the *Rhodnius* genome using BLASTN (40), we designated as *Wolbachia* introgression (WI) regions as those with >80% identities. We selected *Rhodnius* genes present in these regions along with all others having a *Wolbachia* protein as the best hit in a BLASTP analysis (40), which relied upon the bacterial subset from the nr database. We discarded gene models with <40% identity and match length <50% of their own length. The remaining gene group was named horizontally transferred genes (HTG) and follow to codon frequency analysis (see below). We analyzed tRNA genes identified in WI regions using tRNAdb (58) and then used tRNAscan (59) to find similar tRNAs and identify anticodon sequences. We downloaded protein-coding sequences (CDS) for *Wolbachia* from NCBI, and we then manually removed operons and out-of-frame sequences. We controlled for redundancy using CD-Hit (60) at 100% identity.

The annotated *R. prolixus* genome data included *Wolbachia* genomic introgressions (WI) that spanned between a few hundred base pairs to >200,000bp. We identified ninety-six scaffolds carrying *Wolbachia* DNA sequences. At least four large WI segments were integrated into *R. prolixus* scaffolds (207,766bp on scaffold GL563182; 129,267bp on GL562490; 85,810bp on GL562249; and 57,228bp on GL561342). These regions encode the majority of the 31 predicted protein-coding HTG along with two tRNA genes (Tables D1.1-3 in Dataset), which are complimentary to TCA, one of the three most used serine codons in *Wolbachia* (Table D1.4 in Dataset). Expressed sequence tag libraries for *Rhodnius* (5) confirmed the transcription of eight genes from HTG (Tables D1.1-3 in Dataset). Results from experiments using polymerase chain reaction confirmed an additional gene model (RPRC000661), while also demonstrating that a predicted intron within this gene does not exist.

Wolbachia genomes undergo frequent rearrangements due to the high number of transposable elements and repeat regions (61, 62). Within *R. prolixus* HTG, we identified a pair of transposases (RPRC000742 and RPRC000770), a reverse transcriptase (RPRC000723) and several DNA recombination and repair enzymes (two DNA mismatch repair MutL proteins [RPRC010818, RPRC011745], one holiday junction helicase [RPRC004559], and one DNA polymerase I [RPRC006597]). These findings suggest that considerable machinery for transposing, recombining, and repairing the host DNA were necessary for successfully transferring *Wolbachia* genes to *Rhodnius*.

We calculated codon-usage frequency for the HTG as for coding backgrounds from *Rhodnius* (RB), *Wolbachia* (WB) using custom PERL scripts. We manually clustered codon-usage frequencies, and we also utilized Expander Tool (63) to perform both basic clustering with the CLICK algorithm and hierarchical clustering. Manual clustering of codons was based on the frequency difference (%) between HTG comparing to RB and WB. For example, a codon HTG frequency was considered similar to WB if its difference to WB was smaller than ± 5% and to RB was higher than ± 5%. Alternatively, codons presenting frequency values for HTG that were in between WB and RB and differed >±5% from both, were considered "middle". All other codons presenting frequency values that were not close or in between RB and WB background were not labeled.

We calculated codon usage among the 25 horizontally transferred genes (HTG) and the coding backgrounds from *Rhodnius* (RB) and *Wolbachia* (WB). With these three groups, we performed codon usage analysis and clustering to demonstrate that seven codons have frequencies in HTG close to WB and another seven close to RB. Meanwhile 31 codons displayed frequency values to HTG between WB and RB frequencies but closer to RB values than to WB ones (Table D1.4 in Dataset). Basic and hierarchical unsupervised clustering grouped 22 and 36 codons, respectively, with intermediate usage frequency (Figure A2 in Appendix), but in some cases codons with differences <5% were included. The considerable number of codons with intermediate frequencies (Figure A2 in Appendix and Table D1.4 in Dataset) suggests ongoing adaptation to the natural tRNA abundance of the insect.

Micro-RNA prediction and RNAi machinery

We predicted micro-RNA (miRNA) precursors and mature miRNA sequences. Initially, we retrieved sequences that form hairpin-like structures using Einverted (64) and BLASTN (28) tools. These sequences were variable in length (60-110 bp) so they were filtered with minimal free energy (MFE),

GC content (30-65%), mature sequence homology, protein coding genes, noncoding RNAs, and miPred classifier. The first filter calculated MFE using RNAfold within the Vienna RNA Package (65) according to the following parameters: -20 kcal/mol RNA secondary folding energy threshold and with the options "-p -d2 -noLP". To retrieve conserved miRNAs, no more than four mismatches were accepted in whole mature miRNA sequences with 0 mismatches in the seed region (2–8bp). The sequences were compared with *R. prolixus* transcripts to remove those sequences similar to known protein-coding sequences. All non-coding RNAs (i.e., rRNA, snRNA, SL RNA, SRP, tRNAs, and RNase P) were eliminated using Rfam v11.0 (66), and finally, we used miPred to obtain real miRNA precursors by removing pseudo and not real precursor miRNAs (67).

We analyzed a set of structural characteristics and thermodynamic parameters for all pre-miRNAs (i.e., Minimal Free Energy (MFE), Adjusted Minimal Free Energy (AMFE), Minimal Free Energy Index (MFEI), length, A content, U content, C content, G content, GC content, AU content, GC ratio, AU ratio, Minimal Free Energy of the thermodynamic ensemble (MFEE), ensemble diversity, and frequency of the MFE structure in the ensemble). The parameter adjusted MFE (AMFE) was defined as the MFE of a 100bp sequence. The minimal folding free energy index (MFEI) was calculated by the following equation: MFEI=(AMFE×100)/(G%+C%) (68). We measured the diversity, MFE, and frequency of the ensemble using RNAfold as well as MFE of the secondary structures; we measured the GC content and other structural characteristics using Perl scripts. We aligned the pre-miRNA sequences using ClustalX 2.0 (69) and RNAalifold (65), using an adjusted gap opening parameter (22.50) and an adjusted gap extension (0.83), and we generated the mature miRNA sequence logos using WebLogo 2.8.2 (70). We predicted the target genes for *R. prolixus* miRNAs using the software miRanda (71) through an analysis of the 3'UTR sequence available in the consensus gene prediction. Our adjusted parameters while running miRanda used 100% complementarity between the seed region of each rpr-miRNA gene and the 3′UTR sequence targeted. We retrieved the 3'UTR sequences of each gene structure (N=4532) using a custom Perl script based on information obtained in the gene annotation file. Most of the miRNA sequences targeted at least one 3′UTR sequence and a number of miRNAs had multiple target sites within the same target gene. Several 3′UTR sequences were targeted by more than one mature miRNA (Tables D1.10-11 in Dataset). We identified conserved, mature miRNAs and their precursors using an integrated approach that identified clustered miRNAs, duplicated miRNAs, intronic miRNAs, and intergenic miRNAs.

We discovered 65 conserved, precursor miRNAs and 87 mature miRNAs in *R. prolixus*. These numbers were similar to those from *An. gambiae*, which, according to miRBase v20.0, possessed 67 pre-miRNAs and 65 mature miRNAs. The structural and thermodynamic characteristics of the *R. prolixus* miRNAs were comparable with those from other species (Tables D1.6-7 in Dataset). We found sixteen pre-miRNAs distributed across intronic regions (*rpr-miR-9a-2*, *rpr-miR-278*, *rpr-miR-375*, *rpr-miR-2796*, *rpr-miR-9c*, *rpr-miR-190*, *rpr-miR-2788*, *rpr-miR-306*, *rpr-miR-9b*, *rpr-miR-971*, *rpr-miR-2a-1*, *rpr-miR-2c*, *rpr-miR-13b*, *rpr-miR-13a*, *rpr-miR-2a-2*, and *rpr-miR-71*), and we found 49 pre-miRNAs distributed across intergenic regions (Tables D1.8 in Dataset). We identified a conserved miRNA (*rpr-miR-190*) in the 43rd intron of a putative talin gene (RPRC007435), which was found in similar locations across Deuterostomes (72) and Protostomes. All mature miRNAs shared general miRNA characteristics. For instance, the nucleotide uracil was present in the first position of the 5' stem due to its functional role in RISC complex recognition for RNA-induced silencing (Figure A3A in Appendix and Table D1.9 in Dataset).

We identified 22 miRNAs in eight cluster-gene structures (1: *rpr-miR-100* and *rpr-let-7* [*miR-let-7/100a*]; 2: *rpr-miR-12*, *rpr-miR-3477*, and *rpr-miR-283*, [*miR-283/3477/12*]; 3: *rpr-miR-275* and *rpr-miR-305*, [*miR-305/275*]; 4: *rpr-miR-2a-1*, *rpr-miR-2c*, *rpr-miR-13b*, *rpr-miR-13a*, *rpr-miR-2a-2*, and *rpr-miR-71* [*miR-71/2a-2/13a/13b/2c/2a-1*; (Figure A3B in Appendix)]; 5: *rpr-miR-87b* and *rpr-miR-87a* [*miR-87a/87b*]; 6: *rpr-miR-92b* and *rpr-miR-92a* [*miR-92a/92b*]; 7: *rpr-miR-9c*, *rpr-miR-306*, and *rpr-miR-9b* [*miR-9b/306/9c*]; and 8: *rpr-miR-iab-8* and *rpr-miR-iab-4* [*miR-iab-8/iab-4*]). We localized a cluster (*rpr-miR-71/2a-2/13a/13b/2c/2a-1*) to the ninth intron of a gene (RPRC004331) that encodes a putative phosphatase-four-like protein. *C. elegans* possessed a similar cluster (*rpr-miR-71/2*) in the fifth intron of a similar gene (PPFR-1; NCBI Gene ID: 172761). We identified other conserved clusters in the *R. prolixus* genome, including one known from *Ap. mellifera* and *Nasonia vitripennis* (*miR-283/3477/12*), as well as two overlapping clusters that were located on opposite strands (*rpr-miR-iab-4* [+]; *rpr-miR-iab-8* [-]), as previously described in *Dr. melanogaster* (73).

We predicted putative target genes (N=804) for all mature miRNAs (N=87). We next analyzed these target genes regarding their function using GO annotation information and found that most have catalytic, binding, and transporter activity (Figure A3C in Appendix). We found that one cluster (*rpr-miR-124-3p*) targets a predicted Rho-associated protein kinase (RPRC007732-RA). Since *miR-124* was

known to regulate the Rho-associated protein kinase 1 (ROCK1) (74), our finding corroborates this association between *miR-124* and *ROCK* mRNA.

Several studies have demonstrated the role of miRNAs in Huntington's disease (75), including those located in the *Hox* gene clusters, such as *miR-10*. Although we found *miR-10* outside the *Hox* gene cluster and instead localized it to a different scaffold (GL562219), we discovered that the cluster *rpr-miR-10-3p* targeted a Huntington-like gene (RPRC006430-RA). This finding corroborates the conservation of the miRNA-target pair despite the absence of Huntington's dysfunction in *R. prolixus* (76).

RNA interference (RNAi) is a post-transcriptional gene silencing mechanism triggered by double stranded RNAs (dsRNAs), which results in degradation of a target messenger RNA (mRNA) in a sequence-specific manner (77). RNAi can be also triggered by microRNAs (miRNAs), which are small non-coding RNAs ~22bp in length (78). In *Dr. melanogaster*, several genes are associated with RNAi; *Dcr-1* is related to miRNA processing, *Dcr-2* is related to the production of siRNAs from long dsRNAs (79, 80), and *Drosha* participates in the processing of primary miRNAs (81-83). We identified these three genes in the *R. prolixus* genome (Table D1.5 in Dataset). Other molecules participate in RNAi pathways, including proteins with dsRNA binding domains (dsRBD) that act as bridges between Dcr and RISC proteins. In *Dr. melanogaster*, Loquascious (*Loqs*), *R2D2*, and *Pasha* form associations with *Dcr-1*, *Dcr-2*, and *Drosha* proteins, respectively (81, 84-87); we identified all of these genes in the *R. prolixus* genome (Table D1.5 in Dataset).

When we considered the RISC complex, we identified coding sequences in the *R. prolixus* genome for the genes *Ago1* and *Ago2a* (Table D1.5 in Dataset), although we failed to identify others, such as *Wago* and component 3 promoter of RISC (*C3PO*). Among the genes from the PIWI pathway, we identified core components, including *Ago3*, *PIWI*, *Armitage* and *Spindle E* (Table D1.5 in Dataset), but we failed to identify sequence for *Aubergine*.

Nucleic acid receptors or cell membrane-spanning proteins promote the uptake of RNA molecules. In *C. elegans*, *sid-1* codes for a transmembrane dsRNA transporter that is key for systemic RNAi (88-90). A homologous gene for *sid-1* was reported in *Trib. castaneum* (91), however, as in *Dr. melanogaster* (92), we found no homologous gene in *R. prolixus*. RNA-dependent RNA polymerases (RdRPs) are enzymes that mediate the amplification of the RNAi signal in plants and *C. elegans* (93, 94), and

although they potentially determine the duration of RNAi silencing, RdRPs have not been identified in the genomes of any insect species. An RdRP-associated protein, *Drosophila* elongator subunit 1 (D-elp1), was discovered in *Dr. melanogaster* (95), but we failed to identify any homologous genes for RdRPs and *D-elp-1* in *R. prolixus* despite the presence of prolonged RNAi (96).

Selenoprotein machinery

Selenoproteins contain selenocysteine (Sec), an unusual amino acid inserted through the recoding of a UGA codon (normally a stop). We identified selenoprotein genes in using the dedicated gene annotation pipelines Selenoprofiles (97) and Seblastian (98). The SECIS elements, a characteristic secondary structure present in the 3'UTR of selenoprotein genes mRNA, were predicted with SECISearch3. We identified tRNA-Sec using tRNAscan-SE (99). We performed multiple protein sequence alignments using T-Coffee (100), and we reconstructed maximum likelihood trees using the best fitting evolutionary model (101). For this analysis, we used sequences *C. elegans* (TRXR-1, WBGene00015553 and TRXR-2, WBGene00014028), *P. humanus* (PHUM103080 and PHUM369770), *D. melanogaster* (Trxr-2, FBgn0037170 and Trxr-1, FBgn0020653) and *Ac. pisum* (ACYPI064401). We also predicted TRXs in other Paraneoptera genomes downloaded from NCBI: *Diaphorina citri* (GCA_000475195.1), *Halyomorpha halys* (GCA_000696795.1), *Cimex lectularius* (GCA_000648675.1), *Frankliniella occidentalis* (GCA_000697945.1), *Pachypsylla venusta* (GCA_000695645.1), and *Oncopeltus fasciatus* (GCA_000696205.1).

We found a pair of selenoprotein genes in the *R. prolixus* genome, *sps2* (RPRC009014) and *GPx* (RPRC011108), coding for selenophosphate synthase and glutathione peroxidase, respectively. We also detected the complete set of genes necessary for Sec biosynthesis and insertion ("Sec machinery" – Table D1.13 in Dataset). While selenophosphate synthase is a selenoprotein, it is also part of Sec machinery, since selenophosphate is necessary for Sec production. We also identified selenophosphate synthase 1 (*sps1*; RPRC008322), a gene similar to *sps2* but unlikely to be related to selenoprotein synthesis (102). Specifically, *R. prolixus sps1* has an in-frame UGA which is believed to be translated without insertion of Sec, as previously reported for honey bee (103).

Although we identified selenoproteins in *P. humanus* (human body louse), including thioredoxin reductases (TRs), methionine-R-sulphoxide reductases (SelRs), and selenoprotein K (SelK), none were found with a Sec-TGA codon in *R. prolixus*. We detected a cysteine-containing paralogous sequence of

Sec-SelR as well as two cysteine-based TR genes (RPRC014349 and RPRC006341). RPRC014349 displayed evidence of a recent Sec to cysteine conversion, according to a phylogenetic analysis of Paraneoptera genomes that placed one of the *R. prolixus* cysteine-TR proteins with the Sec-TR proteins (Figure 3A), including those from *C. lectularius* (common bed bug), a closely related taxon. This observation supports the hypothesis that the Sec-TGA codon was converted to a cysteine (TGC) following divergence from a common ancestor. Examining the local genomic neighborhood (11 gene window) of RPRC014349 and its orthologs from *C. lecturarius* (CLEC012062, CLEC002932 and CLEC011810) and *P. humanus* (PHUM103080) revealed no conserved gene arrangements, i.e. local synteny has not been maintained among these three species. Interestingly, one of the *C. lecturarius* orthologs (CLEC011810) appears to be an intron-less gene, which could be a consequence of a retrotransposition event that also could explain the absence of synteny within this gene block. The neighborhood (11 gene window) of the second *R. prolixus* TR gene (RPRC006341) and its orthologs from *C. lecturarius* (CLEC004976), *A. pisum* (ACYP004117), *P. humanus* (PHUM369770) and *D. melanogaster* (FBng0037170) showed a single syntenic relation with only 3 orthologous genes (RPRC006305-CLEC004977-ACYP086307) that were maintained as neighbors of the TR gene. We detected expression of the TR gene (RPRC014349) in expressed sequence tags from *R. prolixus*, thereby providing further support to cysteine conversion. These results, together with the finding of a Sec-GPx in *R. prolixus* (see Main Text), indicate that insects originally possessed a more extensive repertoire of selenoproteins, and that Sec loss and replacement by cysteine occurred independently in different species. It is possible that redox metabolism organization in insects reduced the selective benefits that maintain the higher number of selenoproteins found in vertebrates, as previously suggested (103).

Immune pathways

In the *R. prolixus* genome, we failed to discover several key genes of the IMD pathway (e.g., *IMD*, *Fadd*, *Dredd*, and *Caspar*). In contrast to the inactive IMD pathway in *Ac. pisum*, however, we identified particular components of the pathway, including *PGRPs* (classical IMD receptors) and a *Relish* homologue, *rpRelish* (Accession KP129556), suggesting that the pathway might exist in *R. prolixus,* albeit with alternate activation mechanisms. The gene *rpRelish* was therefore manually assembled using contigs from a previously published transcriptome (5). For confirmation, *rpRelish* was cloned and sequenced. The CDS of *rpRelish* contained 2190bp and translated a protein that measured

715 amino acids in length (M.W. 81638.28 Da). The protein displayed typical Rel Homology (RHD) along with IPT domains on the aminoterminal region and five consecutive ankyrin repeats at the carboxiterminal (Figure 2A).

A blood meal increases the population of the gut endosymbiont (*Rhodococcus rhodnii*), which represents the only bacteria present in *Rhodnius* midgut (104). We hypothesized that insects with increased intestinal microbiota following a blood meal would be immunologically challenged and would display increased *rpRelish* transcription. To test this, we extracted total RNA from tissues in all conditions using the TRIZOL reagent (Invitrogen), following the manufacturer's instructions, and we used real-time PCR (qRT-PCR) to assess the transcript abundance and silencing efficiency of the genes of interest. As hypothesized, *rpRelish* transcript expression increased in the midgut and fat body 24h and 72h after a blood meal, respectively (Figure 2B-C). To determine if *Relish* and the IMD pathway were immunologically relevant, we silenced the pathway by injecting *rpRelish* double-stranded RNA (dsRNA) (Figure A6B in Appendix). We used a T7 Megascript kit (*Ambion Inc.*) to generate and purify dsRNA from PCR amplified genes following the manufacturer's instructions. We injected the dsRNA in sterile water into the thorax of adult females, and after three days post dsRNA injection, we provided insects with rabbit blood. After seven days (i.e., four days after blood meal), we measured the expression of defensin A [GenBank: AAO74624.1], lysozyme-A [GenBank: ABX11553.1] and lysozyme-B [GenBank: ABX11554.1] in the midgut. Defensin A expression decreased significantly and hence, defensin A may be modulated by *rpRelish*. On the other hand, transcription of lysozyme-A increased, suggesting that the IMD pathway does not control this antimicrobial peptide in *R. prolixus* (Figure 2D).

Next, we dissected, homogenized, and plated the three sections of intestinal tract from *rpRelish* knockdown insects on BHI-Agar broth, and we analyzed the microbiota of the digestive tract by counting Colony Forming Unit (CFU) after plating homogenates of the different sections of the midgut. Insects injected with *rpRelish* dsRNA possessed an increased bacterial population in the anterior and posterior midgut. The results for the hindgut were not significant, although the results were in the same, hypothesized direction (Figure 2E-G). Together the data support the hypothesis that the IMD pathway is active in *Rhodnius* and is likely responsible for controlling the population of *Rhodococcus rhodnii*.

Trypanosoma cruzi Dm28c clone was grown in a liver infusion tryptose (LIT) culture medium (5.0 g/l liver infusion broth, 5.0 g/l tryptose, 4.0 g/l NaCl, 8.0 g/l Na2HPO4, 0.4 g/l KC, 2.0 g/l glucose, 10.0 mg/l hemin and 100.0 ml/l heat-inactivated fetal bovine serum, pH 7.2) at 28^{o}C, and the epimastigotes were obtained from the log-growth phase and quantified using a neubauer chamber. Insects were fed with heparinized (2.5units/ml) rabbit blood containing 1x10^{7} epimastigotes/ml through a latex membrane feeding apparatus. Blood was previously centrifuged at 2,000 × g to separate the serum from the erythrocytes. Serum was inactivated (56^{o}C for 45 minutes) and the erythrocytes were washed three times in phosphate-buffered saline (PBS, pH 7.2), before the blood was reconstituted. The epimastigotes were then washed one time in PBS and added to the reconstituted blood. Only fully engorged insects were used in experiments.

DNA extraction was performed with Cetyltrimethyl ammonium bromide (CTAB) due to selective precipitation of DNA and elimination of hemin/hemozoin and polysaccharides. After infection with parasites expressing constitutively GFP, anterior midgut, posterior midgut and hindgut were extracted from cold-anesthetized insects at the indicated time points post-infection. Each gut section was homogenized separately in 1 ml of Solution A (1.5% CTAB; 2M NaCl; 10mM EDTA; 100mM sodium acetate, pH 4.6). After homogenization, a 100-µl aliquot was transferred to a 2-ml microfuge tube containing 900µl of Solution B (Solution A containing 10µg salmon sperm DNA as DNA carrier, 10ng of plasmid pLew82 (ble) and 125µg RNAse A per each 900-µl aliquot) and vortexed for two minutes. The samples were then heated for 30 minutes at 65^{o}C for RNAse A activity and 500µl of chloroform was added to the samples and heated for two hours at 65^{o}C. Following this step, the samples were incubated for 10 minutes at 25^{o}C and centrifuged at 9,000 × g at 25^{o}C. After transferring a 600µl aliquot of the aqueous phase to a new tube, 300µl of chloroform were added to the samples and vortexed for two minutes. Then the samples were incubated at 25^{o}C for ten minutes and centrifuged at 20,000 × g for ten minutes at room temperature. A 400-µl aliquot of the aqueous phase of each sample was precipitated at -20^{o}C for one hour after the addition of 1µl of 20mg/ml glycogen solution (Sigma-Aldrich) (or salmon sperm DNA (ssDNA) that also provided comparable results), 40µl of 3M sodium acetate (pH 5.2) and 1ml of ice-cold ethanol. Samples were then centrifuged at 20,000 × g at 4^{o}C for ten minutes. The DNA pellet was washed twice with cold 70% ethanol, dried at room temperature and re-suspended in 50µl of nuclease-free water.

T. cruzi DNA standard were obtained from either parasite-free whole gut homogenates extracted from insects on day 7 post feeding or rabbit blood to which 10^7 parasites were added and serially diluted 10-fold with nuclease-free water containing 10µg/ml salmon sperm DNA (ssDNA) to cover a range between either 10^5 and 0.001 or 2.5 x 10^6 and 2.5 parasite equivalents for Dm28c and CL Brener lineages, respectively per 5µl of sample added to the reaction mixture.

The standard calibration curve for the internal loading control (pLew82) containing the *ble* resistance gene was performed using the same reconstituted samples as carried out for the parasite calibration curves; insect gut homogenates or rabbit blood samples were spiked with 10ng of plasmid and serially diluted 10-fold with nuclease-free water containing 10µg/ml salmon sperm DNA (ssDNA) to cover a range between 10 and 0.01ng of plasmid. All samples were extracted in the presence of 10ng of pLew82 plasmid used as heterologous internal standard. The standard calibration curve for the internal loading control was used to determine the percentage of DNA recovery. Quantitative PCR was performed on a StepOnePlus real-time PCR system (Applied Biosystems) using the Power SYBR Green PCR master mix (Applied Biosystems) in a final volume of 15µl. PCR reactions contained 5µl of DNA sample and 500nM of *T. cruzi* repeat DNA-specific primers (105), which amplify a 182-bp amplicon from a tandemly repeated satellite DNA. Specific *T. cruzi* DNA-oligonucleotides were TcFw, (5'-GCTCTTGCCCACAMGGGTGC-3'), where M = A or C, and TcRv, (5'-CCAAGCAGCGGATAGTTCAGG-3'). In parallel, we used as loading control reactions containing 5µl of DNA sample and 500nM of oligonucleotides designed to PCR amplify a 148-bp fragment of the Sh *ble* gene of the plasmid pLew82 using the following set of primers: BleFw, (5'-CAAGTTGACCAGTGCCGTTC-3') and BleRv, (5'-GCTGATGAACAGGGTCACG-3'). qPCR conditions used for both pairs of primers were the following: ten minutes at 95°C followed by 40 cycles of fifteen seconds at 95°C and fifteen seconds at 60°C. The amplification step was followed by a melting curve to assure a simple product was amplified. The data were analyzed with the StepOne software v2.3. Negative controls for both primer pairs consisted of a reaction with no DNA added. For each primer set, the efficiency of amplification was determined using the following formula: Efficiency (E) = -1 + 10 (−1/slope). We derived qPCR experimental protocols from the Minimum Information Required for Publication of Quantitative Real-Time PCR Experiments (MIQE) Guidelines (106).

Gene family identification and curation

We followed a general approach to identify and curate *R. prolixus* sequences by searching for orthologous sequences in other genomes, including *Ac. pisum*, *Cimex lectularius, Pediculus humanus, N. vitripennis, Ap. mellifera, Bomyix mori, T. castaneum, Ae. aegypti, An. gambiae, Glossina morsitans* and *D. melanogaster*. These genomes served as queries for BLASTP (40) and TBLASTN (40) searches on the official and alternative sets of predicted proteins for *R. prolixus* (described above). We also implemented conserved domain searches using HMMER and Pfam database (107). Homologous sequences from *Rhodnius* were manually curated using BLAST searches with nr (108), uniprot (24) and Gene Ontology (109) databases. We performed conserved domain analyses with *hmmsearch* (110) and Pfam (107) databases. All analyses were carried out in FAT tool (111).

We identified and manually curated most of the major metazoan cell signaling pathways related to development and metabolism (Table D1.16 in Dataset). The nutritional (FOXO, TOR), neurogenesis (Notch), and embryonic development (Wnt) pathways were conserved in the *Rhodnius* genome. Furthermore, the complete Hedgehog (Hh) pathway was also present, including Hh itself and its receptor, Patched (Ptc). Although the above pathways were detected in the *Rhodnius* genome, we were unable to identify the full components of other pathways. For instance, while we identified calcium-mediated signaling genes, including those that encode receptors found in mitochondria, plasma, endoplasmic reticulum, and sarcoplasmic reticulum, we did not find sphingosine kinases or phospholamban. We did not find hypoxia-inducible factors (HIFs), which respond to changes in available oxygen. Likewise, we did not identify key regulatory elements of this pathway, including the oxygen-detecting gene, hypoxia-inducible factor prolyl hydroxylase (*PHD*) or *E1A*/*CBP*, which target HIFs for degradation or for gene transcription (112). The protein kinase Hippo (*hpo*) is a core enzyme within a regulatory pathway that controls organ size. It acts to inhibit the transcriptional co-activator Yorkie (*Yki*) and its proliferative, anti-differentiation, and anti-apoptotic transcriptional program. We did not find several elements of this pathway, including *hpo*, Salvador *(Sav)*, and *Yki.* At the same time, we did identify several anti-apoptotic and pro-proliferation genes, including *kibra*, *fj*, and *dally* (113).

Tyrosine kinases (TKs) contain domains that phosphorylate tyrosines on their targets (114), and the *Rhodnius* tyrosine kinome (Table D1.17 in Dataset) represented the smallest described to date; for comparison Diptera has 30 TKs and *H. sapiens* has 90 TK members distributed among 30 subfamilies (115, 116). Upon searching the *Rhodniu*s genome, we identified a complete major signaling pathway, commonly activated during embryogenesis, that connects extracellular developmental ligands to

MAPKs through receptor tyrosine kinases. Yet, we failed to detect a total of eight TK subfamilies that were present in the insect kinomes of *Dr. melanogaster* and *An. gambiae* (ACK, ARK, PDGFR, RET, DmCG3277, FRK, ROS, NGFR) that could be specific to Diptera. We also found that the torso-like (Tsl) and Gurken (grk) pathways lacked the torso tyrosine kinase receptor (Tor) and the ligand grk itself. Finally, *R. prolixus* putatively possessed only a single receptor that was identified as homologous to the four mammalian ErbB proteins that regulate apicobasal polarization (117).

We also investigated the presence of urea cycle enzymes in *R. prolixus*, and interestingly, we found only argininosuccinate synthetase (RPRC003927) (Table D1.35 in Dataset). This result is similar to the discovery in *Ac. pisum* (118), where the complete absence of the pathway presented a striking contrast to other insects. Previous biochemical data (119) suggested that only a small amount of urea was found in *R. prolixus* urine just after blood ingestion, in an amount that was approximate to the urea content of the blood meal. That none of the enzymes involved in uric acid degradation to ammonia and carbon dioxide were found in *R. prolixus* reinforces the notion that the organism lacks urea production and has potentially adapted to a diet rich in amino acids as a result.

Next, we examined genes in the *R. prolixus* genome that correspond with the following areas of biology: transcription factors, defensins, detox enzymes, juvenile hormones, neuropeptides, behavior/sensory/memory, circadian clock, heme metabolism, chemoreceptors and their accessory proteins, proteases, salivary proteins, cuticle development, energy and lipid metabolism, and amino acid synthesis. Different approaches were used in particular cases and are explained as follows.

Transcription factors: Lineage-specific expansions (LSEs) and divergence of TFs appear to have played a potential role in diversification of animal body plans (120, 121). Hence, we investigated the LSEs of TFs and lineage-specific diversification of TF complements in relation to the emergence of extreme morphological and behavioral disparity in the hemipterans *R. prolixus* and *Ac. pisum*. In addition to these two hemipterans we also queried genomes from 16 other arthropod species for the comparative analysis: *Trib. castaneum* (Tcas), *Dr. melanogaster* (Dmel), *Ap. mellifera* (Amel), *Daphnia pulex* (Dpul), *Heliconius melpomene* (Hmel), *Ixodes scapularis* (Isca), *An. gambiae* (Agam), *Camponotus floridanus* (Cflo), *Nasonia vitripennis* (Nvit), *Bombyx mori* (Bmor), *Dendroctonus ponderosae* (Dpon), *Harpegnathos saltator* (Hsal), *Pediculus humanus* (Phum), *Zootermopsis nevadensis* (Znev), *Mesobuthus martensii* (Mmar) and *Stegodyphus mimosarum* (Smim). We manually selected from the Pfam database a panel of HMMs for 300 diagnostic domains found in eukaryotic

TFs, including the DNA-binding domains that contact specific target sequences. We further manually classified these domains into specific and basal/general transcription factors (TFs) and chromatin proteins (CP). We supplemented these with additional PSI-BLAST profiles for domains not found in Pfam. Using these HMMs we scanned both sets of predicted proteins from the *R. prolixus* genome (alternative and consensus, described above) and clustered the predicted TF/CPs using BLASTCLUST with thresholds set to remove redundancy resulting from the use of two independent gene predictions. The TFs/CPs were grouped into families/clusters using Markov Clustering (MCL) and each cluster was analyzed for its protein domain architectures (122).

We discovered a recent expansion of the helix-turn-helix DNA-binding domain (i.e., the Pipsqueak [Psq] domain; PFAM: HTH_psq) that resulted from independent rounds of proliferation in *Ac. pisum* (84 genes) and *R. prolixus* (36 genes). These expansions occurred through transposon-mediated gene duplication, given that in several genes the Psq domain was still fused with the integrase catalytic domains of the parent transposon, especially in *Ac. pisum*. Among the majority of versions *in R. prolixus*, however, the associated transposase domains had degenerated, suggesting that a subset of the DNA-binding Psq domains were reused as specific TFs. Due to the important role of Psq TFs in regulating developmental patterning in *Drosophila* it is possible that differential LSEs of these proteins in the two hemipterans have roles in their morphological and/or behavioral differences; thus, they offer themselves as potential candidates for future experimental analysis. We also detected variable counts for THAP; *Ac. pisum* possessed over 430 genes encoding THAP domains while *R. prolixus* possessed only three of these genes. The expansion in *Ac. pisum* was primarily due to copies borne by a P-element-like transposon. In the case of SAZ (also called MADF), *A. pisum* again displayed an expansion with 178 copies relative to the 33 SAZ genes found in *R. prolixus*. Similarly, we found evidence that SAZ expansion in *Ac. pisum* was likely mediated by transposons, many of which appeared inactive in *R. prolixus*.

This analysis demonstrated that closely related species displayed notable LSEs and also TF extinctions (e.g., THAP). In particular, the association with transposon expansions suggests the proliferation of these mobile elements during evolutionary radiations in insects might have acted as a source for novel TFs and possibly facilitated the emergence of new adaptations. Expanded TF families may take remarkably different routes, as shown in Psq, SAZ and THAP. While the former two domains were independently expanded and retained in both taxa, THAP is largely eroded in R. *prolixus*. Taken together, these results suggest that evolutionary processes might preferentially purge neutral or

disadvantageous gene families in particular lineages during periods of increased selective pressure and ultimately act to reshape the TF repertoire (122).

It is very important to mention here that Helix-turn-helix and Zinc finger (C2CH type) motifs that were identified as putative LSRs, are present in many different transcription factor families and a simple gene count couldn't be used to identify a reliable LSR. These conserved domains were analyzed in detail as described above and completely scrutinized (122), and then there are no LSRs in Helix-turn-helix and Zinc finger domain containing families (Table D1.30 I Dataset).

Defensins: Defensins likely play an important role in protection from microbial organisms. We retrieved all sequences representing defensins from the gene family clustering analysis generated by OrthoMCL (described above). We identified Group G560 by automatic annotation, and we added unclustered sequences if they contained the conserved domain (PF01097). We aligned all defensins using MAFFT (123) and we constructed a phylogenetic tree using RAxML (124) with 500 bootstrap interactions (Figure A5 in Appendix). We excluded partial sequences (RPRC012334 and RPRC012261) from the alignment and tree-building analyses.

Among the nine insects, *R. prolixus* displayed the greatest number (N=11) of defensin genes, ten of which were arranged in tandem on scaffold GL563087 (Table D1.12 and Figure A8 in Appendix). Seven defensins were in the forward strand (RPRC012180, RPRC012259, RPRC012177, RPRC012182, RPRC004803, RPRC012186, and RPRC012261) and three were in the reverse (RPRC012183, RPRC012184, and RPRC012185) (Figure A8 in Appendix). A single defensin gene (RPRC012334) was located on scaffold GL562086. Previous work (125) in *R. prolixus* identified three defensin genes (*defensin A* [AAO74624]; *defensin B* [AAO74625]; and *defensin C* [AAO74626]) and these correspond to RPRC012185, RPRC004803, and RPRC012184, respectively. As in other insects (126), the presence of a single tandem cluster in *R. prolixus* suggests that the repertoire of defensins arose by recent gene duplication events. The other eight insects possessed variable numbers of defensin genes, from N=7 in *N. vitripennis* to N=0 in *Ac. pisum*.

Detox enzymes: Cytosolic glutathione s-transferases (GSTs) have two domains necessary for their enzymatic activity (N- and C-terminal). Based on phylogenetic placement of the N-terminal sequences, we classified all fourteen *R. prolixus* GSTs into five classes, with seven belonging to the sigma class, four to theta, and one each to omega, delta, and zeta. We also found one microsomal GST. These

findings are similar to other Hymenoptera species but different from other hemipterans, including *Ac. pisum*, according to previously sequenced genomes (127-129). Nevertheless, analyses of transcriptomes from three other Hemiptera species (all planthoppers) found similar numbers compared with *R. prolixus* (130, 131). In summary, most hemipterans and hymenopterans have higher gene numbers of the sigma class while the epsilon and delta classes are much more numerous in Diptera, Coleoptera and Lepidoptera.

Based on their function and phylogenetic relationships, insect carboxylesterases (CCEs) are divided into three broad classes: detoxification/dietary; pheromone/hormone processing; and neural/developmental, and these, in turn, are divided into smaller, more specific clades (128, 132-134). Apart from the pheromone/hormone processing class, where seven genes were found in *R. prolixus*, the number of genes found in the other classes (N=19 for detoxification/dietary and N=12 for neural/developmental) mirrors those for other vectors studied. Regarding the neural/developmental class, we found a pair of acetylcholinesterase genes (*Ace-1* [RPRC000482] and *Ace-2* [RPRC003013]), four neuroligins, two glutactins, a neurotactin, and a gliotactin. We found expansions of *R. prolixus* genes related to pheromones and hormones processing clustered in OrthoMCL cluster G540 (RPRC010231, RPRC014625, RPRC014623, RPRC010262, RPRC010265, RPRC010259, RPRC010267, RPRC010268, RPRC014627, RPRC010260, RPRC010261, RPRC007702, RPRC010228, RPRC007700, RPRC010223, RPRC010226, RPRC010224, RPRC010230, RPRC008235, RPRC010263, RPRC003566, RPRC001239, RPRC003564, RPRC003565, RPRC007703, RPRC008217, RPRC001592, RPRC004680, NVIT156541966, GB52052) and G12349 (RPRC007696, RPRC007698-PA) and most of them are located at the same scaffolds - Table D1.14 in Dataset and (135). We also found a phosphotriesterase (PTE) gene (RPRC012395) with a complete active site, a substrate-binding pocket, and a homodimer interacting site.

We identified 88 cytochrome P450 (CYP) genes, some of which were grouped into seven CYP clans, each with up to eight putative functional genes or gene fragments. We also consistently observed specific expansions of CYP clans, many of them displayed in tandem organization - Table D1.15 in Dataset and (135). Research in other taxa demonstrated that expansions were most often observed in clans associated with environmental interactions (CYP3 and CYP4 clans), but not in subfamilies associated with core functions in development and physiology (mitochondrial and CYP2 clans) (136). In *R. prolixus*, the CYP3 clan was comprised of 50 genes, including many novel families (CYP3084-

CYP3092 and CYP3096). We also observed a specific expansion in the CYP4 clan (27 genes), with the new family CYP3093 as the major contributor. This expansion contained putatively functional genes (*CYP3093A1-A10*) along with eighteen gene fragments. Under the context described above, the insect CYP3 and CYP4 clans potentially play important roles in xenobiotic detoxification and their expansions might drive the potential to acquire resistance to many chemical insecticides. Although similar CYP expansions were reported for mouse (CYP2C) (137), filamentous fungi (CYP68) (138), and crustaceans (CYP370) (139), the insect CYP expansions are well-described, including in red flour beetle (CYP6BQ) (140), the jewel wasp (CYP4AB) (141), mosquitoes (CYP6 and CYP9) (142), and *Drosophila* (143). Among insects, the honey bee (CYP6AS) possessed the lowest number of CYP genes (N=46) (127). Our counts for *R. prolixus* depict a smaller expansion than mosquitoes and beetles, but larger than *Ac. pisum*.

Juvenile hormones: The juvenile hormones (JH) are involved in reproduction, caste determination, behavior, stress response, diapause, and several polyphenisms (144), yet understanding the mode of action of JH at the molecular level has been a major challenge in insect biology. While the general features of JH biosynthesis are conserved in most insects, our understanding of JH biosynthesis improves as organisms with different physiological frameworks are examined, given the clear diversity surrounding JH homologs, the order of the final enzymatic steps, and the role of allatoregulators. There are factors that can stimulate (allatotropins) or inhibit (allatostatins) Corpus allatum (CA) activity (145). In different insect species and at different stages of development, these regulatory factors may include three types of inhibitory allatostatins (AST), at least one type of stimulatory allatotropin (AT), insulin, and perhaps additional neuropeptides, such as ecdysis triggering hormone (ETH) and short neuropeptide F ($_{S}$NPF) (145-147). We therefore identified orthologous sequences of the genes encoding these allatoregulators and their receptors in the genome of *R. prolixus*, as well as the sequences that likely contribute to enzymes involved in the late steps of JH biosynthesis (Table D1.21 in Dataset).

Neuropeptides: We detected 37 neuropeptide precursor genes or hormone genes, which is a comparable result relative to the numbers found in *Bombyx mori* (N=37), *Tribolium castaneum* (N=47), *Drosophila melanogaster* (N=31), *Anopheles gambiae* (N=32), and *Apis mellifera* (N=36) (148-152). Expression patterns were previously confirmed by transcriptomic and peptidomic methods (153-158); hence, we were confident with the annotation of the neuropeptide precursor genes in the *Rhodnius*

genome. We detected splice variants in several *R. prolixus* neuropeptide precursor genes, including ATS B (2 isoforms), CT-like DH (5 isoforms), CAP_{2B} (2 isoforms), CZ (3 isoforms), CCH (2 isoforms), CCAP (3 isoforms), IRP-1 (2 isoforms), CCH-related ITP (2 isoforms), and OK (3 isoforms). This increased the number of neuropeptide-encoding transcripts from 37 to 52 (159). Relative to other insects, some of the *R. prolixus* neuropeptides, such as falps, kinins, SIF amide, sulfakinin, and myosuppressin, displayed unique sequences in the characteristic domains of their core peptides. We did not detect the arginine-vasopressin-like peptide, prothoracicotropic hormone, nor sex peptide, which were previously described in other insect genomes, yet we cannot rule out gaps in the genome assembly or highly divergent sequences that went undetected in our homology search.

Behavior control, sensory function and memory formation: We improved gene models using gene prediction programs such as GeneWise (160), Augustus (161) and Fgenesh+ (162). In the case of the pickpocket (PPK) and transient receptor potential cation channel (TRP) protein families, the sequences of each family were aligned using MEGA 4.0 (163) to detect problematic gene regions and to refine some gene models. After alignment, each new protein included in these alignments was used as a query in iterative searches to find new members of these families. This process was repeated until new sequences did not show any evidence of prediction inconsistencies in the set of these gene models. The final *R. prolixus* gene models have been created using Artemis (164) and BioEdit 7.0.5.3 (165) programs. The behavior genes identified for *R. prolixus* and its main characteristics are shown in the Table D1.32 in Dataset).

Circadian clock: We found orthologous sequences for many *Drosophila* clock genes in the genome of *R. prolixus* (Table D1.33 in Dataset). We also identified specific kinases that play key roles in the cyclic stability and nuclear entry of *Drosophila* clock proteins (166-171), including doubletime (*bdt*), shaggy (*sgg*), casein kinase 2 (*ck2*), nemo (*nmo*), protein phosphatase 1 (*PP1*), and protein phosphatase 2a (*PP2a*). Finally, we identified orthologous sequences for the E3 ubiquitin ligase supernumerary limbs (*slmb*) and F-box protein jetlag (*jet*), which degrade PER and TIM, respectively, in the proteasome (172, 173). One particular result suggests that the circadian clock in *R. prolixus* is more similar to that found in bees and beetles than in *Drosophila* (174), since we detected cryptochrome-2 (CRY-2), but not the blue-light photoreceptor cryptochrome (CRY) (175), a transcriptional repressor which is involved in light-resetting of the clock.

Iron and heme-metabolism: Ferritins are ubiquous iron binding proteins involved in iron storage, transport, and antioxidant protection. These multimeric proteins are composed of two subunits, heavy and light chain homologs (HCH and LCH, respectively), in which the former comprise the ferroxidase center responsible for the oxidation of Fe^{2+} and the latter are involved in iron nucleation (176). In *R. prolixus*, we identified five genes that encode HCH and three that encode LCH subunits (Table D1.31 in Dataset). These genes transcribed secreted proteins except for a cytosolic protein that does not present a signal peptide (RPRC013830) and a putative mitochondrial ferritin (RPRC009359). Whole genome analyses of *Dr. melanogaster, An. gambiae and Ap. mellifera* revealed that each taxon encodes two secreted ferritin polypeptides (i.e., one for each subunit), but our analysis of *R. prolixus* revealed six diferent genes encoded for different secreted HCH and LCH subunits (RPRC007320, RPRC009256, RPRC012024 for HCH and RPRC000395, RPRC009258, RPRC012023 for LCH) (Table D1.31 in Dataset). These genes were localized in different regions of the genome and clustered in pairs (one HCH and one LCH) in a head-to-head position, as previously described for other insects (176). This organization suggests that gene expression is transcriptionally coordinated, ensuring that transcript production will match subunit stoichiometry. After gene expression, intracellular iron levels also have the potential to regulate ferritins. For instance, in cytosolic conditions of low iron, IRP (iron responsive binding protein) binds to a stem-loop structure found in the 5´unstranslated region (UTR) of ferritin transcripts, called the iron-responsive element (IRE), and sterically inhibits ferritin translation. This phenomenon is reversible when an iron atom binds to IRP. In most insects, IREs are found only in HCH subunits, and as such, we discovered IREs in the 5´UTR of both secreted HCH transcripts (RPRC007320 and RPRC009256). We also found two genes that transcribe IRP-like proteins in *R. prolixus* (RPRC001246 and RPRC012271). The latter encodes the mitochondrial IRP that possibly has aconitase activity and participates of the tricarboxylic acid cycle pathway, and thus is not involved in regulation of gene expression in the cytosol. The presence of the former gene, as well as IRE-containing transcripts, supports a conserved mechanism for translational control by iron avalability.

Transferrins comprise soluble, monomeric proteins that bind iron with extremely high affinity and have an essential role in iron distribution (177). We identified at least six genes in the *R. prolixus* genome as members of the transferrin superfamily (Table D1.31 in Dataset), including the ortholog (RPRC10050) of hemolymphatic iron-binding transferrin (Trf1). The gene structure was typical of other transferrins, with a large number of introns and a predicted signal peptide for secretion. The remaining genes were typical of membrane-associated melanoferritins since they displayed a large number of introns and a

predicted GPI-anchor domain, however, we were unable to identify transcripts encoded by these genes in the transcriptomes derived from digestive and whole body libraries (178).

In order to understand the mechanisms involved with intracellular homeostasis and heme biosynthesis, we explored the genes involved in iron transport across membranes between cells and organelles. In mammals, a protein named divalent metal transporter-1 (DMT-1) imports cellular iron from intestinal lumen into enterocytes. In this process, diet-derived Fe^{3+} atoms are reduced by duodenal cytochrome-b before transport to the cytosol by DMT-1 (179). *Dr. melanogaster* possesses a single homologous gene to DMT-1 named Malvolio (DmMLV). DmMLV was shown to be highly expressed in the anterior and posterior regions of fly midgut, and mutation of this gene caused depletion of iron stores in the intestine (180). *R. prolixus* possesses two paralogous genes (RPRC012012 and RPRC006515) with high similarity to DmMLV, and the proteins encoded by both genes displayed conserved transmembrane and NMRAMP-like domains. Zinc and iron regulated transporter proteins (ZIP) are found across different cell structures to move both zinc and iron across membranes. At least eight putative ZIP proteins were described in *Dr. melanogaster* (181), with most associated with an intracellular flux of zinc (e.g., dZIP1, dZIP2, and dZIP7). Recently, dZIP13 was described as an iron exporter, since it was associated with pathways involved with iron excretion by secreted ferritins (182). We identified seven genes that encode members of the ZIP family in *R. prolixus* (RPRC00118; RPRC013358 and its paralog RPRC013359; RPRC005556 and its paralog RPRC003454; RPRC009050; RPRC002967). A pair of genes (RPRC009050 and RPRC002967) displayed similarity to dZIP13. Orthologs with similarity to dZIP1 and dZIP3 (RPRC013358 and RPRC013359, respectively), were tandemly arranged in the same direction, suggesting a potential gene duplication event. Mitoferrin (Mrfn) is a member of the mitochondrial solute carrier family and acts to supply the mitochondria matrix with iron required for heme synthesis as well as for the assembly of the iron-sulfur clusters that comprise components for a variety of proteins involved in energy metabolism pathways. As in *Drosophila*, we found a single putative mitoferrin (RPC002819) in the *R. prolixus* genome. Most living organisms synthesize heme, with the exception of some pathogenic bacteria (183), nematodes such as *C. elegans* (184), and the cattle tick *Riphicephalus microplus* (185). We identified all genes that comprise the heme biosynthesis pathway in *R. prolixus* (Table D1.31 in Dataset), supporting previous conclusions that heme is obtained not only from the diet but also by *de novo* synthesis (186).

After host blood is digested, hematophagous insects contend with large amounts of released heme in the lumen of their midguts, and a number of protective strategies against heme-induced damage are known, including the enzymatic degradation of heme catalyzed by heme oxygenase (HO). In *R. prolixus*, it was shown that heme is degraded by an unique pathway that requires the addition of two gluthathione molecules to the heme molecule before breakage of the porphyrin ring (187). In *Dr. melanogaster*, a recombinant HO was structurally characterized (188), and accordingly, we identified a single gene encoding a HO in *R. prolixus* (RPRC006832) with all conserved residues necessary for heme interaction and degradation. Although heme degradation as catalyzed by HO is well understood, the mechanisms by which heme molecules are transported from the lumen into epithelial cells remain unknown. In vertebrates, a transmembrane protein receptor named feline leukemia virus subgroup C (FLCRV-1) was characterized as a cell-suface heme exporter (189), and experimental evidence suggests that FLVCR-2, a highly conserved homolog of FLVCR-1, acts as a heme importer (190). An ortholog of FLVCR was described in *Drosophila* and its inactivation in clock neurons altered individuals' circadian rhythm (191), a process that is known to be modulated by heme in mammals (192). We identified a highly conserved FLVCR ortholog in *R. prolixus* (RPRC015407).

Hemolymphatic proteins bind heme molecules originating in the midgut in order to impair their prooxidant activity (193, 194). One such protein, *Rhodnius* heme binding protein (RHBP; Genbank:AAM11678) transports heme either to pericardial cells for detoxification or to groowing oocytes, where heme is provided for embryogenesis (195-197). RHBP is transcribed by a gene (RPRC004408) that is putatively related to a pair of paralogous odorant binding proteins (OBPs) (RPC000194 and RPRC000560). RHBP is synthetized by fat body in *R. prolixus* during all developmental stages (198), and there is no evidence that RHBP synthesis occurs in cells typically involved with olfation. Futhermore, mass spectrometry analysis revealed that heme is the only ligand found in purified RHBP (198). Thus, apart from the presence of two highly conserved OBP paralogs and a predicted protein domain typical of OBP superfamily, it seems that RHBP evolved to perform a different function, distinct from binding odorants.

Chemoreceptors: The family of odorant receptors (OR) mediates most of insect olfaction (199, 200), while additional contributions are mediated from a subset of the distantly related gustatory receptor (GR) family proteins, such as the carbon dioxide receptors in flies (201-203), or the unrelated ionotropic receptor proteins (IR), which likely evolved from ionotropic glutamate receptors involved in

synaptic transmission (204). With three transmembrane domains comprising a cation channel and an external ligand-binding domain, IRs are larger than ORs or GRs, and they function as obligate heterodimers, with usually two and sometimes three different proteins. While some of these IRs are highly conserved and implicated in olfaction, there are more derived receptors that are implicated in gustation. Like the ORs, and probably many GRs, divergent IRs function in complexes with some of the more highly conserved proteins, specifically IR8a and/or IR25a (205).

We manually annotated the OR, GR, and IR gene families using methods described previously (206, 207). Briefly, TBLASTN (40) searches were performed using aphid and louse proteins as queries, and matching regions were manually assembled into gene models with the assistance of intron splice site predictions using the "Splice Site Prediction by Neural Network" server at the Berkeley Drosophila Genome Project (http://www.fruitfly.org/seq_tools/splice.html). Iterative searches were conducted with each new *Rhodnius* protein as a query until no new genes were identified in each major subfamily or lineage. All of the *R. prolixus* genes and encoded proteins are detailed in Tables D1.24-26 in Dataset. Similar to other draft genome assemblies, some gene models included short gaps or errors that could be repaired with raw reads, but gaps in long repetitive regions could not be repaired. Occasionally, partial gene models were created when genes spanned more than one scaffold, with no support other than the similarity of the coding sequence with homologous genes (Tables D1.24-26 in Dataset). We adopted a cutoff size of 200 amino acids in order to include pseudogenes in the analysis (i.e., 600bp is roughly half the length of a typical insect odorant receptor (ORs) gene or gustatory receptor (GRs) gene). All *Rhodnius*, *Acyrthosiphon*, and *Pediculus* proteins in each family, as well as select other insect GRs and all *Dr. melanogaster* IRs, were aligned in Clustal X v2.0 (69) using default settings, and problematic gene models and pseudogenes were refined based on these alignments. Poorly aligned regions, represented by columns containing almost only gaps, were trimmed before the phylogenetic analysis. We retained regions of uncertain alignment between highly divergent proteins since they provide important information for relationships within subfamilies. We implemented phylogenetic analyses using RAxML (124) with 500 bootstrap interactions (Figures A11-A13 in Appendix).

Among insects, the OR family ranges from ten genes in *Pediculus humanus* (208) to 400 in *Pogonomyrmex barbatus* (209). The OR family consisted of 79 genes in the only other sequenced hemipteroid insect, *Ac. pisum* (210). In addition, *Dr. melanogaster* possessed 60 OR genes that were spatially distributed within the genome (211); only a small proportion were found in small tandem

arrays. Yet tandem arrays were more typical in other insects, especially those with large repertoires, from which it was inferred that larger repertoires resulted from retained gene duplication events generated by unequal crossing over (212). We explored 106 apparently intact *R. prolixus* OR proteins, and of these, nine were missing N- or C-termini or internal exons and their functionality remains uncertain. In addition, the start codons for a large set of genes (OR62-100) were putatively located in a short upstream exon that we could not confidently identify. Twenty gene fragments were not included due to size and incompleteness, but might represent intact genes. The OR automated gene modeling accessed all available insect ORs in GenBank for comparative information, and it succeeded in building at least partial gene models for 79 of these 106 genes. Only five were precisely correct, however, including the highly conserved Orco protein (RPRC000476-PA). All others required at least one change, while 21 new gene models were generated (not including pseudogenes or those requiring repair of assembly gaps or joins across scaffolds) (Tables D1.24-26 in Dataset). Since gene expression occurred at low levels in only a few cells, transcripts were poorly represented in the whole-body transcriptome library. Nevertheless, our manually built gene models were considered reliable because they included multiple genes in gene subfamilies that displayed the same gene structures (Tables D1.24-26 in Dataset). As expected, there was a single conserved ortholog of Orco (DmOR83b) that shared 59% amino acid identity with both ApOrco and PhOrco. There were no other simple orthologous relationships between *Rhodnius*, *Acyrthosiphon*, and *Pediculus* ORs, except perhaps for the RpOR101/102 proteins with the PhOR3a/b proteins (Figure A11 in Appendix). Instead, as is common for these rapidly evolving proteins in such divergent taxa, there were differential gene subfamily expansions. For comparison, the aphid ORs consisted of three ancient lineages (ApOR2-4), and two relatively recent expansions (ApOR5-13 and 14-79) (210), while the louse ORs were reduced to a set of eleven proteins, with three divergent lineages represented by single genes (PhOR2-4), and the rest represented by a small cluster of relatively old proteins (208) (Figure A11 in Appendix). The *Rhodnius* ORs also contained several old divergent lineages containing single genes (e.g., OR1 [RPRC000579], OR2 [RPRC001689-PA], OR103-105 [RPRC000059; RPRC000120, and RPRC000083]; OR107 [RPRC001726-PA]; and OR112 [RPRC000555]) (Figure A8 in Appendix). But, the vast majority of ORs were in small, expanded subfamilies, some reasonably old, such as OR43-52 (Figure A8 in Appendix), and some very young, such as OR58-87 (Figure A8 in Appendix). Many of the latter are in a single tandem array in one scaffold (Tables D1.24 in Dataset), while smaller tandem arrays comprised the remainder. These *Rhodnius* OR gene subfamily expansions displayed

variable gene structures regarding the introns within the coding regions (Tables D1.24 in Dataset). RpOR1-52 shared a structure characterized by a long first exon, followed by a phase 2 intron, and then four short exons separated by phase 0 introns, which appeared to correspond to the widely present final three phase 0 introns in other insect OR genes (213). OR33 (RPRC000235) and OR52 (RPRC000201) displayed additional idiosyncratic introns interrupting this first long exon. All the remaining genes showed multiple additional introns interrupting the first exon, with up to a maximum of ten introns in OR106 (RPRC000371). The set of 2-0-0-0 introns was also shared with the louse and aphid OR genes, however the aphid genes also commonly possessed earlier introns. Finally, because these ORs were so divergent from both the aphid and louse ORs, and indeed from all other insect ORs, expression studies focused on the antennae will hopefully provide clues to their possible ligands and reveal which ORs are highly expressed and/or sex-specific.

Among insects, the GR family ranges in number from six genes that encode eight proteins in the human body louse (208) to 215 genes that encode 245 proteins in the flour beetle (140). The RpGR gene set consisted of 28 gene models, which encoded 30 proteins. This gene number was smaller than that of most other insects except for *Pediculus humanus* (208), *Ap. mellifera* (212), *Ceratosolen solmsi* (214), and *Glossina morsitans* (215). There were no obvious pseudogenes, although a few fragments were apparent in the genome. Two genes displayed alternative splicing, presenting a pattern similar to several GRs in flies and some other insects, with long first exons alternatively spliced into three shared short C-terminal exons. These particular models remain hypothetical, however, due to the absence of transcriptome evidence. As some of these proteins were very divergent after TBLASTN searches, we added a PSI-BLASTP search of the automated annotations with two iterations, revealing one more highly divergent GR4 protein (RPRC000056). The GR automated gene modeling accessed all available insect GRs in GenBank for comparative information, and it succeeded in building at least partial gene models for fifteen genes (54%). While only one gene model was precisely correct, all others required at least one manual correction, allowing for an additional thirteen suitable gene models (Tables D1.25 in Dataset). Although there were no ESTs for these GRs in the available transcriptome data, the basic gene structure for the entire RpGR set displayed a long initial exon, followed by three short C-terminal exons separated by three phase 0 introns. There were several exceptions: GR1-4 (RPRC001549; RPRC001795; RPRC002023 and RPRC000056) possessed an additional intron splitting the long first exon; GR26-28 (RPRC000368; RPRC000290 and RPRC000068) showed two such introns; and GR20 (RPRC000451) lost all three introns. There were no obvious members of the sugar receptor subfamily

(represented by AmGR1/2 and ApGR1-6 in Figure A12 in Appendix), and we also did not find members of the highly conserved carbon dioxide receptor subfamily (GR21a and 63a in *Dr. melanogaster* and TcGR1-3 in *Trib. castaneum*) (Figure A12 in Appendix). Instead, the *Rhodnius* GRs consist of a highly divergent protein (GR2) and an expanded subfamily unique to *Rhodnius* (GR3-28). The long branches to the divergent lineage and even within this subfamily were similar to the *Apis* and *Pediculus* proteins, in contrast to most of the aphid GRs, which, as several recently expanded gene subfamilies, were putatively derived as a result of positive selection (210). The more recent lineages in the *Rhodnius* subfamily, which putatively encode bitter taste receptors, were located in a tandem array (GR5-11).

We identified thirty-three IRs in the *Rhodnius* genome, with no observed pseudogenes. We used preliminary information from an antennal transcriptome to improve several of these, especially at their N-termini. Our automated predictions generated gene models for 23 IRs, and three were identical to manually curated gene models. We generated ten new gene models that were not in the set of automated predictions (Tables D1.26 in Dataset). Our naming convention followed that employed for the termite genome (216), and the improved pea aphid and human body louse IR models reported therein were utilized here. In addition to the expected highly conserved orthologs of DmIR25a (RPRC000589) and 8a (RPRC000349), which have 60-73% and 45-61% identity to their orthologs, respectively, we found single orthologs of the highly conserved DmIR21a (RPRC001826), 40a (RPRC000383), 68a (RPRC000328), 76b (RPRC000469), and 93a (RPRC008486) lineages that were also present in both the pea aphid and human body louse as single orthologs. *Rhodnius* possessed three copies homologous to the DmIR41a gene, which appeared lost from the aphid, while the IR75 genes showed their greatest expansion in *Rhodnius* in the chemoreceptor superfamily, with sixteen genes. Apart from these findings, the IRs in *Rhodnius* displayed diversification comparable to the pea aphid and body louse IRs, with seven divergent proteins (RpIR101-107) that clustered with some of the pea aphid and body louse IRs (Figure A13 in Appendix). This is in considerable contrast to *Drosophila*, where the IRs expanded and diversified in two subfamilies, as well as the termite *Zootermopsis nevadensis*, which displayed a major expansion of divergent IRs (216).

Chemoreception accessory proteins: We retrieved odorant binding protein (OBP) and chemosensory protein (CSP) sequences from GenBank and from the literature (217) for species covering seven insect orders, including Hemiptera (*Dr. melanogaster*, *An. gambiae*, *Bo. mori*, *Trib. castaneum*, *Ap. mellifera*,

Pediculus humanus, *Ac*. *pisum*, and *Locusta migratoria*). These sequences were queried against the *R*. *prolixus* genome and transcriptome using a variety of search algorithms such as tBLASTn (40), BLASTP (40), and Genewise (160). These alignments were used to identify the transcripts and the automated gene models for OBPs and CSPs, and the gene models were manually validated or corrected considering their alignment with OBPs and CSPs transcript sequences of *R*. *prolixus*. Candidate *R*. *prolixus* OBPs and CSPs were numbered, when possible, according to their closest *Ac*. *pisum* homologues identified in the phylogenetic analysis. Consecutive numbers were assigned to OBP and CSP sequences present on the same scaffold. Mature (without signal peptide) OBP and CSP protein sequences from *R*. *prolixus* (this study), *P*. *humanus* (218), and *Ac*. *pisum* (217, 219) were aligned using MAFFT v6 (123) with advanced settings (E-INS-i with BLOSUM62 scoring matrix, 1000 maxiterate and offset 0.1). We visually inspected and manually adjusted these alignments, as needed, using BioEdit v7.05.3 (165). After removing major gaps and too short OBP sequences (less than 50 AA), we retained 48 and 40 sequences for OBP and CSP alignments, respectively. To select best-fit models of amino acid evolution, we used ProtTest3 v.3.2 (220) and identified the model in the candidate list with the smallest Akaike Information Criterion, Bayesian Information Criterion score, and Decision Theory Criterion. The MtMam and MtArt models were retained for OBP and CSP protein evolution, respectively, and used for the Maximum likelihood tree reconstructions performed using PhyML v3.0 (221) with 500 bootstrap replicates. We created tree representations using the iTOL web server (222).

We identified 27 putative OBPs and nineteen putative CSPs in the *R*. *prolixus* genome (Tables D1.27 in Dataset), and most exhibited the characteristic features of OBPs and CSPs, such as the presence of a signal peptide and the six (OBPs) or four (CSPs) conserved cysteine pattern. Interestingly, *R*. *prolixus* exhibited a higher number of OBPs and CSPs than the other Paraneoptera (*Ac*. *pisum* [N=18 OBPs; N=13 CSPs]; *P*. *humanus* [N=5 OBPs; N=7 CSPs]) or Hemiptera (*Adelphocoris lineolatus* [N=16 OBPs]; *Apolygus lucorum* [N=13 OBPs]) for which such genes have been identified to date (208, 217, 219, 223, 224). Compared to other insect orders, such as Diptera and Lepidoptera, Hemiptera possess fewer OBP genes (e.g., *Dr. melanogaster* [N=52]; *Bo. mori* [N=46]) (225, 226). At the same time, the number of CSP genes is higher than that observed in Diptera (e.g., *Dr. melanogaster* [N=4]) (226), but smaller than in Lepidoptera (*Bo*. *mori* [N=24]; *Heliconius melpomene* [N=33]; *Danaus plexippus* [N=34]) (227, 228). Reduced numbers of OBP and CSP genes in some species may be the result of highly specialized ecologies and/or parasitic lifestyles (229, 230), however, our *R*. *prolixus* data

revealed a higher number of CSPs than in the two other hematophagous insects for which CSPs were annotated (*An*. *gambiae* [N=8]; *P*. *humanus* [N=7]). Several *R*. *prolixus* OBPs and CSPs were organized in large clusters, where nine OBPs (OBP6-OBP14) and six CSPs (CSP1-CSP6) localized along the same scaffold (Figure A14A in Appendix).

Phylogenetic analyses revealed expansions of *R*. *prolixus* OBPs and CSPs characterized by divergent clades comprising multiple protein copies with high bootstrap support. For instance, we noticed CSPs from some clades were clustered in same scaffolds, including CSP1-6, CSP12-16 and CSP10, 11 and 19 (Figure A14B in Appendix). The CSP phylogeny was robust for most clades; seven distinct clades were retained, in general agreement with the phylogenetic relationships within *Ac*. *pisum* (N=7) and *P*. *humanus* (N=5) (Figure A14B in Appendix). The protein sequences for OBPs were less conserved than for CSPs. Nevertheless, we observed three expansions within the former category, including copies localized in the same scaffold (OBP9-11; OBP18-20; OBP22 and OBP26-27) (Figure A14C in Appendix). Moreover, two clades with two copies each were well supported (OBP1-2; OBP6 and OBP17). These multiple copies were monophyletic and could be considered paralogous since some of them were localized to the same scaffold. Conversely, some CSPs (e.g., CSP7-8, CSP16) were found as single copies with phylogenetic affinities to CSPs of other species and hence, were more likely orthologous. We observed the same for OBPs (OBP1-2; OBP4, OBP5, OBP6 and OBP17).

Proteases: Peptidases (E.C. 3.4) have several metabolic roles, including protein and peptide turnover, hormone processing, protein secretion and trafficking, immune system activation, and apoptosis (231). In hematophagous bugs, peptidases in the posterior midgut are putatively involved with the essential digestion of ingested blood proteins (e.g., hemoglobin, fibrinogen and other plasma proteins) (232). We therefore hypothesized that evolutionary pressure for efficient blood digestion promoted the expansion of peptidase gene families in the *R. prolixus* genome. This process could be related to the recruitment of specific genes for blood digestion or to broaden the arsenal of gut hydrolases for recognition and cleaving of different peptides. In contrast to other Hemiptera insects, which use trypsin and chymotrypsin as major digestive proteases, cysteine (cathepsin L-like) and aspartic (cathepsin-D like) peptidases (233) comprise the main endopeptidase activities in *R*. *prolixus*. Proteases are classified by their action pattern (exo- or endopeptidases), mechanism of catalysis (cysteine, serine, aspartic, metallo, and threonine peptidases) and by structural similarities in families (234). Using these classifications, we found 433 genes belonging to 71 families containing a protease PFAM signature in

the genome of *R. prolixus* (Table D1.29 in Dataset). All genes were representatives of peptidase families (e.g., aspartic acid, cysteine, metallo, asparagine, serine, and threonine), and a subsequent comparison among homologous gene sequences between different insects (*R. prolixus*, *Ac. pisum*, *An. gambiae* and *Dr. melanogaster*) suggested that protease gene acquisition and gene family expansion in *R. prolixus* occurred in families A1 (D-like cathepsins), C2 (calpains), M17 (aminopeptidases) and S29 (hepacivirin). Protease gene expansion likely occurred through either gene duplication (A1, C2, M17) or horizontal transmission from bacteria (S29). In the former, protease genes found in *R. prolixus* were homologous to other insect proteases, but displayed higher similarity to *R. prolixus* protein sequences. Alternatively, S29 proteases in *R. prolixus* were homologous to bacterial proteases, and we found no homologous insect proteases. In these cases, the bacteria of origin were related to the genera *Pantoea*, which commonly occur in the gut of Hemipterans (235).

Salivary proteins: Anticlotting agents that interfere with platelet aggregation and vasoconstrictor responses have been described in the saliva or salivary gland homogenates of blood sucking animals (236), including *Rhodnius* (237, 238). *Rhodnius* saliva contained anti-platelet, anti-histaminic, anti-thromboxane, and anti-serotonin activity (i.e., vasoconstrictors released by activated platelets and mast cells) (239), as well as specialized hemoproteins known as nitrophorins that carry the vasodilator nitric oxide (240, 241). Although NO-carrying nitrophorins are thought to be exclusive to *Rhodnius*, they belong to the lipocalin family, whose members were also found in other triatomine genera (239). A previous analysis of the *R. prolixus* salivary transcriptome revealed expansions of the lipocalin family including the nitrophorin and triabin clades (242). At the same time, previous analyses of salivary transcriptomes for *Triatoma brasiliensis* (243)*, T. infestans* (244), *T. dimidiata* (245), *T. rubida* (246), and *T. matogrossensis* (247) purported to find proteins that were specific to *Triatoma.* When we sought to identify salivary proteins similar to those of *Triatoma* that were not previously identified in *Rhodnius,* our predictions for *R. prolixus* included putative homologs for triatox (RPRC006445-PA) and a Kazal-domain containing peptide (RPRC009790-PA). The latter was similar to a vasodilator named Vasotab that was found in the salivary glands of tabanids flies and triatomines (248). While many gene predictions are fragmented due to the nature of the assembly, we discovered twelve novel members of the nitrophorin clade (Figure 3B). In addition, many salivary proteins displayed tandem representation, including nitrophorins (RPRC013026, RPRC000380; RPRC000072, RPRC000367; RPRC000284, RPRC000061); triabin-like proteins (RPRC012954-PA and PB, and RPRC000275-PA); and lipocalins (RPRC015421, RPRC015426, RPRC015425 and RPRC015420) (Table D1.28 in

Dataset). This organization suggests adaptation by gene duplication. These proteins showed higher expression in the salivary glands compared with other tissues (178), thereby supporting their associations with blood feeding adaptations.

The saliva of *Rhodnius* also contains salivary apyrase, an enzyme that prevents platelet aggregation. While salivary apyrase was first described in *Rhodnius* (249, 250), it was never molecularly characterized, despite its subsequent characterization in mosquitoes (251), bed bugs (252), sand flies (253, 254) and *Triatoma infestans* (255).While salivary apyrases in mosquitos and *Triatoma* were shown to depend on either Ca^{++} or Mg^{++} for activity (256), those of bed bugs and sand flies belong to a novel class of enzymes first discovered in *Cimex* that showed strict dependence on Ca^{++} (252, 257). By running a rpsblast search using the pfam06079 motif, we discovered that many organisms possessed only one gene coding for this enzyme, including *Daphnia pulex*, *Pediculus humanus*, *Triboleum castaneum*, *Ac. pisum, Atta cephalotes*, *Solenopsis invicta*, *Bo. mori*, *Ceratitis capitata*, *Nasonia vitripennis*, *Dr. melanogaster*, *Ae. aegypti*, and *Culex quinquefasciatus*, however, we found three such matches in *R. prolixus*. One match appeared complete (RPRC000189-PA) and was most similar to *P. humanus*. A second match (RPRC000276-PA) was most similar to apyrase in *Cimex*. The third apyrase (RPRC003476-PA) contained a single exon and was likely an artifact of the assembly. The consensus phylogenetic tree (Figure A15 in Appendix) showed strong bootstrap support for a clade of the Hemiptera proteins, and the placement of the *Rhodnius* proteins were indicative of a relatively recent gene duplication event.

Sodium channels: *R. prolixus* ability of filtering out the water from the blood meal producing a hypoosmotic urine is well know and was firstly described many years ago (258). This process is carried out in the malpighian tubules (MT) where the ion movements start with an apical vacuolar-type proton-ATPase pump followed by the secondary cation/proton transporters (259). These cation/H^+ exchangers are typically sensitive to amiloride (260) but recently, Paluzzi, Yeung and O'Donnell (261) have shown that amiloride only blocked 5-hydroxytryptamine (5-HT) stimulated secretion in *R. prolixus* MT, but it didn't block the corticotropin releasing factor-related peptide (RhoprCRF) stimulated secretion. They proposed that this inhibition could be a consequence of the know amiloride antagonism of 5-HT receptors and that it is in deed a relatively weak inhibitor of insect MT sodium/proton channels. We identified a LSR of the gene family of amiloride-sensitive sodium channels, represented by the domain (IPR001873). The phylogenetic tree of the proteins containing sodium channel amiloride-sensitive

domain (IPR001873) from *R. prolixus* and the other 15 genomes used in this manuscript (Figure A16 in Appendix) showed that *R. prolixus* don't have the expansions that are present in almost all others genomes analyzed and also that some monophyletic clades don't have any *R. prolixus* sequence, an evidence of gene lost. This LSR could be related to the limited cation transport to the MT lumen producing the hypoosmotic unine and also to the unexpected amiloride insensitiveness (261) but further analyses are needed to elucidate this topic.

Cuticle: The variation of the mechanical properties of the cuticle is critical in determining the body shape of insects. In *Rhodnius*, this is particularly evident during feeding, when the cuticle is expanded to allow blood ingestion. We found that many genes related to the attachment, metabolism, and sclerotization of the cuticle comprise different gene expansions. First, we annotated 117 cuticle protein (CP) encoding genes from nine families (Table D1.22 in Dataset), and detected gene expansions in several families. For instance, in the CPR family, we found 26 gene expansions that contained 43 members (Figures A7- A8 in Appendix). We detected large expansions within one particular subfamily (RR-2) (Figure A8 in Appendix), with one containing seven genes. We also detected gene expansions in the CPF, Tweedle, CPLCP (Figure A9 in Appendix) and CPAP families (Figure A10 in Appendix). Whereas nine of ten members of the CPAP1 gene family were absent in several other arthropods, including *Ac. pisum*, *Trib. castaneum* (262), we found seven members in *R. prolixus*. Next, we annotated fifteen sequences in *R. prolixus* that putatively code for FAR genes. Six of these comprised a *R. prolixus*-exclusive FAR expansion (Table D1.23 in Dataset), including a tandem (RPRC013997, RPRC014002, RPRC013998, RPRC014004) along scaffold GL562731 along with two additional genes (RPRC000880 and RPRC006662). Furthermore, four laccase genes form a pair of *Rhodnius*-specific expansions. One gene family putatively related to chitin acetylation (proteins containing acyltransferase 3 domain - ACT3- IPR002656) showed a LSR (Table D1.30 in Dataset). The phylogenetic tree of the ACT3 proteins (Figure A17 in Appendix) from *R. prolixus* and the other 15 species used for comparison in this manuscript confirmed the absence of ACT3 gene expansions in *R. prolixus*, which happened in almost all others genomes analyzed. The tree showed also ACT3 lost, highlighted by some monophyletic clades that don't have any *R. prolixus* sequence. Proteins containing ACT3 domain are mainly known from bacteria and are involved in acetylation of many substrates including peptidoglicans and lipo-chitin-oligosaccharides to prevent polymer digestion/degradation (263, 264). In *R. prolixus* this LSR could be related to avoiding the over-hardening of the midgut or

cuticle chitin polymer. Experiments should seek to elucidate the roles that these species-specific expansions and reductions have in the sclerotization of the *R. prolixus* cuticle.

At the same time, we failed to detect expansions in other gene families. For instance, although several expansions of the CPLCP family were reported in mosquitoes and flies (265), we detected no expansion in *R. prolixus* with regards to the thirteen annotated genes. We next explored dumpy and dusky, since these are members of a family of transmembrane proteins that control the properties of the overlying cuticle. Dumpy is a gigantic fibrillar protein that mediates mechanical maintenance at tension sites of cuticle–epidermal cell attachment (266), and dusky is involved in the interactions between the apical membrane, the cytoskeleton, and the forming cuticle (267). We identified a pair of genes coding for dumpy and seven genes that coded for dusky (Table D1.22 in Dataset). Recently the presence of chitin in the midgut was demonstrated (268) and we also annotated nine chitinase-like genes in *R. prolixus* (Table D1.22 in Dataset). This relatively low number might be related to incomplete metamorphosis, as suggested for *Ac. pisum,* another hemimetabolous insect.

Glycolysis, gluconeogenesis and pentose-phosphate pathways: As in other insects (269), we identified genes from sugar central energy metabolism pathways, some of which had multiple copies (Table D1.34 in Dataset). We did not detect genes coding for carotenoid biosynthesis, as reported for the aphid *Ac. pisum* (270).

Lipid metabolism: Products from digestion in arthropods, such as metabolites from the blood, contribute to lipid stores and are used to sustain diverse metabolic activities, such as flight, cuticle production and oogenesis. In *R. prolixus*, triacylglycerol synthesis occurs by the glycerol-3-phosphate (G3P) pathway, and in accordance with other studies (271), we failed to identify genes for the monoacylglycerol (MAG) pathway, which is present in mammals. A rate-limiting step in the G3P pathway is a reaction catalyzed by G3P acyltransferase (GPAT), and in mammals, four isoforms of GPAT are encoded by independent genes: two mitochondrial (GPAT1 and 2) and two localized in the endoplasmic reticulum (GPAT3 and 4) (272). We identified a single-copy gene (RPAL013444) that matched the mitochondrial GPAT isoforms, as previously reported (273), as well as a putative second isoform (RPAL016510) that was similar to the mammalian microsomal GPATs (274).

We identified additional components of glycerolipid and sterol synthesis. Among mammals, for instance, dihydroxyacetone-phosphate aciltransferase (DHAPAT) likely contributes to glycerolipid

synthesis via the acylation of dihydroxyacetone-P. We discovered a DHAPAT-like protein (RPRC15873) that is homologous to sequences present in other arthropods. In addition, fatty acids for glycerolipid synthesis are either derived from lipids present in the blood meal or synthesized *de novo* through the production of malonyl-CoA. Malonyl-CoA has a dual role, as a substrate of fatty acid synthase during fatty acid synthesis (275) and as an inhibitor of the mitochondrial transport system for fatty acid oxidation (276). This molecule is catalyzed by acetyl-CoA carboxylase (ACC), for which a single-copy was found (RPRC14457), in accordance with other arthropods. Unlike mammals, insects cannot synthesize sterols from acetyl-CoA (277), so the absorption and intracellular transport of sterols are of special interest. In *Dr. melanogaster*, for instance, mutations in *NPC1A*, which is involved in transporting intracellular cholesterol, caused early larval lethality (278). As in the mosquito, *Culex quinquefasciatus* (279), *R. prolixus* possessed an *NPC1A* duplication (RPRC000135 and RPRC000480). We also explored the regulatory mechanisms of lipid metabolism by identifying two genes (RPRC000287 and RPRC000382) for phosphoethanolamine cytidylyltransferase (PECT), which takes part in the synthesis of phosphatidylethanolamine (PE) (280). PE regulates the processing and activation of the sterol responsive element binding protein (SREBP) (281, 282) and can induce the expression of genes involved in fatty acid synthesis.

Lipids, such as diacylglycerol (DAG), also serve as signaling molecules. To regulate its intracellular levels, DAG kinases (DGKs) convert DAG to phosphatidic acid (PA). We identified three DGK sequences (RPRC14703, RPRC08287 and RPRC15770) that transcribe different isoforms (β, θ, and η), in accordance with mammalian classification (283). An additional DGK sequence (RPRC13417) exhibited similarities to *Drosophila* eye-specific DGK, in which modifications contributed to retinal degeneration (284). Other signaling molecules, including lysophospholipids and acylglycerol kinase (RPRC02130), which generates lysophosphatidic acid (LPA) and PA, were shown to enhance the transmission of *T. cruzi* by *R. prolixus* (285).

Amino acid synthesis: A curated set of genes coding for the essential amino acid (EAA) synthesis pathway enzymes in *Saccharomyces cerevisiae* was built according to a published gene set (286), which mined literature data as well as KEGG, KOG, GenBank and other databases. Another curated dataset was built for the enzymes coding for the de novo pathways of non-essential amino acids (NEAA) biosynthesis. Here, we used the fruit fly as model organism due to its evolutionary proximity to the organisms analyzed here. BLAST was run against each dataset and the *R. prolixus* genome. We

mainly ran tBLASTn (40) analyses using protein queries against translated versions of the genome. Other BLAST searches were performed in the NCBI platform to confirm results obtained by manual curation. Although we used standard default e-values, our results indicated that an e-value of $10e^{-05}$ would be suitable for more automatic analysis. We calculated coverage information for proteins from each model organism aligned to *R. prolixus* gene models. We used the Needleman-Wunsch algorithm (287), as implemented in the EMBOSS package (64), to align the sequences and we developed a custom script to parse the output and calculate the coverage.

Regarding EAA biosynthesis in *R. prolixus*, most genes were concordant with other animal genomes. This includes the ten genes conserved in metazoans, even though their partners for EAA biosynthetic pathways are deleted. A single exception existed for paralogous genes that encode the branched-chain aminotransferase. Most animals possess two copies: a cytosolic version and another targeted to the mitochondria, both of which are encoded in the nuclear genome. Genome annotation in Hemiptera, however, including our annotation for *R. prolixus*, showed that all hemipterans possess only the cytosolic version of the enzyme. Furthermore, when comparing the presence/absence matrix for EAA enzymes found among *R. prolixus* and metazoans, we found four exceptions that are known to be absent in metazoans but were predicted as genes in the *R. prolixus* genome (RPRC008868, RPRC006919, RPRC012609, RPRC009489). These proteins were highly similar to bacterial genes encoding the prokaryotic version of the enzymes, and although the sequences were possibly laterally transferred to *R. prolixus*, we cannot exclude the possibility that portions of the genome were misassembled with contaminated sequence data derived from bacteria. Despite this, one of these genes, phospho-2-dehydro-3-deoxyheptonate aldolase (RPRC008868), was predicted in pea aphid (118) and *Apis mellifera* (288), but not for other insect genomes. Finally, we used NEAA's from *Dr. melanogaster* to search for *Rhodnius* orthologs. Although the taxa diverged over 350 million years ago (289), we expected general conservation of enzymes essential for metabolism as a result of negative selective pressures. As expected, most enzymes participating in the anabolism of NEAA were found in the genome (Table D1.36 in Dataset).

Reference

1. Monteiro FA*, et al.* (2003) Molecular phylogeography of the Amazonian Chagas disease vectors *Rhodnius prolixus* and *R. robustus*. *Mol Ecol* 12(4):997-1006.
2. Pavan MG*, et al.* (2013) A nuclear single-nucleotide polymorphism (SNP) potentially useful for the separation of *Rhodnius prolixus* from members of the *Rhodnius robustus* cryptic species complex (Hemiptera: Reduviidae). *Infect Genet Evol* 14:426-433.
3. Pavan MG , Monteiro FA (2007) A multiplex PCR assay that separates *Rhodnius prolixus* from members of the *Rhodnius robustus* cryptic species complex (Hemiptera: Reduviidae). *Tropical Medicine & International Health* 12(6):751-758.
4. Justi SA, Noireau F, Cortez MR, Monteiro FA (2010) Infestation of peridomestic *Attalea phalerata* palms by *Rhodnius stali,* a vector of *Trypanosoma cruzi* in the Alto Beni, Bolivia. *Tropical Medicine & International Health* 15(6):727-732.
5. Ribeiro JMC*, et al.* (2014) An insight into the transcriptome of the digestive tract of the bloodsucking bug, *Rhodnius prolixus*. *PLoS Negl Trop Dis* 8(1):e2594-e2594.
6. Otto TD, Gomes LHF, Alves-Ferreira M, de Miranda AB, Degrave WM (2008) ReRep: Computational detection of repetitive sequences in genome survey sequences (GSS). *BMC Bioinformatics* 9(1):366-366.
7. Abubucker S*, et al.* (2008) The canine hookworm genome: Analysis and classification of *Ancylostoma caninum* survey sequences. *Mol Biochem Parasitol* 157(2):187-192.
8. Tatusova TA , Madden TL (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* 174(2):247-250.
9. Hoskins RA*, et al.* (2002) Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. *Genome Biol* 3(12):research0085.0081-0085.0016.
10. Carvalho AB*, et al.* (2003) Y chromosome and other heterochromatic sequences of the *Drosophila melanogaster* genome: How far can we go? *Genetica* 117(2-3):227-237.
11. Skaletsky H*, et al.* (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423(6942):825-837.
12. Hall AB*, et al.* (2013) Six novel Y chromosome genes in *Anopheles* mosquitoes discovered by independently sequencing males and females. *BMC Genomics* 14:273-273.
13. Carvalho AB , Clark AG (2013) Efficient identification of Y chromosome sequences in the human and *Drosophila* genomes. *Genome Res* 23(11):1894-1907.
14. Krzywinski J, Nusskern DR, Kern MK, Besansky NJ (2004) Isolation and characterization of Y chromosome sequences from the African malaria mosquito *Anopheles gambiae*. *Genetics* 166(3):1291-1302.
15. Carvalho AB, Koerich LB, Clark AG (2009) Origin and evolution of Y chromosomes: *Drosophila* tales. *Trends Genet* 25(6):270-277.
16. Lahn BT , Page DC (1997) Functional coherence of the human Y chromosome. *Science* 278(5338):675-680.
17. Holt RA*, et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298(5591):129-149.
18. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinformatics* 21(Suppl 1):i351-i358.
19. Bao Z , Eddy SR (2002) Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res* 12(8):1269-1276.
20. Smit AFA, Hubley R, Green P (2010) RepeatMasker Open-3.0.
21. Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics* 5:59-59.
22. Stanke M, Schöffmann O, Morgenstern B, Waack S (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 7:62-62.
23. Slater GSC , Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31-31.
24. UniProt C (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 42(Database issue):D191-D198.
25. Curwen V*, et al.* (2004) The Ensembl automatic gene annotation system. *Genome Res* 14(5):942-950.
26. Blanco E , Abril JF (2009) Computational gene annotation in new genome assemblies using GeneID. *Bioinformatics for DNA Sequence Analysis*, ed Posada D (Humana Press, New York), Vol 537, pp 243-261.
27. Blanco E, Parra G, Guigo R (2007) Using geneid to identify genes. *Curr Protoc Bioinformatics* Chapter 4:Unit 4.3-4.3.28.

28. Camacho C*, et al.* (2009) BLAST+: Architecture and applications. *BMC Bioinformatics* 10:421-421.
29. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* 8(9):967-974.
30. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat Meth* 8(10):785-786.
31. Ribeiro JMC, Topalis P, Louis C (2004) AnoXcel: an *Anopheles gambiae* protein database. *Insect Mol Biol* 13(5):449-457.
32. Leidenroth A , Hewitt JE (2010) A family history of *DUX4*: Phylogenetic analysis of *DUXA*, *B*, *C* and *Duxbl* reveals the ancestral *DUX* gene. *BMC Evolutionary Biology* 10:364.
33. Consortium THGS (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443:931-949.
34. Nielsen H, Brunak S, von Heijne G (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng* 12(1):3-9.
35. Sonnhammer ELL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, ed Glasgow J (AAAI Press),papers3://publication/uuid/820B2DB0-5097-41D3-9225-F0EF7DC43604, pp 175-182.
36. Hansen JE*, et al.* (1998) NetOglyc: Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconj J* 15(2):115-130.
37. Duckert P, Brunak S, Blom N (2004) Prediction of proprotein convertase cleavage sites. *Protein Eng Des Sel* 17(1):107-112.
38. Fischer S*, et al.* (2011) Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr Protoc Bioinformatics* papers3://publication/doi/10.1002/0471250953.bi0612s35:6.12 11-19.
39. Jones P*, et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9):1236-1240.
40. Altschul SF*, et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(0):17.
41. Jurka J*, et al.* (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110(1-4):462-467.
42. Tubío JMC, Naveira H, Costas J (2005) Structural and evolutionary analyses of the Ty3/gypsy group of LTR retrotransposons in the genome of *Anopheles gambiae*. *Mol Biol Evol* 22(1):29-39.
43. Tubío JMC*, et al.* (2011) Evolutionary dynamics of the Ty3/gypsy LTR retrotransposons in the genome of *Anopheles gambiae*. *PLoS One* 6(1):e16328-e16328.
44. Tu Z (2001) Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. *Proc Natl Acad Sci USA* 98(4):1699-1704.
45. Higgins DG , Sharp PM (1988) CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene* 73(1):237-244.
46. Kent WJ (2002) BLAT - the BLAST-like alignment tool. *Genome Res* 12(4):656-664.
47. Gilbert C, Schaack S, Pace JK, 2nd, Brindley PJ, Feschotte C (2010) A role for host-parasite interactions in the horizontal transfer of transposons across phyla. *Nature* 464(7293):1347-1350.
48. O'Neill SL , Karr TL (1990) Bidirectional incompatibility between conspecific populations of Drosophila simulans. *Nature* 348(6297):178-180.
49. Rousset F, Vautrin D, Solignac M (1992) Molecular identification of Wolbachia, the agent of cytoplasmic incompatibility in Drosophila simulans, and variability in relation with host mitochondrial types. *Proceedings Biological sciences / The Royal Society* 247(1320):163-168.
50. Stouthamer R, Breeuwer JA, Hurst GD (1999) Wolbachia pipientis: microbial manipulator of arthropod reproduction. *Annual review of microbiology* 53:71-102.
51. Werren JH, Baldo L, Clark ME (2008) Wolbachia: master manipulators of invertebrate biology. *Nature reviews Microbiology* 6(10):741-751.
52. Zug R , Hammerstein P (2012) Still a host of hosts for Wolbachia: analysis of recent data suggests that 40% of terrestrial arthropod species are infected. *PLoS One* 7(6):e38544.
53. Brelsfoard C*, et al.* (2014) Presence of extensive *Wolbachia* symbiont insertions discovered in the genome of its host *Glossina morsitans morsitans*. *PLoS Negl Trop Dis* 8(4):e2728-e2728.
54. Klasson L, Kambris Z, Cook PE, Walker T, Sinkins SP (2009) Horizontal gene transfer between Wolbachia and the mosquito Aedes aegypti. *BMC Genomics* 10:33.
55. Nikoh N*, et al.* (2008) Wolbachia genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes. *Genome Res* 18(2):272-280.

56. Kondo N, Nikoh N, Ijichi N, Shimada M, Fukatsu T (2002) Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect. *Proc Natl Acad Sci U S A* 99(22):14280-14285.
57. Woolfit M, Iturbe-Ormaetxe I, McGraw EA, O'Neill SL (2009) An ancient horizontal gene transfer between mosquito and the endosymbiotic bacterium Wolbachia pipientis. *Mol Biol Evol* 26(2):367-374.
58. Jühling F*, et al.* (2009) tRNAdb 2009: Compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res* 37(Database issue):D159-D162.
59. Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* 33(Web Server issue):W686-W689.
60. Li W , Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658-1659.
61. Brownlie JC , O'Neill SL (2005) Wolbachia genomes: insights into an intracellular lifestyle. *Current biology : CB* 15(13):R507-509.
62. Foster JM*, et al.* (2009) The Wolbachia endosymbiont of Brugia malayi has an active phosphoglycerate mutase: a candidate target for anti-filarial therapies. *Parasitology research* 104(5):1047-1052.
63. Ulitsky I*, et al.* (2010) Expander: From expression microarrays to networks and functions. *Nat Protoc* 5(2):303-322.
64. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16(6):276-277.
65. Hofacker IL (2009) RNA secondary structure analysis using the Vienna RNA package. *Curr Protoc Bioinformatics* papers3://publication/doi/10.1002/0471250953.bi1202s26:12.12.11-12.12.16.
66. Burge SW*, et al.* (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* 41(Database issue):D226-D232.
67. Jiang P*, et al.* (2007) MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res* 35(Web Server issue):W339-W344.
68. Zhang B, Pan X, Cannon CH, Cobb GP, Anderson TA (2006) Conservation and divergence of plant microRNA genes. *Plant J* 46(2):243-259.
69. Larkin MA*, et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21):2947-2948.
70. Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004) WebLogo: A sequence logo generator. *Genome Res* 14(6):1188-1190.
71. Enright AJ*, et al.* (2003) MicroRNA targets in Drosophila. *Genome Biol* 5(1):R1.
72. Campo-Paysaa F, Semon M, Cameron RA, Peterson KJ, Schubert M (2011) microRNA complements in deuterostomes: origin and evolution of microRNAs. *Evolution & development* 13(1):15-27.
73. Burge SW*, et al.* (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* 41(Database issue):D226-232.
74. An L, Liu Y, Wu A, Guan Y (2013) MicroRNA-124 inhibits migration and invasion by down-regulating ROCK1 in glioma. *PLoS One* 8(7):e69478.
75. Marti E*, et al.* (2010) A myriad of miRNA variants in control and Huntington's disease brain regions detected by massively parallel sequencing. *Nucleic Acids Res* 38(20):7219-7235.
76. Hoss AG*, et al.* (2014) MicroRNAs located in the Hox gene clusters are implicated in Huntington's disease pathogenesis. *PLoS Genet* 10(2):e1004188.
77. Fire A*, et al.* (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391(6669):806-811.
78. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116(2):281-297.
79. Batista TM , Marques JT (2011) RNAi pathways in parasitic protists and worms. *Journal of proteomics* 74(9):1504-1514.
80. Lee YS*, et al.* (2004) Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell* 117(1):69-81.
81. Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature* 432(7014):231-235.
82. Gregory RI*, et al.* (2004) The Microprocessor complex mediates the genesis of microRNAs. *Nature* 432(7014):235-240.
83. Han J*, et al.* (2004) The Drosha-DGCR8 complex in primary microRNA processing. *Genes & development* 18(24):3016-3027.
84. Forstemann K*, et al.* (2005) Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. *PLoS Biol* 3(7):e236.
85. Leuschner PJ, Obernosterer G, Martinez J (2005) MicroRNAs: Loquacious speaks out. *Current biology : CB* 15(15):R603-605.
86. Saito K, Ishizuka A, Siomi H, Siomi MC (2005) Processing of pre-microRNAs by the Dicer-1-Loquacious complex in Drosophila cells. *PLoS Biol* 3(7):e235.

87. Liu Q, *et al.* (2003) R2D2, a bridge between the initiation and effector steps of the Drosophila RNAi pathway. *Science* 301(5641):1921-1925.
88. Feinberg EH , Hunter CP (2003) Transport of dsRNA into cells by the transmembrane protein SID-1. *Science* 301(5639):1545-1547.
89. May RC , Plasterk RH (2005) RNA interference spreading in C. elegans. *Methods Enzymol* 392:308-315.
90. Winston WM, Molodowitch C, Hunter CP (2002) Systemic RNAi in C. elegans requires the putative transmembrane protein SID-1. *Science* 295(5564):2456-2459.
91. Tomoyasu Y, *et al.* (2008) Exploring systemic RNA interference in insects: a genome-wide survey for RNAi genes in Tribolium. *Genome Biol* 9(1):R10.
92. Roignant JY, *et al.* (2003) Absence of transitive and systemic pathways allows cell-specific and isoform-specific RNAi in Drosophila. *RNA* 9(3):299-308.
93. Sijen T, *et al.* (2001) On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell* 107(4):465-476.
94. Sijen T, Steiner FA, Thijssen KL, Plasterk RH (2007) Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. *Science* 315(5809):244-247.
95. Lipardi C , Paterson BM (2010) Identification of an RNA-dependent RNA polymerase in Drosophila establishes a common theme in RNA silencing. *Fly* 4(1):30-35.
96. Paim RM, Araujo RN, Lehane MJ, Gontijo NF, Pereira MH (2013) Long-term effects and parental RNAi in the blood feeder Rhodnius prolixus (Hemiptera; Reduviidae). *Insect Biochemistry & Molecular Biology* 43(11):1015-1020.
97. Mariotti M , Guigo R (2010) Selenoprofiles: Profile-based scanning of eukaryotic genome sequences for selenoprotein genes. *J Gerontol* 26(21):2656-2663.
98. Mariotti M, Lobanov AV, Guigo R, Gladyshev VN (2013) SECISearch3 and Seblastian: New tools for prediction of SECIS elements and selenoproteins. *Nucleic Acids Res* 41(15):e149-e149.
99. Lowe TM , Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25(5):0955-0964.
100. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302(1):205-217.
101. Huerta-Cepas J, *et al.* (2011) PhylomeDB v3.0: An expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res* 39(D1):D556-D560.
102. Lobanov AV, Hatfield DL, Gladyshev VN (2008) Selenoproteinless animals: selenophosphate synthetase SPS1 functions in a pathway unrelated to selenocysteine biosynthesis. *Protein science : a publication of the Protein Society* 17(1):176-182.
103. Chapple CE , Guigo R (2008) Relaxation of selective constraints causes independent selenoprotein extinction in insect genomes. *PLoS One* 3(8):e2968.
104. Eichler S , Schaub GA (2002) Development of symbionts in triatomine bugs and the effects of infections with trypanosomatids. *Exp Parasitol* 100(1):17-27.
105. Cummings KL , Tarleton RL (2003) Rapid quantitation of Trypanosoma cruzi in host tissue by real-time PCR. *Mol Biochem Parasitol* 129(1):53-59.
106. Bustin SA, *et al.* (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical chemistry* 55(4):611-622.
107. Finn RD, *et al.* (2014) Pfam: The protein families database. *Nucleic Acids Res* 42(D1):D222-D230.
108. Coordinators NR (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 43(Database issue):D6-D17.
109. Ashburner M, *et al.* (2000) Gene ontology: Tool for the unification of biology. *Nat Genet* 25(1):25-29.
110. Eddy SR (2011) Accelerated profile HMM Searches. *PLoS Comput Biol* 7(10):e1002195-e1002195.
111. Seabra-Junior ES, Souza EM, Mesquita RD (2011) IFRJ INPI 11083-6.
112. Acevedo JM, Centanin L, Dekanty A, Wappner P (2010) Oxygen sensing in Drosophila: multiple isoforms of the prolyl hydroxylase fatiga have different capacity to regulate HIFalpha/Sima. *PLoS One* 5(8):e12390.
113. Avruch J, *et al.* (2012) Protein kinases of the Hippo pathway: regulation and substrates. *Seminars in cell & developmental biology* 23(7):770-784.
114. Lim WA , Pawson T (2010) Phosphotyrosine signaling: evolving a new cellular communication system. *Cell* 142(5):661-667.
115. Shiu S-H , Li W-H (2004) Origins, lineage-specific expansions, and multiple losses of tyrosine kinases in eukaryotes. *Mol Biol Evol* 21(5):828-840.
116. Blume-Jensen P , Hunter T (2001) Oncogenic kinase signalling. *Nature* 411(6835):355-365.

117. Casci T , Freeman M (1999) Control of EGF receptor signalling: lessons from fruitflies. *Cancer metastasis reviews* 18(2):181-201.
118. Anonymous (2010) Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol* 8(2):e1000313.
119. Wigglesworth VB (1931) The physiology of excretion in a blood sucking insect, Rhodnius prolixus (Hemiptera, Reduviidae). *J Exp Biol* 8:411-427.
120. Lespinet O, Wolf YI, Koonin EV, Aravind L (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res* 12(7):1048-1059.
121. Aravind L, Anantharaman V, Venancio TM (2009) Apprehending multicellularity: regulatory networks, genomics, and evolution. *Birth defects research Part C, Embryo today : reviews* 87(2):143-164.
122. Vidal NM, Grazziotin AL, Iyer LM, Aravind L, Venancio TM (2015) Transcription factors, chromatin proteins and the diversification of Hemiptera. *Insect Biochem Mol Biol* 10.1016/j.ibmb.2015.07.001.
123. Katoh K , Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 9(4):286-298.
124. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *J Gerontol* 30(9):1312-1313.
125. Lopez L, Morales G, Ursic R, Wolff M, Lowenberger C (2003) Isolation and characterization of a novel insect defensin from Rhodnius prolixus, a vector of Chagas disease. *Insect Biochemistry & Molecular Biology* 33(4):439-447.
126. Dassanayake RS, Silva Gunawardene YI, Tobe SS (2007) Evolutionary selective trends of insect/mosquito antimicrobial defensin peptides containing cysteine-stabilized alpha/beta motifs. *Peptides* 28(1):62-75.
127. Claudianos C*, et al.* (2006) A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. *Insect Mol Biol* 15(5):615-636.
128. Oakeshott JG*, et al.* (2010) Metabolic enzymers associated with xenobiotic and chemosensory responses in *Nasonia vitripennis*. *Insect Mol Biol* 1:147-163.
129. Ramsey JS*, et al.* (2010) Comparative analysis of detoxification enzymes in *Acyrthosiphon pisum* and *Myzus persicae*. *Insect Mol Biol* 19 Suppl 2:155-164.
130. Shi H*, et al.* (2012) Glutathione S-transferase (GST) genes in the red flour beetle, *Tribolium castaneum*, and comparative analysis with five additional insects. *Genomics* 100(5):327-335.
131. Zhou WW*, et al.* (2013) Genomic insights into the glutathione S-transferase gene family of two rice planthoppers, *Nilaparvata lugens* (Stal) and *Sogatella furcifera* (Horvath) (Hemiptera: Delphacidae). *PLoS One* 8(2):e56604.
132. Montella IR, Schama R, Valle D (2012) The classification of esterases: an important gene family involved in insecticide resistance - A review. *Memórias do Instituto Oswaldo Cruz* 107(4):437-449.
133. Oakeshott JG, Claudianos C, Campbell PM, Newcomb RD, Russell RJ (2005) Biochemical genetics and genomics of insect esterases. *Comprehensive molecular insect science - pharmacology*, eds Gilbert LI, Latrou K, Gill SS (Elsevier, Oxford), Vol 5, pp 309-381.
134. Tsubota T , Shiotsuki T (2010) Genomic and phylogenetic analysis of insect carboxyl/cholinesterase genes. *Journal of Pesticide Science* 35(3):310-314.
135. Schama R*, et al.* (2015) *Rhodnius prolixus* supergene families of enzymes potentially associated with insecticide resistance. *Insect Biochem Mol Biol* 10.1016/j.ibmb.2015.06.005.
136. Thomas JH (2007) Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. *PLoS Genet* 3(5):e67.
137. Nelson DR*, et al.* (2004) Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. *Pharmacogenetics* 14(1):1-18.
138. Deng J, Carbone I, Dean RA (2007) The evolutionary history of cytochrome P450 genes in four filamentous Ascomycetes. *BMC Evol Biol* 7:30.
139. Baldwin WS, Marko PB, Nelson DR (2009) The cytochrome P450 (CYP) gene superfamily in Daphnia pulex. *BMC Genomics* 10:169.
140. Tribolium Genome Sequencing C*, et al.* (2008) The genome of the model beetle and pest *Tribolium castaneum*. *Nature* 452(7190):949-955.
141. Werren JH*, et al.* (2010) Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science* 327(5963):343-348.
142. Strode C*, et al.* (2008) Genomic analysis of detoxification genes in the mosquito Aedes aegypti. *Insect Biochemistry & Molecular Biology* 38(1):113-123.
143. Tijet N, Helvig C, Feyereisen R (2001) The cytochrome P450 gene superfamily in Drosophila melanogaster: annotation, intron-exon organization and phylogeny. *Gene* 262(1-2):189-198.

144. Nijhout FH (1994) *Insect Hormones* (Princeton University Press, N.J.).
145. Goodman W , Cusson M (2012) The juvenile hormones. ed Endocrinology I (Academic Press, London), pp 310-365.
146. Weaver RJ , Audsley N (2009) Neuropeptide regulators of juvenile hormone synthesis: structures, functions, distribution, and unanswered questions. *Annals of the New York Academy of Sciences* 1163:316-329.
147. Noriega FG (2014) Juvenile hormones biosynthesis in insects: What is new, what do we know, what questions remain? *International Scholarly Research Notices* doi.org/10.1155/2014/967361.
148. Hewes RS , Taghert PH (2001) Neuropeptides and neuropeptide receptors in the Drosophila melanogaster genome. *Genome Res* 11(6):1126-1142.
149. Hummon AB*, et al.* (2006) From the genome to the proteome: uncovering peptides in the Apis brain. *Science* 314(5799):647-649.
150. Li B*, et al.* (2008) Genomics, transcriptomics, and peptidomics of neuropeptides and protein hormones in the red flour beetle Tribolium castaneum. *Genome Res* 18(1):113-122.
151. Riehle MA, Garczynski SF, Crim JW, Hill CA, Brown MR (2002) Neuropeptides and peptide hormones in Anopheles gambiae. *Science* 298(5591):172-175.
152. Roller L*, et al.* (2008) The unique evolution of neuropeptide genes in the silkworm Bombyx mori. *Insect Biochemistry & Molecular Biology* 38(12):1147-1157.
153. Ons S, Richter F, Urlaub H, Rivera-Pomar R (2009) The neuropeptidome of *Rhodnius prolixus* brain. *Proteomics* 9(3):788-792.
154. Ons S, Sterkel M, Diambra L, Urlaub H, Rivera-Pomar R (2011) Neuropeptide precursor gene discovery in the Chagas disease vector Rhodnius prolixus. *Insect Mol Biol* 20(1):29-44.
155. Orchard I, Lee do H, da Silva R, Lange AB (2011) The Proctolin Gene and Biological Effects of Proctolin in the Blood-Feeding Bug, Rhodnius prolixus. *Front Endocrinol (Lausanne)* 2:59.
156. Paluzzi JP , Orchard I (2010) A second gene encodes the anti-diuretic hormone in the insect, Rhodnius prolixus. *Mol Cell Endocrinol* 317(1-2):53-63.
157. Paluzzi JP, Russell WK, Nachman RJ, Orchard I (2008) Isolation, cloning, and expression mapping of a gene encoding an antidiuretic hormone and other CAPA-related peptides in the disease vector, Rhodnius prolixus. *Endocrinology* 149(9):4638-4646.
158. Te Brugge VA, Schooley DA, Orchard I (2008) Amino acid sequence and biological activity of a calcitonin-like diuretic hormone (DH31) from Rhodnius prolixus. *Journal of Experimental Biology* 211(Pt 3):382-390.
159. Ons S*, et al.* (2015) Identification of G protein coupled receptors for opsines and neurohormones in Rhodnius prolixus. Genomic and transcriptomic analysis. *Insect Biochem Mol Biol* 10.1016/j.ibmb.2015.05.003.
160. Birney E, Clamp M, Durbin RR (2004) GeneWise and Genomewise. *Genome Res* 14(5):988-995.
161. Stanke M , Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19 Suppl 2:ii215-ii225.
162. Salamov AA , Solovyev VV (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res* 10(4):516-522.
163. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* 24(8):1596-1599.
164. Rutherford K*, et al.* (2000) Artemis: Sequence visualization and annotation. *Bioinformatics* 16(10):944-945.
165. Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41:95-98.
166. Price JL*, et al.* (1998) *Double*-time is a novel *Drosophila* clock gene that regulates PERIOD protein accumulation. *Cell* 94(1):83-95.
167. Martinek S, Inonog S, Manoukian AS, Young MW (2001) A role for the segment polarity gene *shaggy* /GSK-3 in the *Drosophila* circadian clock. *Cell* 105(6):769-779.
168. Lin J-M*, et al.* (2002) A role for *casein kinase 2α* in the *Drosophila* circadian clock. *Nature* 420(6917):816-820.
169. Sathyanarayanan S, Zheng X, Xiao R, Sehgal A (2004) Posttranslational regulation of *Drosophila* PERIOD protein by protein phosphatase 2A. *Cell* 116(4):603-615.
170. Chiu JC, Ko HW, Edery I (2011) *NEMO/NLK* phosphorylates *PERIOD* to initiate a time-delay phosphorylation circuit that sets circadian clock speed. *Cell* 145(3):357-370.
171. Yu W, Houl JH, Hardin PE (2011) NEMO kinase contributes to core Period determination by slowing the pace of the *Drosophila* circadian oscillator. *Curr Biol* 21(9):756-761.
172. Grima B*, et al.* (2002) The F-box protein slimb controls the levels of clock proteins period and timeless. *Nature* 420(6912):178-182.

173. Koh K, Zheng X, Sehgal A (2006) JETLAG resets the *Drosophila* circadian clock by promoting light-induced degradation of TIMELESS. *Science* 312(5781):1809-1812.
174. Yuan Q, Metterville D, Briscoe AD, Reppert SM (2007) Insect cryptochromes: gene duplication and loss define diverse ways to construct insect circadian clocks. *Mol Biol Evol* 24(4):948-955.
175. Ceriani MF*, et al.* (1999) Light-dependent sequestration of *TIMELESS* by *CRYPTOCHROME*. *Science* 285(5427):553-556.
176. Dunkov B , Georgieva T (2006) Insect iron binding proteins: insights from the genomes. *Insect Biochemistry & Molecular Biology* 36(4):300-309.
177. Nichol H, Law JH, Winzerling JJ (2002) Iron metabolism in insects. *Annu Rev Entomol* 47:535-559.
178. Ribeiro JM*, et al.* (2014) An Insight into the Transcriptome of the Digestive Tract of the Bloodsucking Bug, Rhodnius prolixus. *PLoS Negl Trop Dis* 8(1):e2594.
179. McKie AT*, et al.* (2001) An iron-regulated ferric reductase associated with the absorption of dietary iron. *Science* 291(5509):1755-1759.
180. Bettedi L, Aslam MF, Szular J, Mandilaras K, Missirlis F (2011) Iron depletion in the intestines of Malvolio mutant flies does not occur in the absence of a multicopper oxidase. *Journal of Experimental Biology* 214(Pt 6):971-978.
181. Qin Q, Wang X, Zhou B (2013) Functional studies of Drosophila zinc transporters reveal the mechanism for dietary zinc absorption and regulation. *BMC biology* 11:101.
182. Xiao G, Wan Z, Fan Q, Tang X, Zhou B (2014) The metal transporter ZIP13 supplies iron into the secretory pathway in Drosophila melanogaster. *eLife* 3:e03191.
183. White DC , Granick S (1963) Hemin Biosynthesis in Haemophilus. *Journal of bacteriology* 85:842-850.
184. Rao AU, Carta LK, Lesuisse E, Hamza I (2005) Lack of heme synthesis in a free-living eukaryote. *Proc Natl Acad Sci U S A* 102(12):4270-4275.
185. Braz GR, Coelho HS, Masuda H, Oliveira PL (1999) A missing metabolic pathway in the cattle tick Boophilus microplus. *Current biology : CB* 9(13):703-706.
186. Braz GR, Abreu L, Masuda H, Oliveira PL (2001) Heme biosynthesis and oogenesis in the blood-sucking bug, Rhodnius prolixus. *Insect Biochemistry & Molecular Biology* 31(4-5):359-364.
187. Paiva-Silva GO*, et al.* (2006) A heme-degradation pathway in a blood-sucking insect. *Proc Natl Acad Sci U S A* 103(21):8030-8035.
188. Zhang X, Sato M, Sasahara M, Migita CT, Yoshida T (2004) Unique features of recombinant heme oxygenase of Drosophila melanogaster compared with those of other heme oxygenases studied. *European journal of biochemistry / FEBS* 271(9):1713-1724.
189. Quigley JG*, et al.* (2004) Identification of a human heme exporter that is essential for erythropoiesis. *Cell* 118(6):757-766.
190. Duffy SP*, et al.* (2010) The Fowler syndrome-associated protein FLVCR2 is an importer of heme. *Molecular and cellular biology* 30(22):5318-5324.
191. Mandilaras K , Missirlis F (2012) Genes for iron metabolism influence circadian rhythms in Drosophila melanogaster. *Metallomics : integrated biometal science* 4(9):928-936.
192. Kaasik K , Lee CC (2004) Reciprocal regulation of haem biosynthesis and the circadian clock in mammals. *Nature* 430(6998):467-471.
193. Oliveira PL*, et al.* (1995) A heme-binding protein from hemolymph and oocytes of the blood-sucking insect, Rhodnius prolixus. Isolation and characterization. *J Biol Chem* 270(18):10897-10901.
194. Dansa-Petretski M, Ribeiro JM, Atella GC, Masuda H, Oliveira PL (1995) Antioxidant role of Rhodnius prolixus heme-binding protein. Protection against heme-induced lipid peroxidation. *J Biol Chem* 270(18):10893-10896.
195. Machado EA, Oliveira PL, Moreira MF, de Souza W, Masuda H (1998) Uptake of Rhodnius heme-binding protein (RHBP) by the ovary of Rhodnius prolixus. *Archives of insect biochemistry and physiology* 39(4):133-143.
196. Braz GR, Moreira MF, Masuda H, Oliveira PL (2002) Rhodnius heme-binding protein (RHBP) is a heme source for embryonic development in the blood-sucking bug Rhodnius prolixus (Hemiptera, Reduviidae). *Insect Biochemistry & Molecular Biology* 32(4):361-367.
197. Walter-Nuno AB*, et al.* (2013) Silencing of maternal heme-binding protein causes embryonic mitochondrial dysfunction and impairs embryogenesis in the blood sucking insect Rhodnius prolixus. *J Biol Chem* 288(41):29323-29332.
198. Paiva-Silva GO*, et al.* (2002) On the biosynthesis of Rhodnius prolixus heme-binding protein. *Insect Biochemistry & Molecular Biology* 32(11):1533-1541.
199. Su CY, Menuz K, Carlson JR (2009) Olfactory perception: receptors, cells, and circuits. *Cell* 139(1):45-59.

200. Touhara K , Vosshall LB (2009) Sensing odorants and pheromones with chemosensory receptors. *Annu Rev Physiol* 71:307-332.
201. Jones W, Cayirlioglu P, Kadow I, Vosshall L (2007) Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*. *Nature* 445(7123):86-90.
202. Kwon JY, Dahanukar A, Weiss LA, Carlson JR (2007) The molecular basis of CO2 reception in Drosophila. *Proc Natl Acad Sci USA* 104(9):3574-3578.
203. Lu T*, et al.* (2007) Odor coding in the maxillary palp of the malaria vector mosquito *Anopheles gambiae*. *Current biology : CB* 17(18):1533-1544.
204. Croset V*, et al.* (2010) Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *PLoS Genet* 6(8).
205. Abuin L*, et al.* (2011) Functional architecture of olfactory ionotropic glutamate receptors. *Neuron* 69(1):44-60.
206. Robertson HM, Warr CG, Carlson JR (2003) Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 100 (Supplement 2):14537-14542.
207. Robertson HM , Wanner KW (2006) The chemoreceptor superfamily in the honey bee, *Apis mellifera*: Expansion of the odorant, but not gustatory, receptor family. *Genome Res* 16(11):1395-1403.
208. Kirkness E*, et al.* (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc Natl Acad Sci U S A* 107(27):12168-12173.
209. Smith CR*, et al.* (2011) Draft genome of the red harvester ant Pogonomyrmex barbatus. *Proc Natl Acad Sci U S A* 108(14):5667-5672.
210. Smadja C, Shi P, Butlin R, Robertson H (2009) Large gene family expansions and adaptive evolution for odorant and gustatory receptors in the pea aphid, Acyrthosiphon pisum. *Mol Biol Evol* 26(9):2073-2086.
211. Robertson H, Warr C, Carlson J (2003) Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America* 100 Suppl 2:14537-14542.
212. Robertson H , Wanner K (2006) The chemoreceptor superfamily in the honey bee, Apis mellifera: expansion of the odorant, but not gustatory, receptor family. *Genome research* 16(11):1395-1403.
213. Robertson HM, Warr C, Carlson J (2003) Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America* 100 Suppl 2:14537-14542.
214. Xiao J-H*, et al.* (2013) Obligate mutualism within a host drives the extreme specialization of a fig wasp genome. *Genome Biol* 14(12):R141.
215. Obiero GF*, et al.* (2014) Odorant and gustatory receptors in the tsetse fly *Glossina morsitans morsitans*. *PLoS Negl Trop Dis* 8(4):e2663.
216. Terrapon N*, et al.* (2014) Molecular traces of alternative social organization in a termite genome. *Nature communications* 5:3636.
217. Vieira FG , Rozas J (2011) Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: Origin and evolutionary history of the chemosensory system. *Genome Biol Evol* 3(0):476-490.
218. Kirkness EF*, et al.* (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proceedings of the National Academy of Sciences of the United States of America* 107(27):12168-12173.
219. Zhou JJ*, et al.* (2010) Genome annotation and comparative analyses of the odorant-binding proteins and chemosensory proteins in the pea aphid *Acyrthosiphon pisum*. *Insect Mol Biol* 19 Suppl 2:113-122.
220. Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: Fast selection of best-fit models of protein evolution. *Bioinformatics* 27(8):1164-1165.
221. Guindon S*, et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* 59(3):307-321.
222. Letunic I , Bork P (2007) Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* 23(1):127-128.
223. Hua JF*, et al.* (2012) Identification and Binding Characterization of Three Odorant Binding Proteins and One Chemosensory Protein from *Apolygus lucorum* (Meyer-Dur). *J Chem Ecol* 38(9):1163-1170.
224. Gu SH*, et al.* (2011) Identification and tissue distribution of odorant binding protein genes in the lucerne plant bug *Adelphocoris lineolatus* (Goeze). *Insect Biochem Mol Biol* 41(4):254-263.
225. Gong DP, Zhang HJ, Zhao P, Xia QY, Xiang ZH (2009) The Odorant Binding Protein gene family from the genome of silkworm, *Bombyx mori. BMC Genomics* 10:332.
226. Vieira FG , Rozas J (2011) Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. *Genome Biol Evol* 3:476-490.

227. Gong DP*, et al.* (2007) Identification and expression pattern of the chemosensory protein gene family in the silkworm, *Bombyx mori*. *Insect Biochem Mol Biol* 37(3):266-277.
228. Consortium THG (2012) Islands of divergence underlie adaptive radiation in a butterfly genome. *Nature*.
229. Kirkness EF*, et al.* (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc Natl Acad Sci U S A* 107(27):12168-12173.
230. Zhou JJ*, et al.* (2010) Genome annotation and comparative analyses of the odorant-binding proteins and chemosensory proteins in the pea aphid Acyrthosiphon pisum. *Insect Mol Biol* 19 Suppl 2:113-122.
231. Barrett AJ, Rawlings ND, Woessner JF (1998) *Handbook of proteolytic enzymes* (Academic Press, San Diego) pp xxix, 1666 p.
232. Terra WR , Ferreira C (2005) Biochemistry of digestion. *Comprehensive Molecular Insect Science, Volume 4: Biochemistry and Molecular Biology*, eds Gilbert LI, Latrou K, Gill SS (Elsevier, Amsterdan), Vol 4, pp 171-224.
233. Terra WR, Ferreira C, Garcia ES (1988) Origin, distribution, properties and functions of the major *Rhodnius prolixus* midgut hydrolases. *Insect Biochemistry* 18(5):423-434.
234. Rawlings ND, Waller M, Barrett AJ, Bateman A (2014) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res* 42(Database issue):D503-509.
235. Prado SS , Almeida RP (2009) Phylogenetic placement of pentatomid stink bug gut symbionts. *Current microbiology* 58(1):64-69.
236. Cornwall JW , Patton WS (1914) Some observations on the salivary secretion of the common blood-sucking insects and ticks. *Indian J Med Res* 2:569-593.
237. Hellmann K , Hawkins RI (1964) Anticoagulant and fibrinolytic activities from *Rhodnius prolixus* Stahl. *Nature (London)* 201:1008-1009.
238. Hellmann K , Hawkins RI (1965) Prolixin-S and Prolixin-G: two anticoagulants from *Rhodnius prolixus* Stahl. *Nature (London)* 207:265-267.
239. Ribeiro JMC, Assumpcao TCF, Francischetti IMB (2012) An insight into the sialomes of Bloodsucking Heteroptera. *Psyche* 2012(Article ID 470436):16 pages.
240. Ribeiro JMC, Hazzard JMH, Nussenzveig RH, Champagne D, Walker FA (1993) Reversible binding of nitric oxide by a salivary nitrosylhemeprotein from the blood sucking bug, *Rhodnius prolixus*. *Science* 260:539-541.
241. Ribeiro JMC, Gonzales R, Marinotti O (1990) A salivary vasodilator in the blood sucking bug, *Rhodnius prolixus*. *Br J Pharmacol* 101:932-936.
242. Ribeiro JM*, et al.* (2004) Exploring the sialome of the blood-sucking bug Rhodnius prolixus. *Insect Biochem Mol Biol* 34(1):61-79.
243. Santos A*, et al.* (2007) The sialotranscriptome of the blood-sucking bug *Triatoma brasiliensis* (Hemiptera, Triatominae). *Insect Biochem Mol Biol* 37(7):702-712.
244. Assumpcao TC*, et al.* (2008) An insight into the sialome of the blood-sucking bug *Triatoma infestans,* a vector of Chagas' disease. *Insect Biochem Mol Biol* 38(2):213-232.
245. Kato H*, et al.* (2010) A repertoire of the dominant transcripts from the salivary glands of the blood-sucking bug, *Triatoma dimidiata,* a vector of Chagas disease. *Infect Genet Evol* 10(2):184-191.
246. Ribeiro JM, Assumpcao TC, Pham VM, Francischetti IM, Reisenman CE (2012) An insight into the sialotranscriptome of *Triatoma rubida* (Hemiptera: Heteroptera). *Journal of medical entomology* 49(3):563-572.
247. Assumpcao TC*, et al.* (2012) An insight into the sialotranscriptome of *Triatoma matogrossensis*, a kissing bug associated with fogo selvagem in South America. *The American journal of tropical medicine and hygiene* 86(6):1005-1014.
248. Takac P*, et al.* (2006) Vasotab, a vasoactive peptide from horse fly *Hybomitra bimaculata* (Diptera, Tabanidae) salivary glands. *J Exp Biol* 209(Pt 2):343-352.
249. Ribeiro JMC , Garcia ES (1980) The salivary and crop apyrase activity of *Rhodnius prolixus*. *J Insect Physiol* 26:303-307.
250. Sarkis JJ, Guimaraes JA, Ribeiro JM (1986) Salivary apyrase of Rhodnius prolixus. Kinetics and purification. *Biochem J* 233(3):885-891.
251. Champagne DE, Nussenzveig RH, Ribeiro JM (1995) Purification, partial characterization, and cloning of nitric oxide-carrying heme proteins (nitrophorins) from salivary glands of the blood-sucking insect *Rhodnius prolixus*. *J Biol Chem* 270(15):8691-8695.
252. Valenzuela JG, Charlab R, Galperin MY, Ribeiro JM (1998) Purification, cloning, and expression of an apyrase from the bed bug Cimex lectularius. A new type of nucleotide-binding enzyme. *J Biol Chem* 273(46):30583-30590.
253. Ribeiro JMC, Modi GB, Resh RB (1989) Salivary apyrase activity of some old world phlebotomine sand flies. *Insect Biochem* 19:409-412.

254. Valenzuela JG, Belkaid Y, Rowton E, Ribeiro JM (2001) The salivary apyrase of the blood-sucking sand fly Phlebotomus papatasi belongs to the novel Cimex family of apyrases. *Journal of Experimental Biology* 204(Pt 2):229-237.
255. Faudry E*, et al.* (2004) *Triatoma infestans* apyrases belong to the 5'-nucleotidase family. *J Biol Chem* 279(19):19607-19613.
256. Ribeiro JM*, et al.* (1998) Role of salivary antihemostatic components in blood feeding by triatomine bugs (Heteroptera) *Journal of medical entomology* 35(4):599-610.
257. Valenzuela JG, Chuffe, O.M. and Ribeiro, J.M.C. (1996) Apyrase and anti-platelet activities from the salivary glands of the bed bug *Cimex lectularius*. *Insect Biochem Mol Biol* 21:557-562.
258. Maddrell SHP , Phillips JE (1975) Secretion of hypoosmotic fluid by the lower Malpighian tubules of *Rhodnius prolixus*
. *J Exp Biol* 61:357-377.
259. Ianowski JP , O'Donnell MJ (2006) Electrochemical gradients for Na+, K+, Cl- and H+ across the apical membrane in Malpighian (renal) tubule cells of Rhodnius prolixus. *J Exp Biol* 209(Pt 10):1964-1975.
260. Kleyman TR , Cragoe EJ, Jr. (1988) Amiloride and its analogs as tools in the study of ion transport. *The Journal of membrane biology* 105(1):1-21.
261. Paluzzi JP, Yeung C, O'Donnell MJ (2013) Investigations of the signaling cascade involved in diuretic hormone stimulation of Malpighian tubule fluid secretion in Rhodnius prolixus. *Journal of insect physiology* 59(12):1179-1185.
262. Jasrapuria S, Specht CA, Kramer KJ, Beeman RW, Muthukrishnan S (2012) Gene families of cuticular proteins analogous to peritrophins (CPAPs) in Tribolium castaneum have diverse functions. *PLoS One* 7(11):e49844.
263. Bera A, Biswas R, Herbert S, Gotz F (2006) The presence of peptidoglycan O-acetyltransferase in various staphylococcal species correlates with lysozyme resistance and pathogenicity. *Infect Immun* 74(8):4598-4604.
264. Herbert S*, et al.* (2007) Molecular basis of resistance to muramidase and cationic antimicrobial peptide activity of lysozyme in staphylococci. *PLoS pathogens* 3(7):e102.
265. Willis JH (2010) Structural cuticular proteins from arthropods: annotation, nomenclature, and sequence characteristics in the genomics era. *Insect Biochemistry & Molecular Biology* 40(3):189-204.
266. Wilkin MB*, et al.* (2000) Drosophila dumpy is a gigantic extracellular protein required to maintain tension at epidermal-cuticle attachment sites. *Current biology : CB* 10(10):559-567.
267. Roch F, Alonso CR, Akam M (2003) Drosophila miniature and dusky encode ZP proteins required for cytoskeletal reorganisation during wing morphogenesis. *Journal of cell science* 116(Pt 7):1199-1207.
268. Alvarenga ES*, et al.* (2015) Chitin is a component of the Rhodnius prolixus midgut. *Insect Biochem Mol Biol* 10.1016/j.ibmb.2015.04.003.
269. Nation JL (2008) *Insect physiology and biochemistry* (CRC Press/Taylor & Francis, Boca Raton) 2nd Ed pp ix, 544 p., 547 p. of plates.
270. Moran NA , Jarvik T (2010) Lateral transfer of genes from fungi underlies carotenoid production in aphids. *Science* 328(5978):624-627.
271. Alves-Bezerra M , Gondim KC (2012) Triacylglycerol biosynthesis occurs via the glycerol-3-phosphate pathway in the insect *Rhodnius prolixus*. *Biochimica et Biophysica Acta* 1821(12):1462-1471.
272. Takeuchi K , Reue K (2009) Biochemistry, physiology, and genetics of GPAT, AGPAT, and lipin enzymes in triglyceride synthesis. *American Journal of Physiology Endocrinology and Metabolism* 286(6):E1195-E1209.
273. Alves-Bezerra M , Gondim KC (2012) Triacylglycerol biosynthesis occurs via the glycerol-3-phosphate pathway in the insect Rhodnius prolixus. *Biochim Biophys Acta* 1821(12):1462-1471.
274. Alves-Bezerra M*, et al.* (2015) Adipokinetic hormone receptor gene identification and its role in triacylglycerol metabolism in the blood-sucking insect *Rhodnius prolixus*. *Insect Biochem Mol Biol* 10.1016/j.ibmb.2015.06.013.
275. Wakil SJ (1989) Fatty acid synthase, a proficient multifunctional enzyme. *Biochemistry* 28(11):4523-4530.
276. McGarry JD, Mannaerts GP, Foster DW (1977) A possible role for malonyl-CoA in the regulation of hepatic fatty acid oxidation and ketogenesis. *Journal of Clinical Investigation* 60(1):265-270.
277. Clark AJ , Block K (1959) The absence of sterol synthesis in insects. *J Biol Chem* 234:2578-2582.
278. Fluegel ML, Parker TJ, Pallanck LJ (2006) Mutations of a *Drosophila* NPC1 gene confer sterol and ecdysone metabolic defects. *Genetics* 172(1):185-196.
279. Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV (2011) OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res* 39(Database issue):D282-D288.

280. Bakovic M, Fullerton MD, Michel V (2007) Metabolic and molecular aspects of ethanolamine phospholipid biosynthesis: the role of CTP:phosphoethanolamine cytidylyltransferase (Pcyt2). *Biochemistry and Cell Biology* 85(3):283-230.
281. Dobrosotskaya IY, Seegmiller AC, Brown MS, Goldstein JL, Rawson RB (2002) Regulation of SREBP processing and membrane lipid production by phospholipids in *Drosophila*. *Science (New York, NY)* 296(5569):879-883.
282. Lim HY, Wang W, Wessells RJ, Ocorr K, Bodmer R (2011) Phospholipid homeostasis regulates lipid metabolism and cardiac function through SREBP signaling in *Drosophila*. *Genes & development* 25(5):189-200.
283. Mérida I, Avila-Flores A, Merino E (2008) Diacylglycerol kinases: at the hub of cell signalling. *Biochem J* 409(1):1-18.
284. Pilz A, Schaap D, Hunt D, Fitzgibbon J (1995) Chromosomal localization of three mouse diacylglycerol kinase (DAGK) genes: genes sharing sequence homology to the *Drosophila* retinal degeneration A (*rdgA*) gene. *Genomics* 26(3):599-601.
285. Mesquita RD*, et al.* (2008) *Trypanosoma cruzi* infection is enhanced by vector saliva through immunosuppressant mechanisms mediated by lysophosphatidylcholine. *Infection and Immunity* 76(12):5543-5552.
286. Guedes RLM*, et al.* (2011) Amino acids biosynthesis and nitrogen assimilation pathways: A great genomic deletion during eukaryotes evolution. *BMC Genomics* 12 Suppl 4:S2-S2.
287. Needleman SB , Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48(3):443-453.
288. Anonymous (2006) Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* 443(7114):931-949.
289. Misof B*, et al.* (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346(6210):763-767.