doi:10.1038/nature10640

Table of Contents

S1. Strain selection and DNA preparation

The London strain of *T. urticae* (Koch) originates from a wild-collected *T. urticae* population from the Vineland region (Ontario, Canada). The taxonomic status of the collected species was determined by the Agriculture Canada taxonomy division in Ottawa. Large-scale continuous propagation of the collected population was performed using controlled environment insectaries at the Agriculture and Agri-Food Canada Research Institute in London, Ontario. The strain used for DNA isolation (London strain) in the genome sequencing project was developed from an isofemale line that underwent 8 generations of sib-mating (to maximize homozygosity). Animals were mass reared on the bean plants in growth chambers at 27 °C with a 16:8 hour (light/dark) photoperiod. Plants were washed in 0.1% Triton X-100 detergent solution in 2-liter beakers to release all spider mite life stages. Adult spider mites, nymphs, larvae and eggs were filtered through a series of fine sieves to isolate the pure egg fraction. Eggs were collected in a 1.5 mL Eppendorf tube, treated with bleach solution (to remove plant tissue and possible microbial contaminants) and prepared for DNA extraction. Embryos were ground in a glass tissue grinder and DNA extraction was performed using the QIAGEN Blood&cell culture DNA kit (Midi column #13433) according to manufacturer's protocol. For sequencing (Section S2), the isolated DNA was provided to the Department of Energy Joint Genome Institute (DOE-JGI) (Walnut Creek, California, USA). The preparation of the DNA solely from the *T. urticae* embryos minimized the contamination from the foreign DNA that often occurs if larvae or adult animals are used (i.e., gut plant material and associated microflora).

S2. Sequencing and annotation

S2.1. Sequencing

All sequencing reads were collected with standard Sanger sequencing protocols on ABI 3730XL capillary sequencing machines at the Department of Energy Joint Genome Institute (DOE-JGI; Walnut Creek, CA, USA). Three different sized libraries were used as templates for the plasmid subclone sequencing process, and both ends were sequenced to give matching paired-end reads. A total of 499,872 reads from a 2.5 kb insert library, 572,253 reads from a 8.5 kb insert library, and 107,424 reads from a 35.5 kb fosmid library were produced.

S2.2. Genome assembly and chromosome anchoring

A total of 1,179,549 reads (see Table S2.2.1 for clone size breakdowns) were assembled using a modified version of Arachne v.20071016[1] with parameters maxcliq1=100, correct1_passes=0 and BINGE_AND_PURGE=True. The resulting output of this assembly was then passed through Rebuilder and SquashOverlaps with parameters to merge adjacent assembled alternative haplotypes, and subsequently run through another complete Arachne assembly process to finalize the assembly. This produced 733 scaffold sequences, with L50 of 3.0 Mb, 44 scaffolds larger than 100 kb, and total scaffold size of 90.2 Mb. Each scaffold was screened against bacterial proteins, organelle sequences and GenBank and removed if found to be a contaminant. Additional scaffolds were removed if they consisted of greater than 95% 24mers that occurred four other times in the scaffolds larger than 50kb or if the scaffold contained only unanchored rRNA sequences. The final assembly contains 640 scaffolds that

cover 89.6 Mb of the genome with a contig L50 of 212.8 kb and a scaffold L50 of 3.0 Mb (see Figure S2.2.1 and Table S2.2.2).

We classified the remaining scaffolds in various bins depending on sequence content. We identified contamination using megablast against Genbank NR and blastp against a set of known microbial proteins. No scaffolds were identified as contamination. We classified additional scaffolds as unanchored rDNA (76), and mitochondrion (1). We also removed 16 scaffolds that were less than 1kb in sequence length. The resulting final statistics are shown in Table S2.2.3.

S2.3. Transcriptome sequencing and characterization

S2.3.1. Overview of RNA-seq and transcriptomes analyses

We generated deep *T. urticae* transcriptome data (Illumina RNA-seq) from the reference London strain to: *i*) provide experimental support for gene models (Section S2.4), and *ii*) to study patterns of gene expression across plant hosts (Section S6.2). For the former, we generated 77 bp Illumina reads from single libraries from four diverse developmental stages to maximally capture the complement of transcribed sequences across development; we also generated libraries from mixed stage RNA from mites grown on bean and *Arabidopsis* to capture host specific transcription. For the latter (*ii*), we generated transcriptome data of shorter read lengths (38 bp), but with triplicate biological replication, for one developmental stage (larvae) on the three host plants (here, our primary goal was to detect differential gene expression by plant host; Section S6.2).

S2.3.2. Generation of long RNA-seq reads for annotation support

RNA from four diverse developmental stages – embryo, larvae, nymph, and adult – was prepared from mites using the Trizol Reagent (Invitrogen, Carlsbad, CA, USA). Staging was performed as previously described under S1, yielding nearly pure samples of embryos and largely pure samples for the other stages. Samples were a mix of males and females to capture male and female patterns of transcription, and were reared on beans (*Phaseolus vulgaris* cv California Red Kidney). In addition, RNA was prepared from mixed stage mites propagated on bean, as well as mites that were grown on bean and transferred to *Arabidopsis* for 24 hours.

For construction of RNA-seq libraries, 10 µg of total RNA per sample was used as input for polyA selection of mRNA and for downstream library construction steps with the Illuminam RNA-seq Prep kit as described by the manufacturer (Illumina, San Diego, CA, USA). Following library construction, concentrations and library insert sizes were evaluated with a Bioanalyzer 2100 with DNA 1000 Kits (Agilent, Santa Clara, CA, USA; in all cases, the estimated library sizes agreed closely with that expected from the gel size selection that was part of the library preparation [~150-300 bp]). Sequencing was performed according to Illumina's protocols on an Illumina Genome Analyzer II instrument to generate single end reads of length 77 bp. Two flowcell lanes were used to sequence each of the samples prepared from RNA from the four developmental stages (~18-26 x 10^6 reads were generated per library, Table S2.3.1). For the pilot feeding experiment, each of two biological replicates for mites reared on beans and two different *Arabidopsis* accessions (Bla-2 and Kondara) were sequenced on a flowecell lane to yield ~50x 10^6 reads (bean- ~ 18x 10^6, Bla-2- ~17 x 10^6 and Kondara- ~15 x 10^6 reads).

S2.3.3. Expression quantification – developmental stages

For quantifying gene expression for the developmental samples, we found that trimming the 77 bp reads markedly increased the percentage of alignable reads (this resulted from the removal of low quality sequence at the 3' end of the reads, a known issue with the Illumina method). After evaluating several trimming possibilities, we applied a hard trim at 60 bp, and further discarded reads of low quality at 60 bp as assessed with the CLC Workbench trimming algorithm. Our rationale was that reads of 60 bp in length were long enough to allow most reads to map uniquely while retaining a high count of alignable reads important for gene quantification (see Figure S2.3.1).

For gene expression level quantification, all of the trimmed reads were aligned to the genome with the Bowtie/Tophat software package[2]. To do this, we used the annotation and aligned reads with the following parameters: minimum anchor length of 5 bases (-a 5), minimum intron size of 5 bases (-i 5), maximum intron size of 10 kb (-I 10000), Solexa quality scores (--solexa-quals), unique mappings (-g 1), and three segment mismatches (--segment-mismatches 3). Note that reads that aligned to multiple gene models or genomic locations (repetitive reads) were excluded as specified by the unique mappings flag set to 1. To aid in read alignment, we used the genome annotation as a prior for read alignment as specified in Bowtie/Tophat. A summary of the read mapping results by library can be found in Table S2.3.1.

From the resulting SAM (Sequence Alignment/Map) alignment files, we directly calculated gene expression levels as read counts normalized per kilobase of exon per million reads mapped (RPKM values)[3]. Only reads mapping to annotated exonic regions were included. Additionally, due to the prevalence of overlapping gene models, we designed our pipeline to mask exonic sequences in regions of gene overlaps (in these regions, as the strand of transcription cannot be assessed using our library preparation methods, read assignments to a given gene cannot be determined).

S2.3.4. Differential expression across developmental stages

To assess differential expression among developmental stages, we quantile normalized the RPKM values to correct for varying distributions of RNA-seq read abundances between samples[4]. This extra normalization was important to allow comparisons among stages where a small number of genes accounted for a large fraction of the mapped reads (and thus could globally inflate differential gene expression estimation). For instance, in adults, ~30% of reads were derived from a mere 10 highly expressed genes, many of which are likely associated with reproduction (Figures S2.3.2 and S2.3.3; Tables S2.3.2, S2.3.3a-f and S2.3.4). As the stage specific data was produced from single samples, we used the Z test of Kal *et al.*[5] to examine differential gene expression (we used the CLC Genomics Workbench software package implementation of this test). Briefly, Kal's Z test[5] uses the relative proportions of expression of a gene compared to the whole dataset, and an approximation of the negative binomial distribution. This allows differential expression to be tested between samples in which no direct variance estimates can be estimated due to a lack of biological replicates. For downstream analyses, we defined as differentially expressed genes for which the false discovery rate (FDR) corrected p-value was ≤ 0.05 and for which the fold change ≥ 2 (either up- or down-regulated). FDR estimates were conducted according to the methods of Benjamini and Hochberg[6] as implemented in CLC Genomics Workbench. A fold-change cutoff of 2 was used to limit the number of spurious gene expression differences likely to arise

with a lack of biological replicates (again, the stage data was primarily generated to support the genome annotation, see Section S2.4).

S2.3.5. Hypergeometric tests on GO annotations – developmental stages

We used hypergeometric tests to determine whether given GO categories[7] were significantly over or under-represented among differentially expressed genes using the CLC Genomics Workbench, which uses a method similar to the unconditional GOstats test of Falcon and Gentleman[8]. Classes of genes with p-values < 0.05 were deemed to be over or under-represented. Enriched genes can be found for given comparisons in summary tables Table S2.3.3a-f.

S2.4. Gene prediction and annotation

S2.4.1. Annotation of protein coding genes

Annotation of the *T. urticae* genome was done using the gene prediction platform EuGene. This gene prediction platform is designed to be able to integrate many different sources of extrinsic evidence as well as *ab initio* prediction results. For intrinsic gene prediction, different software modules need to be trained and a number of parameters need to be estimated[9]. For instance, splice sites were identified using the SpliceMachine[10] signal sensor components trained specifically on *T. urticae* data. To this end, a 'positive' set of 2690GT donor and 1455AG acceptor sites was constructed in windows of 402 bp (200 bp up- and downstream of either the donor or the acceptor site), see Figure S2.4.4. The negative set consisted of windows of the same size but with 23,905 GT and 23,854 AG dinucleotides known not to be splice sites as based on EST alignments. These windows were thus exclusively derived from either exon or intron regions. The SpliceMachine models gave specificity scores of 84.3% for donor and 65.2% for acceptor.

The content sensor used by EuGene to recognize coding sequences is an interpolated Markov model that was trained on 15,887 *T. urticae* conserved coding regions that were collected genome-wide using BLASTX on proteins in SWISSprot. For the non-coding part of the IMM, we extracted 6697 introns based upon the spliced alignments obtained by ESTs. We only considered introns that were confirmed by at least six ESTs. The same intron data was used to extract donor and acceptor sites (see above). Training EuGene also requires the estimation of scaling parameters from known *T. urticae* genes within their genomic context. As such, 211 genomic *T. urticae* sequences that each contained abutting genes were constructed and used to train EuGene. After training, we obtained a sensitivity of 93% and specificity 81.4% on a set of 212 manually curated genes.

For extrinsic annotation, the following data sources were integrated into our annotation system: 1) protein sets from the latest Flybase[11] release, other arthropods and Uniprot-Swissprot; and 2) ESTs generated from the *T. urticae* for four different developmental stages as well as a large number of Illumina RNA-seq reads resulting from different developmental stages and several feeding experiments performed on different hosts (*Arabidopsis* and bean). The fully integrated annotation obtained with EuGene yielded 18,414 protein encoding nuclear genes (Table S2.4.1). Following gene prediction and manual annotation, RNA-seq reads were aligned back to predicted gene models with Bowtie/Tophat to assess expression. Of the 18,414 predicted protein-encoding genes, 8243 genes were supported by ESTs, while

14,545 genes were supported by RNA-seq data; 11,433 had a protein homologue in the non-redundant protein database (Figures S2.4.1 and S2.4.2).

To compare the genome features (genes, exons and introns) with reference insect genomes a custom Perl script was used, parsing the GFF3 and genomic fasta files describing the genomes of *D.melanogaster* (FlyBase-r5.29[11]), *T.castaneum* (Tcas2.0[12]) and *T.urticae*. The genome features of all three organisms are presented in the Table S2.4.1.

S2.4.2. Annotation of rRNAs, tRNAs and other non-coding RNAs

A total of 139 tRNAs were predicted using the program t-RNAscan SE[13]. For rRNAs the RNAmmer program was used to predict 56 rRNA sequences, using a cutoff score value linked to the rRNA type[14]. Small nuclear and nucleolar RNAs were annotated using the infernal software[15]. See Table S2.4.2 for a summary of non-coding RNAs in *T. urticae*.

S2.5. Resequencing of the Montpellier strain

To examine polymorphism in *T. urticae*, and to identify SNPs and small indels for use in future genetic mapping and population genetic studies, we sequenced a second strain of *T. urticae*. This strain was reared on detached bean leaves (cv. 'Contandor') in the laboratory of Dr. Maria Navajas from May 2003 to June 2008 (20 generations per year) with moderate population sizes in Montpellier, France (we subsequently refer to this as the Montpellier strain). DNA was prepared as for the sequencing of the London reference strain, and paired-end sequence data (2 x 35 bases, mean insert size ~104 bp/mode = 100 bp) was produced with the Illumina method (on an Illumina Genome Analyzer) at the DOE-JGI. A total of 43,444,913 million paired-end reads were produced corresponding to about 34-fold genome coverage (Table S2.5.1).

S2.5.1. MAQ mapping and SNP prediction

The Mapping and Assembly with Quality (MAQ) package was used for SNP and small indel prediction[16]. Paired-end reads were mapped using a maximum insert size of 800 bp and all other parameters were set to their default values. SNP and indel predictions were filtered by requiring a minimum read depth of 3 reads, a minimum consensus quality of 40, a maximum read quality of at least 60 for reads covering a candidate SNP position, a window size of 5 bases for masking SNPs adjacent to indels, and the number of haplotypes in the pool set to 2. Homozygous and heterozygous SNPs were separated using a simple shell command in which positions where MAQ called the consensus base as a degenerate nucleotide (where MAQ was unable to probabilistically define the exact base to be called) was classified as heterozygous while the remaining SNPs were classified as homozygous. Homozygous and heterozygous indels were filtered based on the frequency of reads supporting the indel; indels with > 75% support were classified as homozygous indels while the remaining indels were designated as heterozygous calls.

We found 401,949 SNPs in the Montpellier strain with more heterozygous SNPs (250,186) than homozygous SNPs (151,763; Table S2.5.1) likely reflecting the fact that sequencing was conducted on a pooled population of mites. When examining the patterns of homozygosity and heterozygosity across the genome, we observed genomic regions enriched in homozygous SNPs indicating regions that have/are becoming fixed during the inbreeding of the Montpellier strain (Figure S2.5.1). The continued high levels of heterozygosity likely indicate

only partial inbreeding of this strain. We called 67,957 deletions (31,254 and 36,703 homozygous and heterozygous, respectively) of sizes ranging from 1 to 15 bases as well as 72,694 insertions (38,649 and 34,045 homozygous and heterozygous, respectively) of 1 to 6 bases in length (Table S2.5.1 and Figure S2.5.4a).

S2.5.2. Annotation and impact of SNPs and small indels

SNP and indel effects were analyzed relative to the London reference annotation. For SNPs in coding sequences, we determined if the SNP changed the amino acid sequence or not (synonymous or nonsynonymous), and where SNPs changed the protein sequence, we determined the change (*i.e.,* missense changes, or in some cases premature stops; Tables S2.5.1, S2.5.2 and S2.5.3 and Figure S2.5.2). We examined the genome for major effect SNPs, herein defined as SNPs likely to have significant effects on gene structure. We found 132 changes causing premature termination codons (Table S2.5.3) that were slightly enriched at the 3' end of genes (Figure S2.5.2) in addition to 46 changes to the annotated start methionine residue and 47 SNPs that extended the protein coding potential of the annotated gene. These may reflect pseudogenization events in the Montpellier strain or could simply represent mis-annotated gene models as well as some level of false polymorphism discovery. Examining polymorphism levels across gene families revealed extensive variation in synonymous/nonsynonymous substition levels likely reflecting varying degrees of selective constraint on gene functions (Figure S2.5.3). Continued follow-up work on more mite strains will be important to further inform molecular evolutionary patterns across genes and gene families. Moreover, we assessed the distribution of indels in coding sequences by length (for instance, to see if frame-shifting indels were less represented than indels that were a multiple of three; Figure S2.5.4 b and c). There is a strong trend for coding indels to be enriched in multiples of 3 (aside from 1 bp indels expected to be the most common indel mutation type) that result in maintenance of the reading frame, regardless of whether predictions were for homozygous or heterozygous variants.

S3. Genome organization and characteristics

S3.1. Chromosomal features

S3.1.1. Holocentric chromosomes

In monocentric eukaryotes, kinetochores – the organelles that attach chromosomes to the spindle during mitosis and meiosis and to allow their segregation – assemble on a distinct region of each chromosome, the centromere. In contrast, holocentric species have diffuse kinetochores that form along the entire length of their chromosomes. *T. urticae* has such holocentric chromosomes, a feature of acariform mites[17] shared with other chelicerates and insect taxa. Although no chromosome-wide assembly of the genome is available yet for *T. urticae*, we searched the genome sequences for possible signs of holocentric chromosomal features. The cell division process, and the chromosome features linked to it, is best documented in monocentric species, with the nematode *C. elegans* being the model holocentric species[18]. No specific local DNA features that control kinetochore assembly have been identified in *C. elegans*, thus no candidates exist to identify particular sequence features for kinetochore assembly in *T. urticae*. However, chromosome compaction during mitosis, which involves the condensin complexes, differs between monocentric species and holocentric chromosomes in that holocentric chromosomes are predicted to be more rigid[18].

This structural difference correlates with the absence of several proteins from the condensin complex I in the *C. elegans* genome[19]. While condensin I is dispensable for chromosome compaction, its interaction with chromatin at the end of prophase compaction stabilizes anaphase chromosomes, allowing these to better resist the spindle forces[20]. We searched the *T. urticae* genome for the condensin I complex subunits missing in *C. elegans* (CAP-D2/CND1, CAP-G/CND3, CAP-H/CND2) and found clear homologues of these three proteins (CAP-D2: tetur01g08780, CAP-G: tetur17g00190, CAP-H: tetur01g05980). This argues against the absence of these proteins as a generic feature of holocentric chromosomes.

S3.1.2. Telomeres

Chromosome sequence motifs that mediate meiotic pairing and synapsis by binding specific Zn-finger proteins have recently been identified in *C. elegans*[21]. These motifs contain a minimal sequence consensus TTGGC, a close relative to the *C. elegans* telomeric repeat, TTAGGC. We searched for telomeric repeats in *T. urticae*, using the arthropod telomeric repeat (TTAGG)n found in most insects, and as short stretches in *I. ricinus*[22] as bait. No large repeats were uncovered, however many short repeats (ranging from n=4 to n=76) were found dispersed all over the genome, with a tendency to local clustering (Table S3.1.1, Figure S3.2.1). The occurrence of such repeats is not accidental, since replacing the arthropod TTAGG pentamer by the TTAGGG vertebrate hexamer one, returns a negative result, as did searches with the TTAGG pentamers where a single position has been changed, except for TTGGG which returns a few very short clusters which are falling inside or at the border of the TTAGG repeats we found.

The TTAGG repeats found close to a scaffold border (sca-45, sca-52, sca-54) were identified as putative telomeres. Close inspection and curation of sca-45 revealed a remarkable modular structure as expected for telomeres. The putative sca-45 telomere is a mosaic of short TTAGG repeats interrupted by non-LTR retrotransposons closely related to TRAS (Telomeric Repeat-Specific). All TRAS proteins, except one, are intact and likely functional copies, with a CDS split on two reading frames interrupted by +1 frame shift (or as two separate proteins). All the TTAGG repeats are on the same strand while the TRAS gag-pol CDS is located on the opposite strand. This structure is unexpectedly similar to the one observed in the silkworm *Bombyx mori* and in other Lepidoptera[23]. In addition to the intertwined TRAS retrotransposons and TTAGG repeats, the three putative telomere share one or several copies of a cluster of two genes in tandem, dubbed TelA and TelB. These proteins have no identifiable homologue in the databases but harbor both a unique C2H2-type Zinc finger motif CxxxxCx(Y,F)x14HxxxxH suggesting an interaction with DNA or RNA, further supported by a nuclear subcellular localization predicted by PSORT. These proteins do not occur elsewhere in the genome, suggesting a telomere-related function. In addition to these common features, other transposons and a few other protein-encoding genes are found in this area. Of particular interest are two BIR domain_IAP repeat-containing genes on sca-45 (yeast BIR domains are involved in regulating cytokinesis) and an archeal-type SMC chromosome segregation protein on sca-54.

S3.1.3. Interstitial telomeric repeats

The many other short TTAGG repeats presumably occur outside telomeric regions. Such interstitial telomeric repeats (ITR) are often seen as chromosomal aberrations[24] or as a mechanism to resist irradiation[25]. They have been experimentally observed in vertebrates and in a few other cases using FISH or PCR. We searched the fully sequenced arthropod genomes

and found these sequences to be commonplace. Their occurrence has probably been under-evaluated[26]. The largest *T. urticae* ITR cluster found in the middle of sca-13 (1083942…1085080) is located inside the second intron of tetur13g02620 encoding NAGLT1c, one of three tandem sodium-dependent glucose transporter genes. Interestingly, this intron has been invaded by a Merlin family DNA-II transposon. This transposon is likely active as both its transposase (1082632..1083591), and its Terminal Inverted Repeats [(TIR): 5'TIR, 1081781..1082128 and 3'TIR, complement (1085127..1085473)] as well as a clear-cut 8 bp Target-Site Duplication TAGTACAT[27] are conserved. The ITR is entirely located between the TIRs, suggesting that this ITR can be moved within the genome, allowing evolutionary plasticity for this ITR whatever its function. In further support for this point, another ITR is located in between two Merlin TIRs on sca-677. Two recent publications[28,29] conclude that some proteins bound to human telomeric repeats also bind ITRs, and could thereby regulate nearby gene transcription and impact long range organization of the chromosome. Our observations of ITRs in the *T. urticae* holocentric genome suggest that the proposed role of ITRs in regulation and chromosome organization is not unique to vertebrates, but is likely universal among eukaryotes.

S3.2. Annotation of transposable elements

An iterative-blast search approach was used to identify putative transposable elements (TE) in the genome of *T. urticae*. To this end, tblastn was used to search for homologs (at the amino acid level) in RepBase. Since there are no chelicerate-specific transposable elements present in RepBase, we relaxed the parameters in order to collect protein domains typically associated with TEs. To obtain close-to-full-length ORFs, the obtained hits were then extended as far as possible on the same reading frame until a stop codon was reached. This process was iteratively repeated with the collected ORFs to find more divergent transposon sequences that RepBase was unable to find. In parallel, LTR_seq[30] was run and regions between LTR-couples were screened using HMMs for TE-specific domains. In addition, RepeatScout and RepeatClass[31] were used, the latter to classify the already collected set of TEs.

The obtained set of TEs was then used as a library for *T. urticae* TEs (Table S3.2.1**)**. Further cleaning was done based on the presence/absence/order of TE-domains and by collapsing (nearly) identical copies. This trimmed set was then clustered into families using TribeMCL[32]. Based on this procedure, LTR, LINE, TIR and Helitron elements were readily identified. Mavericks (Polintons) were much harder to clearly delineate since they are composed of 5-10 ORFs spread over 10-20Kb terminated by inverted repeats (TIRs) of 50 to a few 100 bps, and have a ~6 bp target site duplication and a 5'-AG and a 3'-TC terminus. However, in the case of *T. urticae*, the TIRs are quite long and can be clearly identified. Distribution of the two most abundant TE families (Gypsy and Mariner) along the ten longest scaffolds in windows of 100kb is shown in Figure S3.2.1.

S3.3. Annotation of microsatellites

All perfect microsatellites (SSRs) were identified that have a minimum number of repeat units depending on the size of the repeat pattern. These thresholds ensure that slippage is actually possible and that mutations by replication slippage are likely to happen. Those are thus not "just" repeats but rapidly mutating microsatellites. Applied thresholds were:

1. mono-nucleotide repeats: 12 repeats, i.e. 12 bp repeat
2. di-nucleotide repeats: 6 repeats, i.e. 12 bp repeat

3. tri-nucleotide repeats: 5 repeats, i.e. 15 bp repeat
4. tetra-nucleotide repeats: 4 repeats, i.e. 16 bp repeat
5. penta-nucleotide repeats: 4 repeats, i.e. 20 bp repeat
6. hexa-nucleotide repeats: 4 repeats, i.e. 24bp repeat

S3.3.1 Microsatellites on a whole-genome scale

The spider mite genome is the smallest sequenced arthropod genome to date. Since repeat content of genomes typically scale with genome size, it is expected to find a relatively low abundance of microsatellites which clearly is the case (Figure S3.3.1A). However, when these numbers are normalized with the genome size, the differences, in particular between tick and spider mite, become less pronounced (Figure S3.3.1B). This means that spider mite has relatively few microsatellites, correlated to small size of the genome. Microsatellite density is still among the lowest for arthropods, but remains similar to *Tribolium castaneum* (tcas) and *Aedes aegypti* (agyp). Across the 12 *Drosophila* species, 68% of microsatellites are intergenic and only 28% and 4% occur in introns and exons, respectively. In spider mite, 30% and 8% of microsatellites are located in introns and exons, respectively. A striking feature of spider mite is its unique profile of microsatellite classes (p1 to p6). Mono-nucleotide repeats are virtually non-existent, and even di-nucleotide repeats, normally the most abundant type of microsatellite, are found markedly less often than tri-nucleotides.

446/640 scaffolds (70%) do not contain any microsatellites (Figures S3.3.2 and S3.3.3). Across the remaining scaffolds, microsatellite average proportion per scaffold is low on average (0.43%), although there are a few extreme outliers.

S3.4. U2-type and U12-type Introns

Two types of spliceosomal introns occur in eukaryotes: U2-type introns spliced through the major spliceosome pathway and U12-type introns spliced through the minor spliceosome pathway[33-35]. The U12 introns are rare (from 621 in humans to 19 in *D. melanogaster*) and are essentially found in a unique subset of genes, among which are many genes involved in developmental and neuronal functions. U12 introns are completely absent from some taxa and have been lost repeatedly in multiple invertebrate lineages[36]. Based on a search for conserved donor and branch-point patterns, and also a comparison to U12 introns in orthologs of the 38 U12-harboring genes from honeybee selected from U12DB[37], the genome of *T.urticae* was found to contain only three U12-type introns (Table S3.4.1). This is the lowest number ever observed in a eukaryote to date. All three U12 introns in spider mite have AT-AC borders. One of the three U12 introns is conserved within both invertebrates and vertebrates and a second one only within invertebrates. Interestingly, the third one is a novel U12 intron in every respect, meaning that it occupies a novel position. In 14 cases, U12 introns found in honeybee are changed into U2 introns in *T. urticae* while in the remaining 22 cases the introns have been lost. This loss of U12 introns is not a generic chelicerate-specific feature as the homologs of the honeybee U12-harboring genes in the tick *Ixodes scapularis* most often harbor U12-type introns as well. Indeed, of the 26 honeybee U12 intron-containing genes (totaling 27 U12 introns) for which a homolog can be found in *I. scapularis*, 22 are also found as U12-containing in the tick genes totaling at least 26 U12 introns, with four tick genes having more U12 introns than honeybee. Additionally, the 38 *T. urticae* genes analyzed contain a much lower number of introns (U2 and U12) than their homologs in honeybee (x 2.9) and tick (x 4.1).

The major and minor spliceosomes have many components in common and a few specific features for intron recognition and splicing, including snRNAs. The genes encoding the major spliceosome snRNAs U1, U2 and U6 (2 copies each), U4, and U5 have all been found in the genome of *T. urticae*. For the minor spliceosome, all snRNA genes have been found in single copy (U11, U12, U5 and U6atac) except for U4atac, which has not been found, perhaps because it is too divergent. Besides snRNAs, seven U11/U12 snRNP proteins have also been shown to be unique to the minor spliceosome in humans[35]. While up to six are found in insects as well as in the tick *I. scapularis*, only one, a truncated version of the 65K snRNP protein, was found in *T. urticae*. This reduction/truncation of U12 splicing partners in *T. utricae* suggests that the species is in the process of losing U12-type splicing ability, having found a specific way to splice the few remaining U12 introns. This is not a generic feature of chelicerates, because the tick genome harbors many more U12 introns and has a full repertoire of U12-specific spliceosomal proteins. This therefore raises the possibility that selection on genome size (or another unknown force) is driving the loss of U2 and U12 introns with the associated loss of genes usually required for U12 splicing.

S4. Small RNAs

S4.1. RNA sample preparation and small RNA sequencing

We sequenced three different Illumina small RNA libraries from *T. urticae* that were generated with RNA from *i*) embryos, *ii*) a 1:1 mix of nymph and larvae, and *iii*) adults. Briefly, the small RNA sequencing (smRNA-seq) libraries were constructed from total RNA isolated using Trizol reagent (Carlsbad, CA). Bar coded adapter sequences were ligated onto the small RNAs for sequencing, allowing multiple samples to be run in a lane. The library construction and sequencing was performed by Fasteris SA (Geneva, Switzerland). Read lengths of 38 bp, after barcode trimming, were generated on the Illumina GAII platform. Bar-code and adapter sequences were removed from the reads prior to downstream analyses (performed by Fasteris SA).

A total of more than 10.6 million reads were sequenced, corresponding to 2.4 million unique sequences (Table S4.1.1). We further filtered the reads for low-complexity repeat sequences, and kept only reads having at least three counts, with a length ≥ 18nt, and that mapped perfectly to at least one location in the *Tetranychus* genome. Finally, we filtered out small RNAs corresponding to rRNAs, tRNAs and a number of non-coding RNAs, resulting in a final set of 226,829 unique small RNAs that mapped to 676,266 different loci in the genome. The number of unique small RNA counts per size category shows a peak at 21 and 26 nt respectively, for all developmental stages (Figure S4.1.1). The number of unique small RNA sequences is higher in embryos, larvae and nymphal developmental stages compared to adults (Figure S4.1.1). The two peaks at 21 and 26 nt correspond to siRNAs plus miRNAs and piRNAs (see also S4.2), respectively, similar to what is observed in *Drosophila*[38].

Endogenous siRNAs and piRNAs have been characterized as a defense mechanism against transposable elements in various organisms[39]. To assess whether this is likely in *T. urticae*, we performed a randomization test where small RNAs mapped to transposable elements are compared to small RNAs mapped to random segments of the genome. We followed a procedure defined in a previous study[40]. We used a set of 12,933 high quality annotated transposable elements (TE) in *T. urticae*. TE coordinates were transformed into segments of 250 nt, overlapping by 125 nt. A total of 10,000 bins were selected at random from this pool.

This set was compared with a genome wide set of 1000 sets, each containing 10,000 genome bins randomly selected, having the same size as the TE bins. We further separated the small RNAs in two sets, one for sRNAs having a length between 20 and 25 nt (siRNAs), and one for sRNAs having a length between 26 and 28 nt (putative piRNAs). A total of 29,728 siRNAs were matching the TE set, versus an average of 11,127 ± 1,477 siRNAs for the genome sets, defining a z-score of 12.6. For the piRNAs, a total of 20,145 piRNAs were matching the TE set, versus an average of 4782 ± 286 for the genome sets, thus defining a z-score of 53. Since both categories show very high z-scores (12.6 and 53, respectively) corresponding to extremely small p-values ($p \leq 0.00001$), this suggests that both siRNAs and piRNAs are very likely associated to the silencing of transposons in *Tetranychus*.

S4.2. Characterization of piRNAs

Earlier studies have found that piRNAs are widely distributed in animal genomes, including mammals, but also in insects such as *Drosophila*[38,41,42]. In both mammals and *Drosophila*, distinguishing features of piRNAs include *i*) longer length relative to siRNAs or miRNAs (piRNAs are about 24-28 nt), *ii*) a preponderance of 'U' residues at the first (5') base of the small RNA ('T' residues after cDNA sequencing), and the so-called "ping-pong" signature of sense and antisense reads that is thought to arise as a consequence of the biogenesis of piRNAs[41].

In our analysis of *T. urticae* smRNA lengths, we observed a bimodal length profile, with the second peak in the profile of approximately the same size as expected for piRNAs from other animals (Figure S4.1.1.). We therefore examined first base composition, and searched for potential ping-pong signatures, by small RNA length to provide additional evidence that the larger smRNAs might be piRNAs. As assessed with all smRNAs, we observed a strong preference for 'T' residues for the smRNAs of lengths > 22 nt (this was most pronounced for lengths 24-28 nt, where the T at the first base predominated) (Figure S4.2.1). This expectation is consistent with the expectation that a large fraction of the longer smRNAs are piRNAs.

To provide additional (and more definitive) evidence that the longer smRNAs are analogous to piRNAs, we also tested for the ping-pong signature of sense and antisense overlaps in smRNA alignments to genomic sequences (a peak in sense/antisense overlaps is expected at 10 bp for piRNAs[41,42]). To do this, processed small RNA reads with read lengths 18-28 bases were aligned to *T. urticae* annotated transposon sequences with MAQ[16]. Sequences were mapped with up to 3 mismatches and 10 maximum hits (-C 10) while all other parameters were set to the default values. The resulting mapping file was converted into a *.sam alignment file that was parsed to look for signatures of the ping-pong amplification as a function of smRNA size. To do this, we found the base pair overlap between sense and anti-sense mappings in the alignments. For every mapping, we found all possible 'partner' smRNA (other strand to the 'target' smRNA in question) mappings within 28 bases of the 5' base of the target smRNA (this approach is similar to analysis conducted by Brennecke *et al.*[42]). After identifying all 'partner' mappings, the distance between the 5' end of the target and 5' end of the partner smRNAs were calculated to generate a frequency histogram of base pair overlaps between a given target smRNA and its partner smRNAs. We then tabulated the cumulative overlap profiles by smRNA read lengths. As shown in Figure S4.1.3, an obvious peak in overlap was observed at 10 bp for most of the smRNA lengths, matching the expectation for piRNAs. The magnitude of the ping-pong signature was strongest for the longer smRNAs, especially those of about 24-28 nt, consistent with the above analyses that the longer smRNAs in our sample are analogous to piRNAs in other animals.

S4.3. Annotation of microRNAs

Small RNAs having a length between 18 and 25 nt, and having less than 30 loci on the genome, were selected for further analysis as potential miRNAs. miRNA sequences were predicted using a stringent approach, based on the analysis of secondary structure coupled to the detection of typical Drosha/Dicer cleavage products (see Figure S4.3.1). We then predicted all possible stable secondary structures in the *T. urticae* genome, using the program RNALFold[43] with a sliding window length of 120 nt. We filtered out all the hairpin structures having a free energy ≤ 25 Kcal/mol, retaining a set of 42,919 hairpin sequences for further analysis. The selected small RNAs were then mapped to the hairpin structures and the following criteria were used to define bona fide miRNA structures: the miRNA/miRNA* duplex must be supported by sequenced reads for both the miRNA and the miRNA*; it must be located in a stem region having no more than 5 unpaired bases, and where bulges occur they must not exceed 3 bases; there also must be a differential expression between the miRNA and the miRNA*. The miRNA is defined as the sequence having the highest expression level. The miRNA/miRNA* duplex must display the 2 nt overhangs that are characteristic of the Dicer processing of the stem-loop structure on each side (with a tolerance of 1 nt). All precursor sequences were checked with the randfold program[44] and only those having a p-value ≤ 0.05 were retained. Candidate miRNAs were also checked against a database of *Tetranychus* transposable elements. All candidates showing a high degree of similarity were excluded. All *T. urticae* miRNA sequences were submitted to the miRBase repository[45].

This way, a total of 52 miRNAs were identified in the *T. urticae* genome. Based on the identity of their seed regions (nucleotides 2-7 of the miRNA sequence), the *T. urticae* miRNAs can be grouped into 43 families. We compared the predicted miRNAs to all the annotated arthropod miRNAs contained in the miRBase repository[45] (3,277 sequences covering 27 different species). With no more than two mismatches allowed, the number of miRNAs conserved per species varies from 0 to 20 (Table S4.3.1 and Table S4.3.2). If we consider all the *Tetranychus* miRNAs that have at least one match in the current miRBase arthropod miRNA list, we have a total of 26 conserved sequences (Table S4.3.3). Because there might be novel miRNAs in our set that have not been previously characterized, we also aligned all the *Tetranychus* miRNAs against the completely sequenced arthropod genome of the crustacean *D. pulex* and available sequences of *Ixodes scapularis.* However, this procedure did not reveal more conserved miRNAs than the comparison with the miRBase annotations. The remaining 26 *Tetranychus* miRNAs, that do not match any existing arthropod miRNA sequence, might thus be *T. urticae* or lineage specific (Table S4.3.4).

A few miRNAs (7) are located within introns, in the same orientation as the gene. Within this set, we found at least three miRNAs that are located within very short introns (~55 nt), with the miRNA mature sequence at the border of one of the splice sites (Figure S4.3.1c). Those three sequences might be mirtrons (miRNAs that bypass the Drosha processing step)[46].

We used the RNA hybrid software[47] to predict targets for all miRNAs. The 3' UTR of all *T. urticae* genes were used as target sequences. We kept only miRNA:mRNA duplexes having a p-value ≤ 0.01 and for which all bases located in the seed region (nucleotides 2-8 at the miRNA 5' end) are engaged in Watson-Crick base pairs. This analysis predicted the targets of all 52 miRNAs. We found a total of 3,338 miRNA-target pairs, representing 2,615 target genes (~14% of the total number of genes). There are on average 64 predicted targets per miRNA.

We have also analyzed the expression fold change for all miRNAs after normalization of the expression values. The counts for each miRNA and for each developmental stage were normalized by the number of occurrences for a given miRNA in the genome and were scaled according to each library size. The expression values were then log2 transformed for all the values above or equal to one, all other values being declared as missing data, to avoid the introduction of spurious values. Log fold changes are then calculated for each pair combination of the three developmental stages (embryo, larvae and adult). We found that 49 miRNAs have a fold change ≥ 2 between two developmental stages (Table S4.3.5). There are slightly more miRNAs differentially expressed during the embryo-larvae and larvae-adult stages (34 and 36, respectively) compared to the embryo-adult stages (30). This result might indicate a more prominent role for miRNAs during the early stages of development in *T. urticae*. Taken together, our results reveal a rich and diverse set of small RNAs in the *T. urticae* genome.

S5. Comparative genomics

S5.1. Dollo analysis

S5.1.1 Construction of the datasets and dollo analysis

For the dollo analysis, a selection was made of different arthropods and outgroups. For arthropod reference organisms, we downloaded the predicted protein sequences of *Drosophila melanogaster* (Flybase, FlyBase-r5.29 from 26-Jun-2010), *Tribolium castaneum* (BCM, annotation Tcas2.0 from 16-May-2008), *Nasonia vitripennis* (BCM, annotation Nvit_1.0 from 15-Dec-2008) and *Daphnia pulex* (JGI, version 1.1, Filtered Models). The protein sequences of *Homo sapiens* (CCDS: 13-Dec-2010 on genome build HsGRCh37.1) and *Nematostella vectensis* (JGI, version 1.0, Filtered Models) were downloaded as outgroup organisms. In cases where alternative splice variants were detected, only the longest transcript was used.

N. vitripennis: Baylor College of Medicine, annotation Nvit_1.0 from 15-Dec-2008[48]
http://www.hgsc.bcm.tmc.edu/project-species-i-Nasonia%20vitripennis.hgsc
http://www.hgsc.bcm.tmc.edu/ftp-archive/Nvitripennis/annotation/

T. castaneum: Baylor College of Medicine, annotation Tcas2.0 from 16-May-2008[12]
http://www.hgsc.bcm.tmc.edu/project-species-i-Tribolium%20castaneum.hgsc
http://www.hgsc.bcm.tmc.edu/ftp-archive/Tcastaneum/Tcas2.0/annotations/

D. melanogaster: Flybase, FlyBase-r5.29 from 26-Jun-2010[11]
http://flybase.org/
ftp://ftp.flybase.net/releases/FB2010_06/dmel_r5.29/fasta/

D. pulex: JGI, version 1.1, Filtered Models[49]
http://genome.jgi-psf.org/Dappu1/Dappu1.home.html
http://genome.jgi-psf.org/Dappu1/Dappu1.download.ftp.html

H. sapiens: The Consensus CDS Project. CCDS: 13-Dec-2010 on genome build HsGRCh37.1[50]
http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi

N. vectensis: JGI, version 1.0[51]
http://genome.jgi-psf.org/Nemve1/Nemve1.home.html
http://genome.jgi-psf.org/Nemve1/Nemve1.download.ftp.html

To delineate gene families, a similarity search was performed (all-against-all BLASTP; E-value cutoff E-05) with all proteins of the dataset (seven species, including *T.urticae*). Gene families were constructed with MCLBLASTLINE (Inflation Factor of 2.0) (http://micans.org/mcl/[32,52]) based on the BLASTP analyses. For all gene families, phylogenetic profiles were constructed that reflect the presence or absence of a particular gene family in a particular species.

The parsimonious evolutionary scenario, which included the loss and acquisition of gene families mapped onto the branches of the phylogenetic tree, was inferred by using the DOLLOP program of the PHYLIP package[53]. These results were further used to delineate the minimal gene set for the different ancestral nodes. The DOLLOP program is based on the Dollo parsimony principle, which assumes irreversibility of character loss[54]. To verify the DOLLOP results, a second algorithm named CAFE[55] was used, which uses the number of genes in each family to model the evolution of gene families. In doing so, CAFE not only reports on gains and losses, but it also estimates the most likely number of genes for each family at each of the ancestral nodes of the species tree (data not shown). A gain of a gene family can be considered as a particular case of an expansion for which the gene family was not present in the previous ancestral time point. Gene family loss is again a particular case of a reduction for which an ancestral gene family completely disappears in a descendant species[56].

Genes and gene families were functionally annotated using Gene Ontology (GO)[57]. In a first step, all proteins were annotated by using Blast2GO[58] or Interpro2GO to assign the proteins to GO categories. Proteins mapped to a particular GO category were also explicitly included into all parental categories. GO annotation per family was obtained by listing the GO labels for all of the genes of that family. A weight, equal to the percentage of genes with GO annotation within the same subcategory (molecular function, cellular component, biological process) that carried this label, was attached to all of the GO labels. Only GO labels with a weight greater than 40% were considered as representative for the family. GO labels occurring in fewer than 10 gene families were discarded. The statistical significance of functional GO enrichment was evaluated by using the hypergeometric distribution, whereas multiple hypotheses testing was done by using FDR correction[59].

Using DOLLOP and CAFE, we were able to map the different gene family loss and gain events onto the branches of the phylogenetic tree (see Figures 1 and S5.1.1). In general, both methods predict similar tendencies. Small discrepancies in absolute numbers between DOLLOP and CAFE result from the different underlying assumptions between both methods[56]. To link the loss and gain events with information about the molecular function or biological process the corresponding gene families are involved in, the Gene Ontology (GO) vocabulary was used[57].

S5.2. Gene Family Expansions

For all gene families present in *T. urticae* and at least two other species, the mean gene family size and standard deviation of the phylogenetic profiles were calculated. The matrix of these profiles was transformed into a matrix of *z*-scores to centre and normalize the data.

Subsequently, the profiles of interest were hierarchically clustered by complete linkage clustering based on Pearson correlation values as a distance measure. Clustering and visualization was performed by Genesis[60]. A description was added to all families based on the most frequently occurring *T. urticae* gene descriptions in the corresponding gene family.

Families that are expanded in *T. urticae* but not in most other organisms were extracted. Therefore the gene copy numbers of the gene families present in *T. urticae* and at least two other species were transferred into z-scores. Next, the families with a z-score greater than two in *T. urticae* were extracted, representing families that are significantly expanded in *T. urticae* in comparison with the other organisms. Gene descriptions were added and the z-profiles were visualized using Genesis (see Figures S5.2.1 and S5.2.2).

S5.3. Transcription factors

T. urticae transcription factors (TFs) were predicted with 167 DNA binding domains (DBDs) described in Pfam (v24.0), 38 DNA binding families of 'Superfamily'[61] and transcription-related GO terms. All protein sequences were searched against the Pfam DBD HMMs by pfam_scan.pl with Pfam GA cut off, and scanned against InterPro by IPRscan. Only putative TFs matching a described Pfam DBD, a DNA binding family, and transcription-related GO term, were extracted. However, the collection thus obtained can still contain false positives such as proteins from the basal transcriptional apparatus (DNA polymerases), chromatin alterations, DNA packaging (histones), etc. To delete false positives, all putative TFs were searched against the non-redundant protein database and *D. melanogaster* TFs by BlastP (E-value cutoff e^{-6}) and annotated with the functional description based on the protein homologue with the lowest E-value. If the function description of the TF contained a "keyword" not related to transcription factors (for example: polymerase, histone, splicing factor, etc.) and there was no homology to any of the *D. melanogaster* TFs[62], the TF was discarded. The remaining TFs homologous to *D. melanogaster* TFs were manually checked. Moreover, to obtain true negatives, we manually checked the *T. urticae* proteins not predicted as TFs but homologous to *D. melanogaster* TFs. Finally, gene structures of *T. urticae* TFs were manually curated.

Using this approach, we found a total of 772 TFs in the genome of *T. urticae*, comprising ~4.2% of all *T. urticae* genes. In eukaryotes, approximately 3-5% of all genes usually encode TFs. Of the 772 TFs, 734 TFs are similar to *D. melanogaster* TFs, 33 TFs are specific to arthropods, 8 TFs are specific to insects, and 462 TFs are specific to animals. *T. urticae* TFs are divided into 49 families compared to 50 TF families of *D. melanogaster* (see Figure S5.3.1). BESS TFs appear to be specific to *D. melanogaster* and some other insects while these are missing from *T. urticae*. TFs with the zf-C4 DBD are more expanded in *T. urticae* than in *D. melanogaster*. On the contrary, TFs with the zf-C2H2 DBD are remarkably reduced in *T. urticae* compared *D. melanogaster*.

S5.4. Sex determination

A likely contributing factor to the ability of spider mites to rapidly develop pesticide resistance is their haplo-diploid sex determination[63]. Unfertilized (haploid) eggs will give rise to males while fertilized (diploid) eggs will develop into females. This sex determination mechanism creates a unique system where recessive mutations are exposed to selection directly in males. The system of complementary sex alleles, as described in haplo-diploid hymenopterans[64,65] is unlikely to apply to spider mites. In spider mites, inbreeding of mother-

son matings for many generations does not result in sex aberrants, diploid males, or drastic increase of egg lethality[66,67], and the combination of two identical haploid male chromosomes results in the development of females[68]. Consistent with the model of evolution of sex determination that predicts a higher level of conservation at the base of sex-determining cascades[69-72] the ortholog of *doublesex* (*dsx*) is found in the *T. urticae* genome (see Figure S5.4.1). Upstream genes in the core of sex determination pathway of insects that include *complementary sex determiner, feminizer*, *transformer* and *transformer*-2 were not found. In *Drosophila,* genes which are responsible for determining the relative number of X chromosomes to autosomes and relaying the sex-specific X:A signal to *Sex lethal* (*Sxl*) - *daughterless*, *deadpan*, *scute* and *groucho* - although present in the spider mite genome, are unlikely to be involved in sex determination because the spider mite does not have sex chromosomes. Orthologs of some cofactors such as *hopscotch*, *female lethal d* and *virilizer* that are not sex-specifically expressed but necessary for *Sxl* function sex determination in *Drosophila* are found in the spider mite genome. *Sxl*, however, is absent in the *T. urticae* genome, suggesting that these genes either regulate unknown sex determination genes in the spider mite or are not involved in sex determination.

S5.5. Proteins containing one LY6_UPAR extracellular domain

In the genome of *T. urticae*, we have found 23 small proteins containing one LY6_UPAR extracellular domain, which is known to adopt a characteristic Three Finger Domain (TFD) fold (see Table S5.5.1). The TFD domains are found in membrane GPI-linked glycoproteins playing essential roles in cell adhesion, signalling and lipid metabolism. Comparison of the 23 *T. urticae* TFDs and their homologues in insects (*Apis, Tribolium* and *Drosophila*) and one crustacean (*Daphnia*) has led to the identification of a core of 9 conserved TFD proteins that constitute a signature of all the arthropod genomes analysed in so far. Four of these conserved genes are known to be involved in flies in the formation of Septate Junction cell adhesion structures, suggesting a common ancient role for these proteins in the arthropod lineage.

However, the members of this gene family are particularly prone to duplication, paralogue divergence and functional diversification. Consistently, in *T. urticae*, we find a group of 13 species-specific TFD paralogues that could assume specific roles in the chelicerate lineage. We have also annotated in *T. urticae* the 6 conserved TGFbeta/activin receptor members, whose ligand-binding ectodomain adopts a TFD fold.

S6. Feeding and detoxification

S6.1. Gene families for digestion, detoxification enzymes and transporters

S6.1.1. CYP genes encoding P450 enzymes

The *T. urticae* gene models were searched by tblastn[73] with insect CYP sequences representing the CYP2, 3, 4 and mitochondrial P450 clans. All the models with predicted proteins that included the canonical heme binding sequence were verified manually for the presence of the other key features of P450 enzymes[74] and the gene model corrected whenever necessary (incorrect predictions such as fusions with adjacent genes or fragmentation) or possible (when EST sequences were available). Models that lacked the heme binding sequence or that were incomplete were analyzed further as follows. The 4 kb region of the scaffold surrounding the putative CYP gene was analyzed by blastx[73] against all *T. urticae*

CYP sequences to identify neighbouring exons, resulting in the generation of new gene models. Pseudogenes and gene fragments (detritus exons) were separated from putative full length CYP coding sequences. The latter were aligned by MUSCLE 3.8.31[75] and a tree constructed by PhyML 3.0 on www.phylogeny.fr using the default substitution model assuming an estimated proportion of invariant sites of 0.017 and 4 gamma-distributed rate categories. The gamma shape parameter was estimated directly from the data (gamma=1.689). Reliability for internal branches was assessed using the aLRT test (SH-Like). The four deepest branches of the tree correspond to the four CYP clans found in insects and crustaceans[76] as shown by introducing insect and crustacean sequences without modifying the overall topology of the tree. All sequences were submitted to the P450 nomenclature committee (D. Nelson, Univ. Tennessee) for naming. Four sequences (CYP302A1, CYP307A1, CYP314A1 and CYP315A1) were assigned as the only orthologs of the eponymous P450 of insects and crustacea. Their predicted sequence was verified by RT-PCR and sequencing of their cDNAs, Figure S6.1.1, Table S6.1.1.

S6.1.2. Glutathione-S-transferases (GST)

A HMM profile was constructed using all glutathione-S-transferase (GST) sequences of *Drosophila melanogaster*, *Anopheles gambiae*, *Bombyx mori*, *Apis mellifera*, *Daphnia pulex*, *Caenorhabditis elegans* and *Mus musculus*. This profile was used with the HMMER program (http://hmmer.wustl.edu/) to search against the predicted proteome of *T. urticae*. Next, a tblastn[73] search (E-value cutoff=10^{-4}) of all HMMER hits corresponding to complete GSTs (containing the GST N-and/or C-terminal domain as defined by PFAM domains PF00043 and PF02798 and having a sequence length larger than 200 amino acids) was conducted against the genome of *T. urticae*. This combined approach resulted in the annotation of 32 complete *T. urticae* GSTs. Gene models were refined or created on the basis of homology and EST support, and manually created gene models were confirmed with RT-PCR and sequencing. Protein blast searches of the 32 *T. urticae* GSTs on the NCBI website revealed that one of these sequences showed a very high similarity (E-value= 10^{-42}) to kappa-class GSTs, which are localized in the mitochondria. The remaining 31 cytosolic T. urticae GST proteins were aligned with those of *D. melanogaster*, *A. gambiae*, *A. mellifera* and mu-and delta class GST proteins of Acari (for accession numbers see Table S6.1.3) using MUSCLE version 3.8.31[75]. Model selection was done with ProtTest 1.4[77] and according to the Akaike information criterion the LG+I+G model was optimum for phylogenetic analysis. Finally, a maximum likelihood analysis was performed using Treefinder[78], bootstrapping with 1000 pseudoreplicates. The resulting tree was midpoint-rooted (Tables S6.1.3[79,80] and S6.1.4, Figure S6.1.2).

S6.1.3. Carboxyl/cholinesterases (CCE)

A HMM profile was constructed using all carboxyl/cholinesterases (CCEs) of *D. melanogaster*, *A. gambiae* and *A. mellifera*. This profile was used with the HMMER program (http://hmmer.wustl.edu/) to search against the predicted proteome of *T. urticae*. Additionally, a tblastn[73] search (E-value cutoff=10^{-4}) with the acetylcholinesterase gene of *T. urticae* as query, was conducted against the genome of *T. urticae*. Pseudogenes and gene fragments (17) were separated from putative full length CCEs and gene models were refined on the basis of homology, EST and RT-PCR support. A final selection of 71 CCEs from *T. urticae* were aligned with a representative set of CCE sequences of *D. melanogaster*, *A. gambiae* and *A. mellifera* using MUSCLE version 3.8.31[75]. The resulting alignment was trimmed at both ends according to Claudianos *et al.*[81]. Model selection was performed with ProtTest 1.4[77] and

according to the Akaike information criterion the WAG+I+G+F model was optimum for phylogenetic reconstruction. Finally, a maximum likelihood analysis was performed using Treefinder[78], bootstrapping with 500 pseudoreplicates. The resulting tree was midpoint rooted, Figure S6.1.3 and Tables S6.1.5 and S6.1.6[80,82].

S6.1.4. ABC transporters (ABC)

ABC transporters were identified in a similar way as for *Daphnia pulex*[83]. Briefly, tblastn searches[73] were performed on the *T. urticae* genome sequence assembly using the highly conserved nucleotide binding domain (NBD)[84] of *D. melanogaster* ABC proteins as queries. One search was carried out per subfamily (ABC-B and ABC-C), each using the sequence of the NBD of the representative *Drosophila* protein (B: CG3879 (mdr49); C: CG9270). If the *Drosophila* transporter had two NBDs, the N-terminal domain was used. All hits with an E-value less than 10^{-4} were withdrawn for analysis and gene models were refined or created on the basis of homology and EST support. To assign putative *T. urticae* ABC genes to B and C subfamilies, the NBDs of the gene models were extracted using the ScanProsite facility (ExPASy), with predicted protein sequences and the Prosite profile PS50893. NBDs were then subjected (after aligning with MUSCLE 3.8.31[75]) to a phylogenetic analysis to confirm their position within classes together with NBDs of *D. melanogaster* and human transporters, using neighbour joining (JTT-model, pairwise deletion, default parameters) and bootstrapping with 1000 replicates in the program package MEGA4[86]. NBDs of *D. melanogaster* and human ABC transporters formed subfamily specific clusters, with separate groupings for the N- and C-terminal NBDs of full transporters. This allowed unequivocal assignment of *T. urticae* ABC-C and ABC-B transporters. The subfamily assignment was further confirmed by protein BLAST analyses of the manually corrected models on the National Center for Biotechnology Information website. A set of 39 ABC-C and 4 ABC-B (2 half and 2 full transporters) were identified (Table S6.1.7, Table S6.1.8[83,87]).

S6.1.5. Intradiol ring-cleavage dioxygenases (ID-RCD)

A tblastn-search (E-value=10^{-4}) was conducted against the *T. urticae* genome using the protein sequence of protocatechuate dioxygenase of *Streptomyces avermitilis* (GenBank accession number NP_826938) as query. Pseudogenes (1) were separated from putative full length intradiol ring-cleavage dioxygenases (containing the cd03457 domain in the NCBI conserved domain database (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) and gene models were refined on the basis of homology, EST and RT-PCR support. This resulted in the annotation of 16 complete *T. urticae* intradiol dioxygenases (Table S6.3.1).

S6.1.6. Peptidases

The *T. urticae* gene models were searched by blastp with aspartic, serine and cysteine peptidase sequences in a recurrent way. First, a complete amino acid insect sequence from data banks corresponding to a protein of the family was used. Then, the retrieved protein sequences of the spider mite were used to search in its own proteome. All the models with predicted proteins were verified manually for the presence of key features of the different peptidases, and the gene model corrected whenever necessary or possible. The same method was used for searching peptidases in the NCBI RefSeq databases for *Tribolium castaneum* (Build 2.1), *Caenorhabditis elegans* (WS190), *Apis mellifera* (Amel 4.0), *Drosophila melanogaster* (Build 5.22), and *Homo sapiens* (Build 37.2). Retrieved sequences were aligned by MUSCLE 3.8.31[75]. Alignments ambiguities and gaps were excluded from phylogenetic

analysis using GBLOCKS version 0.91b[88]. The resulting *T. urticae* gene models are shown in Tables S6.1.9, S6.1.10 and S6.1.11. Trees were constructed by PhyML 3.0[89] on http://www.phylogeny.fr using default parameters. Reliability for internal branches was assessed using the aLRT test (SH-Like). Trees were visualized by MEGA4[86], see Figures S6.1.4 and S6.1.5.

S6.2. Spider mite adaptation response to different host plants

S6.2.1. Preparation of biological material

To examine response of mites to growth on different hosts, we reared mites on beans (*Phaseolus vulgaris* cv 'California Red Kidney', the long standing host for the London strain) for four days prior to sample collection. At four days, bean plants were washed with a solution of 0.03% Tween 20 in water to collect mites (6 trays with ~450 bean plants were used). The suspension was filtered through sieves of decreasing diameter, and eggs were collected on 0.1 mm sieves (~500,000 eggs). To have a single layer of eggs, the eggs were plated and air-dried on Petri dishes (between 20,000 and 30,000 eggs per plate). Plates were sealed with parafilm and incubated at 27 °C under a constant light cycle for 48 h to allow hatching.

S6.2.2. Inoculation of plants

Larvae were collected after hatching and gently deposited one by one on plants using a wet brush. Three different species of plants were used as hosts for the mite larvae (Table S6.2.1): 1) bean (*Phaseolus vulgaris* cv 'California Red Kidney', the standard host for the London strain): 3 replicates x 2 plants (2 plants per pot), 2 weeks old (2 leaves), were inoculated on both leaves with ~600 larvae per plant (~1200 larvae per replicate); 2) tomato (*Solanum lycopersicum*; genotype Heinz 1706): 3 replicates x 2 plants (2 plants per pot), 3 weeks old (2 leaves), were inoculated on both leaves with ~600 larvae per plant (~1200 larvae per replicate); and 3) *Arabidopsis thaliana* (ecotype Bla-2): 3 replicates x 6 plants (6 plants per pot), 3 weeks old (4 leaves well developed), were inoculated on the entire plant with ~600 larvae per plant (~1200 larvae per replicate). All plants used were cultivated under the same conditions used for experiments and for rearing mites (16 hour light/8 hour dark, 27 °C). Following inoculation with larvae, all of the plants were placed at 27 °C under continuous light for 12 hours. After this time, larvae were gently collected one by one with a wet brush, and were transferred to an Eppendorf tube on dry ice. The tube was changed every 30 minutes to liquid nitrogen, and samples were stored at -80 °C until RNA was prepared.

S6.2.3. RNA extraction, library preparation, and cDNA sequencing

Total RNA was prepared by grinding frozen larvae in 600 µl of Trizol (Invitrogen, Carlsbad, CA, USA) and subsequent extraction steps were performed according to the manufacturer's protocol. RNA concentration was estimated using a Nanodrop spectrophotometer (Thermo Scientific, Wilmington, DE, USA), after which RNA samples were sent to Fasteris SA (Geneva, Switzerland) for standard mRNA-seq library construction and sequencing. At Fasteris SA, barcoded libraries were prepared such that four samples could be run per lane. The quality of starting RNA was assessed prior to library construction at Fasteris SA with an Agilent Biolanalyzer (Agilent, Santa Clara, CA, USA). Each library was sequenced with single end methods to give read lengths of 45 base pairs. After trimming the bar code sequences, the usable read lengths for alignment to the genome were 38 bp (Table S6.2.2).

S6.2.4. Analysis of gene expression by host plant

For gene expression quantification for each sample, read alignments were performed using the Bowtie/Tophat software package[2]; the following parameters were used: -a 5, -i 5, -I 10000, --solexa-quals, -g 1, --segment-mismatches 3, -p 5 with the genome annotation used as prior information. As we used the –g 1 flag, we excluded repetitive mappings from the analysis. The resulting SAM file for each sample was parsed as described in section S2.3.3 to give both raw count data and RPKM values[3]. To assess differential gene expression, and to make use of the 3X biological replication per host plant, we used the DESeq R package[90]. Raw read count data was used as input into DESeq and pairwise comparisons were examined between each feeding condition (three comparisons in total: bean-*Arabidopsis*, bean-tomato and *Arabidopsis*-tomato). Genes with a FDR adjusted p-value < 0.05 (Benjamini-Hochberg procedure as implemented in DESeq) were defined as differentially expressed for downstream analyses, Figures S6.2.1 and S6.2.2.

S6.2.5. Hypergeometric tests on GO annotations – Mite feeding experiment

We used hypergeometric tests to determine whether given GO categories[7] were significantly over or under-represented among differentially expressed genes as described in Section 2.3.5[8]. Classes of genes with p-values < 0.05 were deemed to be over or under-represented. Enriched genes can be found for given comparisons in summary tables S6.2.3a-c.

S7. Hormones and neuropeptides

Genes involved in ecdysteroid biosynthesis and methylfarnesoate biosynthetic pathway are listed in Tables S7.1.1 and S7.1.4 respectively.

S7.1. Identification of ponasterone A

About 500 mg of spider mite nymphs (mixed ages) and adult females were homogenized in liquid nitrogen then extracted with methanol. The methanol extract was reduced in vacuo and analyzed by HPLC coupled with enzyme immunoassay (EIA) of the fractions. The HPLC conditions were: silica (250 mm x 4.6 mm, Hypersil) normal phase column eluted with dichloromethane/isopropanol/water, 125:30:1.5 (v:v:v) at a flow rate of 1 ml/min. Fractions were collected every 0.7 min and ecdysteroid quantified by EIA with the ecdysteroid polyclonal antibody L2[91], a conjugate of 20-hydroxyecdysone coupled to peroxidase as enzymatic tracer and tetramethyl benzidine as a color reagent. The fractions with maximal activity eluted with the retention time of authentic ponasterone A were analyzed further by LC-MS. For ecdysteroid identification, high resolution (R = 6000) mass spectrometry (HRMS) was performed with a double focusing MAT 95XP-Trap mass spectrometer (Thermo Finnigan) equipped with a TSQ/SSQ 7000 atmospheric pressure ionization source. Ten microliters of the sample extracts, thermostated at 15 ± 1°C, were injected with a direct HPLC loop injector (Thermo Finnigan) in a methanol/water (50/50, v/v) mobile phase. Prior to each injection, both the syringe and the needle were cleaned with 2 ml of methanol. Polyethylene glycol (PEG600, 30 ng/µl in methanol/water, 50/50 v/v, 0.5% formic acid) was continuously dosed by syringe injection as a reference for HRMS. A total flow of 49.2 µl/min entered the ionization source, operating in electrospray positive ionization (ESI+) mode. The capillary voltage, capillary temperature and the sheath gas (N_2, 99.995% purity) pressure were set at 3.00 kV, 250°C and 4 bar, respectively. The mass analyzer operated in multiple ion detection

(MID) mode, enabling the selective monitoring of the $[M+H]^+$ and $[M+NH_4]^+$ ions of both target analytes, with a maximum accurate mass deviation of 10 ppm. Data processing was done with XCalibur V1.4 software incorporated in the instrument. In agreement with Li *et al.*[92], ponasterone A was positively identified by its characteristic m/z at 465.3 and 482.4, corresponding to the protonated molecular ion $[M+H]^+$ and its ammonium adduct $[M+NH_4]^+$, respectively. We did not observe 20-hydroxyecdysone at its characteristic m/z at 481.3 and 498.3 of a 20-hydroxyecdysone standard. The structure of ponasterone A is given in Suppl. Figure S7.1.1.

S7.2. Identification of methyl farnesoate (MF)

All the mevalonate pathway enzymes involved in the early steps of JH biosynthesis[93]were identified in T. urticae including a putative JH acid O-methyltransferase gene (Table S7.1.4). However, we found no ortholog for CYP15A1, the enzyme that introduces the signature epoxide of insect juvenile hormones[94]. We expanded our genomic analysis with biochemical detection by LC-MS that showed the presence of methyl farnesoate (MF) in extracts of spider mites but not in the bean host plant extract.
About 500 mg of spider mite nymphs (mixed ages) and fresh host bean plant material were homogenized in liquid nitrogen and then extracted with hexane. The hexane extract was dried in vacuo and then redissolved in methanol for analysis by LC-MS. For MF and JH-III identification, high resolution (R = 6000) mass spectrometry (HRMS) was performed with a double focusing MAT 95XP-Trap mass spectrometer (Thermo Finnigan) equipped with a TSQ/SSQ 7000 atmospheric pressure ionization source. Ten microliters of the sample extracts, thermostated at 15±1°C, were injected with a HPLC loop injector (Thermo Finnigan) in a methanol/water (50/50, v/v; 1 mM ammonium acetate) mobile phase. Prior to each injection, both the syringe and the needle were cleaned with 2 ml of methanol. Polyethylene glycol (PEG300, 6 ng/µl in methanol/water, 50/50 v/v; 1 mM ammonium acetate) was continuously dosed by syringe injection as a reference for HRMS. A total flow of 46.4 µl/min entered the ionization source, operating in electrospray positive ionization (ESI+) mode. The capillary voltage, capillary temperature and the sheath gas (N_2, 99.995% purity) pressure were set at 3.00 kV, 250°C and 4 bar, respectively. The mass analyzer operated in multiple ion detection (MID) mode, enabling the selective monitoring of the $[M+H]^+$ and $[M+NH_4]^+$ ions of both target analytes, with a maximum accurate mass deviation of 10 ppm. Data processing was completed with XCalibur V1.4 software incorporated in the instrument. MF was positively identified in the spider mite extracts by its characteristic m/z at 251.2 and 268.2, corresponding to the protonated molecular ion $[M+H]^+$ and its ammonium adduct $[M+NH_4]^+$, respectively, while in the host bean plant extract, no MF was found. We did not observe JH-III in the spider mites extract at the characteristic m/z at 267.2 and 284.1 of a JH-III standard. These results suggest that, as in crustaceans, methyl farnesoate (MF) may be the final product of the pathway rather than JH. The function of methyl farnesoate in spider mite physiology is unknown.

S7.3. Nuclear receptors

We found 30 nuclear receptor (NR) genes in the genome, spanning the different subfamilies N0-N6 (Tables S7.1.2 and S7.1.3). This number is higher than the 19-22 NRs found in insect genomes, but closer to the 25 NRs of *D. pulex*. HR38 (NR4A4) was found as a tandem duplicate, and there were also two paralogs of USP, the arthropod RXR homolog. The NR1J group represented in insects by the single HR96 receptor and in vertebrates by PXR and CAR was expanded to eight genes throughout the genome. Although the precise function of insect

HR96 is still unclear, experiments in *D. melanogaster* suggest that DHR96 plays a role in the response to xenobiotics, as do the xenosensors PXR and CAR in vertebrates[95].

Neuropeptide encoding genes were searched as described previously[96,97] and are shown in Table S7.1.5.

S8. Hox complex

S8.1. Isolation of T.urticae *Hox complex*

We isolated putative Hox genes in *T. urticae* by performing BLAST (blastn and blastp), using *Drosophila melanogaster* and *Tribolium castaneum* Hox nucleotide and protein sequences curated from NCBI as queries against the *T. urticae* genome, coding sequence, and proteome datasets. E-values of $<10^{-5}$ were taken to represent significant hits of similarity (Table S8.1.1). Best reciprocal blastp matches, using *T. urticae* Hox sequences as queries against the NCBI Genbank non-redundant protein sequence database, further reinforced identification of putative *T. urticae* Hox genes. We further validated BLAST hits by aligning the homeodomain and flanking regions of Hox proteins of *T. urticae*, *D. melanogaster*, *T. castaneum*, and several other species (curated from NCBI)[98] using the alignment program MAFFT v6[99] (shown in Fig. S8.1.1) as performed by de Rosa *et al.*[98] to highlight conserved peptide motifs located outside of the homeodomain unique to individual Hox protein sequences.

To further confirm the identity of the *T. urticae* Hox genes, phylogenetic analysis utilizing neighbor-joining and maximum likelihood methods in the Phylip software package. *T. urticae*, *D. melanogaster*, and *T. castaneum* homeodomain (plus 6 amino acids flanking the N-terminus of the homeodomain and 18 flanking the C-terminus of the homeodomain) sequences were aligned with MAFFT v6, bootstrapped using SEQBOOT (100 replicates), and run through PROML. A consensus tree was constructed using the program CONSENSE. Scaffold number, position and inter-genic distances of the *T. urticae* Hox orthologs were identified based on previously completed computational gene prediction annotations as described in Section S2.4. The Hox genes *labial* and *proboscipedia* localized to scaffold 11. The other 8 Hox homologs localized to scaffold 20. Performing blastp with a generic homeodomain sequence did not uncover any additional homeodomain sequences within or adjacent to these two clusters of HOX genes. Based on these results, the *T. urticae* Hox cluster was reconstructed as shown in the main Figure 4a.

S8.2. VP SEM and in situ *hybridization experiments*

We have performed whole mount *in situ* hybridization of spider mite embryos according to previously published methodology[100]. For Variable Pressure Scanning Electron Microscopy (VP SEM) spider mites were mounted laterally on conductive adhesive pads and imaged using Hitachi S-3400 VP SEM instrument using back-scatter detector, 6 Pa pressure, 30 KV accelerating voltage and 5 mm working distance.

S9. Silk

The *T. urticae* fibroin genes were isolated by performing a prior analysis of 81 sequences called fibroin or spidroin available at NCBI. None of those sequences are complete and

represent either the C-term or N-term region with a portion of a repeat region. The repeats are characteristically rich in G/A/S and those features where used to screen the *T.urticae* genome. Doing so, using TBLASTN with filters turned off, we discovered 17 genes in the *T. urticae* genome containing fibroin repeats, Table S9.1.1. Given the unusual codon usage due to the repeats, most of the *T. urticae* fibroin genes were either not predicted by EuGene or incorrectly predicted.

All putative *T. urticae* fibroin proteins have high G, A and S amino acids organized in patterns similar to the one shown in Figure S9.1.1. They have unusually high serine content (27-39% of total amino acid content) relative to fibroins known from other species. Glycine content ranges from 17-29%, and alanine and asparagine vary between 8-20% of the total amino acids in silk proteins. *T. urticae* silk proteins share the general properties of charged N- and C- termini, repeated glycine-serine or glycine-serine-alanine fibroin motifs interspersed with additional asparagine residues and sometimes interrupted by spacers. This organization is crucial for the correct assembly and it maximizes protein-protein interactions in the processes of silk formation[101]. The expression pattern of the fibroin gene in embryos, larvae, nymphs and adults is shown in Figure S9.1.2. Many of the putative fibroin genes are expressed in the moving stages of *T. urticae*, representing candidate genes for the silk protein production.

S9.1. Mechanical characteristics of spider mite silk

Silicon trenches with depths of 1 μm and widths of 3.5–12 μm were produced with photolithography and reactive ion etching at the Western Nanofabrication Facility (London, Canada) using a custom-designed mask produced at the University of Alberta NanoFab facility (Edmonton, Canada). Adult and larval *T. urticae* spider mites were allowed to deposit fibers while walking on the silicon substrates bearing trenches. We performed three-point bending tests at the center of suspended segments of fibers with an atomic force microscope (MultiMode with Nanoscope IIIa controller, Digital Instruments). Measurements were made with NP-S silicon nitride cantilevers (Veeco, Plainview, NY) and the cantilevers used for adult and larval fibers had spring constants of 0.347 ± 0.009 N/m and 0.083 ± 0.003 N/m, respectively, as determined by the thermal noise method[102-105]. A total of 10 measurements were made at eight unique locations on four adult fibers and a total of seven measurements were made at three unique locations on three larval fibers. All measurements were performed at the ambient temperature of the instrument at approximately 32°C and at relative humidities of 14–30%. Curves of cantilever deflection versus vertical scanner position were acquired while deforming the fibers. The cantilever deflection was calibrated using reference curves acquired while pushing into the substrate, where the cantilever deflection is known to equal the scanner displacement. Cantilever deflection was converted to fiber restoring force by multiplying by the cantilever spring constant, and the fiber displacement was determined from the difference between scanner displacement and cantilever deflection, yielding curves of fiber restoring force versus fiber displacement which are readily modeled using static beam theory.

Mechanical properties were extracted by modeling the fibers as linear elastic cylinders with clamped boundary conditions. Heidelberg *et al*. showed that in cases where the fiber displacement is comparable to or larger than the fiber diameter, which is the case here, tension due to elongation of the fiber must be considered[106]. In addition to this deformation-induced tension, we find that the deposited silk fibers exhibit a significant initial tension that was accounted for. Elastic moduli were extracted by fitting the modified model to the elastic portion of restoring force versus fiber displacement curves. Comparison of the elastic modulus

of adult and larval spider mite silk to other silks and common materials is shown in Table S9.1.2.

S10. Immunity and RNAi

S10.1. Immunity

Spider mites form dense colonies that may favor pathogen transmission including fungal, bacterial and viral infections[107]. Pathogen invasions in arthropods trigger an innate immune response organized in three main pathways – Toll, Imd/Jnk, and JAK-STAT – that activate antimicrobial defense peptide[108]. In addition, RNAi, phagocytic cells, production of reactive oxygen species and nitrogen metabolites, and melanization pathways are also used in defense against invaders.

We performed tblastn searches of the *T. urticae* genome using 126 *Drosophila* genes described to be involved in cellular and/or humoral immune response (www.flybase.org). With this approach we uncovered 64 genes with putative defensive roles against pathogens in *T. urticae* (Table S10.1.1). Discounting all paralogs from this quantification, we have scanned the genome for roughly 90 different genes used by *Drosophila* in immunity of which we recovered less than half (41/86). This search was validated using a cut-off of e^{-6} and by, *a posteriori*, recovering the expected *Drosophila* gene as a first hit when blasting its genome with the *T. urticae* predicted protein (blastp). This procedure is highly conservative and we expect few false positives. On the other hand, we may expect false negatives particularly in two gene classes, the serine proteases (involved in recognition, mostly) and the small Anti Microbial Peptides (AMP).

Using the same procedure as above, and mining the specialized literature[109,110] we could establish the same dataset for the crustacean *Daphnia pulex* and the basal insect *Acyrthosiphon pisum* (Table S10.1.2).

Components of the Toll and Imd pathways, central elements of the anti-bacterial and anti-fungal responses of *D. melanogaster*, were readily identified in *T. urticae*. However, neither pathway was complete (Fig. S10.1.1). Missing from all pathways were components known to be essential for effective immune signaling in *D. melanogaster*. The extracellular serine proteases (*e.g. Necrotic* or *Spatzle-Processing Enzyme*), *tube* and most of the Anti-Microbial Peptide (AMP) effector gene orthologs could not be identified. We found a single ortholog for the PGRP receptor family and none for the GNBP family, presumed central players in insect pathogen recognition. In addition, we have identified a large number of lysozyme-coding genes, cystatin and many lectin-containing genes with putative roles in humoral immunity. This humoral artillery is reminiscent of the one found in the basal chelicerate *Limulus* (horseshoe crab)[111]. No putative phenoloxidase (PO) was found and indeed the hemocyanins precursors for PO in horseshoe crabs were lacking[112].

Thus, the *T. urticae* genome harbors only a fraction of the described *D. melanogaster* immunity genes. A reduction in immunity related genes was also found in the aphid[110] and the honeybee genome[113], suggesting that the low number of *Drosophila* immunity genes in other arthropod genomes is a general and ancestral feature. Particularly, the incipient nature of the IMD pathway from the root of arthopods may not be a specific adaptation to Gram negative symbionts. The repertoire of immunity genes found in *T. urticae* is consistent with a pattern

emerging from comparative studies in invertebrate immunity suggesting invertebrates use diverse solutions to build an immune response, likely driven by specific life-histories[114].

S10.2. RNA interference

Eukaryotic small RNA-mediated silencing is characterized by the interactions of three fundamental sets of molecules: small RNAs, Dicer enzymes and Argonaute proteins. Dicer enzymes and their co-factors excise the small RNAs from their precursors while Argonaute proteins (coupled with their small RNA cargos) execute silencing functions[115]. We found two *T. urticae* genes homologous to *Drosophila* Dicer enzymes, with a high degree of similarity (Table S10.2.1). Those two genes have the PAZ and RNAse III domains, which play central roles in the excision of small RNAs, as well as the other typical domains (Figure S10.2.1). We also found homologs of the *Drosophila* RNAse III DGCR8/Pasha and Drosha enzymes, which are involved in the release of pre-miRNA precursors from transcripts (Table S10.2.1). However, we did not find *T. urticae* homologs for the *Drosophila* Dicer-2 co-factors R2D2 or C3PO, but there are two clear homologs of the co-factor Loquacious.

Small RNAs guide RNA silencing effector complexes, like the RNA induced silencing complex (RISC) or the RNA induced initiation of transcriptional silencing complex (RITS). At the heart of those effector complexes lie members of the Argonaute family of proteins. There are three major clades in the superfamily: the Piwi subgroup that binds piRNAs, the Argonaute subgroup that associates with siRNAs and miRNAs, and a third group (WAGO) that is currently known only in nematodes[116]. Most eukaryotic species have multiple Argonaute genes. For example, there are five, eight and 27 Argonaute paralogs in *Drosophila*, humans and *C. elegans*, respectively. In *T. urticae*, we found seven Argonaute and seven Piwi homologs, but no WAGO members (Table S10.2.1).

RNA dependent RNA polymerase (RdRP) proteins are required for the amplification of some silencing triggers[117]. RdRPs are not required for all RNA silencing pathways and they are not present in all organisms. For example, no RdRP homologs are found in mammals[117]. There are no RdRP homologs in flies, but it was recently shown that the Drosophila RNA polymerase II elongator complex subunit I (D-elp1) is a non-canonical RdRP[118]. Interestingly, we found five clear RdRP homologs in the T. urticae genome (Table S10.2.1) where they might play a role in the amplification of silencing signals. The presence of RdRPs in the spider mite genome may represent an ancestral condition since they were also identified in a basal chordate, Branchiostoma (Cephalochordata) as well as in C. elegans, plants, fungi and protists[119]. Nonetheless, no orthologs of RdRP could be identified in Crustaceans nor in the insect, *Acyrthosiphon pisum* (Table S10.2.2). Both qualitatively and quantitatively, the RNAi genetic machinery found in *T. urticae* is reminiscent of other Arthropoda, from *Daphnia* and aphids (Table S10.2.2) to *Tribolium* and *Apis*[109] with the exception of the Argonaut gene family that has markedly expanded in the spider mite lineage. This expansion correlates with the spider mite´s unique feature of possessing RdRP genes. Whether this correlation underlies important functional and adaptive processes await testing.

It is noteworthy that experimental gene silencing by means of RNAi was previously successfully demonstrated in *T. urticae* through maternal injection of dsRNA and siRNA targeting *T. urticae Distal-less* resulting in induction of embryonic limb truncation phenotypes as well as the fusion of leg segments[120] in agreement with the conserved *Dll* function in Arthropoda[121,122].

S11. Chelicerate cuticular proteins

S11.1. Proteins with the pfam0379 domain

About 70% of sequenced insect proteins associated with cuticle contain the "chitin_bind_4" domain (pfam0379)[123]; this domain is now widely used to identify cuticle proteins. Proteins containing pfam0379 can be classified depending on the "RR" motif they contain. In insects, about 1/3 of cuticle proteins (CPs) contain the RR-1 motif, and have been isolated from flexible cuticle, whereas about 2/3 contain the RR-2 motif and have been associated with hard cuticle. Glycine and alanine are major amino acids in all pfam0379-containing CPs; for instance, in *Anopheles gambiae* these two amino acids represent about 20% of the CP residues[124]. Both histidine and lysine are involved in sclerotization, and RR-2 genes have more histidine than do RR-1 CPs, consistent with the association of RR-2 CPs with hard cuticle.

We found 45 CPs in *T. urticae* using the search term "cuticle" (Table S11.1.1). Of these 45, two had no known motif, one had the cl02629 superfamily motif (which includes the Chitin-binding domain type 2 and overlapping pfam01607 domain), whereas the remaining 42 contained the pfam0379 motif. Interestingly, all but one of the latter belonged to the RR-2 class. It is striking that only one RR-1 containing CP was found, in contrast to insect CPs where this motif occurs in about 1/3 of all CPs. The very low abundance of RR-1 CPs may be a chelicerate-specific feature; of the chelicerate CPs from the cuticle protein database (http://biophysics.biol.uoa.gr/cuticleDB), no RR-1 containing CPs was found: of the 15 CP sequences, six included an RR-2 motif, whereas nine included no discernable RR motif. BLAST analyses of the *T. urticae* CPs showed strongest homology to other chelicerate CPs and much more distant homology to insect CPs. Searching the *T. urticae* proteome for additional CP sequences using *Anopheles gambiae* CP or known chelicerate CP sequences produced only weakly supported genes, none of which were supported following subsequent analyses (BLAST, sequence motifs, etc).

T. urticae CPs of the RR-2 class contain levels of alanine, glycine, proline, lysine, and tyrosine that are similar to those found in pfam0379-containing insect CPs (Table S11.1.2). By contrast, they contained many fewer histidine residues compared to RR-2 insect CPs (only 3.6 % compared to around ~11% for insect RR-2 CPs). A similar situation occurs in aphids and may be a consequence of their mode of feeding or brooding, which is similar to that of *T. urticae*. The RR domain has been shown to interact with chitin; the presence of CPs with lower abundance of histidines suggests that whereas the CPs are able to bind to chitin, they may not form a rigid cuticle through cross-linking with other CPs, as occurs with insect CPs from the RR-2 class.

Recently an experimental and bioinformatics search in *Tribolium castaneum* revealed 18 proteins related to the Peritrophin-A subfamily (ChtBD2=CBM_14=Pfam 01607) as a new group of Cuticular Proteins Analogous to Peritrophins (CPAPs), eight of which are closely related the *gasp/obstructor* family. Obstructor family members are proteins expressed in cuticle forming tissues, whereas members of the Gasp family may be present in other tissues than the peritrophic matrix, particularly where nutrient or gas exchange are important, and/or where invasion by parasites or viruses is possible. We found 15 CPAP's in *T. urticae* using the search term "peritrophin" (Table S11.1.3). Of these, five were homologous to *gasp/obstructor* proteins of *T. castaneum* and *D. melanogaster*.

S12. DNA methylation

Several lines of evidence suggest that there are potentially low levels of genomic DNA methylation in *T. urticae*, or have been in the recent past of the lineage, based upon the following observations: *i*) the presence of genes or pseudogenes with homology to DNA methyltransferases as well as a gene that encodes a putative methyl-CpG binding domain protein in the *T. urticae* genome; *ii*) lower CpG O/E than GpC O/E in the genomic level and in CDSs; and *iii*) the presence of putative methyl-cytosines from the survey of several genes using PCR and sequencing of bisulfite converted genomic DNA.

S12.1. Presence of DNA methylation machinery in T. urticae genome

DNA methyltransferases (DNMTs) perform DNA methylation and all share a conserved catalytic domain, suggesting a common and ancient origin. Studies of mammalian systems have established that different DNMTs undertake distinct functions[125]. For example, human and mouse genomes contain one DNMT1, one DNMT2, and three DNMT3s (DNMT3a/b and DNMT3L). DNMT1 is responsible for maintaining the pattern of DNA methylation through DNA replication and is referred to as the *maintenance* methyltransferase. DNMT3s mediate *de novo* methylation of previously unmethylated cytosines. The role of DNMT2 is still not completely resolved, but recent studies suggest that it may act as a tRNA methyltransferase[126]. We investigated whether *T. urticae* contains a complete repertoire of putative *dnmt*s.

S12.1.1. BLAST search for DNMT homologs in the mite genome

We performed a *blastp* search against the *T. urticae* protein sequences using human, honeybee, and *Nasonia dnmt*s as queries. For *dnmt1*, this analysis found two genic fragments (the predicted adjacent genes tetur13g02110 and tetur13g02100) that are separated by a transposable element insertion. This insertion, along with other sequence features, strongly suggests that these two predicted gene fragments reflect a single ancestral *dnmt1* gene that has become a pseudogene (with multiple accumulating disruptions). Thus, it appears that *T. urticae* lacks an extant functional *dnmt1* gene. Otherwise, we identified a putative *dnmt*3 homolog (tetur16g01020), and a putative *dnmt2* homolog (tetur02g05760) (Table S12.1.1). HMM searches of the above genes (or pseudogene in the case of *dnmt1*) against the Pfam database revealed each to have a nucleobase methyltransferase domain (Table S12.1.2).

In addition to DNMTs, there is a putative gene that encodes proteins containing methyl-CpG binding domains (MBD), which can recognize sites of DNA methylation. In human, there are five Mbd proteins, Mbd1, Mbd2, Mbd3, Mbd4 and MeCP2. It has been proposed that these proteins probably function in transcriptional repression. We used the human Mbd proteins as the queries to search against the *T. urticae* genome. One MBD homolog was found, encoded by tetur05g01540, which has very high identity (50%) and similarity (68%) to the MBD proteins in human. In addition, an HMM search in the Pfam database showed that this protein contains the Methyl-CpG binding domain (E-value = 8.40 $E10^{-20}$; Table S12.1.2).

Given that traditional *dmnt1* activity (maintenance methylation) should be absent in *T. urticae* by analogy to other systems, we wondered if the *de novo* methylation homolog (*dnmt3*) was a pseudogene as well by virtue of incorrect annotation or lack of expression. At least by RNA-seq, however, *dnmt3* (tetur16g01020) is well supported [the mean RPKM value across all developmental stages and host plant contrasts is 65.59, and all 13 annotated exon-exon splicing events are supported by at least one intron spanning (splice derived) RNA-seq read].

As assessed by RNA-seq reads, the *dnmt2* and *mbd* homologs also appeared to be expressed as well, albeit at lower levels (RPKM values of 14.56 and 15.46, respectively).

S12.1.2. Phylogenetic analyses

We performed phylogenetic analyses using the NJ-method for phylogenetic reconstruction[127]. Although we note that the bootstrap values are low, the resulting tree shows the expected grouping of the three *dnmt* families and placement of the *T. urticae* homologs (Table S12.1.3 and Figure S12.1.1).

S12.2. Genome-wide patterns of CpG depletion

DNA methylation in animals predominantly occurs at cytosines followed by guanine, or 'CpG' dinucleotides[128]. Methylated cytosines are highly vulnerable to spontaneous deamination, which causes C to T transitions. Consequently, methylated genomic regions gradually lose CpG dinucleotides. Thus, methylated regions are generally depleted of CpGs[129]. The level of CpG depletion is measured by the observed frequency of CpG dinucleotides normalized by the expected frequency, or CpG O/E (see below). We examined the distribution of CpG O/E in the *T. urticae* genome to infer genomic levels of DNA methylation. As a control, we investigated GpC O/E of the same genomic regions in parallel. Because GpC nucleotide contexts include the same number of G and C residues as a CpG, yet are not subject to (canonical) DNA methylation, GpC O/E is commonly used as a control for the influence of DNA methylation[130,131].

S12.2.1. CpGs are underrepresented in the mite genome

CpG O/E is estimated using the following formula:

$$CpG_{O/E} = \frac{P_{CpG}}{P_C * P_G},$$

where $\boldsymbol{P}_{CpG}$, $\boldsymbol{P}_c$ and $\boldsymbol{P}_G$ are the frequencies of CpG dinucleotides, C nucleotides, and G nucleotides, respectively, estimated from each genomic fragment.

We first investigated the distribution of CpG O/E from the whole genome, by analyzing randomly cut 1000 bp segments. The mean CpG O/E in the whole genome is 0.68, while the mean GpC O/E is 0.97. CpG O/E is significantly lower than GpC O/E (Mann-Whitney test, $P < 10^{-15}$), suggesting a mechanism that is reducing CpG content (or has historically). The level of CpG depletion is approximately 30% in the whole genome. DNA methylation tends to be concentrated in coding sequences in invertebrate genomes[132,133]. We found that coding sequences (CDs) of *T. urticae* exhibit a similar level of CpG depletion as the whole genome. The mean of CDS CpG O/E is 0.69, which is significantly lower than the mean of GpC O/E (0.95) (Figure S12.2.1).

S12.3. Experimental validation of DNA methylation

We investigated the presence of methyl cytosines in *T. urticae* genome using experimental means. Specifically, we used targeted PCR of bisulfite converted genomic DNA. Genomic DNA of *T. urticae* was bisulfite converted using the EpiTect Bisulfite Kit (Qiagen).

Following this conversion, un-methylated cytosines are converted to thymine while methylated cytosines remain as cytosines. We investigated nine loci, based upon the following two criteria: 1) genes with low CpG O/E values and 2) genes that are orthologous to known methylated targets in the honeybee.

We used the Methyl Primer Express version 1.0 (Applied Biosystems) to design primers for targeted PCR. The CpG sites within each amplicon varied from 2 to 7. After PCR reactions, we transformed the PCR products into *E. coli* cells by using the TOPO TA Cloning Kit (Invitrogen). For each candidate gene, we picked eight colonies for sequencing on average.

The conversion rate, estimated from non-CpG cytosines, was 99.5%. We observed 5 methyl-cytosines out of 330 CpG dinucleotides from our experiments. The proportion of non-converted versus converted CpG is (1.54%) three-fold higher than the proportion of non-converted versus converted Cs in non-CpG context (0.5%). This difference is statistically significant, as tested using the Fisher's exact test ($P = 0.03$). These results raise the possibility of a low level of DNA methylation in the *T. urticae* genome. However, given the small (absolute) number of observed conversion events, and the observation that the *T. urticae dmnt1* gene is apparently nonfunctional (see Section S12.1.1), we believe that additional work, *i.e.,* whole-genome bisulfite sequencing by high-throughput approaches, is needed to establish and characterize DNA methylation in *T. urticae*. In this context, ascertaining whether the putatively functional *dnmt3 de novo* methylase has a role in DNA methylation in *T. urticae* will be particularly important.

Supplementary Tables

Table S2.2.1. Genomic libraries included in the *Tetranychus urticae* genome assembly and their respective assembled sequence coverage levels in the final release.

Library Type	Average Insert Size	Read Number	Assembled Sequence Coverage (X)
3kb	2,423	499,872	3.55
8kb	8,534	572,253	3.89
Fosmid	35,551	107,424	0.61
Total		1,179,549	8.05

Table S2.2.2. Summary statistics of the output of the whole genome shotgun assembly before screening and removal of organellar DNA and contaminating scaffolds. The table shows the total number of contigs and total number of assembled base pairs for each set of scaffolds greater than the given size.

Size	Number	Contigs	Scaffold Size	Base pairs	% Non-gap Base pairs
5,000,000	3	185	19,855,029	19,706,793	99.25
2,500,000	11	449	48,910,718	48,583,448	99.33
1,000,000	30	808	78,409,356	77,792,979	99.21
500,000	37	924	83,819,188	83,046,293	99.08
250,000	41	958	85,347,754	84,542,707	99.06
100,000	44	978	85,994,807	85,162,321	99.03
50,000	58	1,093	86,922,775	85,960,868	98.89
25,000	78	1,202	87,611,020	86,547,972	98.79
10,000	187	1,465	89,257,328	88,043,685	98.64
5,000	283	1,603	89,907,304	88,680,067	98.63
2,500	695	2,137	91,395,611	90,137,643	98.62
1,000	717	2,165	91,436,718	90,175,450	98.62
0	733	2,181	91,448,241	90,186,973	98.62

We classified the remaining scaffolds in various bins depending on sequence content. We identified contamination using megablast against Genbank NR and blastp against a set of known microbial proteins. No scaffolds were identified as contamination. We classified additional scaffolds as unanchored rDNA (76), and mitochondrion (1). We also removed 16 scaffolds that were less than 1kb in sequence length.

Table S2.2.3. Final summary: statistics for chromosome scale assembly.

Fold genome coverage	8.05
Total span	89.6 Mb
Scaffold sequence total	90.8 MB
Contig sequence total	89.6 MB (1.3% gap)
Scaffold N/L50	10/3.0 MB
Contig N/L50	120/212.8 KB
Largest scaffold	7,801,961 bp
Average scaffold length	141,899 bp
Number of contigs	2,035
Largest contig	929,118 bp
Average contig length	44,029 bp
N50 (contigs)	120 Kb

Table S2.3.1. Metrics for stage specific RNA sequencing and read alignment.

Stage	Number of reads	Sequence output (bases)	Number of mapped reads[a]	Percent reads mapped[a]	Attributable to gene models	Percent mapped attributable to gene models
Adult	23,032,034	1,750,434,584	8,738,691	37.94	7,191,982	82.30
Embryo	26,816,319	2,038,040,244	13,202,630	49.23	11,829,974	89.60
Larvae	18,216,393	1,384,445,868	9,136,097	50.15	8,124,713	88.93
Nymph	26,551,275	2,017,896,900	11,241,675	42.34	9,822,626	87.38

[a] Reads were mapped uniquely

Table S2.3.2. Number of differentially expressed genes in the 6 pair wise comparisons between developmental stages.

Comparison	Number of differentially expressed genes
Adult vs Embryo	3,582
Adult vs Larvae	2,771
Adult vs Nymph	617
Embryo vs Larvae	1,343
Embryo vs Nymph	2,396
Larvae vs Nymph	673

Genes with a FDR adjusted p-value < 0.05 and a proportions fold change ≥ |2| using the Z test of Kal *et al.*[5] were defined as differentially expressed. The magnitude of differential gene expression is greatest between embryo and adult, which are the two most different developmental stages.

Table S2.3.3 Enriched GO terms among differentially expressed genes between different developmental stages (in pairwise combinations, a-f).

Table S2.3.3a. Enriched GO terms among differentially expressed genes between adults and embryos.

Biological Process

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0006030	chitin metabolic process	31	19	8	11	1.19E-005
GO:0045449	regulation of transcription	195	71	47	24	6.38E-005
GO:0007156	homophilic cell adhesion	15	11	4	7	8.49E-005
GO:0006869	lipid transport	8	6	2	4	3.61E-003
GO:0006979	response to oxidative stress	18	10	4	6	4.30E-003
GO:0006413	translational initiation	12	7	3	4	0.01
GO:0006865	amino acid transport	12	6	3	3	0.05

Cellular Component

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0005576	extracellular region	67	24	16	8	0.01
GO:0016020	membrane	284	80	67	13	0.03
GO:0030130	clathrin coat of trans-Golgi network vesicle	6	4	1	3	0.03
GO:0030132	clathrin coat of coated pit	6	4	1	3	0.03
GO:0016459	myosin complex	4	3	1	2	0.04

Molecular Function

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0042302	structural constituent of cuticle	41	23	10	13	9.03E-006
GO:0008061	chitin binding	31	19	7	12	9.61E-006
GO:0008234	cysteine-type peptidase activity	72	30	17	13	6.11E-004
GO:0004185	serine-type carboxypeptidase activity	5	5	1	4	7.85E-004
GO:0003677	DNA binding	267	85	64	21	1.49E-003
GO:0005319	lipid transporter activity	8	6	2	4	3.32E-003
GO:0003743	translation initiation factor activity	8	6	2	4	3.32E-003
GO:0003755	peptidyl-prolyl cis-trans isomerase activity	6	5	1	4	3.77E-003
GO:0005509	calcium ion binding	80	28	19	9	0.02
GO:0030130	clathrin coat of trans-Golgi network vesicle	6	4	1	3	0.03
GO:0030132	clathrin coat of coated pit	6	4	1	3	0.03
GO:0015171	amino acid transmembrane transporter activity	12	6	3	3	0.04
GO:0004096	catalase activity	4	3	1	2	0.05
GO:0016459	myosin complex	4	3	1	2	0.05

Table S2.3.3b. Enriched GO terms among differentially expressed genes between adults and larvae.

Biological Process

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0006030	chitin metabolic process	31	17	6	11	3.10E-006
GO:0006979	response to oxidative stress	18	10	3	7	3.22E-004
GO:0006869	lipid transport	8	5	1	4	6.08E-003
GO:0006270	DNA replication initiation	6	4	1	3	0.01
GO:0006816	calcium ion transport	9	5	2	3	0.01
GO:0006122	mitochondrial electron transport, ubiquinol to cytochrome c	4	3	1	2	0.02
GO:0045449	regulation of transcription	195	46	35	11	0.02
GO:0007165	signal transduction	47	14	8	6	0.03
GO:0006303	double-strand break repair via nonhomologous end joining	2	2	0	2	0.03
GO:0006166	purine ribonucleoside salvage	2	2	0	2	0.03
GO:0046836	glycolipid transport	2	2	0	2	0.03
GO:0007156	homophilic cell adhesion	15	6	3	3	0.04
GO:0006506	GPI anchor biosynthetic process	8	4	1	3	0.04
GO:0006865	amino acid transport	12	5	2	3	0.05
GO:0006754	ATP biosynthetic process	24	8	4	4	0.05

Cellular Component

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0005576	extracellular region	67	23	13	10	2.17E-003
GO:0016459	myosin complex	4	3	1	2	0.02

Molecular Function

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0008061	chitin binding	31	17	6	11	8.20E-006
GO:0042302	structural constituent of cuticle	41	19	8	11	5.53E-005
GO:0004129	cytochrome-c oxidase activity	12	8	2	6	3.89E-004
GO:0008234	cysteine-type peptidase activity	72	25	14	11	1.05E-003
GO:0008121	ubiquinol-cytochrome-c reductase activity	4	4	1	3	1.29E-003
GO:0004185	serine-type carboxypeptidase activity	5	4	1	3	5.46E-003
GO:0005319	lipid transporter activity	8	5	2	3	8.18E-003
GO:0003677	DNA binding	267	66	51	15	9.58E-003
GO:0005634	nucleus	113	31	21	10	0.02
GO:0008233	peptidase activity	4	3	1	2	0.02
GO:0016459	myosin complex	4	3	1	2	0.02
GO:0004001	adenosine kinase activity	2	2	0	2	0.04
GO:0043565	sequence-specific DNA binding	2	2	0	2	0.04

Table S2.3.3c. Enriched GO terms among differentially expressed genes between adults and nymphs.

Biological Process

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0006030	chitin metabolic process	31	13	1	12	5.00E-011
GO:0006869	lipid transport	8	4	0	4	1.64E-004
GO:0006066	alcohol metabolic process	14	3	1	2	0.02
GO:0016052	carbohydrate catabolic process	1	1	0	1	0.04
GO:0045947	negative regulation of translational initiation	1	1	0	1	0.04
GO:0006469	negative regulation of protein kinase activity	1	1	0	1	0.04
GO:0006801	superoxide metabolic process	9	2	0	2	0.05

Cellular Component

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0005576	extracellular region	67	14	3	11	2.69E-007
GO:0045263	proton-transporting ATP synthase complex, coupling factor F(0)	3	2	0	2	5.29E-003
GO:0005739	mitochondrion	10	3	0	3	7.29E-003
GO:0000275	mitochondrial proton-transporting ATP synthase complex, catalytic core F(1)	1	1	0	1	0.04

Molecular Function

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0042302	structural constituent of cuticle	41	17	2	15	1.07E-012
GO:0008061	chitin binding	31	13	1	12	1.98E-010
GO:0005319	lipid transporter activity	8	4	0	4	2.39E-004
GO:0008083	growth factor activity	8	3	0	3	4.22E-003
GO:0005739	mitochondrion	4	2	0	2	0.01
GO:0004869	cysteine-type endopeptidase inhibitor activity	13	3	1	2	0.02
GO:0005509	calcium ion binding	80	8	4	4	0.03
GO:0016757	transferase activity, transferring glycosyl groups	7	2	0	2	0.04
GO:0019905	syntaxin binding	1	1	0	1	0.04
GO:0005578	proteinaceous extracellular matrix	1	1	0	1	0.04
GO:0004862	cAMP-dependent protein kinase inhibitor activity	1	1	0	1	0.04
GO:0008190	eukaryotic initiation factor 4E binding	1	1	0	1	0.04
GO:0000275	mitochondrial proton-transporting ATP synthase complex, catalytic core F(1)	1	1	0	1	0.04

Table S2.3.3d. Enriched GO terms among differentially expressed genes between embryos and larvae.

Biological Process

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0006030	chitin metabolic process	31	15	3	12	2.14E-007
GO:0055114	oxidation reduction	210	48	23	25	3.31E-007
GO:0006270	DNA replication initiation	6	6	1	5	1.85E-006
GO:0005975	carbohydrate metabolic process	118	27	13	14	1.52E-004
GO:0008152	metabolic process	324	56	36	20	2.87E-004
GO:0006979	response to oxidative stress	18	7	2	5	2.16E-003
GO:0007156	homophilic cell adhesion	15	6	2	4	3.84E-003
GO:0006508	proteolysis	285	42	32	10	0.03
GO:0009166	nucleotide catabolic process	3	2	0	2	0.03
GO:0006096	glycolysis	13	4	1	3	0.05
GO:0006869	lipid transport	8	3	1	2	0.05

Cellular Component

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0005576	extracellular region	67	18	5	13	6.14E-007
GO:0005634	nucleus	276	39	21	18	1.19E-005

Molecular Function

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0042302	structural constituent of cuticle	41	18	4	14	4.67E-009
GO:0008061	chitin binding	31	15	3	12	1.83E-008
GO:0016491	oxidoreductase activity	146	30	14	16	1.76E-005
GO:0004869	cysteine-type endopeptidase inhibitor activity	13	6	1	5	5.96E-004
GO:0008234	cysteine-type peptidase activity	72	16	7	9	6.89E-004
GO:0003824	catalytic activity	141	23	13	10	4.71E-003
GO:0005634	nucleus	113	18	10	8	0.01
GO:0016820	hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	3	2	0	2	0.02
GO:0004499	flavin-containing monooxygenase activity	3	2	0	2	0.02
GO:0004602	glutathione peroxidase activity	3	2	0	2	0.02
GO:0005319	lipid transporter activity	8	3	1	2	0.03
GO:0008271	secondary active sulfate transmembrane transporter activity	4	2	0	2	0.05

Table S2.3.3e. Enriched GO terms among differentially expressed genes between embryos and nymphs.

Biological Process

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0045449	regulation of transcription	195	69	34	35	2.87E-010
GO:0007156	homophilic cell adhesion	15	10	3	7	3.08E-005
GO:0006030	chitin metabolic process	31	14	5	9	2.89E-004
GO:0055114	oxidation reduction	210	54	37	17	1.08E-003
GO:0006869	lipid transport	8	5	1	4	5.55E-003
GO:0008152	metabolic process	324	70	56	14	0.02
GO:0006979	response to oxidative stress	18	7	3	4	0.03
GO:0006072	glycerol-3-phosphate metabolic process	2	2	0	2	0.03
GO:0006355	regulation of transcription, DNA-dependent	151	35	26	9	0.04
GO:0005975	carbohydrate metabolic process	118	28	21	7	0.05

Cellular Component

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0005634	nucleus	276	60	40	20	1.86E-004
GO:0030130	clathrin coat of trans-Golgi network vesicle	6	4	1	3	5.12E-003
GO:0030132	clathrin coat of coated pit	6	4	1	3	5.12E-003
GO:0009331	glycerol-3-phosphate dehydrogenase complex	2	2	0	2	0.02
GO:0005576	extracellular region	67	16	10	6	0.03

Molecular Function

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0003677	DNA binding	267	79	45	34	5.33E-008
GO:0008061	chitin binding	31	14	5	9	2.03E-004
GO:0004869	cysteine-type endopeptidase inhibitor activity	13	8	2	6	3.60E-004
GO:0042302	structural constituent of cuticle	41	16	7	9	5.56E-004
GO:0004104	cholinesterase activity	7	5	1	4	2.08E-003
GO:0004185	serine-type carboxypeptidase activity	5	4	1	3	3.45E-003
GO:0005319	lipid transporter activity	8	5	1	4	4.77E-003
GO:0000166	nucleotide binding	81	23	14	9	5.97E-003
GO:0008234	cysteine-type peptidase activity	72	21	12	9	6.04E-003
GO:0030130	clathrin coat of trans-Golgi network vesicle	6	4	1	3	8.97E-003
GO:0030132	clathrin coat of coated pit	6	4	1	3	8.97E-003
GO:0016491	oxidoreductase activity	146	35	25	10	0.02
GO:0004367	glycerol-3-phosphate dehydrogenase (NAD+) activity	2	2	0	2	0.03
GO:0008083	growth factor activity	8	4	1	3	0.03

Table S2.3.3f. Enriched GO terms among differentially expressed genes between larvae and nymphs.

Biological Process

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0006869	lipid transport	8	4	0	4	1.58E-004
GO:0045449	regulation of transcription	195	17	8	9	1.79E-003
GO:0007156	homophilic cell adhesion	15	4	1	3	2.47E-003
GO:0006506	GPI anchor biosynthetic process	8	2	0	2	0.04
GO:0006836	neurotransmitter transport	19	3	1	2	0.04
GO:0006355	regulation of transcription, DNA-dependent	151	11	6	5	0.04
GO:0006801	superoxide metabolic process	9	2	0	2	0.05

Cellular Component

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0005634	nucleus	276	19	12	7	0.01

Molecular Function

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0003677	DNA binding	267	29	12	17	6.97E-006
GO:0005319	lipid transporter activity	8	4	0	4	2.56E-004
GO:0042302	structural constituent of cuticle	41	6	2	4	9.93E-003
GO:0005634	nucleus	113	10	5	5	0.03
GO:0004104	cholinesterase activity	7	2	0	2	0.04
GO:0019905	syntaxin binding	1	1	0	1	0.05
GO:0008083	growth factor activity	8	2	0	2	0.05

Table S2.3.4. Stage specific gene expression - top 10 highly expressed genes in each stage.

Adult

Tetur ID	Gene name	Gene description	RPKM	Percentage total
tetur03g09300	n/a	Hypothetical protein	41,463.14	7.09
tetur18g03050	n/a	Hypothetical protein	34,571.42	5.91
tetur03g09310	n/a	Vitellogenin	23,600.05	4.04
tetur18g03040	n/a	Hypothetical protein	19,286.27	3.30
tetur18g03030	n/a	TEP1	18,804.52	3.22
tetur05g01730	n/a	Hypothetical protein	8,848.24	1.51
tetur01g02670	n/a	MDL1	7,026.74	1.20
tetur12g01560	n/a	Hypothetical protein	6,574.57	1.12
tetur18g03020	n/a	Hypothetical protein	6,055.06	1.04
tetur516g00020	Vg1	Vitellogenin 1	5,544.08	0.95
Top 10 genes				29.38

Nymph

Tetur ID	Gene name	Gene description	RPKM	Percentage total
tetur03g09300	n/a	Hypothetical protein	15,554.99	2.49
tetur18g03050	n/a	Hypothetical protein	10,780.64	1.72
tetur18g03040	n/a	Hypothetical protein	8,054.70	1.29
tetur18g03030	n/a	TEP1	7,908.19	1.26
tetur03g09310	n/a	Vitellogenin	7,558.92	1.21
tetur05g01730	n/a	Hypothetical protein	7,542.55	1.21
tetur12g01560	GRSP11	Glycine-rich secreted protein Secreted protein with GYG repeats	6,563.54	1.05
tetur39g00530	SSPC1	Small secreted protein, family C	5,615.83	0.90
tetur07g00320	n/a	Hypothetical protein	4,995.46	0.80
tetur03g08300	HCP2	Hypothetical cuticular protein	3,847.52	0.62
Top 10 genes				12.54

Larvae

Tetur ID	Gene name	Gene description	RPKM	Percentage total
tetur07g00320	n/a	Hypothetical protein	11,397.74	1.83
tetur05g01730	n/a	Hypothetical protein	6,852.51	1.10
tetur01g00130	CPR 30	Cuticle protein	6,216.88	1.00
tetur39g00530	SSPC1	Small secreted protein, family C	5,561.30	0.89
tetur05g05670	GRSP3	Glycine-rich secreted protein	5,514.90	0.88
tetur12g01560	n/a	Hypothetical protein	4,541.17	0.73
tetur03g08300	n/a	Hypothetical protein	4,319.67	0.69
tetur02g10770	n/a	AGAP002847-PA	4,256.49	0.68
tetur20g00200	SERP	Secreted protein with 8aa repeat structure	3,705.71	0.59
tetur03g09300	n/a	Hypothetical protein	3,687.74	0.59
Top 10 genes				8.98

Table S2.3.4. (continued) Stage specific gene expression - top 10 highly expressed genes in each stage.

Embryo

Tetur ID	Gene name	Gene description	RPKM	Percentage total
tetur07g00320	n/a	Hypothetical protein	5,368.77	0.97
tetur03g08300	n/a	Hypothetical protein	3,727.15	0.68
tetur11g00600	CPR 19	cuticle protein	3,287.17	0.60
tetur01g00130	CPR 30	cuticle protein	3,272.82	0.59
tetur34g00420	n/a	Hypothetical protein	2,906.56	0.53
tetur26g01570	n/a	histone 1	2,880.41	0.52
tetur04g01580	n/a	Hypothetical protein	2,478.14	0.45
tetur13g04260	n/a	ribosomal protein L41	2,466.95	0.45
tetur04g01610	CPR 51	Cuticle protein	2,444.49	0.44
tetur20g02780	n/a	Cuticular protein similar to Bombyx	2,354.59	0.43
Top 10 genes				5.66

Table S2.4.1. Comparison of genome and annotation statistics for the draft sequence of the spider mite *T. urticae* genome and genomes of *D. melanogaster* and *T. castaneum.*

		T.urticae	%*	*D.melanogaster*	%*	*T.castaneum*	%*
nr. of loci (exons + introns):		18,414		14,861		14,460	
av. length of loci:	nt	2,652		6,328		5,422	
loci density:	nt/gene	4,866		12,160		10,471	
	genes/Mb	205.51		91.53		95.55	
nr. of genes:		18,414		13,353		14,452	
av. length of gene:	nt	1,428		1,506		1,371	
median length of genes	nt	1,138		1,139		1,011	
nr. exons:		70,405		38,648		65,479	
cumul. exon length:	nt	26,292,088	29.34	20,111,395	12.39	19,818,063	13.10
av. length of exon:	nt	374		520		303	
median length of exon:	nt	178		312		195	
longest exon:	nt	45,659[a]		27,510		26,331	
av. nr. exons/gene:		3.82		2.89		4.53	
most exons/gene:		55[b]		81		105	
cumul CDS length:	nt	19,505,397	21.77	17,825,960	10.98	19,784,616	13.07
av. length of CDS:	nt	1,060		1,335		1369	
longest CDS:	nt	54,762[c]		68,916		63354	
shortest CDS:	nt	63[d]		2		60	
%GC of CDS:		37.8		53.2		44.3	
cumul. intron length:	nt	20,681,179	23.08	14,956,521	9.21	34,101,322	22.53
av. length of intron:	nt	400		597		711	
median length of intron:	nt	96		67		53	
nr. of big introns (>20 kb):		36		198		121	
longest intron:	nt	50,833[e]		131,739		98,797	
%GC of intron:		29.7		40.3		32.1	
genome size (scaffolds):	nt	90,815,494		168,736,537		210,566,138	
genome size (contigs):	nt	89,600,102		162,367,812		151,333,735	
largest scaffold:	nt	7,801,961		29,004,656		38,791,480	
av. scaffold length:	nt	141,899		11,249,102		97,394	
number of contigs:		2,035		137		8,828	
largest contig:	nt	929,118		27,905,053		597,263	
av. contig length:	nt	44,030		1,185,167		17,142	
gaps (>50N):		1,395		119		6,660	
percent of the genome involved in protein encoding transcripts (exon + intron)			52.43		21.60		35.63

* percent of total genome
[a] exon: tetur30g00590.4
[b] gene: tetur04g02800
[c] gene: tetur30g00590
[d] 84 genes in total
[e] exon: tetur07g02140.1 - exon: tetur07g02140.2

Table S2.4.2. Number of annotated non-coding RNAs.

miRNA	52 (5,311 nt)
partial	2 (1,753 nt)
pseudo	98 (225,613 nt)
rRNA	133 (34,230 nt)
snoRNA	108 (12,326 nt)
spliceosomal RNA	15 (2,062 nt)
tRNA	139 (10,459 nt)

Table S2.5.1. SNP and INDEL counts by prediction type for the Montpellier strain.

Variant type	Count
Homozygous SNPs	151,763
Heterozygous SNPs	250,186
Homozygous insertions	38,649
Homozygous deletions	31,254
Heterozygous insertions	34,045
Heterozygous deletions	36,703

Table S2.5.2. Non-synonymous and synonymous nucleotide changes between the Montpellier and London reference strain.

	MAQ-all[a]	Homozygous	Heterozygous
non-synonymous	9,103	3,322	5,781
synonymous	21,814	7,875	13,939
K_a/K_s ratio[b]	0.417	0.422	0.415

[a] MAQ-all is the combination of all SNPs (homozygous or heterozygous) called with MAQ.
[b] A simple K_a/K_s ratio was defined as the ratio of nonsynonymous to synonymous changes. This approximation was performed because it was not possible to construct pseudochromosomes from the Montpellier population that was highly heterozygous.

Table S2.5.3. Summary of large-effect SNP changes between the Montpellier and London reference strain.

	MAQ-all[a]	Homozygous	Heterozygous
protein truncating[b]	132	51	81
start Met altering[c]	46	15	31
protein elongating[d]	47	21	26

[a] MAQ-all is the combination of all SNPs (homozygous or heterozygous) called with MAQ.
[b] Protein truncation changes are SNPs that result in a premature stop codon.
[c] Start methionine changes are SNPs that alter the start methionine residue of a predicted gene.
[d] Protein elongating changes are SNPs found in the termination codon of a gene in the annotated London reference strain that result in incorporation of a new amino acid and read-through transcription.

Table S3.1.1. TTAGG repeats in the genome of *T. urticae.*

scaffold	strand	5' border	3' border	size	N	Id. %
1	P	6741410	6741475	66	13	95
1	P	1099767	1099824	58	12	96
1	M	5654973	5654923	51	10	94
1	M	5660392	5660342	51	10	94
1	M	2478560	2478518	43	9	93
2	M	1608262	1608197	66	13	96
2	P	2857883	2857943	61	12	96
2	M	507745	507710	57	11	96
2	M	4615969	4615943	27	5	96
2	P	6418404	6418430	27	5	96
2	M	6833627	6833562	66	13	96
2	P	6545267	6545297	31	6	93
2	M	5367195	5367170	26	5	96
2	P	155751	155770	20	4	100
3	M	3527591	3527392	200	40	100
3	M	3527907	3527748	160	32	99
3	M	3527668	3527619	50	10	100
3	M	3527720	3527696	25	5	96
3	M	4688865	4688808	58	12	96
3	M	2704474	2704420	55	11	90
3	M	507747	507716	32	6	96
3	M	3145931	3145902	30	6	96
3	P	1763334	1763364	31	6	93
3	M	488961	488940	22	4	100
3	P	971400	971425	26	5	96
3	M	945694	945670	25	5	96
3	M	961553	961529	25	5	96
3	M	1025243	1025219	25	5	96
3	M	1026405	1026381	25	5	96
4	M	2758795	2758730	66	13	95
4	M	3192942	3192877	66	13	95
4	P	651873	651915	43	9	93
5	M	2215264	2215214	51	10	96
5	P	3296371	3296393	23	5	100
5	M	3791852	3791827	26	5	96
5	M	1422154	1422130	25	5	96
5	P	1432909	1432933	25	5	96
5	M	1469385	1469361	25	5	96
6	P	1841417	1841560	144	29	99
6	P	1835174	1835284	111	22	99
6	M	1672613	1672538	76	15	96
6	P	1681923	1681980	58	12	96
6	M	1768890	1768847	44	9	95
6	P	1726894	1726931	38	8	92
6	M	1532431	1532398	34	7	100
6	P	2744264	2744291	28	6	96
6	P	3024031	3024055	25	5	96
7	P	14144	14388	245	49	96
7	P	1110876	1111087	212	42	97
7	M	340674	340627	48	10	95
7	P	1548501	1548526	26	5	96
7	P	1553197	1553222	26	5	96
7	P	254972	254996	25	5	96
7	M	400658	400638	21	4	100
8	P	366549	366586	38	8	92
8	P	366937	366965	29	6	96
8	P	366366	366400	35	7	91
8	P	367011	367039	29	6	93
8	M	1056004	1055981	24	5	95
9	P	2723652	2723713	62	12	93
9	M	2645428	2645393	36	7	94
9	M	3601186	3601111	76	15	96
9	P	3386628	3386685	58	12	94
9	M	944629	944606	24	5	100
9	P	983876	983903	28	6	96
9	M	935106	935081	26	5	96
10	M	1390867	1390867	31	6	96
10	P	1292626	1292667	42	8	90
10	P	1338608	1338633	26	5	96
10	P	254337	254363	27	5	96
10	P	1959560	1959586	27	5	96
10	P	274349	274373	25	5	96
10	P	266264	266286	23	5	95
10	P	1545504	1545504	24	5	95
10	M	2653197	2653175	23	5	95
11	P	2495988	2496014	27	5	96
11	P	2473774	2473807	34	7	91
11	M	2465208	2465184	25	5	96
11	P	2479114	2479142	29	6	93
11	M	2493458	2493436	23	5	95

Table S3.1.1. (continued) TTAGG repeats in the genome of *T. urticae*.

scaffold	strand	5' border	3' border	size	N	Id. %
11	M	89343	89312	32	6	90
11	M	89440	89416	25	5	96
12	M	363125	363076	50	10	94
12	P	1571066	1571105	40	8	92
12	M	752494	752460	35	7	94
12	M	781325	781306	20	4	100
12	M	119587	119562	26	5	100
12	M	2301001	2300975	27	5	96
12	M	1756980	1756955	26	5	96
13	P	1084206	1084584	379	76	97
13	P	1084747	1084964	218	44	100
13	P	1084989	1085080	92	18	100
13	P	1083942	1084137	196	39	98
13	P	1084610	1084721	112	22	100
13	M	1082397	1082360	38	8	97
13	P	464906	464988	83	17	97
13	P	1878702	1878744	43	9	93
13	M	1231068	1231044	25	5	96
13	P	1442405	1442428	24	5	95
14	P	2097021	2097042	22	4	100
14	P	2116891	2116919	29	6	100
14	P	2136663	2136691	29	6	100
14	P	1135218	1135257	40	8	92
14	P	1083950	1083975	26	5	96
14	P	1628512	1628548	37	7	97
14	P	1602424	1602448	25	5	96
14	M	1607565	1607541	25	5	96
16	P	1187875	1187917	43	9	93
16	P	1545385	1545409	25	5	96
16	M	318010	317987	24	5	100
16	P	278741	278768	28	6	96
16	P	1850105	1850144	40	8	92
17	P	827468	827503	36	7	97
18	P	1724382	1724421	40	8	92
18	P	214643	214673	31	6	96
19	P	1168579	1168598	20	4	100
19	P	803010	803036	27	5	92
19	P	1512207	1512229	23	5	95
20	P	1468068	1468095	28	6	96
20	P	1462334	1462356	23	5	95
20	M	1416204	1416174	31	6	93
20	M	1275641	1275619	23	5	95
21	P	951578	951603	26	5	96
21	M	722796	722772	25	5	96
21	P	1046944	1046972	29	6	93
22	M	488490	488425	66	13	95
22	P	1228531	1228553	23	5	100
24	P	815806	815834	29	6	96
24	P	816149	816177	29	6	96
24	P	816240	816240	29	6	96
24	P	816453	816481	29	6	93
24	P	557963	557990	28	6	96
26	P	1107129	1107160	32	6	93
27	P	967165	967191	27	5	92
28	P	559699	559722	24	5	95
29	M	493508	493473	36	7	94
29	M	885059	885036	24	5	95
29	P	606898	606924	27	5	92
30	P	747170	747487	318	64	98
32	P	494552	494579	28	6	96
32	M	504912	504890	23	5	95
33	P	506986	507034	49	10	89
33	P	503710	503734	25	5	96
33	M	484372	484348	25	5	96
34	P	36007	36032	26	5	100
34	M	601793	601793	23	5	95
34	P	676599	676599	26	5	96
34	M	663802	663782	21	4	100
34	M	688792	688769	24	5	95
34	P	396888	396910	23	5	95
35	M	81202	81174	29	6	100
37	P	100621	100645	25	5	96
37	M	117108	117084	25	5	96
38	P	281752	281775	24	5	95
43	P	98073	98128	56	11	98
45	M	12507	12214	294	59	100
45	M	46469	46221	249	50	99
45	M	54350	54175	176	35	97
45	M	87123	86972	152	30	97
45	M	65	2	64	13	98

Table S3.1.1. (continued) TTAGG repeats in the genome of *T. urticae*.

scaffold	strand	5' border	3' border	size	N	Id. %
52	P	48797	48889	94	19	97
54	P	40784	41003	220	44	100
54	P	37693	37906	214	43	99
54	P	38303	38356	54	11	100
54	P	40717	40769	53	11	100
54	P	25349	25502	154	31	100
54	P	19560	19604	45	9	100
54	P	19998	20036	39	8	94
113	M	13513	13490	24	5	100
143	P	4738	4760	23	5	95
169	P	2605	2644	40	8	92
213	M	3524	3494	31	6	93
234	P	3664	3683	20	4	100
324	M	3840	3810	31	6	90
389	P	3951	3974	24	5	95
429	M	282	38	245	49	99
503	P	1883	1911	29	6	93
548	M	3043	3005	39	8	100
558	M	140	30	111	22	99
558	M	2866	2779	88	18	100
564	M	2379	2334	46	9	93
576	M	2200	2115	86	17	97
677	M	1956	1929	28	6	96
Total				9893	1979	

Table S3.2.1. Composition of transposable elements (TEs) in the *T. urticae* genome. A TE is regarded as complete when its sequence shows at least 90% coverage in length with a similar TE.

	Total bp TE	% bp TE	No. of TEs	Total bp complete TE	% bp complete TE	No. of complete TEs
Transposable element	**9,089,640**	**10.01**	**13,552**	**5,350,678**	**58.87**	**2,243**
Class I: retrotransposon	5,657,281	6.23	6,738	3,512,918	62.10	1,169
LTR retrotransposon	3,510,815	3.87	3,459	2,343,456	66.75	427
Gypsy	2,827,124	3.11	2,594	1,912,013	67.63	348
Copia	683,691	0.75	865	431,443	63.10	79
Non-LTR retrotransposon	2,146,466	2.36	3,279	1,169,462	54.48	742
LINE	2,146,466	2.36	3,279	1,169,462	54.48	742
L1	1,536,281	1.69	2,853	695,386	45.26	624
CR1	296,083	0.33	222	225,829	76.27	77
R2	201,874	0.22	145	177,115	87.74	24
I	102,101	0.11	53	65,606	64.26	15
LOA	10,127	0.01	6	5,526	54.57	2
Class II: DNA transposon	3,432,025	4.00	6,813	1,837,760	53.55	1,074
TIR	2,290,988	2.52	6,016	1,157,831	50.54	1,002
Tc1-Mariner	1,487,149	1.64	3,983	748,908	50.36	656
PiggyBac	291,499	0.32	661	92,499	31.73	3
Mutator	138,067	0.15	253	103,021	74.62	70
Merlin	115,975	0.13	411	38,953	33.59	46
CACTA	71,309	0.08	41	61,547	86.31	29
hAT	56,031	0.06	99	45,957	82.02	30
MITE	86,975	0.10	473	41,020	47.16	146
P	16,223	0.02	20	13,335	82.20	10
Harbinger	9,383	0.01	14	7,743	82.52	8
IS4EU	2,776	0.00	3	2,776	100.00	3
Pogo	15,601	0.02	58	2,072	13.28	1
Helitron	78,741	0.09	91	53,040	67.36	22
Maverick	1,062,296	1.17	706	626,889	59.01	50
unclassified	334	0.00	1	0	0.00	0

Table S3.4.1. U12-type introns and U12 intron-containing genes in the *T. urticae* genome, as compared with the ones from the insect *Apis mellifera* (reference grey column) and their homologs in the tick *Ixodes scapularis*. NI denotes the total number of introns in the given gene. U12: the U12-type intron(s) in this gene. X: denotes what is present at the position homologous to the honeybee U12-intron, i.e. either a U12-type intron (U12), a U2-type intron (U2) or no intron at all (none). nU12: number of U12 intron in the given tick gene. IM: incomplete gene model. Honeybee has been chosen as reference, as being currently the insect with the highest number of documented U12-type introns. NF: Not found.

Gene	Description	NI	U12	*T. urticae*	NI	U12	X	*I. scapularis*	NI	X	nU12
WLS	Wntless, Sprinter	9	I8	tetur09g04100	4		none	NF		#	
IMP4	U3 small nucleolar ribonucleoprotein	2	I1	tetur26g00720	3		none	EEC04995.1	5	U2	
PTEN	Phosphatidylinositol-3,4,5-trisphosphate 3-phosphatase	6	I1	tetur04g05240	2		U2	EEC15363.1	8	U12	1
TWS	Protein phosphatase PP2A 55 kDa regulatory subunit B	7	I5	tetur21g02310	6		U2	EEC10168.1	7	U12	1
TM41B	SNARE conserved transmembrane protein 41B	3	I3	tetur09g01110	1		none	EEC00250.1	>=4	U12	1
UBCH2	Ubiquitin-conjugating enzyme E2 H	4	I1	tetur11g00580	4		U2	XP_002401453.1		U12	1
CSL4	exosomal core protein CSL4	4	I3	tetur12g01170	3		none	NW_002702032.1	4	U12	1
RNF121	Ring Finger protein 121	6	I6	tetur04g03230	1		none	XP_002414555.1	4	U12	2
INT4	integrator complex, subunit4	6	I5	tetur10g04860	7		none	XM_002416526.1	19	U12	1
ZPR1	Growth-controlling Zinc Finger protein ZPR1	2	I2	tetur08g08339	0		none	XP_002413166.1	0	none	
CYCT	Cyclin T	8	I1	tetur01g09960	9		U2	NF		#	
MLXI	Mlx interactor	14	I13	tetur26g02720	10		U2	NF		#	
IPO4	Importin 4	15	I13	tetur01g09060	15		U2	XP_002402287.1	22	U12	1
SLC12	potassium/chloride symporter	9	I9	tetur13g04280	3		none	XP_002414472	IM	U12	>=2
VPS35	vacuolar protein sorting 35	11	I2	tetur02g05530	0		none	XP_002416036.1	>=12	U12	1
VAC14	endosomal protein of the PI(3,5)P2 regulatory complex	5	I4	tetur24g00560	3		none	EEC01685.1	>=4	U12	1
ZN207	Zinc Finger Protein 207	5	I1	tetur01g13560	2		U2	EEC10946.1	9	U2	
LSM12	Sm-like protein LSM12	2	I1	tetur29g01806	1		none	XP_002406586.1	4	U12	1
UBL5	Ubiquitin-like protein 5	1	I1	tetur25g01660	0		none	EEC14026.1	>=1	U12	1
PPP5C	Protein Phosphatase 5 Catalytic subunit	7	I3	tetur07g06830	1		none	EEC18365.1	14	U12	1
ROSY	Xanthine OxidoReductase	23	I2	NF	#		#	(many)	IM	#	
NAT5	N-acetyltransferase 5	5	I2	tetur06g05330	4		none	XP_002412276.1	5	U12	1
TOM	Tomosyn	21	I2	tetur03g05170	12		none	EEC13715.1	>15	#	

Table S3.4.1. (continued) U12-type introns and U12 intron-containing genes in the *T. urticae* genome, as compared with the ones from the insect *Apis mellifera* (reference grey column) and their homologs in the tick *Ixodes scapularis*. NI denotes the total number of introns in the given gene. U12: the U12-type intron(s) in this gene. X: denotes what is present at the position homologous to the honeybee U12-intron, i.e. either a U12-type intron (U12), a U2-type intron (U2) or no intron at all (none). nU12: number of U12 intron in the given tick gene. IM: incomplete gene model. Honeybee has been chosen as reference, as being currently the insect with the highest number of documented U12-type introns. NF: Not found.

Gene	Description	NI	U12	*T. urticae*	NI	U12	X	*I. scapularis*	NI	X	nU12
SLC9A7	Sodium/hydrogen exchanger 7	11	I7	tetur20g00330	3		none	EEC12213.1	14	U12	2
SPCS3	Microsomal signal peptidase complex subunit 3	4	I2	tetur01g07070	0		none	XP_002403300.1	5	U12	1
TRN1	Transportin 1	14	I1	tetur27g01540	12		none	EEC15287.1	>15	#	
TCPA	T-complex protein 1 subunit alpha	9	I5	tetur18g00250	6		U2	EEC18843.1	0	none	
TSP97E	Tetraspanin 97E	5	I2	tetur01g13050	3		U2	NF	#	#	
XPO7	Exportin 7	9	I8	tetur02g04890	13		U2	EEC05625.1	19	U12	2
RPB7	DNA-directed RNA polymerases I, II, and III 7kDa polypeptide	IM	I1	tetur01g10590	2		U2	EEC19574.1	3	U12	1
MCTS1	Cap-binding translation enhancer	5	I4	tetur06g06080	3		none	EEC06754.1	>4	U2	
INPP5	Inositol polyphosphate-5-phosphatase	17	I15	tetur08g06610	9		none	NF	#	#	
PHB2	Prohibitin-2	5	I3	tetur29g00190	5	I4	U12	EEC14254.1	0	#	
CA1D	Voltage-dependent calcium channel type D subunit alpha-1	35	I1/I15	tetur04g01120	26		n/U2	EEC11274.1	33	U12	2
SYX6	Syntaxin 6	6	I2	tetur29g00880	4		U2	XP_002409905.1	6	U12	1
SF3A1	Splicing factor 3 subunit 1	10	I1	tetur03g03590	4		U2	NF	#	#	
ZNF830	Zn Finger Protein 830	3	I1	tetur09g05690	1		none	NF	#	#	
CACNA1G	voltage-gated T-type calcium channel alpha subunit	31	I2	tetur29g01770	22	I1	U12	XP_002401506.1	IM	#	
ADAM10	ADAM metalloprotease-disintegrin	14	none	tetur32g00930	13	I10	U12	XP_002407705.1	13	none	none

Table S4.1.1. Deep sequencing of *T. urticae* small RNAs.

Developmental stage	Number of reads	Number of unique sequences
Embryo	4,356,855	946,825
Nymph + larvae	5,081,468	1,173,195
Adult	1,240,082	334,828
Total	10,678,405	2,454,848

Table S4.3.1. miRNAs conserved between *D. melanogaster* and *T. urticae*. Conservation is based on near perfect match with *D. melanogaster* miRNA sequences with a maximum of 2 mismatches.

T. urticae	*D. melanogaster*
tur-mir-87	dme-miR-87
tur-mir-190	dme-miR-190
tur-mir-184	dme-miR-184
tur-mir-2-2	dme-miR-2a
tur-mir-2-1	dme-miR-2a
tur-mir-7	dme-miR-7
tur-mir-133	dme-miR-133
tur-mir-34	dme-miR-34
tur-mir-10	dme-miR-10
tur-mir-1	dme-miR-1
tur-mir-12a	dme-miR-12
tur-mir-263a	dme-miR-263a
tur-mir-305	dme-miR-305
tur-mir-281	dme-miR-281
tur-mir-276	dme-miR-276a
tur-mir-252	dme-miR-252
tur-mir-124-1	dme-miR-124
tur-mir-124-2	dme-miR-124
tur-mir-137	dme-miR-137

Table S4.3.2. Number of *T. urticae* miRNAs conserved in other arthropod groups.

Arthropod group	Species	Number of miRNAs in miRBase	Number of conserved *T. urticae* miRNAs
Chelicerata, Arachnida, Acari	*Ixodes scapularis*	37	18
Mandibulata, Pancrustacea, Branchiopoda	*Daphnia pulex*	45	19
Mandibulata, Pancrustacea, Insecta	*Aedes aegypti*	101	18
	Anopheles gambiae	67	18
	Apis mellifera	174	19
	Acyrthosiphon pisum	123	17
	Bombyx mori	487	18
	Culex quinquefasciatus	72	16
	Drosophila ananassae	76	16
	Drosophila erecta	81	16
	Drosophila grimshawi	82	16
	Drosophila melanogaster	238	19
	Drosophila mojavensis	71	16
	Drosophila persimilis	75	16
	Drosophila pseudoobscura	211	19
	Drosophila sechellia	78	16
	Drosophila simulans	136	18
	Drosophila virilis	74	16
	Drosophila willistoni	77	16
	Drosophila yakuba	80	15
	Heliconius melpomene	2	0
	Locusta migratoria	7	3
	Nasonia giraulti	32	11
	Nasonia longicornis	28	13
	Nasonia vitripennis	53	18
	Tribolium castaneum	206	20
Mandibulata, Myriapoda, Chilopoda	*Strigamia maritima*	4	1

The arthropod miRNAs sequences were obtained from the miRBase[45] version 17, April 2011.

Table S4.3.3. List of the 26 *T. urticae* miRNAs conserved in other arthropod species.

T. urticae miRNA	Number of matches	*Ixodes scapularis*	*Daphnia pulex*	*Aedes aegypti*	*Anopheles gambiae*	*Apis mellifera*	*Acyrthosiphon pisum*	*Bombyx mori*	*Culex quinquefasciatus*	*Drosophila ananassae*	*D. erecta*	*D. grimshawi*	*D. melanogaster*	*D. mojavensis*	*D. persimilis*	*D. pseudoobscura*	*D. sechellia*	*D. simulans*	*D. virilis*	*D. willistoni*	*D. yakuba*	*Locusta migratoria*	*Nasonia giraulti*	*N. longicornis*	*N. vitripennis*	*Tribolium castaneum*	*Strigamia maritima*
tur-mir-276	25	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	
tur-mir-10	24	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	
tur-mir-2-1	23	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	
tur-mir-133	23	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	
tur-mir-184	23	●		●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	
tur-mir-2-2	23	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	
tur-mir-7	23	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●			●	●	●	
tur-mir-1	23	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●		●	●	
tur-mir-281	22			●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	
tur-mir-124-2	22	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●				●	●	
tur-mir-124-1	22	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●				●	●	
tur-mir-34	21		●	●	●	●		●		●	●	●	●	●	●	●	●	●	●	●	●		●	●	●	●	
tur-mir-305	21	●	●	●	●	●		●	●	●	●	●	●	●	●	●	●		●	●	●		●	●	●		
tur-mir-263a	21	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●					●	
tur-mir-12a	20	●	●	●	●	●			●	●	●	●	●	●	●	●	●	●	●	●				●	●	●	
tur-mir-87	18	●		●			●	●	●	●	●	●	●	●	●	●	●	●	●	●	●					●	
tur-mir-137	15	●	●	●	●	●	●	●	●				●			●		●					●	●	●	●	
tur-mir-190	11			●	●	●	●	●	●				●			●		●							●	●	
tur-mir-317	7	●	●		●	●																	●		●	●	
tur-mir-252	7						●	●	●				●			●		●							●		
tur-mir-263b	6		●			●	●																●	●		●	
tur-mir-279	4	●	●				●																			●	
tur-mir-745	1		●																								
tur-mir-3931	1																										●
tur-mir-92	1		●																								
tur-mir-278	1	●																									

● – denotes match between *T. urticae* miRNA and miRNA from other species.
The arthropod miRNAs sequences were obtained from the miRBase[45] version 17, April 2011.

Table S4.3.4. List of 26 *T. urticae* miRNAs that are not conserved in other arthropods and appear to be *T. urticae* or lineage specific.

T. urticae miRNA
tur-mir-5728-2
tur-mir-5727
tur-mir-5732
tur-mir-5728-1
tur-mir-71-2
tur-mir-5729a
tur-mir-5730
tur-mir-71-1
tur-mir-12b
tur-mir-5737
tur-mir-993a-1
tur-mir-993a-2
tur-mir-993b
tur-mir-5735
tur-mir-307
tur-mir-5736b
tur-mir-5733
tur-mir-5734
tur-mir-5736a
tur-mir-9-1
tur-mir-5731
tur-mir-210-2
tur-mir-5738
tur-mir-5729b
tur-mir-9-2
tur-mir-210-1

Table S4.3.5. Top 20 fold changes higher than 2 between two different developmental stages.

Developmental stages	miRNA ID	Stage 1	Stage 2	Logfold change
Embryo-Larvae	tur-mir-307	1.51	11.07	-9.56
Larvae-Adult	tur-mir-307	11.07	2.81	8.26
Larvae-Adult	tur-mir-12a	12.78	4.64	8.14
Embryo-Larvae	tur-mir-124-2	3.46	11.26	-7.80
Embryo-Larvae	tur-mir-124-1	3.46	11.26	-7.80
Embryo-Adult	tur-mir-92	10.35	2.58	7.77
Larvae-Adult	tur-mir-2-1	11.36	4.46	6.90
Larvae-Adult	tur-mir-2-2	11.36	4.46	6.90
Embryo-Adult	tur-mir-993a-1	10.21	3.58	6.63
Embryo-Adult	tur-mir-993a-2	10.21	3.58	6.63
Larvae-Adult	tur-mir-993a-1	9.48	3.58	5.89
Larvae-Adult	tur-mir-993a-2	9.48	3.58	5.89
Larvae-Adult	tur-mir-34	8.67	3.00	5.67
Larvae-Adult	tur-mir-5729a	7.47	2.00	5.47
Embryo-Larvae	tur-mir-2-1	5.90	11.36	-5.46
Embryo-Larvae	tur-mir-2-2	5.90	11.36	-5.46
Embryo-Larvae	tur-mir-12a	7.36	12.78	-5.43
Embryo-Adult	tur-mir-5736b	0.83	6.23	-5.40
Larvae-Adult	tur-mir-92	7.88	2.58	5.29
Embryo-Adult	tur-mir-5736a	1.68	6.91	-5.23

Table S5.5.1. *T. urticae* genes belonging to the LY6_UPAR family.

T.urticae gene ID	Short name	Fly Orthologue	Fly function
tetur01g02070	TeturLY6_UPAR1		
tetur01g08990	TeturLY6_UPAR2	*bou*	Septate Junctions
tetur02g03340	TeturLY6_UPAR3a		
tetur02g14631	TeturLY6_UPAR3b		
tetur02g05170	TeturLY6_UPAR4		
tetur03g01290	TeturLY6_UPAR5	*crok*	Septate Junctions
tetur05g06110	TeturLY6_UPAR6	*rtv*	Chitin secretion
tetur06g01010	TeturLY6_UPAR7	*sss*	Circadian rythm
tetur07g00260	TeturLY6_UPAR8	CG3955	Unknown
tetur07g05160	TeturLY6_UPAR9		
tetur08g03980	TeturLY6_UPAR10	*cold*	Septate Junctions
tetur10g04470	TeturLY6_UPAR11		
tetur12g03100	TeturLY6_UPAR12		
tetur13g02570	TeturLY6_UPAR13		
tetur14g00310	TeturLY6_UPAR14		
tetur14g00320	TeturLY6_UPAR15		
tetur19g01210	TeturLY6_UPAR16	CG6583	Unknown
tetur20g02630	TeturLY6_UPAR17		
tetur22g02040	TeturLY6_UPAR18		
tetur27g01160	TeturLY6_UPAR19	*crok*	Septate Junctions
tetur27g02450	TeturLY6_UPAR20	*crim*	Septate Junctions
tetur28g02000	TeturLY6_UPAR21	CG31323	Unknown
tetur20g03422	TeturLY6_UPAR22		
tetur13g02090	TeturLY6_UPAR23		

Table S6.1.1. Gene numbers and gene names of P450 genes in *T. urticae.*

Clan 2

T.urticae gene ID	CYP number
tetur02g06650	CYP392B2
tetur02g14020	CYP392A8
tetur02g14330	CYP392A9v2
tetur02g14400	CYP392A10v2
tetur03g00020	CYP392A13v2
tetur03g00830	CYP392A12
tetur03g00970	CYP392A11
tetur03g03950	CYP392C1
tetur03g04990	CYP392D2
tetur03g05000	CYP392D3
tetur03g05010	CYP392D4
tetur03g05030	CYP392D6
tetur03g05070	CYP392D8
tetur03g05540	CYP392E1
tetur03g09941	CYP392A15
tetur03g09961	CYP392D7
tetur06g02400	CYP392E2
tetur06g02820	CYP392E3
tetur06g04520	CYP392A16
tetur07g06410	CYP392A1
tetur07g06460	CYP392A3
tetur07g06480	CYP392A4
tetur08g07950	CYP392A14
tetur08g08050	CYP392A13v1
tetur10g03900	CYP307A1
tetur11g00530	CYP392A6
tetur11g04390	CYP392A5
tetur16g03500	CYP392A7
tetur16g03790	CYP392A10
tetur20g00290	CYP392B3
tetur20g03200	CYP392B1
tetur23g00260	CYP392D1
tetur27g00330	CYP392E6
tetur27g00340	CYP392E7
tetur27g00350	CYP392E8
tetur27g01020	CYP392E9
tetur27g01030	CYP392E10
tetur27g02598	CYP392E4
tetur47g00090	CYP392A9

Clan 3

T.urticae gene ID	CYP number
tetur03g01560	CYP382A1
tetur05g04000	CYP385B1
tetur07g05500	CYP385A1
tetur11g05000	CYP385C2
tetur11g05520	CYP385C4
tetur11g05540	CYP385C3
tetur14g01350	CYP383A1
tetur26g01470	CYP385C1
tetur38g00650	CYP384A1
tetur46g00150	CYP385C4v2
tetur46g00170	CYP385C3v2
tetur602g00010	CYP385B1
	end of scaffold

Clan M

T.urticae gene ID	CYP number
tetur03g03020	CYP314A1
tetur05g02550	CYP302A1
tetur06g05620	CYP315A1
tetur13g02840	CYP381A2
tetur13g02850	CYP381A1

Clan 4

T.urticae gene ID	CYP number
tetur01g00650	CYP4CL1
tetur01g04440	CYP406A1
tetur01g06120	CYP387A2
tetur03g00910	CYP390A1
tetur03g05190	CYP388A1
tetur05g02950	CYP389C2
tetur05g02960	CYP389C3
tetur05g02970	CYP389C4
tetur05g06560	CYP389C12
tetur05g06570	CYP389C11
tetur05g06580	CYP389C10
tetur05g06600	CYP389C9
tetur05g06610	CYP389C8
tetur05g06620	CYP389C7
tetur05g06630	CYP389C6
tetur05g08400	CYP389C5
tetur07g08087	CYP390B1
tetur08g06170	CYP387A1
tetur09g03800	CYP4CF2
tetur11g06070	CYP386A1
tetur20g00830	CYP407A1
tetur25g02050	CYP389A1
tetur25g02060	CYP389B1
tetur34g00510	CYP389C1
tetur36g00920	CYP391A1

Table S6.1.2. Comparison of CYP gene number in Insecta, Crustacea and *T. urticae*.

	Total	CYP2 clan	mitochondrial CYP clan	CYP3 clan	CYP4 clan
Insecta					
Drosophila melanogaster	88	7	11	36	32
Anopheles gambiae	105	10	9	40	46
Aedes aegypti	160	12	9	82	57
Bombyx mori	85	7	12	30	36
Apis mellifera	46	8	6	28	4
Nasonia vitripennis	92	7	7	48	30
Tribolium castaneum	134	8	9	72	45
Acyrthosiphon pisum	64	10	8	23	23
Pediculus humanus	36	8	8	11	9
Crustacea					
Daphnia pulex	75	20	6	12	37
Acari					
Tetranychus urticae	86	48	5	10	23

numbers are derived from Feyereisen[134] and this study.

Table S6.1.3. An overview of GSTs and the number of genes belonging to different subgroups.

GST-family	*D. melanogaster*	*A. gambiae*	*A. mellifera*	*N. vitripennis*	*T. castaneum*	*B. mori*	*H. sapiens*	*T.urticae*
alpha	-	-	-	-	-	-	5	-
delta	11	12	1	5	3	4	-	16
epsilon	14	8	-	-	19	8	-	-
mu	-	-	-	-	-	-	5	12
pi	-	-	-	-	-	-	1	-
omega	5	1	1	2	4	4	2	2
sigma	1	1	4	8	7	2	1	-
theta	4	2	1	3	1	1	2	1
zeta	2	1	1	1	1	2	1	-
unknown	-	3	-	-	-	2	-	-
Total	37	28	8	19	35	23	17	31

numbers are derived from Hayes *et al.*[79], Oakeshott *et al.*[80] and this study.

Table S6.1.4. Gene numbers and accession numbers of GST sequences of *T. urticae, D. melanogaster, A. gambiae, A. mellifera* and Acari.

T. urticae

T.urticae gene ID	Name
tetur01g02230	TuGSTd01
tetur01g02320	TuGSTo01
tetur01g02470	TuGSTd02
tetur01g02480	TuGSTd03
tetur01g02500	TuGSTd04
tetur01g02510	TuGSTd05
tetur03g07920	TuGSTd06
tetur03g09230	TuGSTm01
tetur05g05180	TuGSTm02
tetur05g05190	TuGSTm03
tetur05g05200	TuGSTm04
tetur05g05210	TuGSTm05
tetur05g05220	TuGSTm06
tetur05g05240	TuGSTm07
tetur05g05250	TuGSTm08
tetur05g05260	TuGSTm09
tetur05g05270	TuGSTm10
tetur05g05290	TuGSTm11
tetur05g05300	TuGSTm12
tetur07g02560	TuGSTz01
tetur12g03900	TuGSTo02
tetur22g02300	TuGSTk01
tetur26g01450	TuGSTd07
tetur26g01460	TuGSTd08
tetur26g01490	TuGSTd11
tetur26g01500	TuGSTd12
tetur26g01510	TuGSTd13
tetur26g02801	TuGSTd09
tetur26g02802	TuGSTd10
tetur29g00220	TuGSTd14
tetur31g01330	TuGSTd15
tetur31g01390	TuGSTd16

D. melanogaster

Flybase symbol	Name in tree
CG6662	Dm_CG6662
CG6776	Dm_CG6776
CG6781	Dm_CG6781
CG6673	Dm_CG6673
CG9362	Dm_CG9362
CG9363	Dm_CG9363
GstS1	Dm_GstS1
CG30005	Dm_CG30005
CG30000	Dm_CG30000
CG1702	Dm_CG1702
CG1681	Dm_CG1681
CG4688	Dm_CG4688
CG11784	Dm_CG11784
CG5224	Dm_CG5224
CG16936	Dm_CG16936
CG17639	Dm_CG17639
GstE1	Dm_GstE1
GstE2	Dm_GstE2
GstE3	Dm_GstE3
GstE4	Dm_GstE4
GstE5	Dm_GstE5
GstE6	Dm_GstE6
GstE7	Dm_GstE7
GstE8	Dm_GstE8
GstE9	Dm_GstE9
GstE10	Dm_GstE10
GstD1	Dm_GstD1
GstD2	Dm_GstD2
GstD3	Dm_GstD3
GstD4	Dm_GstD4
GstD5	Dm_GstD5
GstD6	Dm_GstD6
GstD7	Dm_GstD7
GstD8	Dm_GstD8
GstD9	Dm_GstD9
GstD10	Dm_GstD10

A. mellifera

Beebase identifier	Name in tree
GB16959	Am_GSTS1
GB19254	Am_GSTS3
GB14372	Am_GSTS4
GB18045	Am_GSTD1
GB11466	Am_GSTO1
GB12047	Am_GSTT1
GB17672	Am_GSTZ1

A. gambiae

GenBank accession	Name in tree
AAC79992.1	Ag_GST D1
Z71480.1	Ag_GST D2
AF513638.1	Ag_GST D3
AF513635.1	Ag_GST D4
AF513634.1	Ag_GST D5
AF513636.1	Ag_GST D6
AF071161.1	Ag_GST D7
AF316637.1	Ag_GST D8
AY255857.1	Ag_GST D9
AF515527.1	Ag_GST D10
AF513637.1	Ag_GST D11
AF316638.1	Ag_GST D12
AF316635.1	Ag_GST E1
AF316636.1	Ag_GST E2
AY070234.1	Ag_GST E3
AY070254.1	Ag_GST E4
AY070255.1	Ag_GST E5
AY070256.1	Ag_GST E6
AF491816.1	Ag_GST E7
AY070257.1	Ag_GST E8
AY255856.1	Ag_GST O1
AF513639.1	Ag_GST S1
AF515526.1	Ag_GST T1
AF515525.1	Ag_GST T2
AF515522.1	Ag_GST Z1

Acari

GenBank accession	Name in tree
AAX37325.1	Ss_delta_1
AAX37323.1	Ss_delta_2
AAX37328.1	Dp_delta_1
AAO92279.1	Dv_delta_1
ACF35539.1	Dv_delta_2
AAX37327.1	Dp_mu_1
AAX37326.1	Dp_mu_2
AAX37321.1	Ss_mu1
AAD15991.1	Bm_mu1
AAQ74442.1	Ra_mu1
AAF19264.1	Po_mu1
CAQ51257.1	Po_mu2

Table S6.1.5. Gene numbers and accession numbers of CCE sequences of *T. urticae*, *D. melanogaster*, *A. gambiae*, and *A. mellifera*.

T. urticae

T.urticae gene ID	
tetur01g08680	tetur12g04600
tetur01g10740	tetur12g04610
tetur01g10750	tetur13g03700
tetur01g10760	tetur16g02380
tetur01g10800	tetur16g02390
tetur01g10810	tetur16g02410
tetur01g10820	tetur16g02420
tetur01g10830	tetur17g00080
tetur01g14090	tetur17g00300
tetur01g14170	tetur17g00350
tetur01g14180	tetur17g00410
tetur01g16180	tetur17g00750
tetur01g16371	tetur19g00850
tetur02g04030	tetur207g00010
tetur02g06930	tetur20g03250
tetur02g08440	tetur23g00910
tetur02g09870	tetur24g01310
tetur02g10640	tetur25g00710
tetur02g14551	tetur25g01850
tetur03g00310	tetur26g01130
tetur03g02700	tetur26g02780
tetur04g02550	tetur29g00820
tetur04g02770	tetur29g00930
tetur04g06380	tetur29g00940
tetur04g06770	tetur29g00970
tetur04g08480	tetur309g00020
tetur06g03330	tetur30g01290
tetur10g05290	tetur30g01560
tetur11g01500	tetur31g00250
tetur11g01560	tetur31g01560
tetur11g01750	tetur35g00200
tetur11g03760	tetur35g00210
tetur11g05760	tetur37g00330
tetur11g05770	tetur37g00340
tetur11g06371	tetur37g00590
tetur12g03000	

D. melanogaster

Flybase symbol	Name in tree
CG1121	Dm_CG1121
CG12869	Dm_CG12869
CG13772	Dm_CG13772
CG17148	Dm_CG17148
CG17907	Dm_CG17907
CG2505	Dm_CG2505
CG31146	Dm_CG31146
CG3903	Dm_CG3903
CG5397	Dm_CG5397
CG6917	Dm_CG6917
CG9280	Dm_CG9280
CG9287	Dm_CG9287
CG9289	Dm_CG9289
CG9704	Dm_CG9704
CG8425	Dm_CG8425

A. gambiae

GenBank accession	Name in tree
XP_310628.3	Ag_CCEace1o
XP_321792.2	Ag_CCEace3o
XP_315770.4	Ag_CCEae2D
XP_316296.3	Ag_CCEae2F
XP_309020.3	Ag_CCEae3G
XP_556009.2	Ag_CCEbe1C
XP_322067.4	Ag_CCEbe2o
XP_317277.4	Ag_CCEgli11o
XP_311922.2	Ag_CCE7o
XP_309784.4	Ag_CCEglt1H
XP_554880.3	Ag_CCEglt1l
XP_309783.3	Ag_CCEglt2H
XP_317954.4	Ag_CCEglt2l
XP_552325.3	Ag_CCEnrl16o
XP_315662.4	Ag_CCEnrt9o
XP_315860.3	Ag_CCEjhe3E

A. melifera

Beebase identifier	Name in tree
GB14873	Am_GB14873
GB18414	Am_GB18414
GB11064	Am_GB11064
GB16342	Am_GB16342
GB10820	Am_GB10820
GB15327	Am_GB15327
GB12309	Am_GB12309
GB10066	Am_GB10066
GB13939	Am_GB13939
GB18290	Am_GB18290
GB18720	Am_GB18720
GB18836	Am_GB18836
GB19830	Am_GB19830

Table S6.1.6. An overview of CCEs and the number of genes belonging to different subgroups.

CCE-clade	*D. melanogaster*	*A. gambiae*	*A. mellifera*	*N. vitripennis*	*T. castaneum*	*B. mori*	*T. urticae*
Dietary class (Clade A, B & C)	13	16	8	13	26	57	-
Hormone/semiochemical class							
Clade D (integument esterases)	3	0	1	4	2	2	-
Clade E (secreted β esterases)	2	4	2	11	7	2	-
Clade F (dipteran JHE)	3	6	2	2	2	4	-
Clade G (lepidopteran JhE)	-	4	-	-	-	0	-
Clade F' (Crustacean/Acari JhE)	-	-	-	-	-	0	2
Neuro/developmental class							
Clade H (glutactin)	5	10	1	1	1	1	2
Clade I (uncharacterized clade)	1	1	1	1	1	2	0
Clade J (AChEs)	1	2	2	2	2	2	1
Clade K (gliotactin)	1	1	1	1	1	1	1
Clade L (neuroligins)	4	5	5	5	5	3	5
Clade M (neurotactins)	2	2	1	1	2	2	1
Clade J'	-	-	-	-	-	-	34
Clade J"	-	-	-	-	-	-	22
Undetermined	-	-	-	-	-	-	3
Total	35	51	24	41	49	76	71

data are derived from Yu *et al.*[82], Oakeshott *et al.*[80] and this study

Table S6.1.7. Gene numbers of *T. urticae* ABC-transporter sequences belonging to class B and C.

Tetranychus urticae			
ABCC		ABCB	
T. urticae gene ID	Name	*T. urticae* gene ID	Name
tetur01g07880	TuABCC-01	tetur11g04030	TuABCB-mdr1
tetur01g10390	TuABCC-02	tetur11g04040	TuABCB-mdr2
tetur01g15310	TuABCC-03	tetur17g02000	TuABCB-01
tetur01g15330	TuABCC-04	tetur32g01330	TuABCB-02
tetur01g15340	TuABCC-05		
tetur03g02240	TuABCC-06		
tetur03g07460	TuABCC-07		
tetur03g07490	TuABCC-08		
tetur03g07840	TuABCC-09		
tetur03g09800	TuABCC-10		
tetur03g09880	TuABCC-11		
tetur04g04360	TuABCC-12		
tetur04g05540	TuABCC-13		
tetur04g07860	TuABCC-14		
tetur04g07910	TuABCC-15		
tetur05g01110	TuABCC-16		
tetur05g04300	TuABCC-17		
tetur06g00360	TuABCC-18		
tetur06g03510	TuABCC-19		
tetur06g03560	TuABCC-20		
tetur07g04290	TuABCC-21		
tetur07g04410	TuABCC-22		
tetur09g00580	TuABCC-23		
tetur09g00590	TuABCC-24		
tetur09g04610	TuABCC-25		
tetur09g04620	TuABCC-26		
tetur11g02060	TuABCC-27		
tetur11g02120	TuABCC-28		
tetur11g05990	TuABCC-29		
tetur14g02290	TuABCC-30		
tetur14g02310	TuABCC-31		
tetur14g02320	TuABCC-32		
tetur14g02330	TuABCC-33		
tetur16g03480	TuABCC-34		
tetur23g02452	TuABCC-38		
tetur25g01780	TuABCC-35		
tetur40g00010	TuABCC-36		
tetur06g00280	TuABCC-37		
tetur28g01950	TuABCC-39		

Table S6.1.8. An overview of ABC transporters class B and C.

ABC subfamily	*S. cerevisiae*	*C. elegans*	*H. sapiens*	*A. gambiae*	*D. melanogaster*	*D. pulex*	*T. urticae*
B-full	1	14	4	2	4	2	2
B-half	3	9	7	3	4	5	2
C	6	9	12	14	14	7	39

numbers were derived from Sturm *et al.*[83], Roth *et al.*[87] and this study.

Table S6.1.9. Cysteine peptidase genes in *T. urticae*.

C1A (papains, cathepsin L)	C2 (calpains)	C13 (legumains)	C13 (GPI-transamidases)	C14 (caspases)	C50 (separases)
tetur02g08840	tetur06g02730	tetur01g06610	tetur01g10630	tetur02g04840	tetur06g06140
tetur02g14540	tetur11g04940	tetur02g14360		tetur04g08660	
tetur06g02930	tetur12g01370	tetur02g14380		tetur04g08820	
tetur06g03040	tetur16g02530	tetur02g14410		tetur15g02850	
tetur06g03520	tetur43g00180	tetur03g02400			
tetur09g00350		tetur05g04550			
tetur09g00470		tetur05g04700			
tetur09g04400		tetur05g04710			
tetur09g04420		tetur06g02610			
tetur09g04470		tetur06g03540			
tetur10g01680		tetur08g06030			
tetur123g00050		tetur10g00310			
tetur12g01810		tetur16g03650			
tetur12g01820		tetur16g03670			
tetur12g01830		tetur16g03800			
tetur12g01840		tetur28g01760			
tetur12g01850		tetur415g00010			
tetur12g01860		tetur452g00010			
tetur12g04631		tetur60g00020			
tetur13g02490					
tetur16g03680					
tetur16g03770					
tetur23g00050					
tetur23g00860					
tetur23g00880					
tetur23g00890					
tetur23g01290					
tetur25g00650					
tetur31g00700					
tetur01g05230					
tetur01g05480					
tetur01g16463					
tetur01g16473					
tetur02g11420					
tetur03g07930					
tetur03g07950					
tetur03g08010					
tetur03g08030					
tetur03g09997					
tetur06g02580					
tetur08g05010					
tetur08g05020					
tetur08g05030					
tetur08g05310					
tetur08g06770					
tetur09g00570					
tetur09g00600					
tetur24g00270					
tetur24g00280					
tetur28g01390					
tetur28g01400					
tetur28g01420					
tetur28g01430					
tetur55g00100					
tetur55g00120					
tetur55g00130					
tetur04g08380					

Table S6.1.10. Aspartic peptidase genes in *T. urticae*.

T. urticae gene ID	Family
tetur14g03010	A1 (pepsins)
tetur32g01380	A1 (pepsins)

Table S6.1.11. Number of genes belonging to different cysteine peptidase families found in the genomes of several metazoan species.

Cysteine peptideases	*T. urticae*	*D. melanogaster*	*A. mellifera*	*T. castaneum*	*C. elegans*	*H. sapiens*
C1A (papains)	57	10	5	23	27	13
C2 (calpains)	5	4	4	7	10	13
C13 (legumains + GPI)	19+1	0+1	0+1	0+1	1+1	1+1
C14 (caspases)	4	7	4	5	3	10
C50 (separases)	1	1	0	1	2	1
Total	87	23	14	37	44	39

Table S6.2.1. Summary of the mite feeding experiment. Approximately 3500 larvae were deposited on each of the three possible hosts with three biological replicates. The number of larvae recovered and the behaviour/morphology of the larvae are also noted.

Host	Number of larvae deposed	Number of larvae recovered after 12h of feeding	Fitness of larvae
Bean (California Red Kidney)	~3500	1959	Big/green/feeding. All were recovered on leaves (where they were applied). None were found on the stem.
Tomato (Heinz variety)	~3500	2020	Small/translucent/moving (~5% of mites slightly larger). Larvae shape is very similar to when they were applied. The majority of mites not feeding but running. Many were recovered on other leaves and on stems. Some recovered mites were pink suggesting uptake of anthocyanin from feeding.
Arabidopsis thaliana (Bla-2 accession)	~3500	1066	Big/green/feeding. Smaller than on bean and moving more. Most were recovered from young leaves.

Table S6.2.2. Metrics for host-specific RNA sequencing and read alignment. Shown are summaries for each of the biological replicates (marked 1-3) and for the combination of the biological replicates (total).

Sample	Number of reads	Sequence output (bases)	Number of mapped reads	Percent reads mapped	Attributable to gene models	Percent mapped attributable to gene models
bean 1	15,679,989	1,191,679,164	11,259,147	71.81	10,118,991	89.87
bean 2	15,491,432	1,177,348,832	12,394,736	80.01	11,085,862	89.44
bean 3	10,077,336	765,877,536	7,227,328	71.72	6,577,193	91.00
bean (total)	41,248,757	3,134,905,532	30,881,211	74.87	27,782,046	89.96
tomato 1	11,332,749	861,288,924	8,160,273	72.01	7,309,029	89.57
tomato 2	15,112,239	1,148,530,164	11,468,678	75.89	10,216,891	89.09
tomato 3	11,987,642	911,060,792	8,823,748	73.61	7,954,844	90.15
tomato (total)	38,432,630	2920,879,880	28,452,699	74.03	25,480,764	89.55
Arabidopsis 1	16,469,424	1,251,676,224	11,318,375	68.72	10,107,226	89.30
Arabidopsis 2	11,050,464	839,835,264	8,137,571	73.64	7,291,731	89.61
Arabidopsis 3	12,467,252	947,511,152	8,957,065	71.84	7,986,696	89.17
Arabidopsis (total)	39,987,140	3,039,022,640	28,413,011	71.06	25,385,653	89.35

Table S6.2.3. Enriched GO terms among differentially expressed genes between mites reared on different hosts (a, bean-*A. thaliana*; b, bean-tomato, c, *A. thaliana*-tomatoes).

Table S6.2.3a. Enriched GO terms among differentially expressed genes between mites reared on beans and A. thaliana.

Biological Process

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0006030	chitin metabolic process	31	17	6	11	1.95E-005
GO:0007156	homophilic cell adhesion	15	10	3	7	1.19E-004
GO:0008152	metabolic process	324	91	66	25	1.90E-004
GO:0005975	carbohydrate metabolic process	118	36	24	12	4.48E-003
GO:0055114	oxidation reduction	210	58	42	16	4.61E-003
GO:0006508	proteolysis	285	75	58	17	5.25E-003
GO:0051258	protein polymerization	10	6	2	4	6.66E-003
GO:0006412	translation	155	44	31	13	7.77E-003
GO:0006730	one-carbon metabolic process	3	3	1	2	8.24E-003
GO:0051603	proteolysis involved in cellular protein catabolic process	3	3	1	2	8.24E-003
GO:0009166	nucleotide catabolic process	3	3	1	2	8.24E-003
GO:0006869	lipid transport	8	5	2	3	0.01
GO:0006821	chloride transport	2	2	0	2	0.04

Cellular Component

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0005840	ribosome	93	35	16	19	9.93E-007
GO:0005576	extracellular region	67	22	12	10	1.15E-003
GO:0043234	protein complex	10	6	2	4	2.88E-003
GO:0005839	proteasome core complex	3	3	1	2	5.15E-003
GO:0016020	membrane	284	61	49	12	0.03
GO:0015935	small ribosomal subunit	5	3	1	2	0.04

Molecular Function

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0003735	structural constituent of ribosome	94	36	16	20	3.42E-007
GO:0016758	transferase activity, transferring hexosyl groups	86	33	14	19	1.02E-006
GO:0008061	chitin binding	31	17	5	12	1.37E-006
GO:0016491	oxidoreductase activity	146	41	24	17	3.25E-004
GO:0043234	protein complex	10	6	2	4	2.47E-003
GO:0004222	metalloendopeptidase activity	31	12	5	7	2.83E-003
GO:0005319	lipid transporter activity	8	5	1	4	4.67E-003
GO:0015450	P-P-bond-hydrolysis-driven protein transmembrane transporter activity	3	3	1	2	4.68E-003
GO:0004175	endopeptidase activity	3	3	1	2	4.68E-003
GO:0008234	cysteine-type peptidase activity	72	20	12	8	0.01
GO:0005524	ATP binding	21	8	4	4	0.02
GO:0004579	dolichyl-diphosphooligosaccharide-protein glycotransferase activity	4	3	1	2	0.02
GO:0005102	receptor binding	2	2	0	2	0.03
GO:0019904	protein domain specific binding	2	2	0	2	0.03
GO:0015377	cation:chloride symporter activity	2	2	0	2	0.03
GO:0005215	transporter activity	46	13	8	5	0.03
GO:0004185	serine-type carboxypeptidase activity	5	3	1	2	0.04

Table S6.2.3b. Enriched GO terms among differentially expressed genes between mites reared on beans and tomatoes.

Biological Process

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0006412	translation	155	99	48	51	4.86E-012
GO:0006030	chitin metabolic process	31	20	10	10	9.66E-005
GO:0006457	protein folding	39	23	12	11	2.13E-004
GO:0007156	homophilic cell adhesion	15	11	5	6	8.05E-004
GO:0015031	protein transport	25	14	8	6	7.23E-003
GO:0008272	sulfate transport	4	4	1	3	8.82E-003
GO:0006423	cysteinyl-tRNA aminoacylation	4	4	1	3	8.82E-003
GO:0006413	translational initiation	12	8	4	4	0.01
GO:0008152	metabolic process	324	118	99	19	0.01
GO:0051258	protein polymerization	10	7	3	4	0.01
GO:0006508	proteolysis	285	104	87	17	0.02
GO:0051603	proteolysis involved in cellular protein catabolic process	3	3	1	2	0.03
GO:0009166	nucleotide catabolic process	3	3	1	2	0.03
GO:0030163	protein catabolic process	7	5	2	3	0.03

Cellular Component

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0005840	ribosome	93	72	27	45	0
GO:0005737	cytoplasm	88	38	26	12	3.49E-003
GO:0043234	protein complex	10	7	3	4	9.47E-003
GO:0005839	proteasome core complex	3	3	1	2	0.03
GO:0005622	intracellular	318	106	94	12	0.05

Molecular Function

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0003735	structural constituent of ribosome	94	70	26	44	0
GO:0008061	chitin binding	31	20	8	12	1.43E-005
GO:0005840	ribosome	8	7	2	5	6.61E-004
GO:0005524	ATP binding	21	12	6	6	3.57E-003
GO:0005737	cytoplasm	59	26	16	10	3.67E-003
GO:0008271	secondary active sulfate transmembrane transporter activity	4	4	1	3	5.45E-003
GO:0043234	protein complex	10	7	3	4	5.81E-003
GO:0016787	hydrolase activity	55	23	15	8	0.01
GO:0005622	intracellular	9	6	2	4	0.02
GO:0008234	cysteine-type peptidase activity	72	28	20	8	0.02
GO:0015450	P-P-bond-hydrolysis-driven protein transmembrane transporter activity	3	3	1	2	0.02
GO:0004175	endopeptidase activity	3	3	1	2	0.02
GO:0004531	deoxyribonuclease II activity	3	3	1	2	0.02
GO:0004222	metalloendopeptidase activity	31	14	8	6	0.02
GO:0016758	transferase activity, transferring hexosyl groups	86	32	23	9	0.03
GO:0008415	acyltransferase activity	10	6	3	3	0.03
GO:0003743	translation initiation factor activity	8	5	2	3	0.04
GO:0016491	oxidoreductase activity	146	49	40	9	0.05
GO:0003755	peptidyl-prolyl cis-trans isomerase activity	6	4	2	2	0.05

Table S6.2.3c. Enriched GO terms in differentially expressed genes between mites reared on A. thaliana *and tomatoes.*

Biological Process

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0008152	metabolic process	324	65	29	36	2.36E-011
GO:0006030	chitin metabolic process	31	10	3	7	2.27E-004
GO:0005975	carbohydrate metabolic process	118	22	11	11	5.66E-004
GO:0006508	proteolysis	285	41	26	15	9.97E-004
GO:0055114	oxidation reduction	210	27	19	8	0.03

Cellular Component

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0005840	ribosome	93	14	6	8	1.35E-003
GO:0005576	extracellular region	67	10	4	6	7.37E-003
GO:0016023	cytoplasmic membrane-bounded vesicle	5	2	0	2	0.03

Molecular Function

Category	Description	Full set	In subset	Expected in subset	Observed - expected	p-value
GO:0016758	transferase activity, transferring hexosyl groups	86	25	6	19	7.53E-010
GO:0016491	oxidoreductase activity	146	31	11	20	2.98E-008
GO:0008234	cysteine-type peptidase activity	72	17	5	12	1.04E-005
GO:0008061	chitin binding	31	10	2	8	4.19E-005
GO:0003735	structural constituent of ribosome	94	14	7	7	7.68E-003
GO:0004197	cysteine-type endopeptidase activity	28	6	2	4	0.01
GO:0004602	glutathione peroxidase activity	3	2	0	2	0.02
GO:0004579	dolichyl-diphosphooligosaccharide-protein glycotransferase activity	4	2	0	2	0.03
GO:0016023	cytoplasmic membrane-bounded vesicle	5	2	0	2	0.05
GO:0004185	serine-type carboxypeptidase activity	5	2	0	2	0.05

Table S6.3.1. Examples of putative lateral gene transfers in the *T. urticae* genome.

T. urticae gene ID	Description	NCBI Best Hit	NCBI Definition	E-value
tetur16g00990	cobalamine-independent methionine synthase (Met E)	ZP_07900462	5-methyltetrahydropteroyltriglutamate--homocysteine S-methyltransferase [Paenibacillus vortex V453]	0
tetur29g01230	Levanase, beta-D-fructofuranosidase, secreted	YP_004430685	Glycosyl hydrolase family 32 domain protein [Krokinobacter diaphorus 4H-3-7-5]	4E-102
tetur29g01280	Levanase, beta-D-fructofuranosidase, secreted	ZP_06875663	levanase [Bacillus subtilis subsp. spizizenii ATCC 6633]	2E-99
tetur01g11260	carotenoid cyclase/synthase	XP_001943170	PREDICTED: bifunctional enzyme CarRP-like isoform 1 [Acyrthosiphon pisum]	6E-170
tetur01g11270	carotenoid desaturase	XP_001943225	PREDICTED: phytoene dehydrogenase-like [Acyrthosiphon pisum]	0
tetur11g04810	carotenoid desaturase	NP_001171302	carotene dehydrogenase [Acyrthosiphon pisum]	6E-95
tetur11g04820	carotenoid desaturase	NP_001171302	carotene dehydrogenase [Acyrthosiphon pisum]	4E-104
tetur11g04840	carotenoid cyclase/synthase	XP_001943170	PREDICTED: bifunctional enzyme CarRP-like isoform 1 [Acyrthosiphon pisum]	3E-57
tetur01g00490	intradiol ring-cleavage dioxygenase	XP_001228656	hypothetical protein CHGG_02140 [Chaetomium globosum CBS 148.51]	3E-30
tetur04g00150	intradiol ring-cleavage dioxygenase	NP_825411	protocatechuate dioxygenase [Streptomyces avermitilis MA-4680]	3E-12
tetur04g08620	intradiol ring-cleavage dioxygenase	ZP_07603962	intradiol ring-cleavage dioxygenase [Streptomyces violaceusniger Tu 4113]	1E-29
tetur06g00450	intradiol ring-cleavage dioxygenase	XP_003029311	hypothetical protein SCHCODRAFT_39609 [Schizophyllum commune H4-8]	7E-31
tetur06g00460	intradiol ring-cleavage dioxygenase	XP_003029311	hypothetical protein SCHCODRAFT_39609 [Schizophyllum commune H4-8]	4E-31
tetur07g02040	intradiol ring-cleavage dioxygenase	YP_004216807	intradiol ring-cleavage dioxygenase [Acidobacterium sp. MP5ACTX9]	6E-32
tetur07g05930	intradiol ring-cleavage dioxygenase	NP_826938	protocatechuate dioxygenase [Streptomyces avermitilis MA-4680]	1E-31
tetur07g05940	intradiol ring-cleavage dioxygenase	NP_826938	protocatechuate dioxygenase [Streptomyces avermitilis MA-4680]	2E-31
tetur07g06570	intradiol ring-cleavage dioxygenase, pseudogene	YP_004203327	intradiol ring-cleavage dioxygenase [Thermus scotoductus SA-01]	0,35
tetur12g04671	intradiol ring-cleavage dioxygenase	XP_002770542	DEHA2E24376p [Debaryomyces hansenii CBS767]	1E-31
tetur13g04550	intradiol ring-cleavage dioxygenase	EDP47674	extracellular dioxygenase, putative [Aspergillus fumigatus A1163]	4E-30
tetur19g02300	intradiol ring-cleavage dioxygenase	NP_825411	protocatechuate dioxygenase [Streptomyces avermitilis MA-4680]	5E-31
tetur19g03360	intradiol ring-cleavage dioxygenase	NP_825411	protocatechuate dioxygenase [Streptomyces avermitilis MA-4680]	7E-32
tetur20g01160	intradiol ring-cleavage dioxygenase	YP_004139576	intradiol ring-cleavage dioxygenase [Mesorhizobium ciceri biovar biserrulae WSM1271]	2E-32
tetur20g01790	intradiol ring-cleavage dioxygenase	NP_825411	protocatechuate dioxygenase [Streptomyces avermitilis MA-4680]	2E-28
tetur28g01250	intradiol ring-cleavage dioxygenase	XP_003029311	hypothetical protein SCHCODRAFT_39609 [Schizophyllum commune H4-8]	2E-28
tetur44g00140	intradiol ring-cleavage dioxygenase	XP_003052962	hypothetical protein NECHADRAFT_77509 [Nectria haematococca mpVI 77-13-4]	6E-16
tetur28g02430	cyanate hydratase	ZP_01128696	cyanate lyase [Nitrococcus mobilis Nb-231]	2E-28

Table S7.1.1. Genes of the ecdysteroid biosynthesis pathway.

Gene	*T. urticae* gene ID
CYP307A1 ecdysteroid biosynthesis (spook homolog)	tetur10g03900
CYP302A1 ecdysteroid 22-hydroxylase (disembodied ortholog)	tetur05g02550
CYP315A1 ecdysteroid 2-hydroxylase (shadow ortholog)	tetur06g05620
CYP314A1 ecdysteroid 20-hydroxylase (shade ortholog)	tetur03g03020
Rieske domain protein (neverland ortholog)	tetur02g05350
short-chain dehydrogenase/reductase (non glossy/shroud ortholog)	tetur08g02130

Table S7.1.2 The 30 nuclear receptors (NR) genes in *T. urticae.*

T. urticae gene ID	NuReBASE	Name
tetur19g00650	NR1D3	Ecdysone-induced protein 75 (E75)
tetur07g04810	NR1E1	Ecdysone-induced protein 78 (E78)
tetur03g08440	NR1F4	Hormone receptor like in 46 (HR3)
tetur01g15140	NR1H1	Ecdysteroid Receptor (EcR)
tetur34g00750	NR1J1	Hormone receptor 96-like a (HR96-like a)
tetur30g01210	NR1J1	Hormone receptor 96-like b (HR96-like b)
tetur17g03630	NR1J1	Hormone receptor 96-like c (HR96-like c)
tetur01g07820	NR1J1	Hormone receptor 96-like d (HR96-like d)
tetur20g01820	NR1J1	Hormone receptor 96-like e (HR96-like e)
tetur04g03100	NR1J1	Hormone receptor 96-like f (HR96-like f)
tetur11g01960	NR1J1	Hormone receptor 96-like g (HR96-like g)
tetur36g00260	NR1J1	Hormone receptor 96-like h (HR96-like h)
tetur01g10970	NR1M	Hormone Receptor 10 (HR10)
tetur05g04280	NR2A4	Hepatocyte nuclear factor 4 (HNF4)
tetur31g01930	NR2B4	Retinoid-X-Receptor 1 (RXR1)
tetur01g09240	NR2B4	Retinoid-X-Receptor 2 (RXR2)
tetur30g02190	NR2D1	Hormone Receptor 78 (HR78)
tetur08g01210	NR2E2	Tailless (Tll)
tetur03g02550	NR2E6	Photoreceptor specific NR (PNR)
tetur01g11040	NR2E3	Hormone receptor 51 (HR51)
tetur01g02690	NR2E4	Dissatisfaction (DSF)
tetur01g07700	NR2E5	Hormone Receptor 83 (HR83)
tetur04g01460	NR2F3	Seven up (SVP)
tetur28g00490	NR3B4	Estrogen-related receptor (ERR)
tetur10g04690	NR4A4	Hormone receptor like in 38 (HR38)
tetur10g04710	NR4A5	Hormone receptor like in 38 b (HR38b)
tetur08g06490	NR5A3	Fuchi tarazu transcription factor 1 (FTZ-F1)
tetur11g04570	NR5B1	Hormone receptor like in 39 (HR39)
tetur07g00140	NR6A1	Hormone receptor 4 (HR4)
tetur34g00430	NR0A2	Knirps-like (KNR-like)

Table S7.1.3. Nuclear receptors (NRs) gene families in mite genome, spanning the different subfamilies N0-N6.

Subfamily	*T. urticae*	*D. pulex*	*D. melanogaster*
NR1	13	11	5
NR2	10	7	8
NR3	1	1	1
NR4	2	1	1
NR5	1 (x2)	2	2
NR6	1	1	1
NR0	1	2	3
Total	30	25	21

Table S7.1.4. Genes of the methyl farnesoate biosynthetic pathway.

		Gene	*T. urticae* gene ID
Methyl farnesoate biosynthesis pathway	Isoprenoid pathway	AACT	tetur01g04550, tetur01g04530, tetur12g04400
		HMGS	tetur02g14571
		HMGR	tetur05g02770
		MevK	tetur06g06460
		MevPK	tetur04g03460
		MevPPD	tetur28g00120
		IPPI	tetur04g01280
		FPPS	tetur05g07470
	Farnesoic acid O-methyltransferase gene		tetur13g03250

Table S7.1.5. *T. urticae* neuropeptide genes.

Gene name	*T. urticae* gene ID
allatotropin	tetur19g01800
allatostatin A	tetur06g00870
allatostatin B	tetur10g00930
allatostatin C	tetur24g00120
allatostatin CC	tetur15g02010
bursicon A	tetur02g05570
bursicon B1	tetur28g01650
bursicon B2	tetur08g03250
CCAP, partial	tetur35g01676
CHH-1	tetur08g01900
CHH-2	tetur07g06870
CHH-3	tetur22g01190
CHH-4	tetur06g00590
CHH-5	tetur14g03120
CCHamide-1	tetur18g03711
CCHamide-2	tetur16g03896
DH31	tetur03g08630
eclosion hormone 1	tetur11g04650
eclosion hormone 2	tetur08g04650
eclosion hormone 3	tetur02g09290
ETH	tetur22g02000
elevenin-1	tetur01g07080
elevenin-2	tetur01g07090
GnRH-like-1	tetur19g02360
GnRH-like-2	tetur07g00370
GnRH-like-3	tetur18g02230
GPA2	tetur01g03780
GPB5	tetur01g03770
IMRFamide	tetur11g05070
Insulin 1	tetur04g06150
Insulin 2	tetur10g03320
Insulin 3	tetur08g06450
NPF-1	tetur10g02210
NPF-2	tetur10g02190
orcokinin	tetur04g00100
periviscerokinin	tetur06g00930
proctolin	tetur10g01490
SIFamide	tetur22g02010
sNPF	tetur30g01200
tachykinin	tetur29g01400
vasopressin	tetur09g03810

Table S8.1.1. *T. urticae* Hox genes. Blastp results of *D. melanogaster* Hox sequences queried against spider mite amino acid database.

T. urticae gene ID	Gene name	Max score	E-value
tetur11g05940	*lab*	108	6E-24
tetur11g05920	*pb*	105	5E-23
tetur20g02580	*Dfd*	174	7E-44
tetur20g02540	*Scr*	172	2E-43
tetur20g02520	*ftz -1*	79	3E-15
tetur20g02530	*ftz -2*	79	3E-15
tetur20g02430	*Antp -1*	139	2E-33
tetur20g02440	*Antp -2*	126	9E-30
tetur20g02400	*Ubx*	101	4E-22
-	*abd-A*	-	-
tetur20g02350	*Abd-B*	88	2E-18

Table S9.1.1. The list of putative fibroin genes identified in *T. urticae* genome.

T.urticae gene ID	Expression Support (Y/N)
tetur01g04020	Y
tetur01g15000	Y
tetur01g16320	Y
tetur03g09921	Y
tetur04g08890	Y
tetur08g00010	N
tetur08g00720	N
tetur08g00730	N
tetur12g00140	Y
tetur21g03310	Y
tetur22g02971	Y
tetur28g01030	Y
tetur92g00010	Y
tetur92g00020	N
tetur118g00010	N
tetur238g00010	Y
tetur238g00020	Y

Table S9.1.2. Comparison of the elastic modulus of adult and larval spider mite silk to other silks and common materials.

Material	Elastic modulus (GPa)
Adult *T. urticae* spider mite silk	24±3
Larval *T. urticae* spider mite silk	15±3
Nephila clavipes spider silk[135]	1–10
Bombyx mori silkworm silk[135]	5
Steel[135]	200
Kevlar 49[135]	130
Cotton bale[136]	6

Table S10.1.1. *D. melanogaster* immunity-related genes in *T. urticae.*

Gene name (*Drosophila*)	Gene function/pathway	*T. urticae* gene ID	Blast E-value
PGRPs	Recognition	tetur14g01760	3,00E-15
Ankyrin	IMD	tetur15g02730	0
iap2	IMD	tetur12g01630	1,00E-16
relish	IMD	tetur07g00650	4,00E-60
Sick	IMD	tetur01g13520	2,00E-37
UBC13	IMD	tetur02g09420	2,00E-70
Uev1A	IMD	tetur29g01590	2,00E-61
Basket	JNK/IMD	tetur07g07260	0
Kayak	JNK/IMD	tetur03g03610	1,00E-07
Tak1	JNK/IMD	tetur07g04360	3,00E-64
Hemipterous	JNK	tetur02g11770	4,00E-96
Domeless	JAK/STAT	tetur05g02920	9,00E-11
Hopscotch/JAK	JAK/STAT	tetur02g05790	3,00E-49
Stat92E	JAK/STAT	tetur36g00600	1,00E-88
Cactin	TOLL	tetur21g02750	1,00E-112
Cactus	TOLL	tetur14g01540	3,00E-34
dMyD88	TOLL	tetur11g01630	8,00E-25
Dorsal	TOLL	tetur11g04270	E-110
Gprk2	TOLL	tetur35g00320	0
Grass	TOLL	tetur09g00280	2,00E-36
Nec	TOLL	tetur05g01350	3,00E-36
Pelle	TOLL	tetur01g00840	4,00E-47
	TOLL	tetur03g06520	3,00E-31
Pellino	TOLL	tetur07g04160	1,00E-166
Spatzle	TOLL	tetur11g06410	7,00E-15
Spirit	TOLL	tetur13g03390	2,00E-36
TOLLs	TOLL	tetur34g01180	0
	TOLL	tetur09g04990	0
	TOLL	tetur21g01990	5,00E-37
	TOLL	tetur21g01980	8,00E-28
Mcr	Cellular response/effector	tetur04g01020	1.00E-177
TEP	Cellular response/effector	tetur08g02040	1,00E-155
Draper	Cellular response	tetur04g05840	6.00E-62
		tetur08g04300	2.00E-65

Table S10.1.2. Immunity-related gene number across Arthropoda.

Gene name	Gene function/pathway	*T. urticae*	*D. pulex*	*A. pisum*	*D. melanogaster*
attacins	Effector/AMP	0	0	0	4
cecropins	Effector/AMP	0	0	0	4
defensin	Effector/AMP	0	0	0	1
diptericin	Effector/AMP	0	0	0	2
drosocin	Effector/AMP	0	0	0	1
drosomycin	Effector/AMP	0	0	0	6
metchnikowin	Effector/AMP	0	0	0	1
Mcr	Cellular response/Effector	1	1 to 4	1	1
TEPs	Cellular response/Effector	1	3 to 6	2	4
Turandots	Cellular response/Effector	0	0	0	8
alk	Cellular response	1	1	1	1
croquemort	Cellular response	0	0	1	1
draper	Cellular response	2	1	1	1
Eater	Cellular response	0	0	0	1
Hemese	Cellular response	0	0	0	1
Hemolectin	Cellular response	0	1	1	1
nimA	Cellular response	0	0	0	1
phl	Cellular response	1	1	1	1
pnt	Cellular response	1	1	1	1
pvr	Cellular response	1	1	1	1
rac	Cellular response	1	1	1	2
scavenger-receptor C	Cellular response	4	2	1	4
Serpent	Cellular response	0	0	0	1
PGRPs	Recognition	1	0	0	13
GNBPs	Recognition	0	11	2	3
cactin	TOLL	1	1	1	1
Cactus	TOLL	1	1	1	1
DIF	TOLL	0	0	0	1
dMyD88	TOLL	1	1	1	1
Dorsal	TOLL	1	1	1	1
Gprk2	TOLL	1	1	1	1
grass	TOLL	1	1	0	1
necrotic	TOLL	1	0	0	1
Pelle	TOLL	2	1	1	1
pellino	TOLL	1	1	1	1
Persephone	TOLL	0	0	0	1
serpin 27A	TOLL	0	0	1	1
Spaetzle	TOLL	6	7	6	6
SPE	TOLL	0	0	0	1
spheroide	TOLL	0	0	0	1
sphinx1	TOLL	0	0	0	1
sphinx2	TOLL	0	0	0	1
spirit	TOLL	1	0	0	1
tolls	TOLL	4	7	7	9

Table S10.1.2. (continued) Immunity-related gene number across Arthropoda.

Gene name	Gene function/pathway	*T. urticae*	*D. pulex*	*A. pisum*	*D. melanogaster*
Tube	TOLL	0	0	1	1
traf6 (traf2)	TOLL/JNK	0	0	0	1
Eiger	JNK	0	1	1	1
hemipterous	JNK	1	1	1	1
JRA (d-Jun)	JNK	0	1	1	1
Wengen	JNK	0	1	0	1
Basket/JNK	IMD / JNK	1	1	1	1
dTAB2	IMD / JNK	0	0	0	1
dTAK1	IMD / JNK	1	1	1	1
ird5	IMD / JNK	0	1	1	1
kayak	IMD / JNK	1	1	0	1
kenny	IMD / JNK	0	1	0	1
ankyrin	IMD	1	1	1	1
Caspar	IMD	0	1	1	1
dFADD	IMD	0	0	0	1
dnr1	IMD	0	1	1	1
Dredd	IMD	0	0	0	1
IAP2	IMD	1	1	1	1
IMD	IMD	0	1	0	1
Relish	IMD	1	1	0	1
sick	IMD	1	1	1	1
ubc13	IMD	1	1	1	1
uev1A	IMD	1	1	1	1
Domeless	JAK/STAT	1	1	1	1
Hopscotch/JAK	JAK/STAT	1	1	1	1
STAT92E	JAK/STAT	1	1	1	1
Unpaired3 (Upd3)	JAK/STAT	0	0	0	1

Table S10.2.1. *T. urticae* genes involved in small RNA silencing.

Gene family (organism)	*T. urticae* gene ID	Blast E-value
Dicer (*Drosophila*)	tetur19g00520	1e-122
	tetur07g00990	6e-57
Pasha (*Drosophila*)	tetur36g00220	1e-117
	tetur36g00250	1e-115
Drosha (*Drosophila*)	tetur12g00910	0.0
vig (*Drosophila*)	tetur22g01310	4,00E-15
Loquacious (*Drosophila*)	tetur13g00430	2,00E-24
	tetur13g00410	3,00E-21
Argonaute (*Drosophila*)	tetur20g02910	0.0
	tetur09g00620	1,00E-117
	tetur09g03140	1,00E-95
	tetur02g10560	3,00E-85
	tetur02g10580	4,00E-90
	tetur04g01190	4,00E-99
	tetur02g10570	7,00E-84
Piwi/Aub/Ago3 (*Drosophila*)	tetur06g03300	1,00E-165
	tetur28g00450	1,00E-134
	tetur28g00340	1,00E-131
	tetur06g05580	1,00E-131
	tetur06g05570	1,00E-131
	tetur06g05600	1,00E-125
	tetur17g03380	2,00E-68
Gw182 (*Drosophila*)	tetur05g07970	2,00E-61
	tetur09g00260	1,00E-71
Exportin-5 (*Drosophila*)	tetur02g00520	1,00e-54
	tetur02g00500	3,00e-57
RdRP (*C. elegans*)	tetur02g08750	1,00E-57
	tetur02g08760	7,00E-57
	tetur02g08780	2,00E-56
	tetur02g08810	2,00E-50
	tetur02g08820	3,00E-48

Table S10.2.2. Small RNA silencing gene number across Arthropoda**.**

Gene name	Gene function/pathway	*T. urticae*	*D. pulex*	*A. pisum*	*D. melanogaster*
ARGONAUTE	RNAi	7	2	3	3
Armi	RNAi	1	1	1	1
Aub/piwi/Ago3	RNAi	7	7	>4	2
Dicer	RNAi	2	3	2	3
R2D2	RNAi	0	0	0	1
vig	RNAi	1	1	0	1
RdRP	RNAi	5	0	0	0
Gw182	RNAi	2	1	1	1
Exportin 5	RNAi	2	1	1	1

Table S11.1.1. *T. urticae* cuticle proteins.

T. urticae gene ID	cuticleDB			Probability		Best cleavage site	
	Type	Score	Evalue	Signal peptide	Signal anchor	Probability	Location
tetur01g00130	RR2	26.0	3.00E-08	1.000	0.000	0.813	15..16
tetur01g03940	RR2	12.1	7.90E-06	0.999	0.000	0.994	16..17
tetur01g04380	None	-	-	0.997	0.000	0.987	16..17
tetur01g09260	None	-	-	0.998	0.002	0.897	20..21
tetur01g12820	RR2	32.6	3.00E-10	0.999	0.000	0.539	19..20
tetur01g12830	RR2	31.6	6.20E-10	0.999	0.000	0.710	18..19
tetur01g12840	RR2	32.6	3.00E-10	0.999	0.000	0.539	19..20
tetur01g13650	None	-	-	1.000	0.000	0.867	20..21
tetur01g15250	RR2	13.8	4.80E-06	0.986	0.004	0.609	18..19
tetur02g07000	RR2	12.6	6.90E-06	1.000	0.000	0.996	24..25
tetur02g09530	None	-	-	0.998	0.000	0.631	16..17
tetur03g02980	RR2	39.0	3.70E-12	1.000	0.000	0.644	22..23
tetur04g01610	None	-	-	0.996	0.001	0.842	16..17
tetur05g02790	RR2	20.9	5.60E-07	1.000	0.000	0.594	16..17
tetur05g04530	RR2	21.4	4.90E-07	0.977	0.001	0.903	19..20
tetur06g01680	RR2	35.5	4.00E-11	0.996	0.001	0.663	19..20
tetur07g01030	RR2	15.2	3.20E-06	0.847	0.000	0.842	21..22
tetur09g03360	RR2	5.6	5.70E-05	0.663	0.284	0.557	27..28
tetur09g06230	RR2	46.0	2.90E-14	0.999	0.000	0.575	15..16
tetur10g00330	RR2	28.5	5.30E-09	0.991	0.007	0.502	21..22
tetur10g02300	RR2	16.8	1.90E-06	0.820	0.014	0.363	16..17
tetur11g00600	RR2	25.5	4.20E-08	0.999	0.000	0.724	16..17
tetur11g01330	RR2	38.0	7.20E-12	0.921	0.051	0.718	22..23
tetur11g06050	RR2	28.6	4.80E-09	0.990	0.009	0.970	22..23
tetur12g02060	RR2	31.6	6.20E-10	0.995	0.004	0.840	23..24
tetur12g03140	RR2	38.9	3.90E-12	0.001	0.891	0.001	15..16
tetur15g01740	RR2	25.0	5.90E-08	0.848	0.148	0.841	26..27
tetur15g01760	RR2	12.0	8.40E-06	0.976	0.000	0.974	14..15
tetur15g02250	RR1	15.1	1.20E-07	0.999	0.001	0.967	18..19
tetur18g02430	RR2	43.6	1.50E-13	0.986	0.009	0.856	20..21
tetur19g00760	RR2	17.4	1.60E-06	0.966	0.003	0.668	16..17
tetur19g01870	RR2	35.0	5.70E-11	0.999	0.000	0.842	22..23
tetur19g02860	None	-	-	0.002	0.000	0.001	21..22
tetur19g02870	RR2	33.0	2.30E-10	0.628	0.000	0.481	19..20
tetur19g02890	RR2	31.7	5.70E-10	0.001	0.000	0.001	18..19
tetur19g02910	RR2	31.7	5.70E-10	0.093	0.789	0.050	45..46
tetur19g02920	RR2	36.4	2.20E-11	0.000	0.000	0.000	21..22
tetur19g03000	RR2	38.7	4.60E-12	0.011	0.000	0.005	17..18
tetur19g03020	RR2	37.1	1.30E-11	0.969	0.020	0.637	20..21
tetur21g01110	RR2	23.2	2.10E-07	0.999	0.001	0.906	19..20
tetur23g00980	RR2	32.2	4.00E-10	0.560	0.086	0.414	24..25
tetur23g01280	None	-	-	0.998	0.001	0.432	19..20
tetur26g02180	RR2	13.1	6.00E-06	0.982	0.008	0.812	19..20
tetur32g02090	RR2	26.4	2.30E-08	0.985	0.009	0.394	20..21
tetur43g00340	RR2	37.5	1.00E-11	0.999	0.000	0.934	19..20

cuticleDB Evalue threshold : 1E-4
SignalP.3.0: default parameters (SignalP-HMM probabilities)

Table S11.1.2. Composition of *T. urticae* cuticle proteins of the RR-2 class.

T. urticae gene ID	Length (AA)	Pfam (25.0)	Evalue	# His in Pfam00379	%His	%Lys	%Pro	%Gly	%Ala	%Tyr	%Ala+Gly
tetur01g00130	176	PF00379.17	1.3E-12	0	0.00	1.14	6.82	10.23	18.18	2.84	28.41
tetur01g03940	430	PF00379.17	6.3E-12	1	3.72	12.79	17.90	4.19	5.35	18.84	9.54
tetur01g04380	375	PF00379.17	3.3E-07	1	6.93	10.67	14.13	4.80	3.20	18.13	8.00
tetur01g09260	1421	PF00379.17	3.1E-10	1	1.83	4.50	7.74	4.36	3.38	2.67	7.74
tetur01g12820	189	PF00379.17	8.6E-11	1	1.59	1.59	9.00	9.00	25.40	5.82	34.40
tetur01g12830	157	PF00379.17	7.4E-10	1	1.91	1.27	8.92	9.56	15.29	6.37	24.85
tetur01g12840	189	PF00379.17	8.6E-11	1	1.59	1.59	9.00	9.00	25.40	5.82	34.40
tetur01g15250	680	PF00379.17	9.1E-08	1	0.29	4.56	4.85	24.12	7.50	6.62	31.62
tetur02g07000	312	PF00379.17	2.2E-06	0	6.09	2.24	10.58	12.18	2.56	3.21	14.74
tetur02g09530	135	PF11018.2	0.021	no motif	2.96	2.96	14.82	2.22	34.07	6.67	36.29
tetur03g02980	146	PF00379.17	6.1E-14	2	10.27	7.53	4.80	13.01	8.90	4.80	21.91
tetur04g01610	204	PF11018.2	0.1	no motif	8.33	2.45	10.29	2.45	22.55	3.43	25.00
tetur05g02790	269	PF00379.17	9.1E-13	1	1.12	6.32	14.50	7.81	15.61	17.84	23.42
tetur05g04530	739	PF00379.17	1.6E-16	2	3.25	7.71	8.12	3.65	4.33	5.14	7.98
tetur06g01680	189	PF00379.17	3.1E-14	3	4.23	6.35	10.58	7.94	10.58	13.23	18.52
tetur07g01030	796	PF00379.17	1.2E-14	3	4.52	7.92	11.31	4.15	3.02	5.91	7.17
tetur09g03360	155	PF00379.17	1.8E-10	2	4.52	4.52	3.87	9.68	1.29	7.10	10.97
tetur09g06230	268	PF00379.17	8.4E-09	2	1.12	2.99	11.19	5.60	24.63	9.33	30.23
tetur10g00330	257	PF00379.17	4.9E-11	1	14.40	6.23	3.50	14.01	7.78	7.39	21.79
tetur10g02300	193	PF00379.17	1.8E-12	2	2.59	5.70	11.40	7.25	7.77	16.58	15.02
tetur11g00600	529	PF00379.17	4.9E-12	0	0.19	3.40	12.10	5.86	10.21	1.32	16.07
tetur11g01330	223	PF00379.17	5E-14	0	2.24	8.97	3.14	13.00	5.83	6.73	18.83
tetur11g06050	304	PF00379.17	5.5E-13	3	9.21	8.88	3.29	16.12	3.95	4.61	20.07
tetur12g02060	191	PF00379.17	2.1E-12	1	1.05	0.52	7.85	7.85	15.71	9.95	23.56
tetur12g03140	317	PF00379.17	3.2E-11	0	2.52	10.10	2.84	8.83	9.78	5.99	18.61
tetur15g01740	194	PF00379.17	4.9E-13	2	6.19	10.31	8.25	6.70	6.70	8.76	13.40
tetur15g01760	226	PF00379.17	7.1E-14	4	3.10	6.20	8.41	7.52	6.64	7.08	14.16
tetur15g02250	241	PF00379.17	1.1E-12	2	1.66	7.88	8.71	7.47	8.30	7.05	15.77
tetur18g02430	248	PF00379.17	1.1E-13	0	0.81	3.63	7.26	6.45	8.47	5.24	14.92
tetur19g00760	360	PF00379.17	3.3E-13	0	0.83	1.39	9.17	5.83	18.89	3.06	24.72
tetur19g01870	198	PF00379.17	3.7E-13	2	21.72	6.06	2.02	18.69	6.06	3.54	24.75
tetur19g02860	209	PF00379.17	5.9E-08	1	2.39	4.79	3.35	4.31	10.53	2.87	14.84
tetur19g02870	160	PF00379.17	1.6E-17	0	0.63	3.75	3.13	5.63	9.38	4.38	15.01
tetur19g02890	217	PF00379.17	6.2E-17	0	0.46	4.15	6.91	5.07	12.44	3.69	17.51
tetur19g02910	309	PF00379.17	1.6E-16	0	1.29	3.56	5.18	3.56	11.33	2.91	14.89
tetur19g02920	160	PF00379.17	2.4E-17	0	1.25	3.75	3.75	5.63	6.25	3.75	11.88
tetur19g03000	262	PF00379.17	3E-18	0	0.38	2.29	6.11	5.73	10.31	3.05	16.04
tetur19g03020	197	PF00379.17	7.5E-18	0	0.00	3.55	4.06	5.08	14.72	4.06	19.80
tetur21g01110	192	PF00379.17	4.9E-15	1	8.85	7.81	5.73	10.94	8.33	4.69	19.27
tetur23g00980	451	PF00379.17	1.7E-12	0	6.87	4.44	5.54	7.10	6.43	5.77	13.53
tetur23g01280	154	PF01607.18	1.2E-12	other	2.60	3.90	3.90	3.90	5.84	7.79	9.74
tetur26g02180	251	PF00379.17	3E-13	1	2.39	1.20	8.77	7.57	11.95	5.18	19.52
tetur32g02090	200	PF00379.17	2.8E-13	1	0.50	1.50	11.00	8.50	22.00	8.50	30.50
tetur43g00340	200	PF00379.17	6.7E-13	0	1.50	4.00	9.00	5.50	10.50	5.00	16.00
average	302			1.05	3.63	4.93	7.79	7.91	11.08	6.65	18.99
count	44			41	44	44	44	44	44	44	44
max.	1421			4	21.72	12.79	17.90	24.12	34.07	18.84	36.29
min.	135			0	0.00	0.52	2.02	2.22	1.29	1.32	7.17
Abundance in comparison to *Anopheles* RR-2				low	low	~normal	~normal	~normal	~normal	~normal	

Table S11.1.3. Cuticular proteins analogous to peritrophins in *T. urticae*.

T. urticae gene ID	Definition	Signal protein	Location	Probability
tetur02g07830	cuticular protein analogous to peritrophins 1-A	Yes	20..21	0.890
tetur02g09340	cuticular protein analogous to peritrophins 1-A	No		
tetur03g01610	cuticular protein analogous to peritrophins 1-A	No		
tetur03g08110	cuticular protein analogous to peritrophins 1-A	No		
tetur04g05710	cuticular protein analogous to peritrophins 1-A	No		
tetur05g04610	cuticular protein analogous to peritrophins 1-A	No		
tetur06g01550	cuticular protein analogous to peritrophins 1-A, chitin-binding protein obstructor	Yes	20..21	0.852
tetur07g05090	cuticular protein analogous to peritrophins 1-A	Yes	45..46	0.754
tetur15g01480	cuticular protein analogous to peritrophins 1-A, chitin-binding protein obstructor	Yes	18..19	0.99
tetur19g00240	cuticular protein analogous to peritrophins 1-A, chitin-binding protein gasp	Yes	15..16	0.99
tetur23g01280	cuticular protein analogous to peritrophins 1-A	Yes	19..20	0.99
tetur26g01260	cuticular protein analogous to peritrophins 1-A	No		
tetur27g00090	cuticular protein analogous to peritrophins 1-A, chitin-binding protein obstructor	Yes	20..21	1.0
tetur36g00950	cuticular protein analogous to peritrophins 1-A, chitin-binding protein obstructor	Yes	20..21	0.99
tetur44g00060	cuticular protein analogous to peritrophins 1-A	Yes	19..20	0.99

Table S12.1.1. A subset of tblastn results suggesting the correspondence between certain *T. urticae* CDSs and *dnmt*s. Amel = *Apis mellifera*, Nv = *Nasonia vitripennis*. Using human dnmts provided similar results, except for dnmt2. Using Nv ortholog produced the most significant hit. Tetur13g02100 and Tetur13g02110 correspond both to parts of a *T.urticae* DNMT-1 that was disrupted by a Gypsy-like transposable element.

Subject	Query	E-value	Bit score
tetur13g02100	Amel dnmt1	1.00E-143	506
tetur13g02110	Amel dnmt1	4.00E-36	150
tetur16g01020	Amel dnmt3	8.00E-49	191
tetur02g05760	Nv Dnmt2	3.00E-49	190

Table S12.1.2. HMM search of protein domains: Results of pfam hmm search. tetur13g02100 and tetur13g02110 correspond both to parts of a *T.urticae* DNMT-1 that was disrupted by a Gypsy-like transposable element. We used the hmm search option in the pfam database to determine whether these putative *dnmt* genes harbor methylase domains. All the putative *dnmt*s possess methylase domains.

T. urticae gene ID	Family	Description	Bit score	E-value
tetur13g02100	DNA methylase	C-5 cytosine-specific DNA methylase	169.4	8.50E-50
tetur13g02110	DNMT1-RFD	Cytosine specific DNA methyltransferase	67.8	6.10E-19
tetur16g01020	DNA_methylase	C-5 cytosine-specific DNA methylase	28.7	5.40E-07
tetur02g05760	N6-adenineMlase	Probable N6-adenine methyltransferase	171.1	1.10E-50

Table S12.1.3. Reciprocal BLASTP analysis: Best hits from BLASTP using *T. urticae* putative *dnmt*s as queries. tetur13g02100 and tetur13g02110 correspond both to parts of a *T.urticae* DNMT-1 that was disrupted by a Gypsy-like transposable element.

T.urticae gene ID	Best hits[a]	E-value
	Homo sapiens	
tetur02g05760	NP_777588.1 N-6 adenine-specific DNA methyltransferase 2 (putative) [Homo sapiens]	1.00E-50
tetur13g02100	NP_001124295.1 DNA (cytosine-5-)-methyltransferase 1 isoform a [Homo sapiens]	1.00E-158
tetur13g02110	NP_001370 DNA (cytosine-5-)-methyltransferase 1 isoform b [Homo sapiens]	3.00E-36
tetur16g01020	NP_072046.2 DNA cytosine methyltransferase 3 alpha isoform a	3.00E-74
	Apis mellifera	
tetur02g05760	XP_396376.2 PREDICTED: similar to CG14439-PA	1.40E-01
tetur13g02100	XP_001122269.1 PREDICTED: similar to DNA (cytosine-5-)-methyltransferase 1	4.00E-145
tetur13g02110	XP_001122269.1 PREDICTED: similar to DNA (cytosine-5-)-methyltransferase 1	3.00E-37
tetur16g01020	XP_394274.3 PREDICTED: similar to DNA cytosine-5 methyltransferase 3 beta isoform 2	3.00E-58
	Nasonia vitripennis	
tetur02g05760	NP_001123319.1 N-6 adenine-specific DNA methyltransferase 2	1.00E-53
tetur13g02100	XP_001605635.1 PREDICTED: similar to DNA methyltransferase	4.00E-146
tetur13g02110	XP_001605635.1 PREDICTED: similar to DNA methyltransferase	1.00E-35
tetur16g01020	XP_001599223.1 PREDICTED: similar to DNA cytosine-5 methyltransferase 3B2 [Nasonia vitripennis]	9.00E-60
	Gallus gallus	
tetur02g05760	NP_001006277.1 N-6 adenine-specific DNA methyltransferase 2	7.00E-47
tetur13g02100	NP_996835.1 DNA (cytosine-5-)-methyltransferase 1 [Gallus gallus]	1.00E-157
tetur13g02110	NP_996835.1 DNA (cytosine-5-)-methyltransferase 1 [Gallus gallus]	6.00E-36
tetur16g01020	NP_001020003.1 DNA cytosine methyltransferase 3 alpha [Gallus gallus]	1.00E-70

[a] We performed reciprocal BLASTP using several species proteomes. This analysis confirmed that the putative *dnmt*s from *T. urticae* genome are reciprocal best hits to other *dnmt*s.

Supplementary Figures

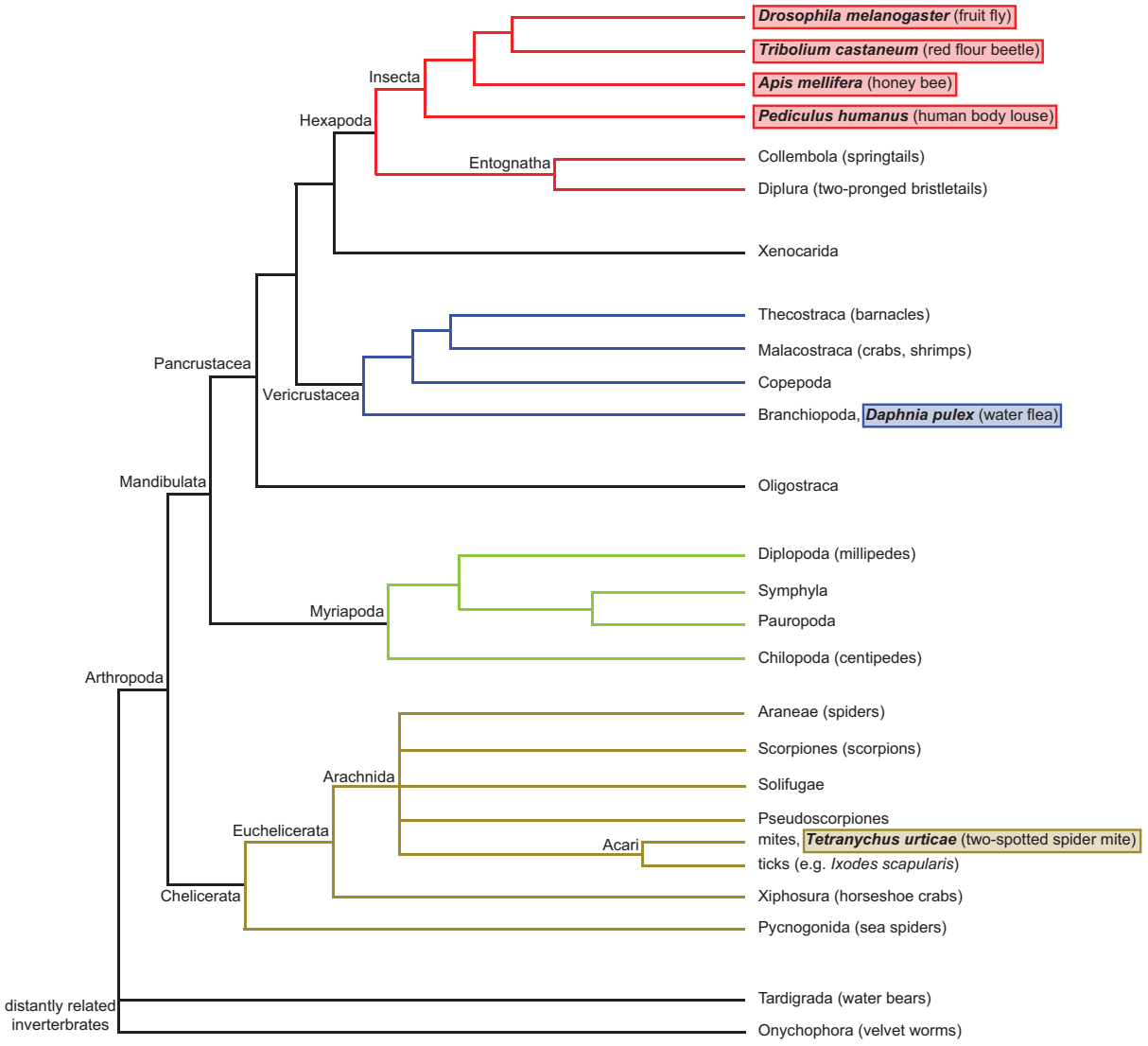


Figure S1.1. Phylogenetic position of the spider mite, *Tetranychus urticae* within the phylum Arthropoda. The tree represents generally accepted arthropod relationships as described in Regier *et al*.[137]. Species and clades mentioned in the text are represented, and fully sequenced genomes are in bold.

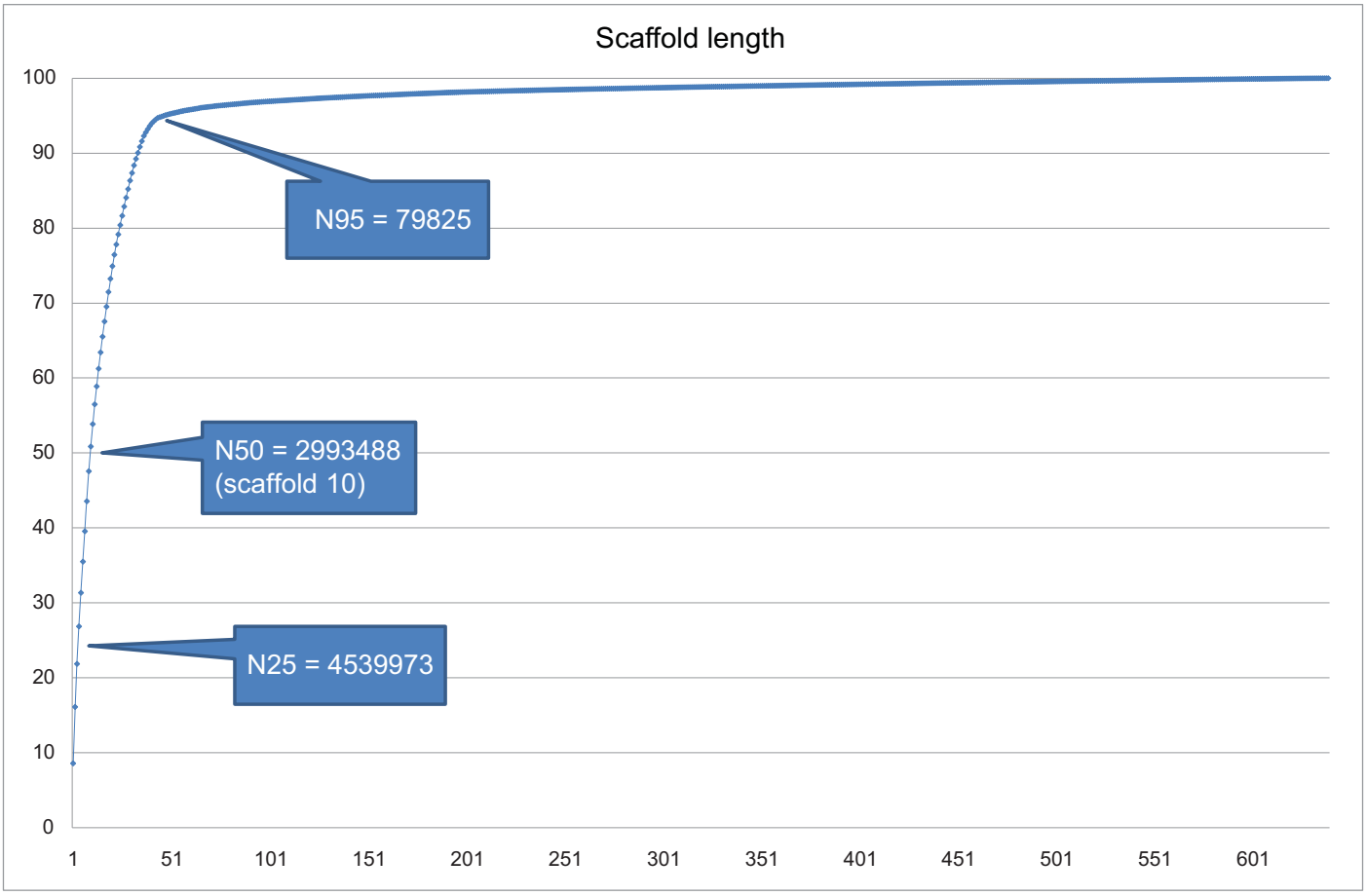


Figure S2.2.1. Scaffold size distribution with N25, N50 and N95.

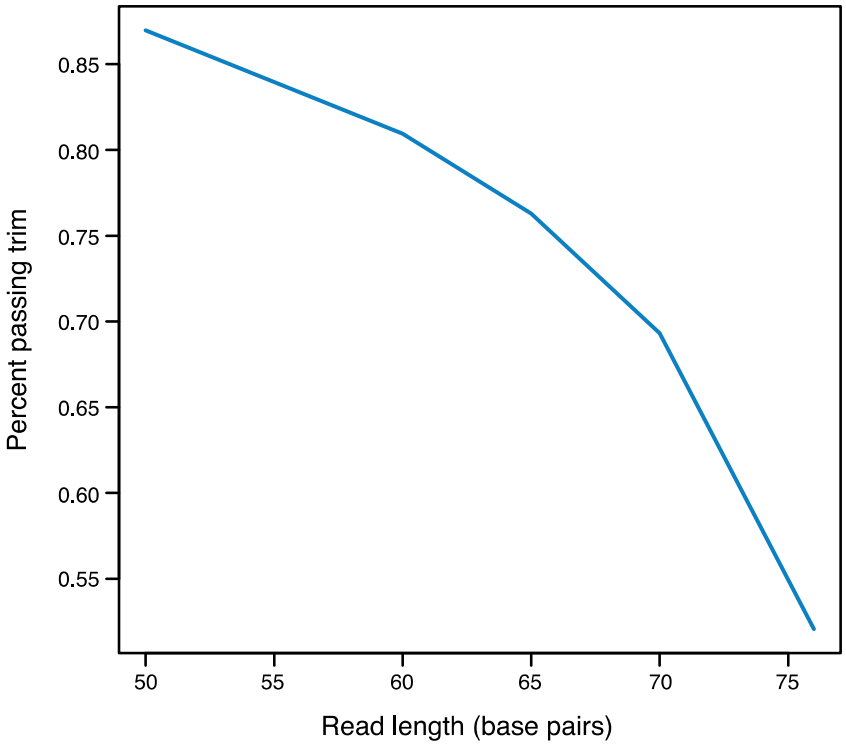


Figure S2.3.1. RNA-seq read length profile after trimming to remove low quality 3' sequences. Reads are from the developmental time course experiment (77 cycles of Illumina sequencing). We selected 60 bases for trimming as an acceptable compromise between maintenance of read length (ability to map uniquely, and across splice sites) and number of reads that passed the trimming filter.

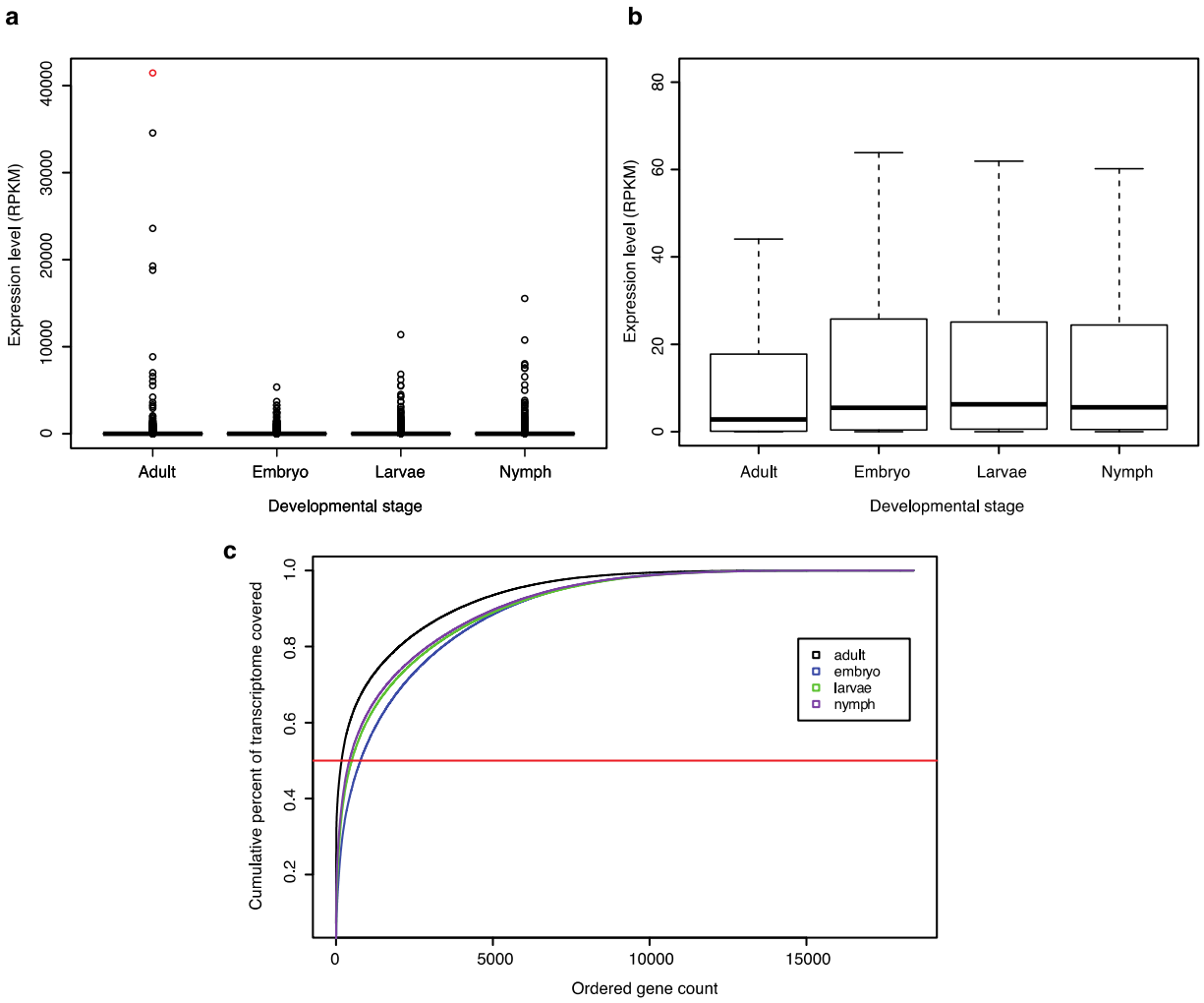


Figure S2.3.2. Distribution of read coverage for spider mite transcripts across stages. A) Boxplot of RPKM distributions for the four developmental stages. Note the extreme outliers for the adult sample. Highlighted in red in the expression of Tetur03g09300 in adults, which accounts for 7% of all mapped reads. B) A closer look (note scale on y-axis) at the boxplot in panel A highlighting the distorting effect of the small number of abundantly expressed genes in adults. The core distribution of RPKM values are noticeably lower compared to the other stages. C). RNA-seq reads tend to follow a power distribution. A few highly expressed genes comprise a large percentage of mapped RNA-seq reads (show is non-quantile normalized data). Note the differences in read distributions between the stages where 50% of all reads are mapped: adult (176 genes), nymph (338 genes), larvae (404 genes), and embryo (924 genes).

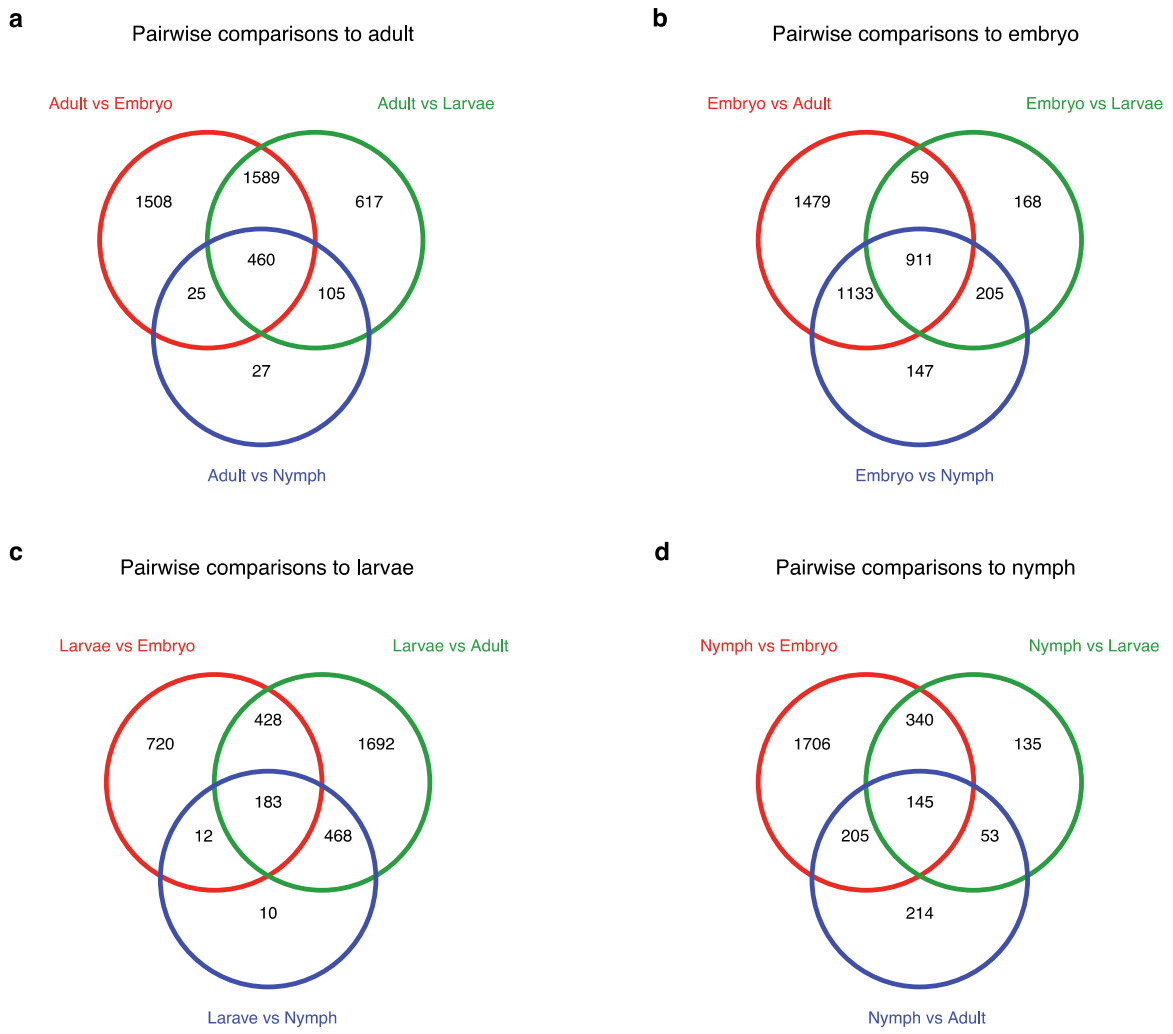


Figure S2.3.3. Venn diagrams showing pairwise comparisons of differentially expressed genes between stages (FDR < 0.05). Differential expression was assessed using the Z test of Kal *et al.*[5] for each of the 6 pairwise comparisons between stages. A) Results for the three comparisons of adult to embryos, larvae or nymphs. The number of shared differentially expressed genes is shown in the overlapping areas. B) Results for the three comparisons of embryo to adults, larvae or nymphs. The number of shared differentially expressed genes is shown in the overlapping areas. C) Results for the three comparisons of larvae to embryos, nymphs or adults. D) Results for the three comparisons of nymphs to embryos, larvae or adults. The number of shared differentially expressed genes is shown in the overlapping areas.

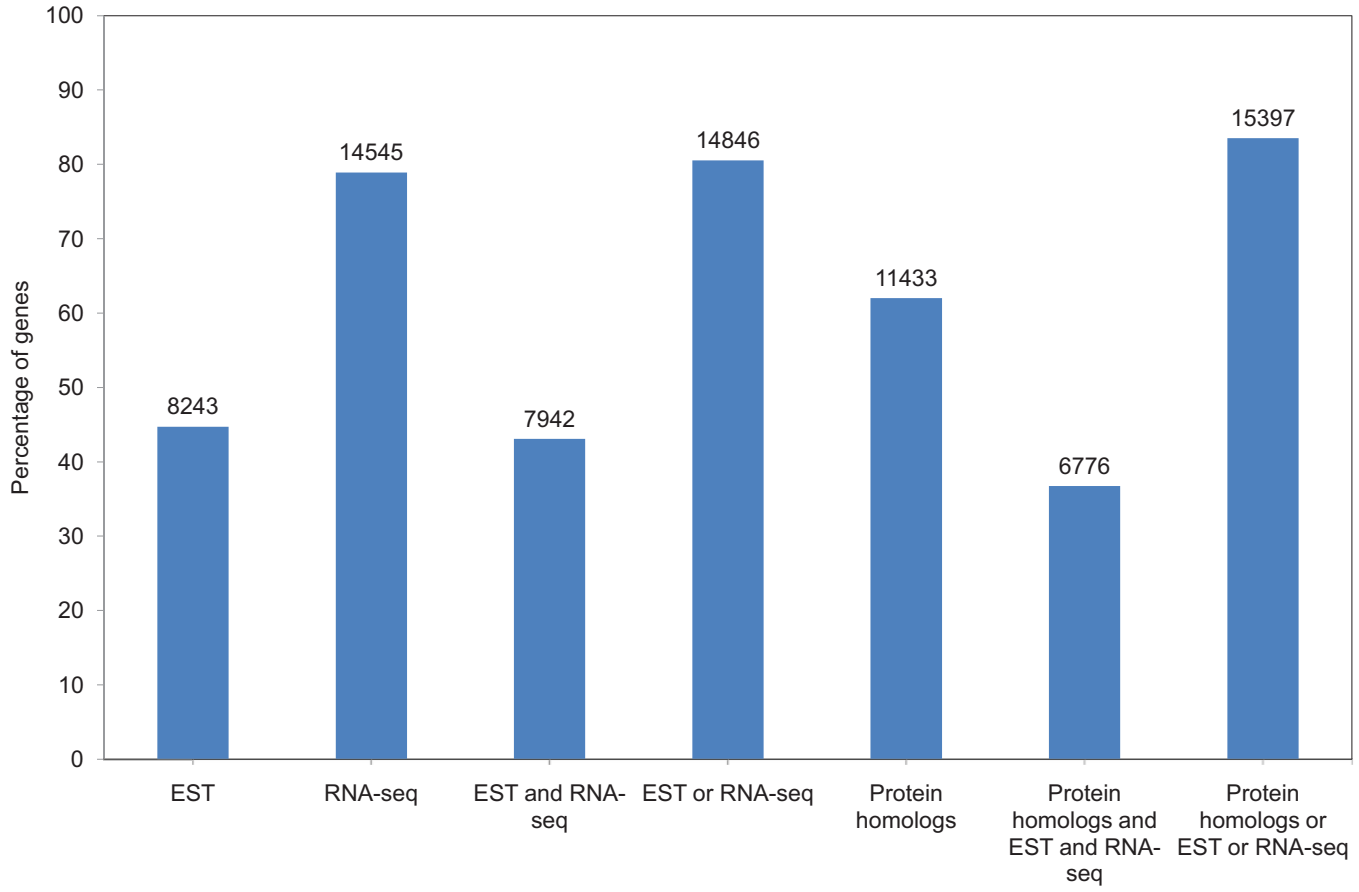


Figure S2.4.1. Predicted *T. urticae* genes supported by protein homologs, ESTs or RNA-seq reads/splice junctions. Protein homologs were determined from BLASTP results of the predicted proteins against a protein database (E-value cutoff e^{-3}). Genes supported by ESTs were identified by GenomeTheader. Genes supported by RNA-seq reads/splice junctions were identified by Bowtie and Tophat. Genes were regarded as supported by RNA-seq reads/splice junctions if having at least three RNA-seq reads/splice junctions aligning to it.

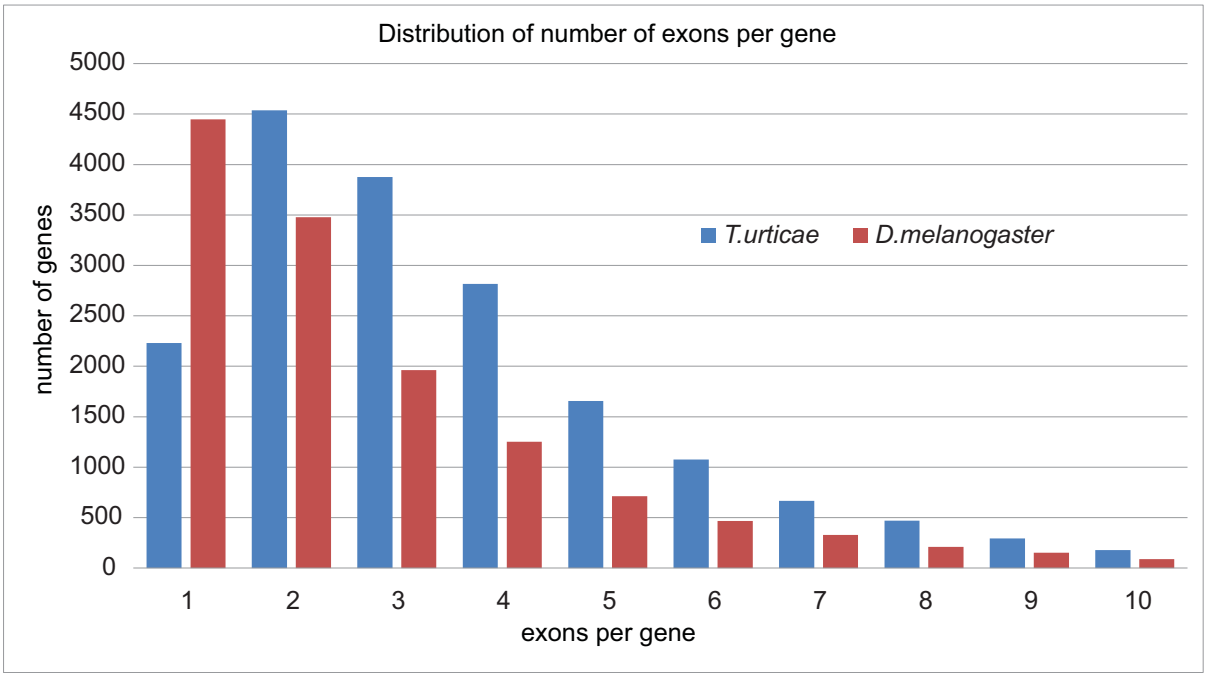


Figure S2.4.2. Distribution of the number of exons per gene (cut-off is 10 exons) for *T. urticae* and *D. melanogaster*. 2966 genes contain no introns. 529 genes contain 11 exons or more.

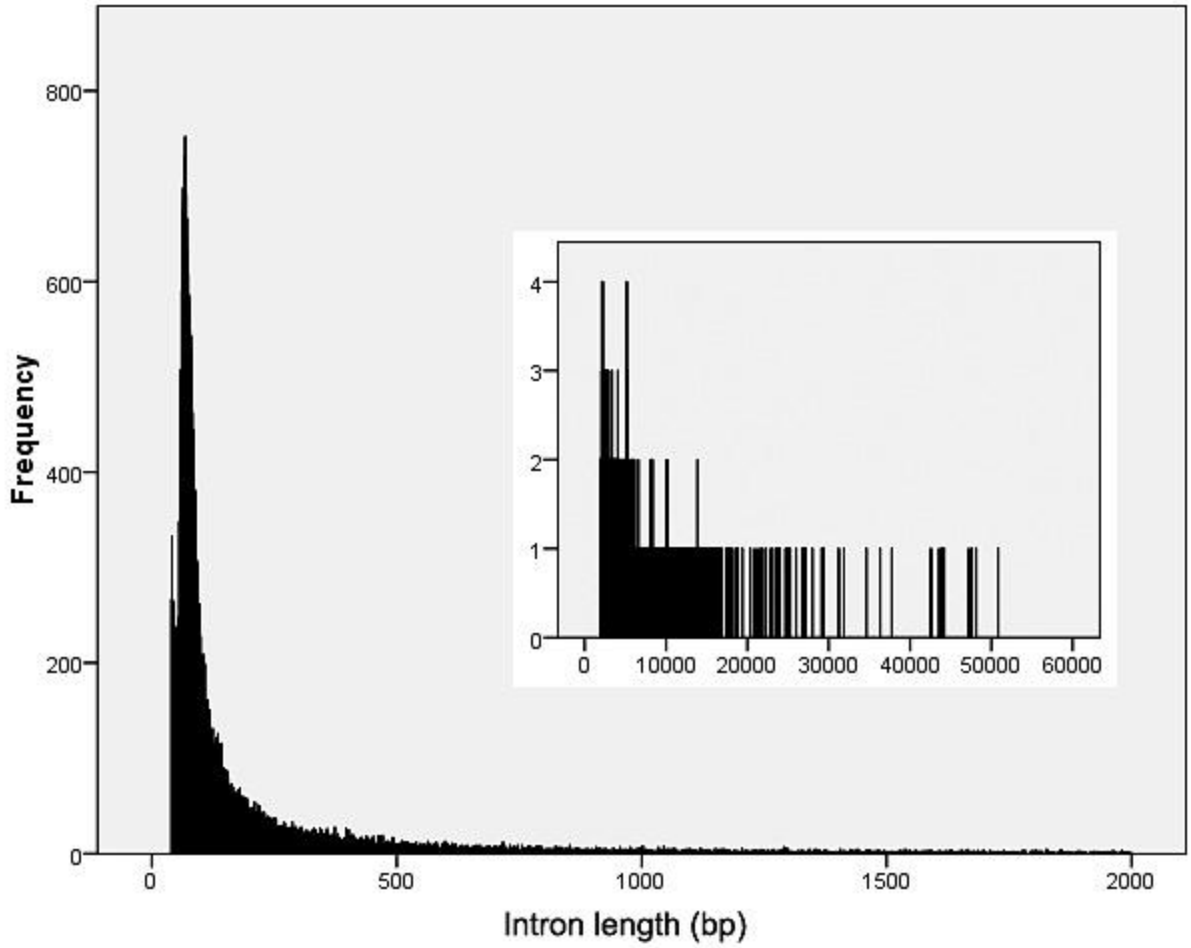


Figure S2.4.3. Intron length distribution for genes of *T. urticae*. The minimum intron size is about 40bp with 267 introns. About 70% of introns have a length between 40bp to 150bp.

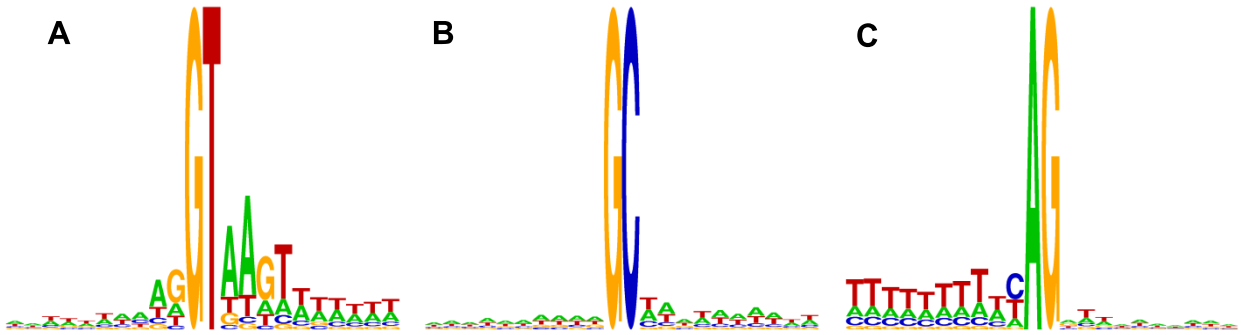


Figure S2.4.4. Sequence logos of donor and acceptor sites. A) Sequence logo of GT donor site (plus 10 bp up- and downstream) created from 43,732 sequences. B) Sequence logo of GC donor sites created from 1,647 sequences. C) Sequence logo of AG acceptor sites created from 45,384 sequences.

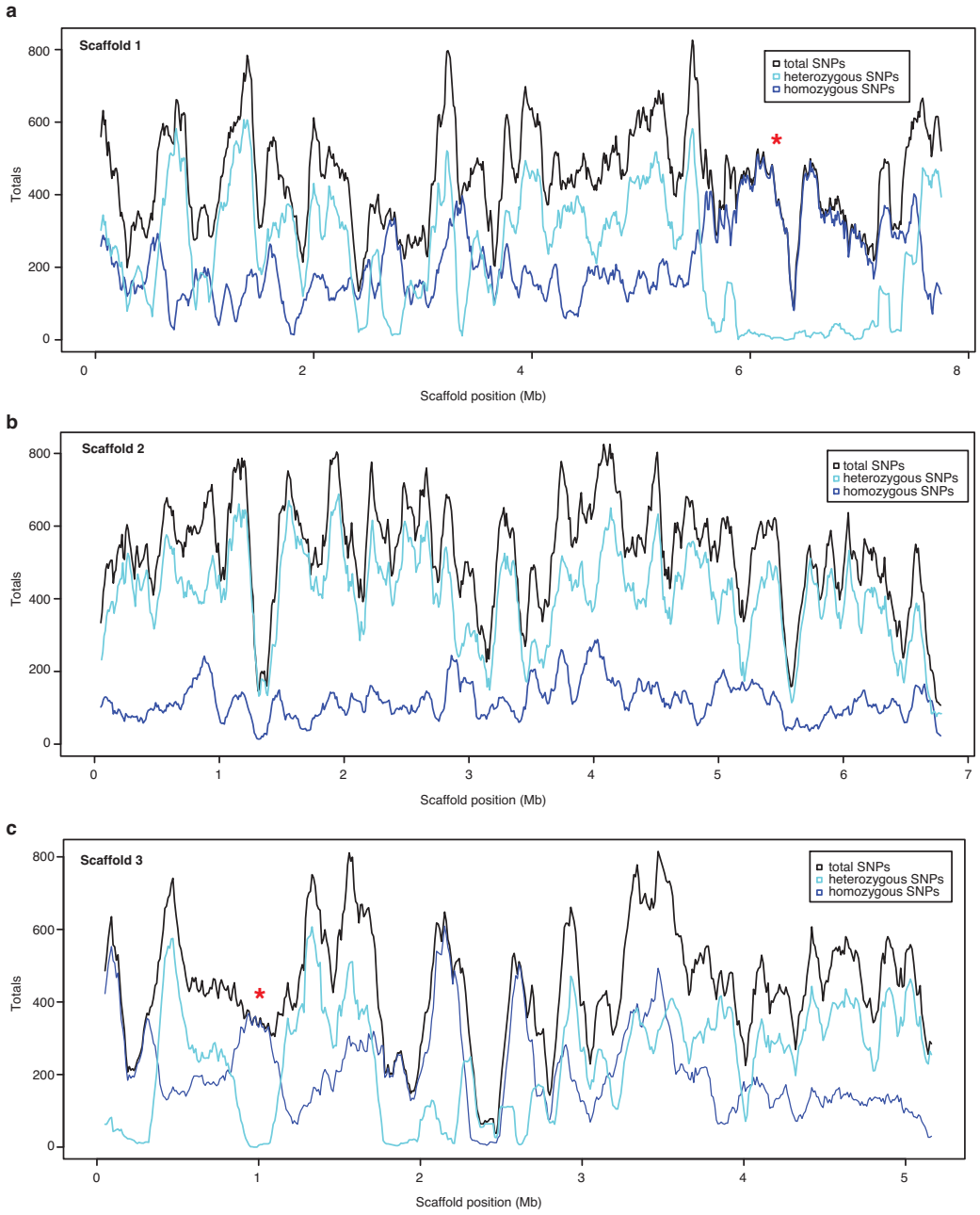


Figure S2.5.1. SNP density for the Montpellier strain across the longest three scaffolds in the mite genome assembly. A) SNP density across scaffold 1 for total SNPs (black), homozygous SNPs (dark blue) and heterozygous SNPs (light blue). B) SNP density across scaffold 2 for total SNPs (black), homozygous SNPs (dark blue) and heterozygous SNPs (light blue). C) SNP density across scaffold 3 for total SNPs (black), homozygous SNPs (dark blue) and heterozygous SNPs (light blue). Areas marked with red asterisk (A & C) indicate genomic regions enriched for homozygous SNPs. In the majority of regions, clear anti-correlation between homozygosity and heterozygosity is apparent, with more heterozygous SNPs). This is generally consistent with partial inbreeding in the Montpellier population. Shown are data from a sliding window analysis with a window size 100 kb and with an offset of 10kb.

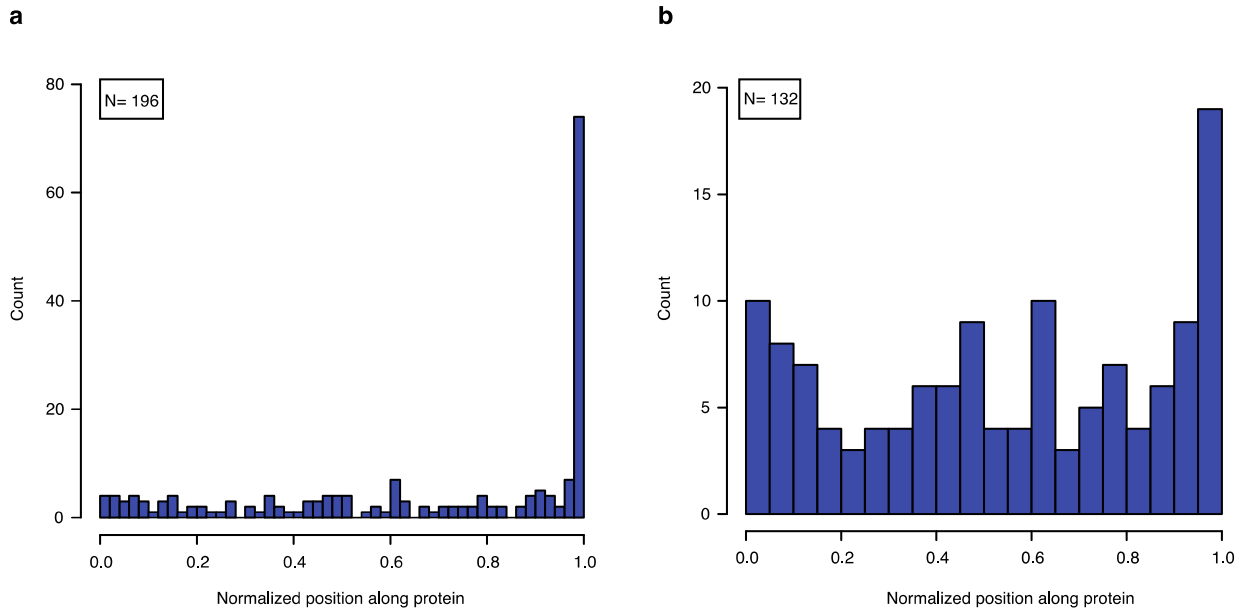
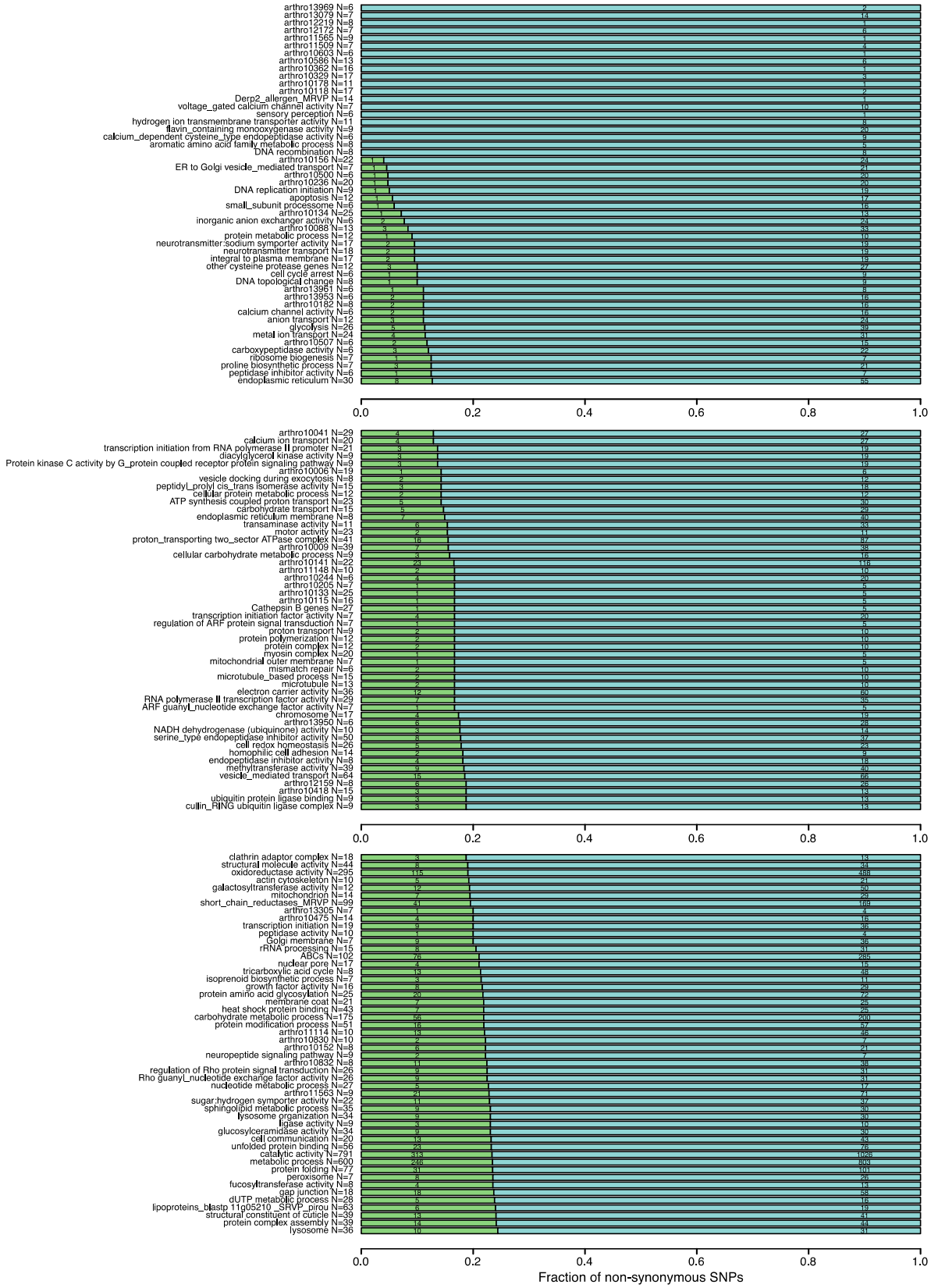
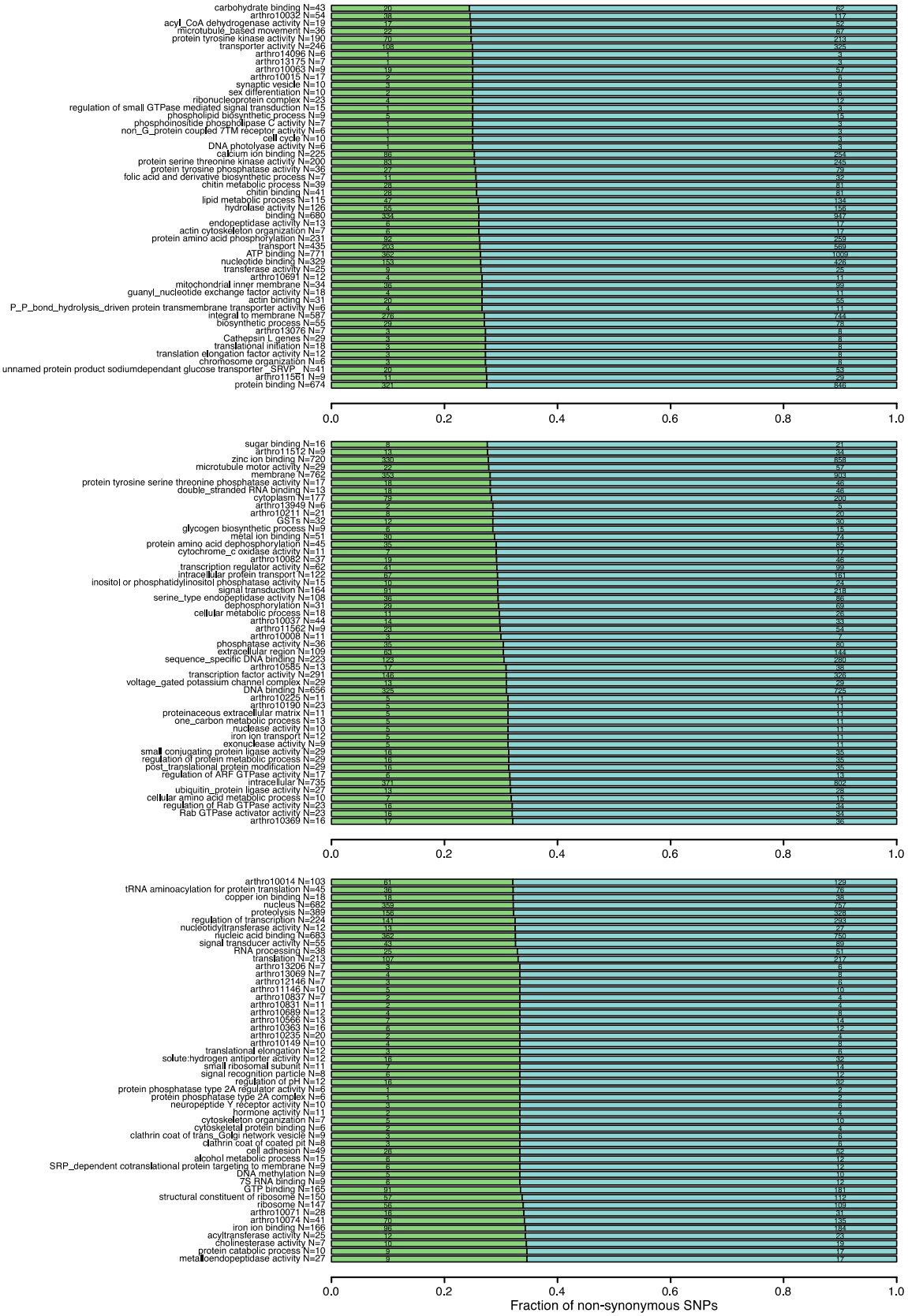
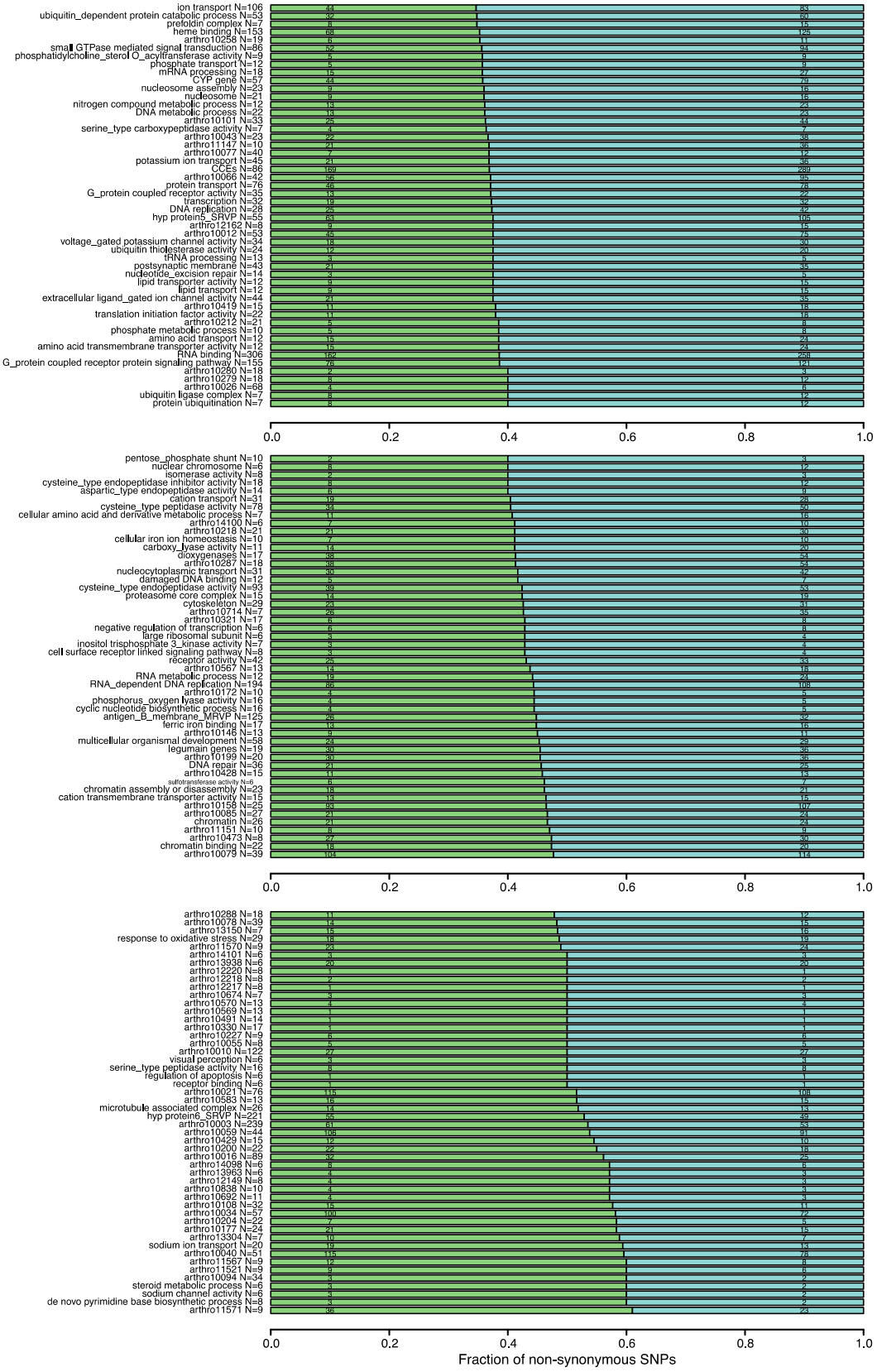


Figure S2.5.2. Density of non-sense mutations within genes in the Montpellier strain of *T. urticae*. A) Positions of non-sense mutations inferred from SNPs predicted by MAQ (*n*=196). The positions of DNA variants generating a stop codon have been normalized by protein length to allow for comparisons across genes. A clear enrichment of synonymous stop site mutations results in a peak at the end of genes. B) Positions of premature stop codons with the removal of synonymous stop site mutations. There is a slight enrichment for premature stops near the end of genes, many of which might be effectively neutral (most of the protein is nonetheless produced).


arthro13969 N=6
arthro13079 N=7
arthro12219 N=8
arthro12172 N=7
arthro11565 N=9
arthro11509 N=7
arthro10603 N=6
arthro10586 N=13
arthro10362 N=16
arthro10329 N=17
arthro10178 N=11
arthro10118 N=17
Derp2_allergen_MRVP N=14
voltage_gated calcium channel activity N=7
sensory perception N=6
hydrogen ion transmembrane transporter activity N=11
flavin_containing monooxygenase activity N=9
calcium_dependent cysteine_type endopeptidase activity N=6
aromatic amino acid family metabolic process N=8
DNA recombination N=8
arthro10156 N=22
ER to Golgi vesicle_mediated transport N=7
arthro10500 N=6
arthro10236 N=20
DNA replication initiation N=9
apoptosis N=12
small_subunit processome N=6
arthro10134 N=25
inorganic anion exchanger activity N=6
arthro10088 N=13
protein metabolic process N=12
neurotransmitter:sodium symporter activity N=17
neurotransmitter transport N=18
integral to plasma membrane N=17
other cysteine protease genes N=12
cell cycle arrest N=6
DNA topological change N=8
arthro13961 N=6
arthro13953 N=6
arthro10182 N=8
calcium channel activity N=6
anion transport N=12
glycolysis N=26
metal ion transport N=24
arthro10507 N=6
carboxypeptidase activity N=6
ribosome biogenesis N=7
proline biosynthetic process N=7
peptidase inhibitor activity N=6
endoplasmic reticulum N=30
0.0
0.2
0.4
0.6
0.8
1.0
arthro10041 N=29
calcium ion transport N=20
transcription initiation from RNA polymerase II promoter N=21
diacylglycerol kinase activity N=9
Protein kinase C activity by G_protein coupled receptor protein signaling pathway N=9
arthro10006 N=19
vesicle docking during exocytosis N=8
peptidyl_prolyl cis_trans isomerase activity N=15
cellular protein metabolic process N=12
ATP synthesis coupled proton transport N=23
carbohydrate transport N=15
endoplasmic reticulum membrane N=8
transaminase activity N=11
motor activity N=23
proton_transporting two_sector ATPase complex N=41
arthro10009 N=39
cellular carbohydrate metabolic process N=9
arthro10141 N=22
arthro11148 N=10
arthro10244 N=6
arthro10205 N=7
arthro10133 N=25
arthro10115 N=16
Cathepsin B genes N=27
transcription initiation factor activity N=7
regulation of ARF protein signal transduction N=7
proton transport N=9
protein polymerization N=12
protein complex N=12
myosin complex N=20
mitochondrial outer membrane N=7
mismatch repair N=6
microtubule_based process N=15
microtubule N=13
electron carrier activity N=36
RNA polymerase II transcription factor activity N=29
ARF guanyl_nucleotide exchange factor activity N=7
chromosome N=17
arthro13950 N=6
NADH dehydrogenase (ubiquinone) activity N=10
serine_type endopeptidase inhibitor activity N=50
cell redox homeostasis N=26
homophilic cell adhesion N=14
endopeptidase inhibitor activity N=8
methyltransferase activity N=39
vesicle_mediated transport N=64
arthro12159 N=8
arthro10418 N=15
ubiquitin protein ligase binding N=9
cullin_RING ubiquitin ligase complex N=9
0.0
0.2
0.4
0.6
0.8
1.0
clathrin adaptor complex N=18
structural molecule activity N=44
oxidoreductase activity N=295
actin cytoskeleton N=10
galactosyltransferase activity N=12
mitochondrion N=14
short_chain_reductases_MRVP N=99
arthro13305 N=7
arthro10475 N=14
transcription initiation N=19
peptidase activity N=10
Golgi membrane N=7
rRNA processing N=15
ABCs N=102
nuclear pore N=17
tricarboxylic acid cycle N=8
isoprenoid biosynthetic process N=7
growth factor activity N=16
protein amino acid glycosylation N=25
membrane coat N=21
heat shock protein binding N=43
carbohydrate metabolic process N=175
protein modification process N=51
arthro11114 N=10
arthro10830 N=10
arthro10152 N=8
neuropeptide signaling pathway N=9
arthro10832 N=8
regulation of Rho protein signal transduction N=26
Rho guanyl_nucleotide exchange factor activity N=26
nucleotide metabolic process N=27
arthro11563 N=9
sugar:hydrogen symporter activity N=22
sphingolipid metabolic process N=35
lysosome organization N=34
ligase activity N=9
glucosylceramidase activity N=34
cell communication N=20
unfolded protein binding N=56
catalytic activity N=791
metabolic process N=600
protein folding N=77
peroxisome N=7
fucosyltransferase activity N=8
gap junction N=18
dUTP metabolic process N=28
lipoproteins_blastp 11g05210 _SRVP_pirou N=63
structural constituent of cuticle N=39
protein complex assembly N=39
lysosome N=36
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of non-synonymous SNPs


carbohydrate binding N=43
arthro10032 N=54
acyl_CoA dehydrogenase activity N=19
microtubule_based movement N=36
protein tyrosine kinase activity N=190
transporter activity N=246
arthro14096 N=6
arthro13175 N=7
arthro10063 N=9
arthro10015 N=17
synaptic vesicle N=10
sex differentiation N=10
ribonucleoprotein complex N=23
regulation of small GTPase mediated signal transduction N=15
phospholipid biosynthetic process N=9
phosphoinositide phospholipase C activity N=7
non_G_protein coupled 7TM receptor activity N=6
cell cycle N=10
DNA photolyase activity N=6
calcium ion binding N=225
protein serine threonine kinase activity N=200
protein tyrosine phosphatase activity N=36
folic acid and derivative biosynthetic process N=7
chitin metabolic process N=39
chitin binding N=41
lipid metabolic process N=115
hydrolase activity N=126
binding N=680
endopeptidase activity N=13
actin cytoskeleton organization N=7
protein amino acid phosphorylation N=231
transport N=435
ATP binding N=771
nucleotide binding N=329
transferase activity N=25
arthro10691 N=12
mitochondrial inner membrane N=34
guanyl_nucleotide exchange factor activity N=18
actin binding N=31
P_P_bond_hydrolysis_driven protein transmembrane transporter activity N=6
integral to membrane N=587
biosynthetic process N=55
arthro13076 N=7
Cathepsin L genes N=29
translational initiation N=18
translation elongation factor activity N=12
chromosome organization N=6
unnamed protein product sodiumdependant glucose transporter _SRVP_ N=41
arthro11561 N=9
protein binding N=674
20 62
38 117
17 52
22 67
70 213
108 329
1 3
1 3
3 9
2 6
4 12
1 3
5 15
1 3
1 3
1 3
1 3
86 254
83 245
27 79
11 32
28 81
28 81
47 134
55 156
334 947
6 17
6 17
92 259
203 569
360 1009
153 426
9 25
4 11
36 99
4 11
20 55
4
276 744
29 78
3 8
3 8
3 8
3 8
3 8
20 53
11 29
321 846
0.0
0.2
0.4
0.6
0.8
1.0
sugar binding N=16
arthro11512 N=9
zinc ion binding N=720
microtubule motor activity N=29
membrane N=762
protein tyrosine serine threonine phosphatase activity N=17
double_stranded RNA binding N=13
cytoplasm N=177
arthro13949 N=6
arthro10211 N=21
GSTs N=32
glycogen biosynthetic process N=9
metal ion binding N=51
protein amino acid dephosphorylation N=45
cytochrome_c oxidase activity N=11
arthro10082 N=37
transcription regulator activity N=62
intracellular protein transport N=122
inositol or phosphatidylinositol phosphatase activity N=15
signal transduction N=164
serine_type endopeptidase activity N=108
dephosphorylation N=31
cellular metabolic process N=18
arthro10037 N=44
arthro11562 N=9
arthro10008 N=11
phosphatase activity N=36
extracellular region N=109
sequence_specific DNA binding N=223
arthro10585 N=13
transcription factor activity N=291
voltage_gated potassium channel complex N=29
DNA binding N=656
arthro10225 N=11
arthro10190 N=23
proteinaceous extracellular matrix N=11
one_carbon metabolic process N=13
nuclease activity N=10
iron ion transport N=12
exonuclease activity N=9
small conjugating protein ligase activity N=29
regulation of protein metabolic process N=29
post_translational protein modification N=29
regulation of ARF GTPase activity N=17
intracellular N=735
ubiquitin_protein ligase activity N=27
cellular amino acid metabolic process N=10
regulation of Rab GTPase activity N=23
Rab GTPase activator activity N=23
arthro10369 N=16
8 21
13 34
330 858
22 57
353 903
18 46
18 46
79 200
2 5
8 20
12 30
6 15
30 74
35 85
7 17
19 46
41 99
67 161
10 24
91 218
36 86
29 69
11 25
14 33
23 54
3 7
35 80
63 144
123 280
17 38
146 328
13 29
325 725
5 11
5 11
5 11
5 11
5 11
5 11
5 11
5 11
16 35
16 35
16 35
6 13
371 802
13 28
7 15
16 34
16 34
17 36
0.0
0.2
0.4
0.6
0.8
1.0
arthro10014 N=103
tRNA aminoacylation for protein translation N=45
copper ion binding N=18
nucleus N=682
proteolysis N=389
regulation of transcription N=224
nucleotidyltransferase activity N=12
nucleic acid binding N=683
signal transducer activity N=55
RNA processing N=38
translation N=213
arthro13206 N=7
arthro13069 N=7
arthro12146 N=7
arthro11146 N=10
arthro10837 N=7
arthro10831 N=11
arthro10689 N=12
arthro10566 N=13
arthro10363 N=16
arthro10235 N=20
arthro10149 N=10
translational elongation N=12
solute:hydrogen antiporter activity N=12
small ribosomal subunit N=11
signal recognition particle N=8
regulation of pH N=12
protein phosphatase type 2A regulator activity N=6
protein phosphatase type 2A complex N=6
neuropeptide Y receptor activity N=10
hormone activity N=11
cytoskeleton organization N=7
cytoskeletal protein binding N=6
clathrin coat of trans_Golgi network vesicle N=9
clathrin coat of coated pit N=8
cell adhesion N=49
alcohol metabolic process N=15
SRP_dependent cotranslational protein targeting to membrane N=9
DNA methylation N=9
7S RNA binding N=9
GTP binding N=165
structural constituent of ribosome N=150
ribosome N=147
arthro10071 N=28
arthro10074 N=41
iron ion binding N=169
acyltransferase activity N=25
cholinesterase activity N=7
protein catabolic process N=10
metalloendopeptidase activity N=27
51 129
36 76
18 38
359 757
156 328
141 293
13 27
382 750
43 89
25 51
107 217
3 6
4 8
3 6
5 10
2 4
2 4
4 8
7 14
6 12
2 4
4 8
4 8
6 12
16 32
7 14
6 12
16 32
1 2
1 2
3 6
2 4
5 10
2 4
3 6
3 6
26 52
6 12
6 12
5 10
6 12
91 181
57 112
56 109
16 31
70 135
86 162
12 23
10 19
9 17
9 17
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of non-synonymous SNPs


ion transport N=106
ubiquitin_dependent protein catabolic process N=53
prefoldin complex N=7
heme binding N=153
arthro10258 N=19
small GTPase mediated signal transduction N=86
phosphatidylcholine_sterol O_acyltransferase activity N=9
phosphate transport N=12
mRNA processing N=18
CYP gene N=57
nucleosome assembly N=23
nucleosome N=21
nitrogen compound metabolic process N=12
DNA metabolic process N=22
arthro10101 N=33
serine_type carboxypeptidase activity N=7
arthro10043 N=23
arthro11147 N=10
arthro10077 N=40
potassium ion transport N=45
CCEs N=86
arthro10066 N=42
protein transport N=76
G_protein coupled receptor activity N=35
transcription N=32
DNA replication N=28
hyp protein5_SRVP N=55
arthro12162 N=8
arthro10012 N=53
voltage_gated potassium channel activity N=34
ubiquitin thiolesterase activity N=24
tRNA processing N=13
postsynaptic membrane N=43
nucleotide_excision repair N=14
lipid transporter activity N=12
lipid transport N=12
extracellular ligand_gated ion channel activity N=44
arthro10419 N=15
translation initiation factor activity N=22
arthro10212 N=21
phosphate metabolic process N=10
amino acid transport N=12
amino acid transmembrane transporter activity N=12
RNA binding N=306
G_protein coupled receptor protein signaling pathway N=155
arthro10280 N=18
arthro10279 N=18
arthro10026 N=68
ubiquitin ligase complex N=7
protein ubiquitination N=7
0.0
0.2
0.4
0.6
0.8
1.0
pentose_phosphate shunt N=10
nuclear chromosome N=6
isomerase activity N=8
cysteine_type endopeptidase inhibitor activity N=18
aspartic_type endopeptidase activity N=14
cation transport N=31
cysteine_type peptidase activity N=78
cellular amino acid and derivative metabolic process N=7
arthro14100 N=6
arthro10218 N=21
cellular iron ion homeostasis N=10
carboxy_lyase activity N=11
dioxygenases N=17
arthro10267 N=18
nucleocytoplasmic transport N=31
damaged DNA binding N=12
cysteine_type endopeptidase activity N=93
proteasome core complex N=15
cytoskeleton N=29
arthro10714 N=7
arthro10321 N=17
negative regulation of transcription N=6
large ribosomal subunit N=6
inositol trisphosphate 3_kinase activity N=7
cell surface receptor linked signaling pathway N=8
receptor activity N=42
arthro10567 N=13
RNA metabolic process N=12
RNA_dependent DNA replication N=194
arthro10172 N=10
phosphorus_oxygen lyase activity N=16
cyclic nucleotide biosynthetic process N=16
antigen_B_membrane_MRVP N=125
ferric iron binding N=17
arthro10148 N=13
multicellular organismal development N=58
legumain genes N=19
arthro10199 N=20
DNA repair N=36
arthro10428 N=15
sulfotransferase activity N=6
chromatin assembly or disassembly N=23
cation transmembrane transporter activity N=15
arthro10158 N=25
arthro10085 N=27
chromatin N=26
arthro11151 N=10
arthro10473 N=8
chromatin binding N=22
arthro10079 N=39
0.0
0.2
0.4
0.6
0.8
1.0
arthro10288 N=18
arthro10078 N=39
arthro13150 N=7
response to oxidative stress N=29
arthro11570 N=9
arthro14101 N=6
arthro13938 N=6
arthro12220 N=8
arthro12218 N=8
arthro12217 N=8
arthro10674 N=7
arthro10570 N=13
arthro10569 N=13
arthro10491 N=14
arthro10330 N=17
arthro10227 N=9
arthro10055 N=8
arthro10010 N=122
visual perception N=6
serine_type peptidase activity N=16
regulation of apoptosis N=6
receptor binding N=6
arthro10021 N=76
arthro10583 N=13
microtubule associated complex N=26
hyp protein6_SRVP N=221
arthro10003 N=239
arthro10059 N=44
arthro10429 N=15
arthro10200 N=22
arthro10016 N=89
arthro14098 N=6
arthro13963 N=6
arthro12149 N=8
arthro10838 N=10
arthro10692 N=11
arthro10108 N=32
arthro10034 N=57
arthro10204 N=22
arthro10177 N=24
arthro13304 N=7
sodium ion transport N=20
arthro10040 N=51
arthro11567 N=9
arthro11521 N=9
arthro10094 N=34
steroid metabolic process N=6
sodium channel activity N=6
de novo pyrimidine base biosynthetic process N=8
arthro11571 N=9
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of non-synonymous SNPs

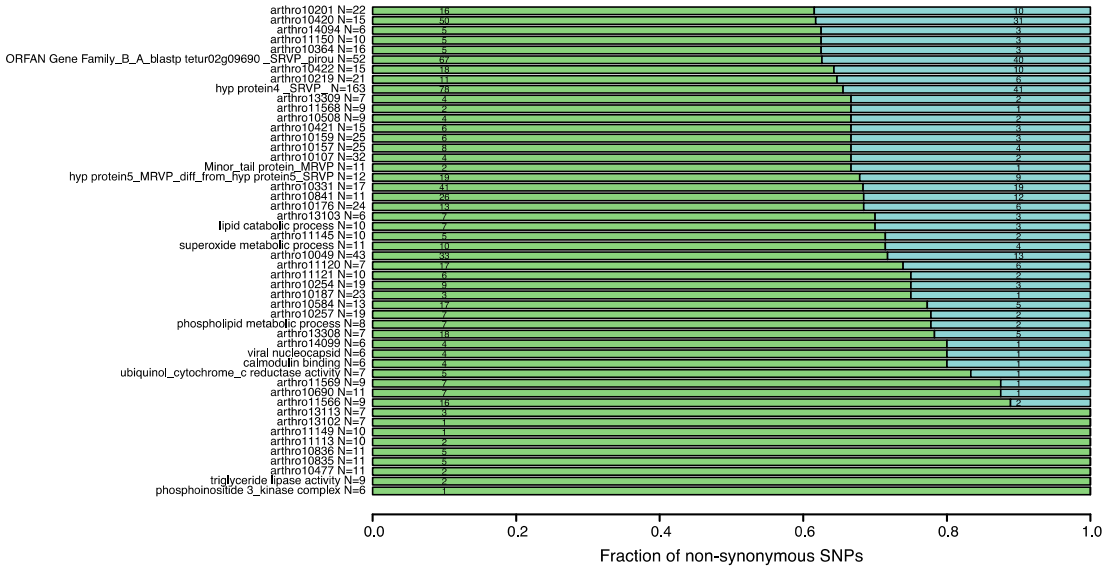


Figure S2.5.3. Numbers and fraction of non-synonymous and synonymous changes in protein families and GO annotated groups. Gene families were either grouped through manual annotation, by GO category or by orthoMCL clustering (IDs with arthro heading; supplementary methods S5.1). Gene families with a minimum of 5 members were included in the analysis. Gene families are sorted on the fraction of non-synonymous SNPs (green) while synonymous SNPs are marked in blue. Totals for non-synonymous and synonymous SNPs as summed across all members of a gene family/GO category are indicated within the horizontal bars in the plot. The family name/GO category and number of representative genes are as indicated at the far left.

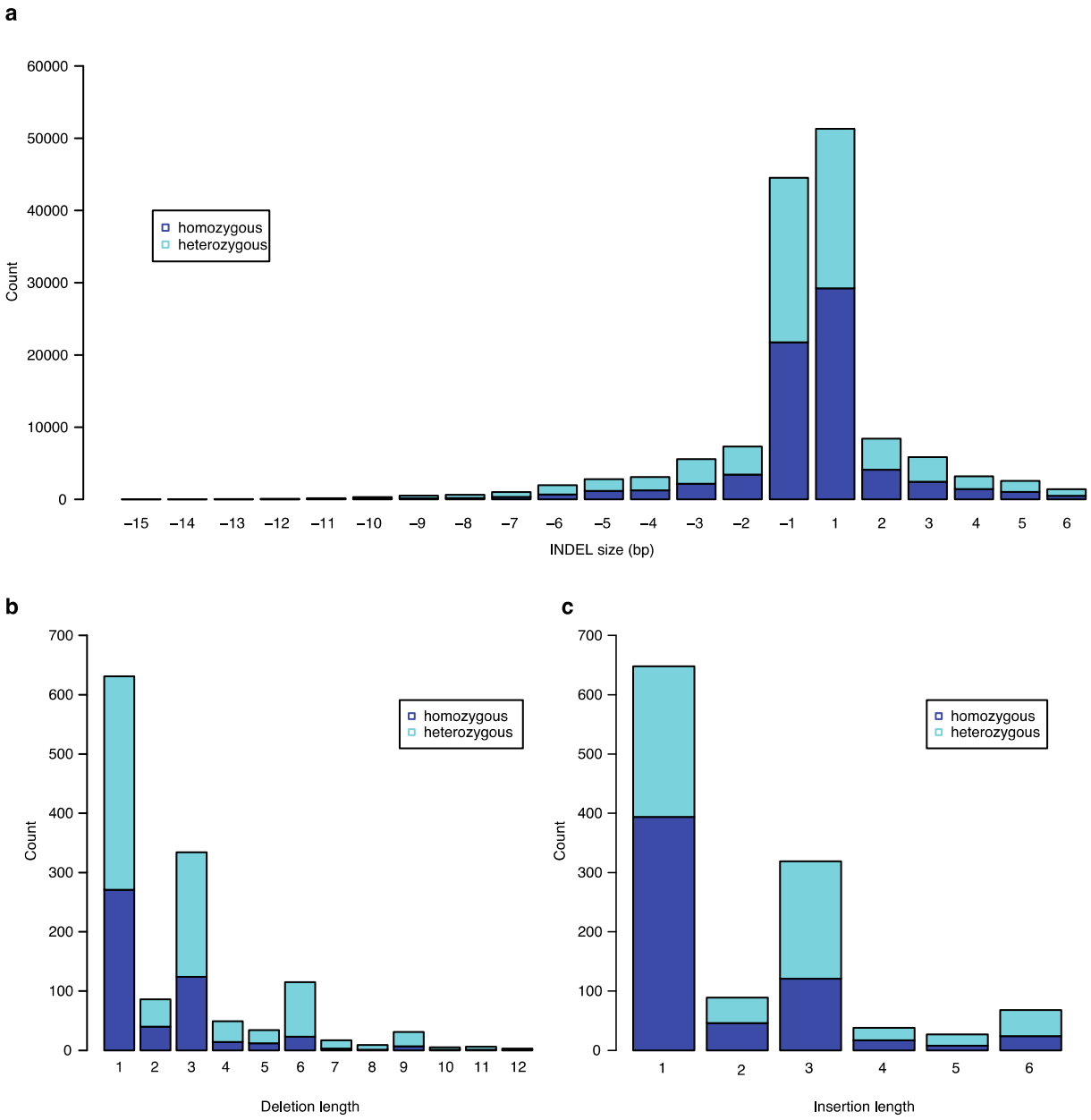


Figure S2.5.4. Characterization of small indels in Montpellier relative to the London reference strain. A) Shown is the size distribution of insertions and deletions predicted by MAQ[16] for both homozygous (blue) and heterozygous (cyan) indels. Due to the algorithmic considerations, MAQ is able to predict longer deletions than insertions, accounting (presumably in large part) for the different profiles between deletions and insertions. B) Frequency in coding sequences for deletions of various lengths. Apart from 1 bp deletions (the most common type of deletion mutation observed across animal and plant genomes) note the enrichment for tri-nucleotide deletions that maintain reading frames (that is, 3, 6 and 9 bp). C) Frequency in the coding sequence of insertions of various lengths. Enrichment for tri-nucleotide insertion sizes is also readily apparent.

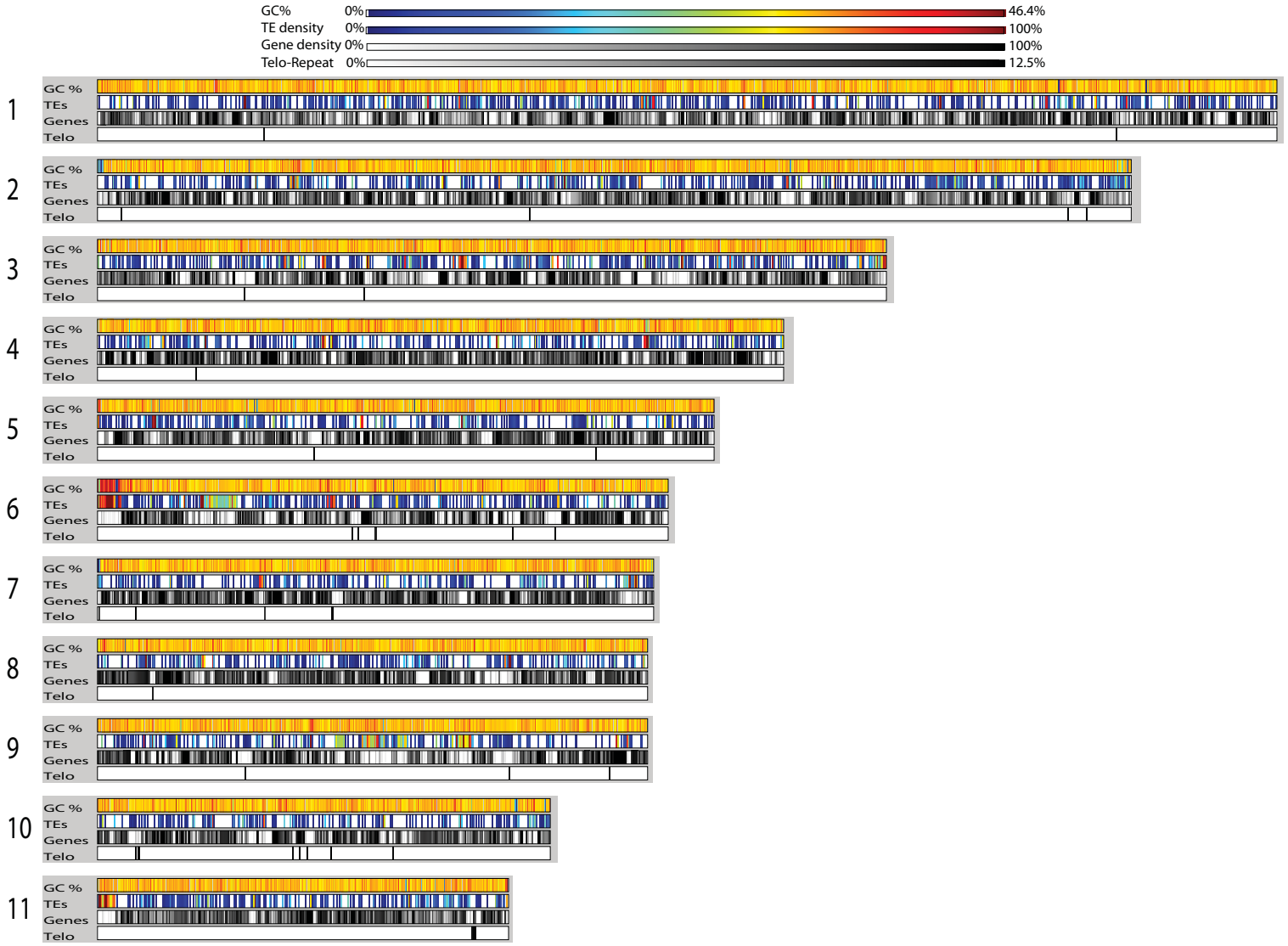


Figure S3.2.1. Distribution of %GC, TE density, gene density and the telomeric-repeats along the 11 longest scaffolds in windows of 4kb. The 11 longest scaffolds make up over 50% of the *T. urticae* genome.

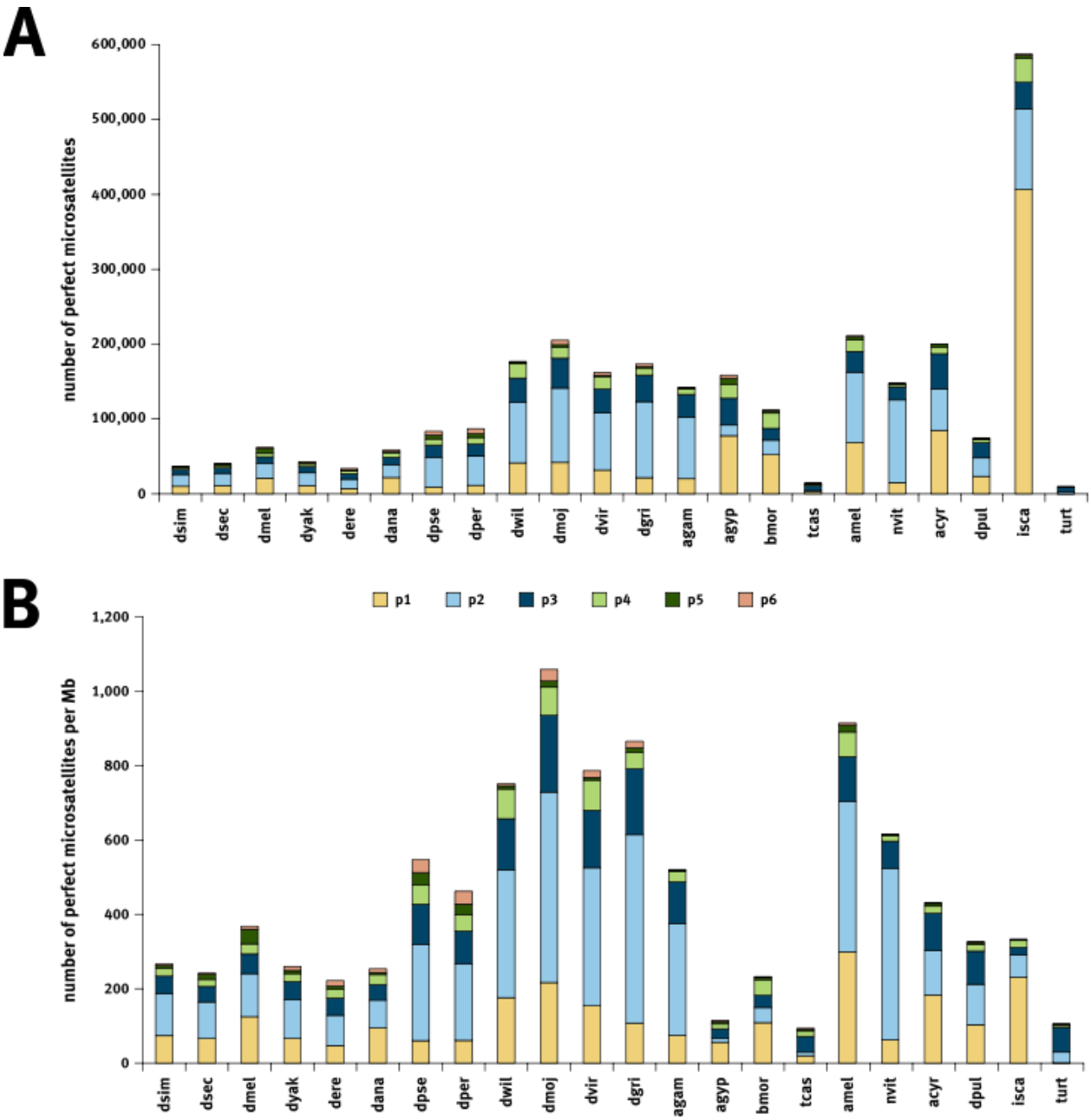


Figure S3.3.1. Comparison of full-genome microsatellite content between spidermite (turt) and other arthropod species. Different classes of microsatellites are distinguished based on the length of the repeat pattern (p1 [mono-nucleotide repeat] to p6 [hexa-nucleotide repeat]). A) Total abundance of microsatellites per genome. B) Microsatellite density, i.e. number of microsatellites per Megabase to adjust for different genome sizes.

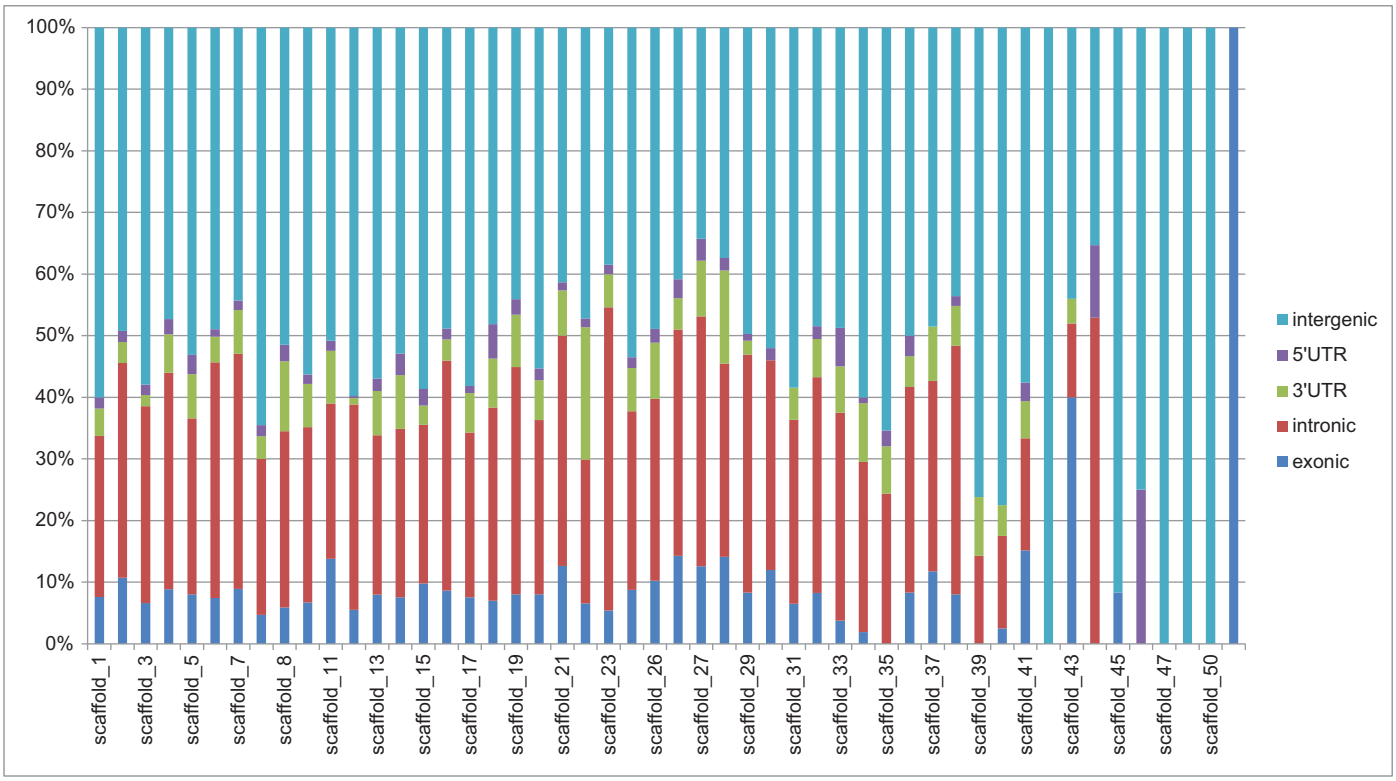


Figure S3.3.2. Microsatellite distribution relative to the annotation. Values are the number of occurrence of microsatellites in intergenic, 5'UTR, 3'UTR, intronic or exonic regions. Data shown for the 50 largest scaffolds.

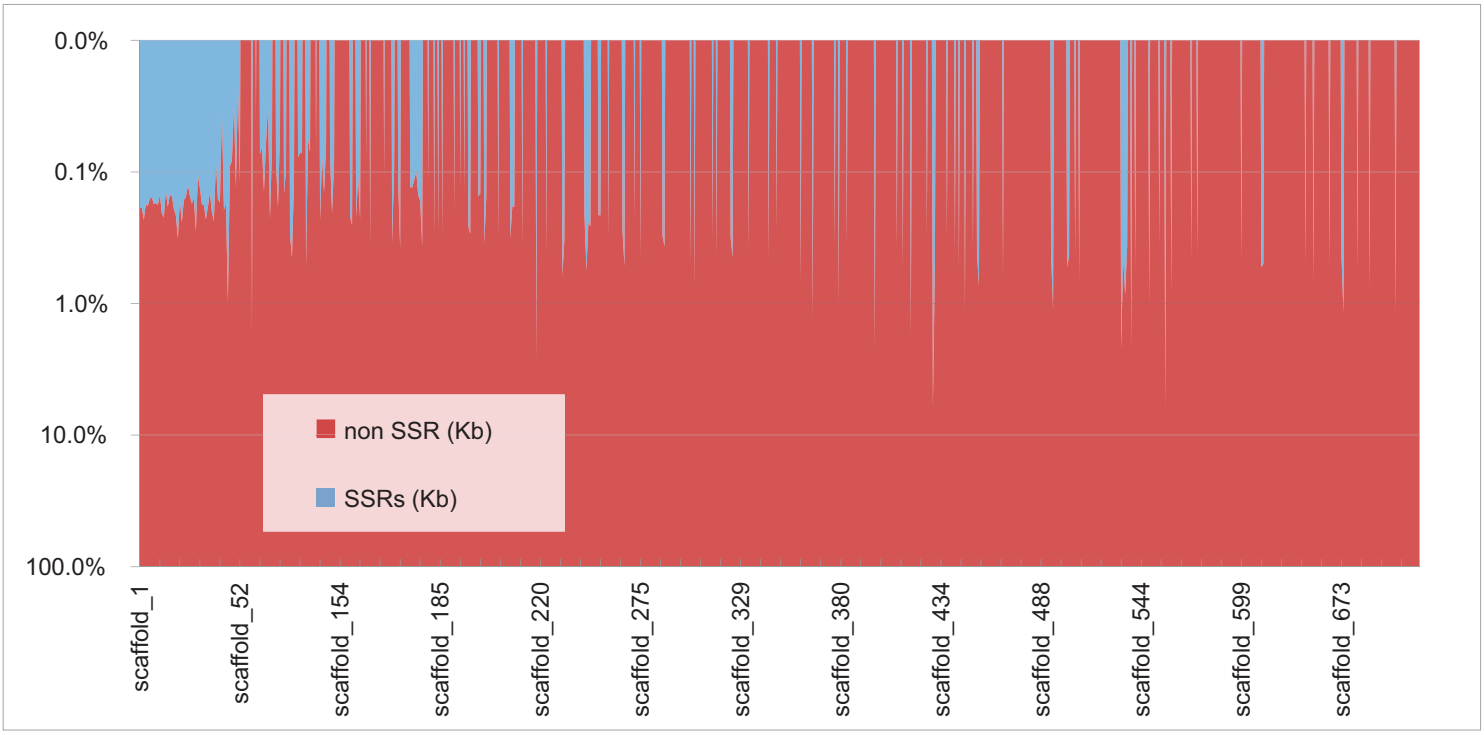


Figure S3.3.3. Microsatellite proportion per scaffold. Microsatellites on average cover less than 0.43% of the scaffolds. Data shown for all 640 scaffolds with a logarithmic (10) scale.

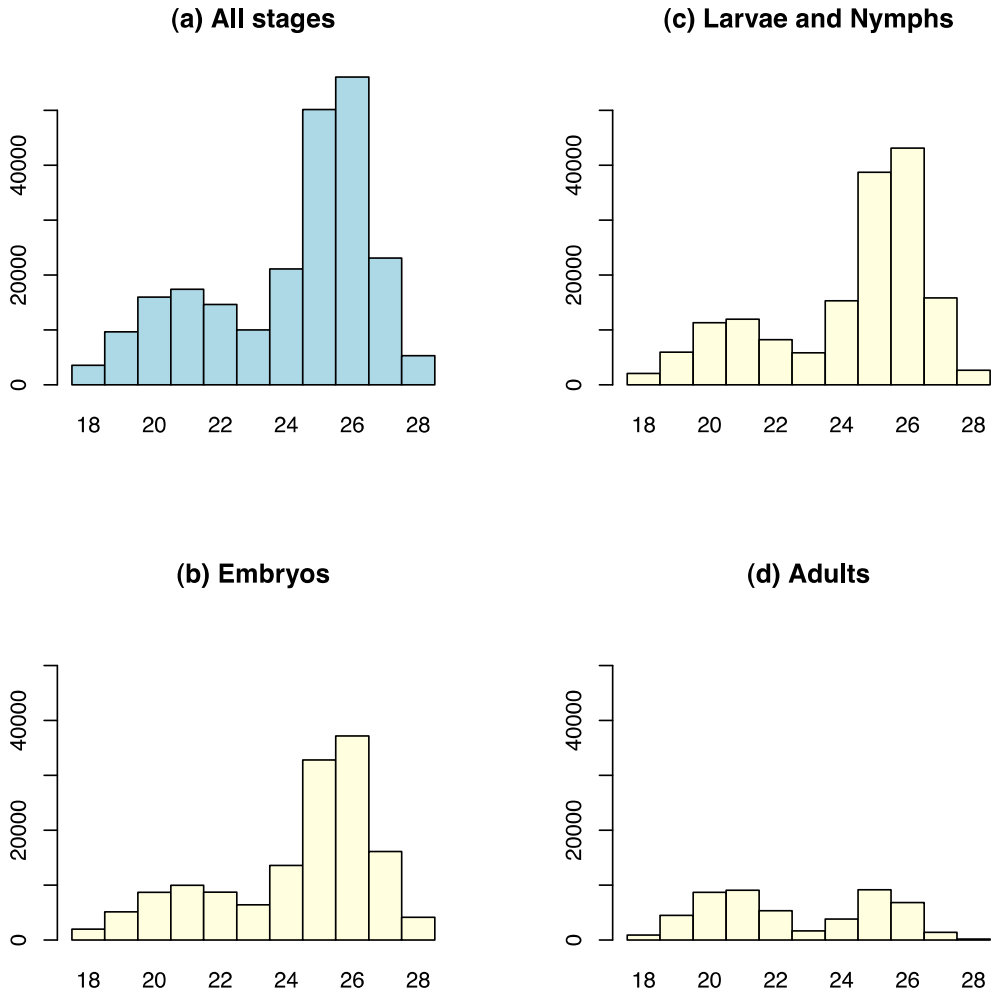


Figure S4.1.1. Number of unique small RNA sequences (y-axis) per size category (x-axis). Shown are analyses for all the developmental stages (a), embryos (b), larvae and nymphs (c) and adults (d).

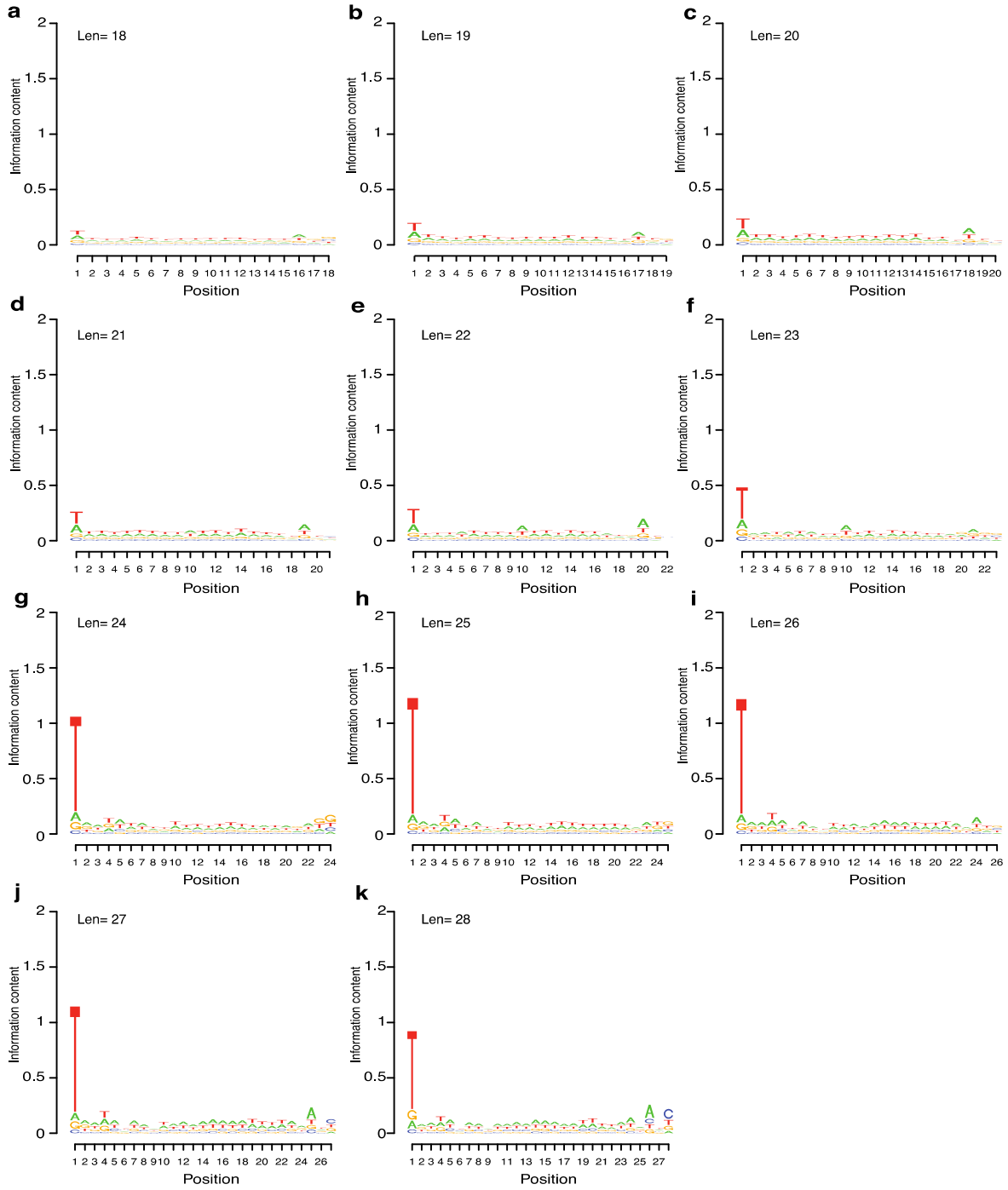


Figure S4.1.2. Sequence logos of small RNAs of sizes 18-28 bases in length. Sequence logo representations of small RNA base composition biases in all small RNAs of 18 (a), 19 (b), 20 (c), 21 (d), 22 (e), 23 (f), 24 (g), 25 (h), 26 (i), 27 (j) and 28 (k) nucleotides in length. Note the strong bias for thymine (uracil in the RNA) residues at position 1 of the larger classes of small RNAs.

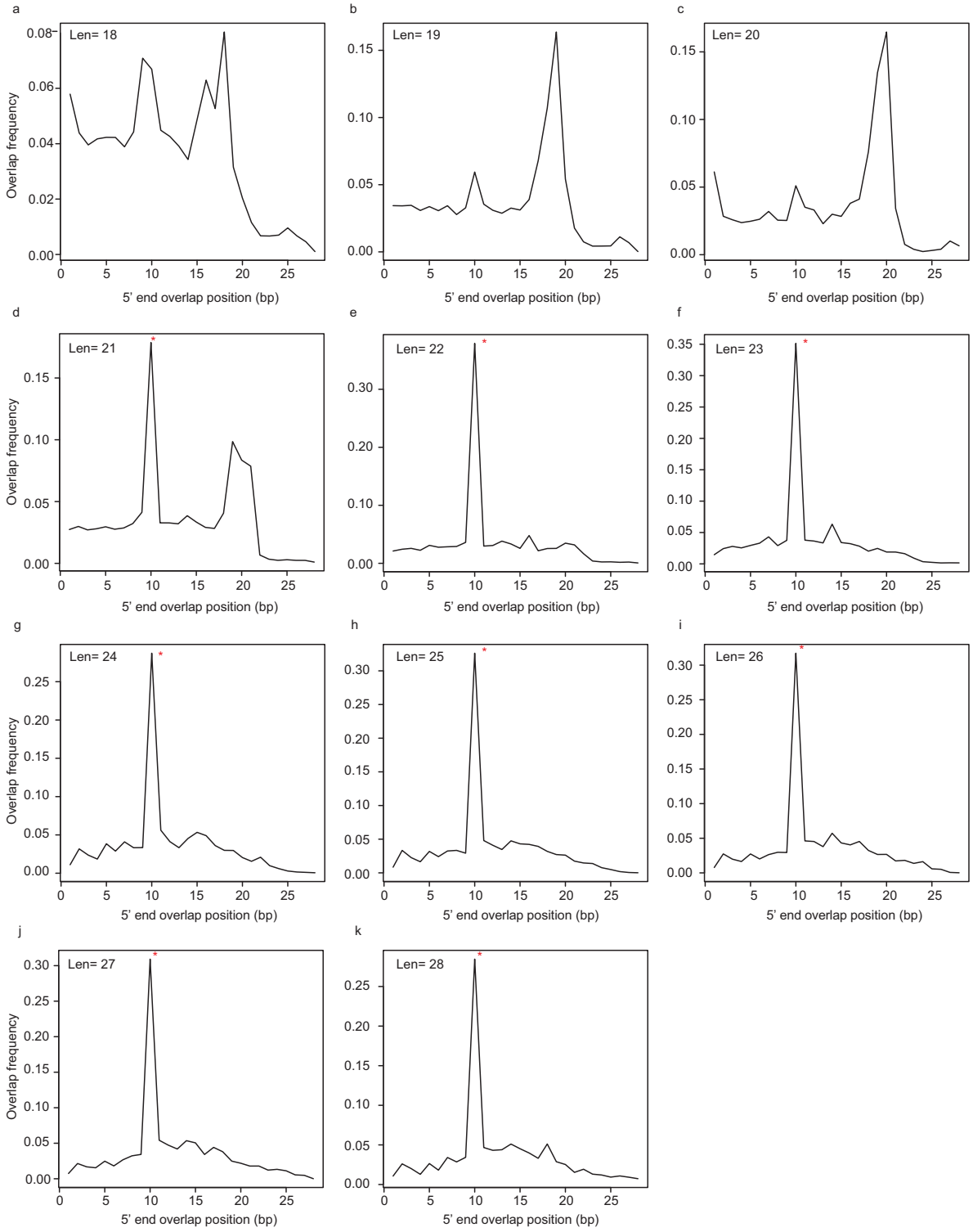


Figure S4.1.3. Ping-pong signatures of piRNA generation based on overlapping small RNA reads to spider mite transposons. Small RNA sequence reads were mapped to a database of transposon sequences. Mapped small RNAs of various lengths 18 (a), 19 (b), 20 (c), 21 (d), 22 (e), 23 (f), 24 (g), 25 (h), 26 (i), 27 (j) and 28 (k) were examined for small RNA partners on the opposite strand that overlapped the sequence read. Plotted is the frequency of the location of the 5' most base from the sequence on the opposing strand. An enrichment of sequence at 10 bp is indicative of ping-pong biogenesis by the piRNA machinery (PIWI/Aub and AGO3)[42].

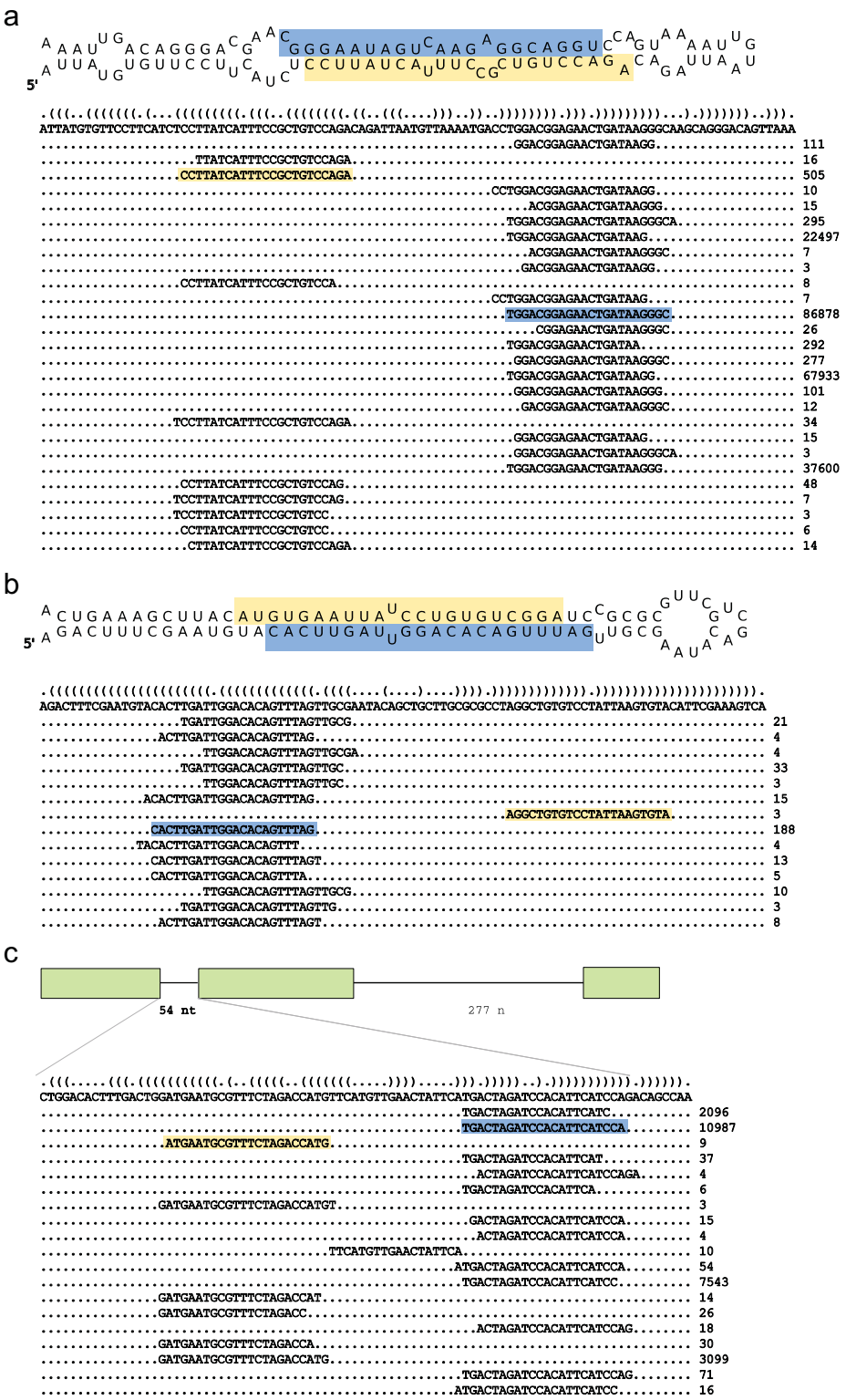


Figure S4.3.1 Three examples of *T. urticae* miRNAs with their corresponding reads from high-throughput sequencing. The miRNA (blue) and miRNA* (yellow) position are indicated, with the predicted secondary structure shown as a graph and in bracket notation. The sum of the counts for each read across the three developmental stages is shown on the right of each figure. A) mir-st21753-r11, a miRNA conserved between *D. melanogaster* (miR-184) and *T. urticae*. B) mir-st26681-r2, a non-conserved miRNA C) mir-st41843-r196, a candidate mirtron, located in the second intron of the protein-coding gene tetur08g05450 (gene structure is represented by the green boxes).

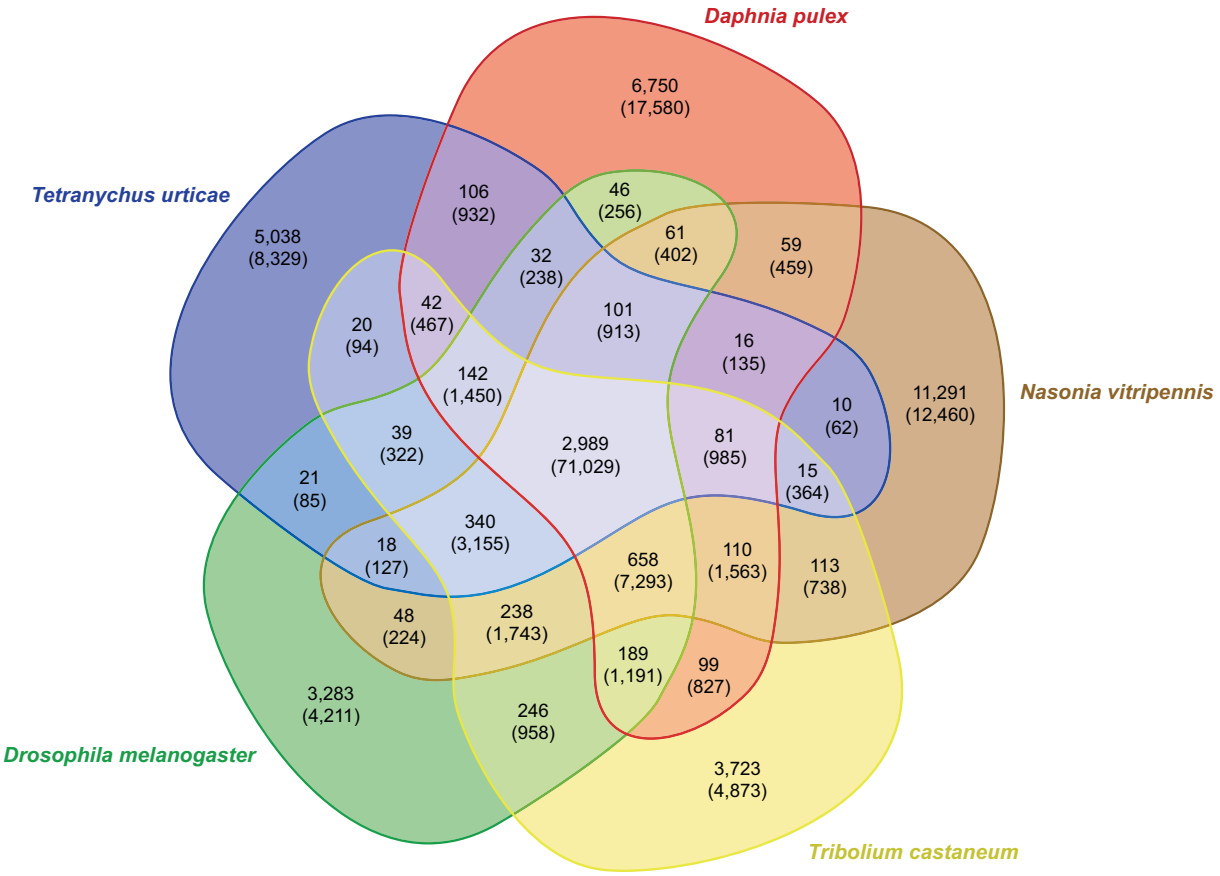


Figure S5.1.1. Gene family history. Venn diagram representation of shared/unique gene families for five compared organisms. The numbers of genes are shown in parentheses.

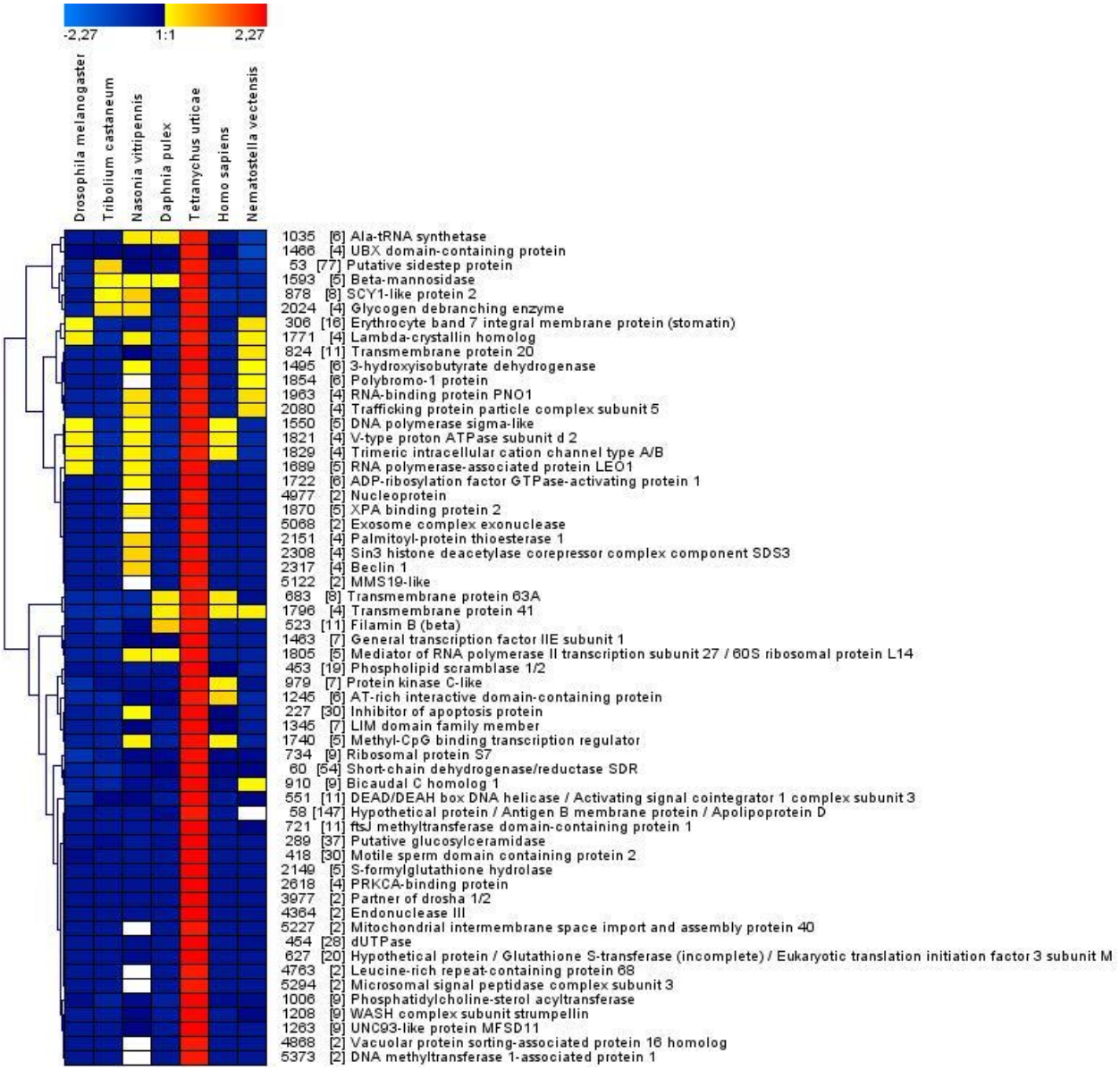


Figure S5.2.1. Hierarchical clustering of the significantly expanded gene families in *T. urticae*. Gene families are shown with a z-score greater than two in *T. urticae* and present in at least two other species. The blue to red scale at the top (based on *z*-scores) shows that, given a certain gene family and organism, the gene family size is substantially smaller resp. larger than the mean gene family size (in the organisms listed at the top). Hence, red blocks reflect gene family expansions.

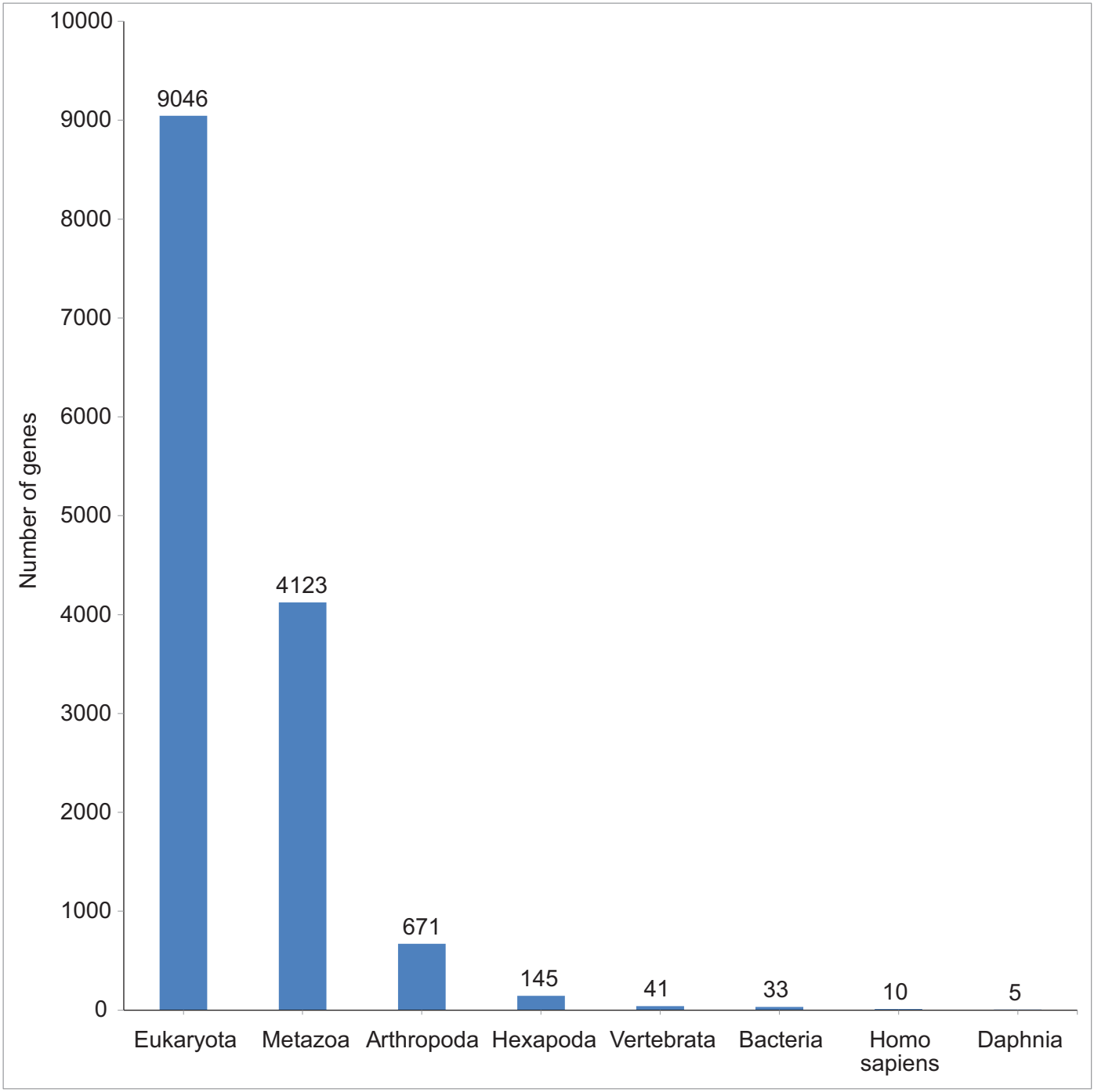


Figure S5.2.2. *T. urticae* genes with clade specific homology hits. All hits with E-value cutoff e^{-3} were assigned a clade name based on NCBI taxonomy.

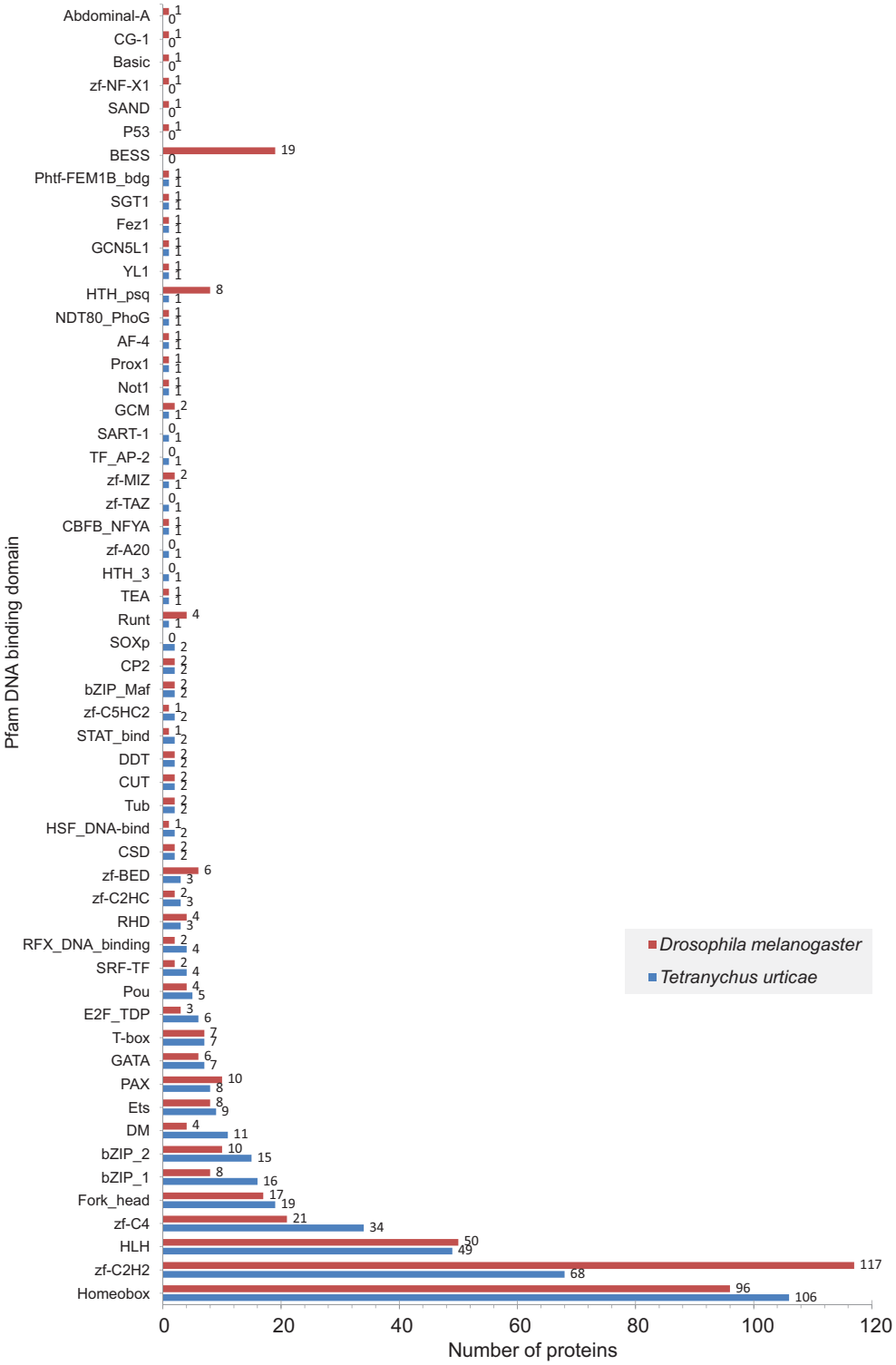


Figure S5.3.1. Transcription factor families in *T. urticae* and *D. melanogaster*.

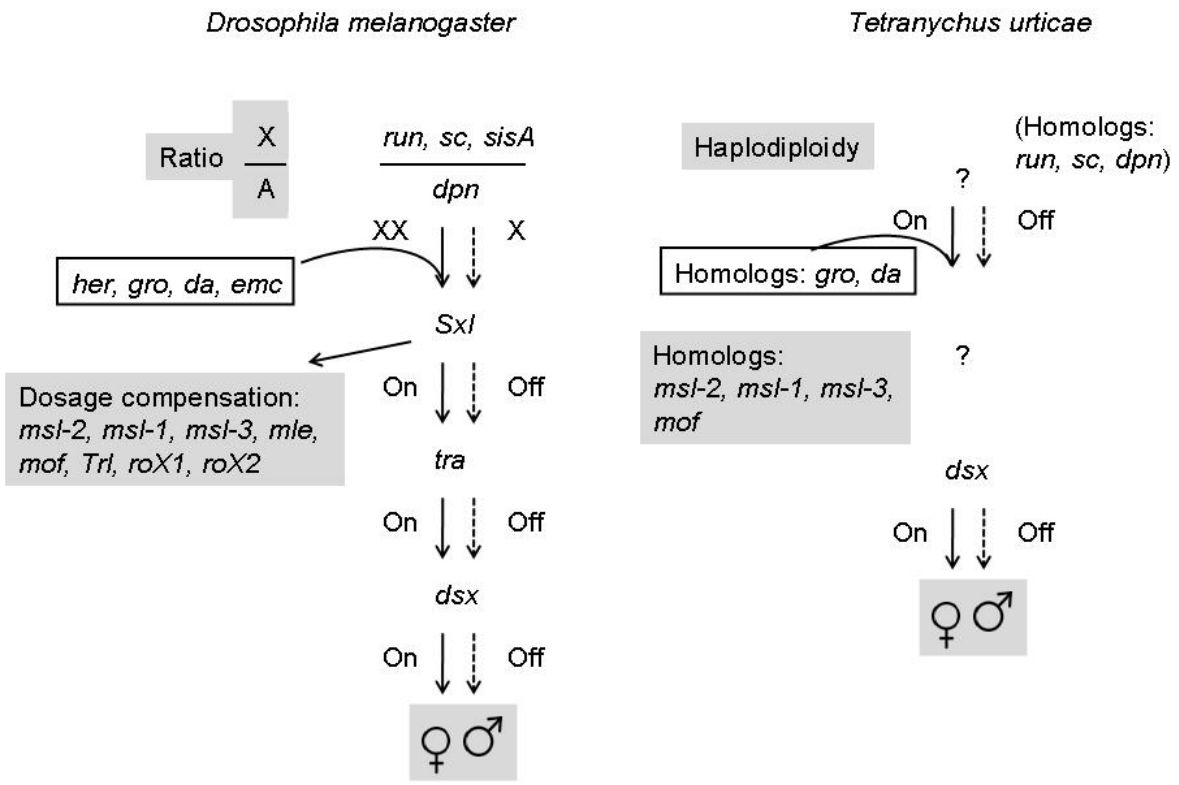


Figure S5.4.1. Sex determination pathways in *Drosophila* and *T. urticae*. Sex in *T. urticae* is determined by the ploidy, where diploid eggs develop into females and haploid eggs become males. The initial ploidy-counting mechanism is not known in *T. urticae*. Despite the lack of sex chromosomes, putative orthologs of the genes in *Drosophila* that assess the sex chromosome to autosome ratio are present in the spider mite - *scute* (*sc*) and *deadpan* (*dpn*), while *Sisterless A* (*sisA*) and *runt* (*run*) are missing. Genes that are involved with the transfer of the sex-specific X:A signal to the next downstream gene, *Sex lethal* (*Sxl*), have orthologs (*groucho, daughterless*) in the spider mite but no *extra-macrochaetae* (*emc*) or *hermaphrodite* (*her*) homologs are present. Despite lack of X-chromosome dosage compensation in haplodiploids, homologs of *Drosophila* dosage compensation genes are found in the spider mite genome. The most downstream gene of the pathway in *Drosophila*, *dsx*, is present in the *T. urticae*.

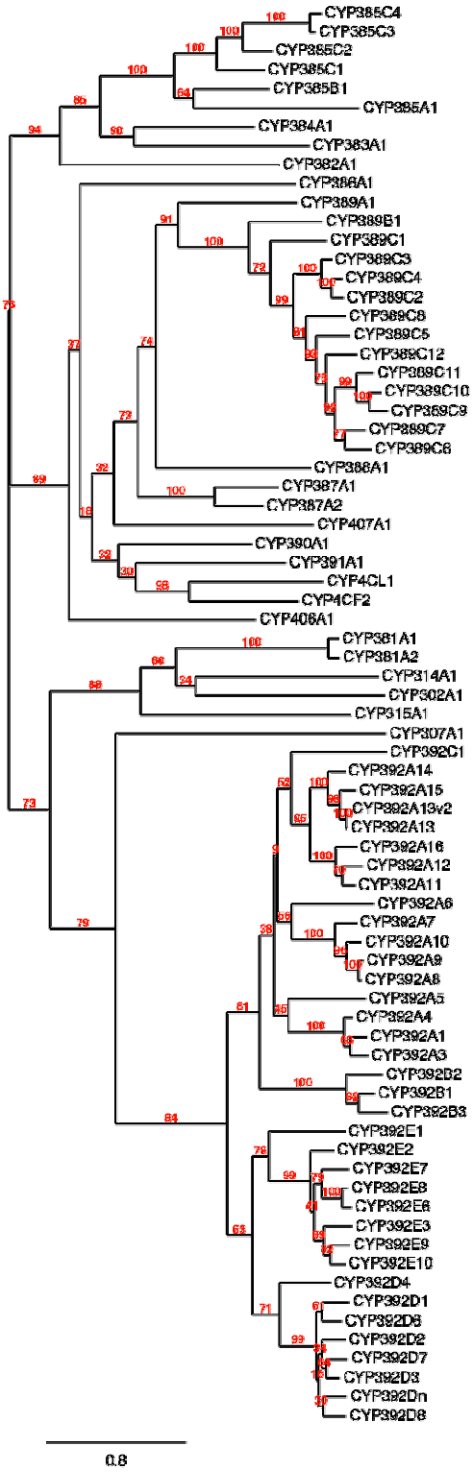


Figure S6.1.1. Phylogenetic analysis of CYP proteins.

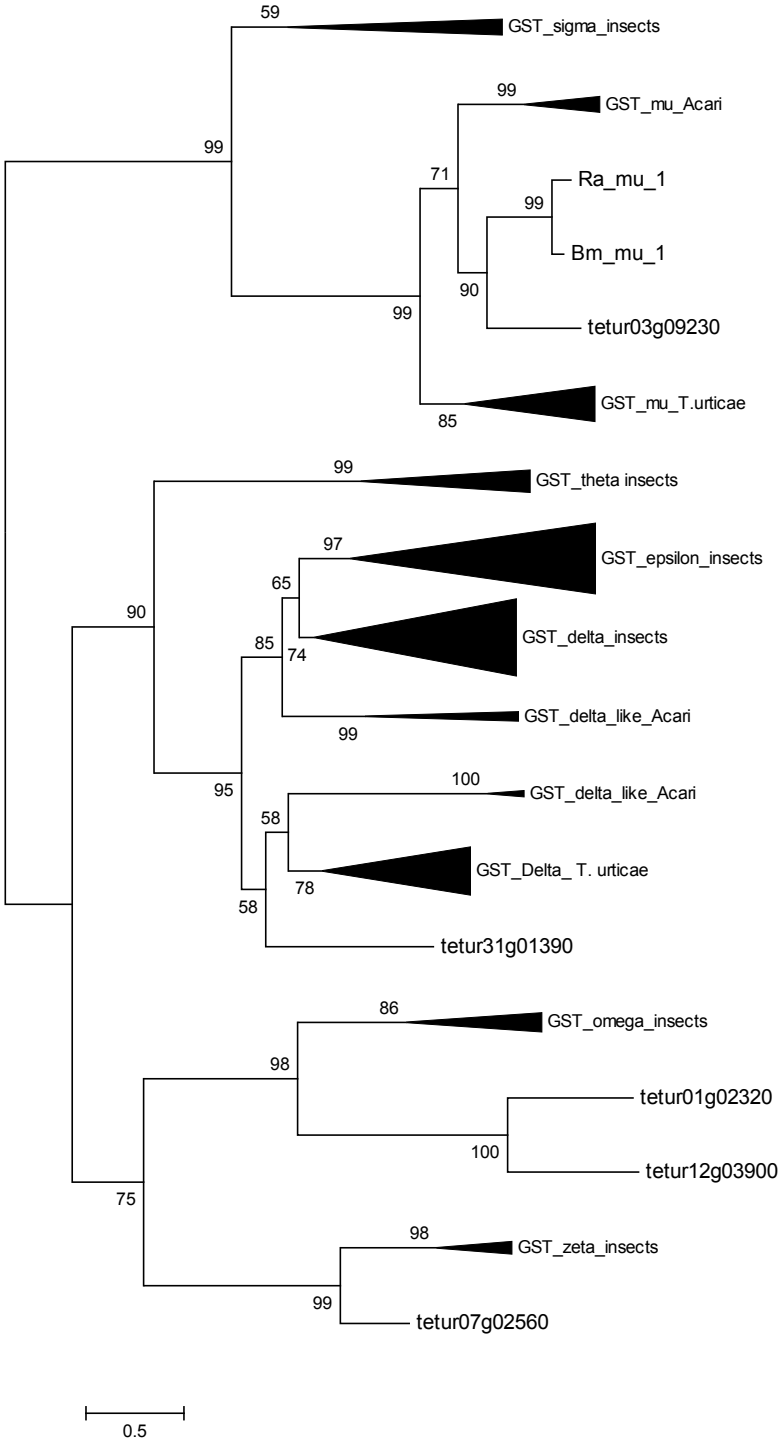


Figure S6.1.2. Phylogenetic analysis of 31 cytosolic *T. urticae* GST proteins with those of *D. melanogaster*, *A. gambiae*, *A. mellifera* and mu-and delta class GST proteins of Acari. Tree was generated using a maximum likelihood approach[78], bootstrapping with 1000 pseudoreplicates. The resulting tree was midpoint-rooted.

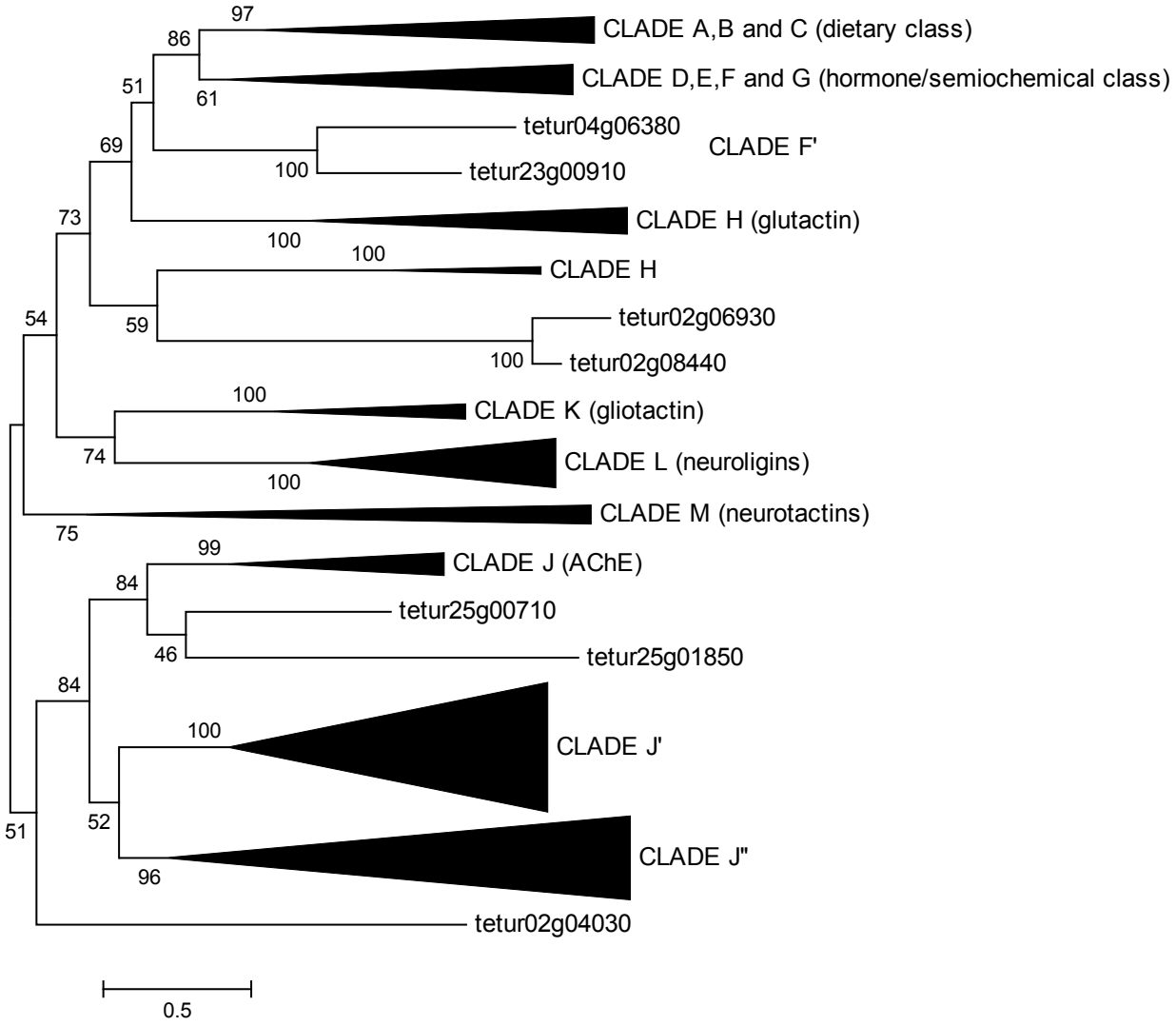


Figure S6.1.3. Phylogenetic analysis of 71 CCEs from *T. urticae* with a representative set of CCE sequences of *D. melanogaster*, *A. gambiae* and *A. mellifera*. Tree was generated using a maximum likelihood approach[78], bootstrapping with 500 pseudoreplicates. The resulting tree was midpoint rooted.

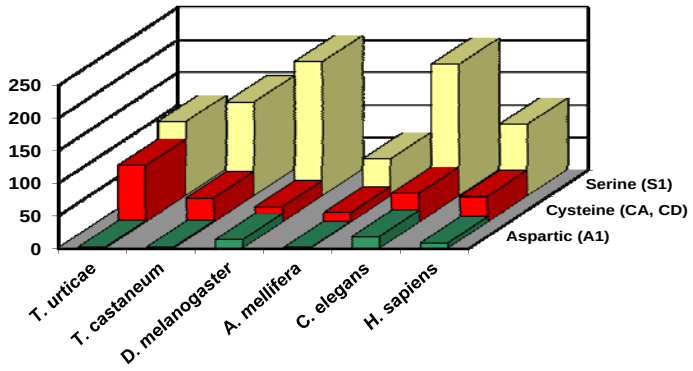


Figure S6.1.4. Number of genes belonging to different peptidase groups found in the genomes of several metazoan species.

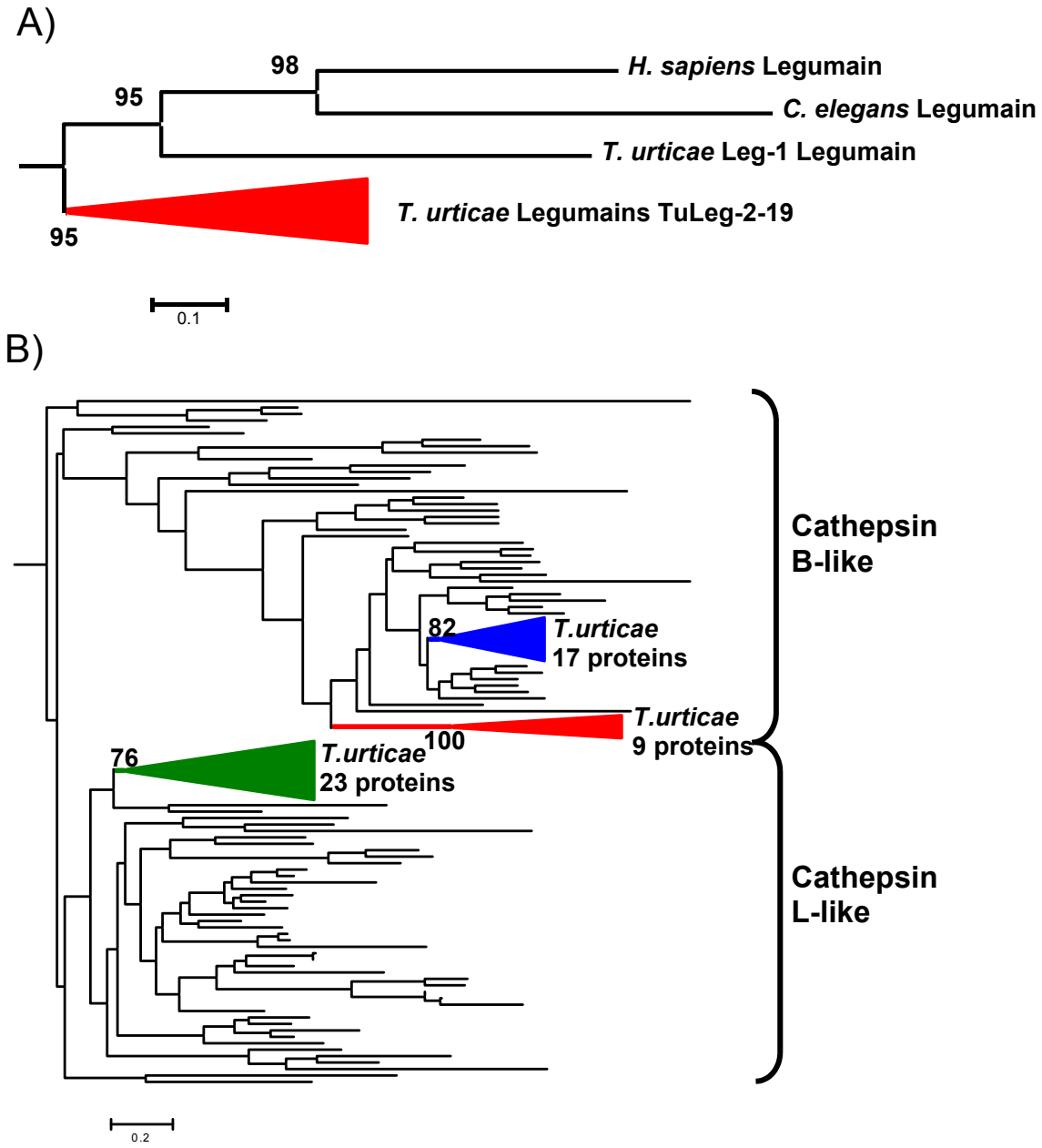


Figure S6.1.5. Phylogenetic trees C13 legumains A) and C1A papains B) from the metazoan species selected. Coloured triangles show the gene proliferations found in *T. urticae*

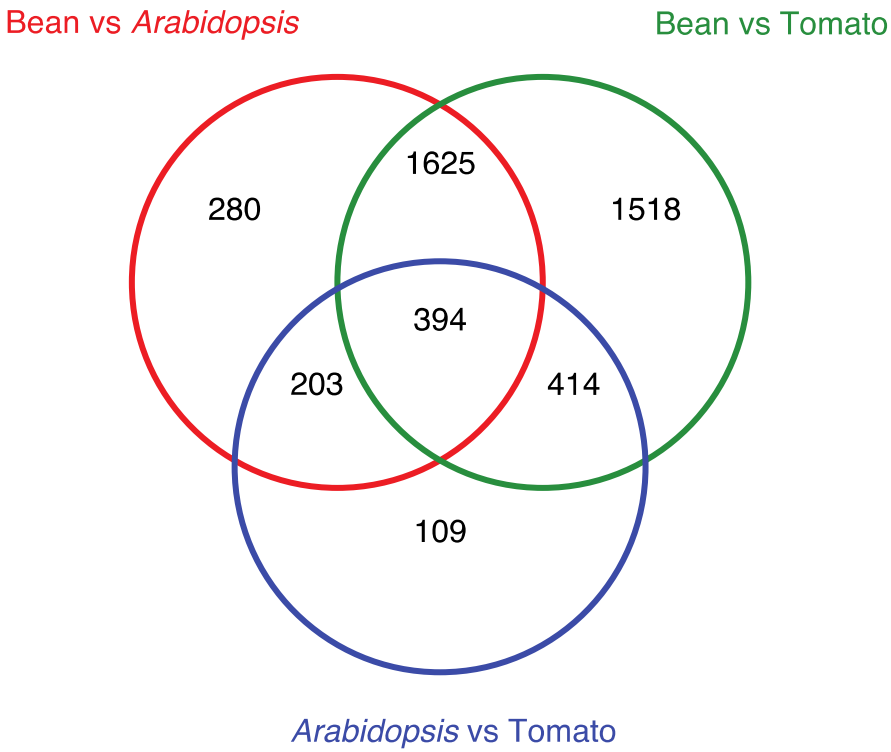


Figure S6.2.1. Overlap among differentially expressed genes from pariwise comparisons of mites on different plant hosts. Differential gene expression of mites reared on either beans, *Arabidopsis* or tomatoes are shown. Differentially expressed gene sets were compared for each of the possible pairwise comparisons (n=3). Note the overall similarity between *Arabidopsis* and tomato (blue) while comparisons to bean (the preferred host for the London strain; red and green) highlight much larger changes in gene expression (mites were reared on beans in the laboratory prior to addition to the three hosts under controlled environmental conditions; see Section 6.2).

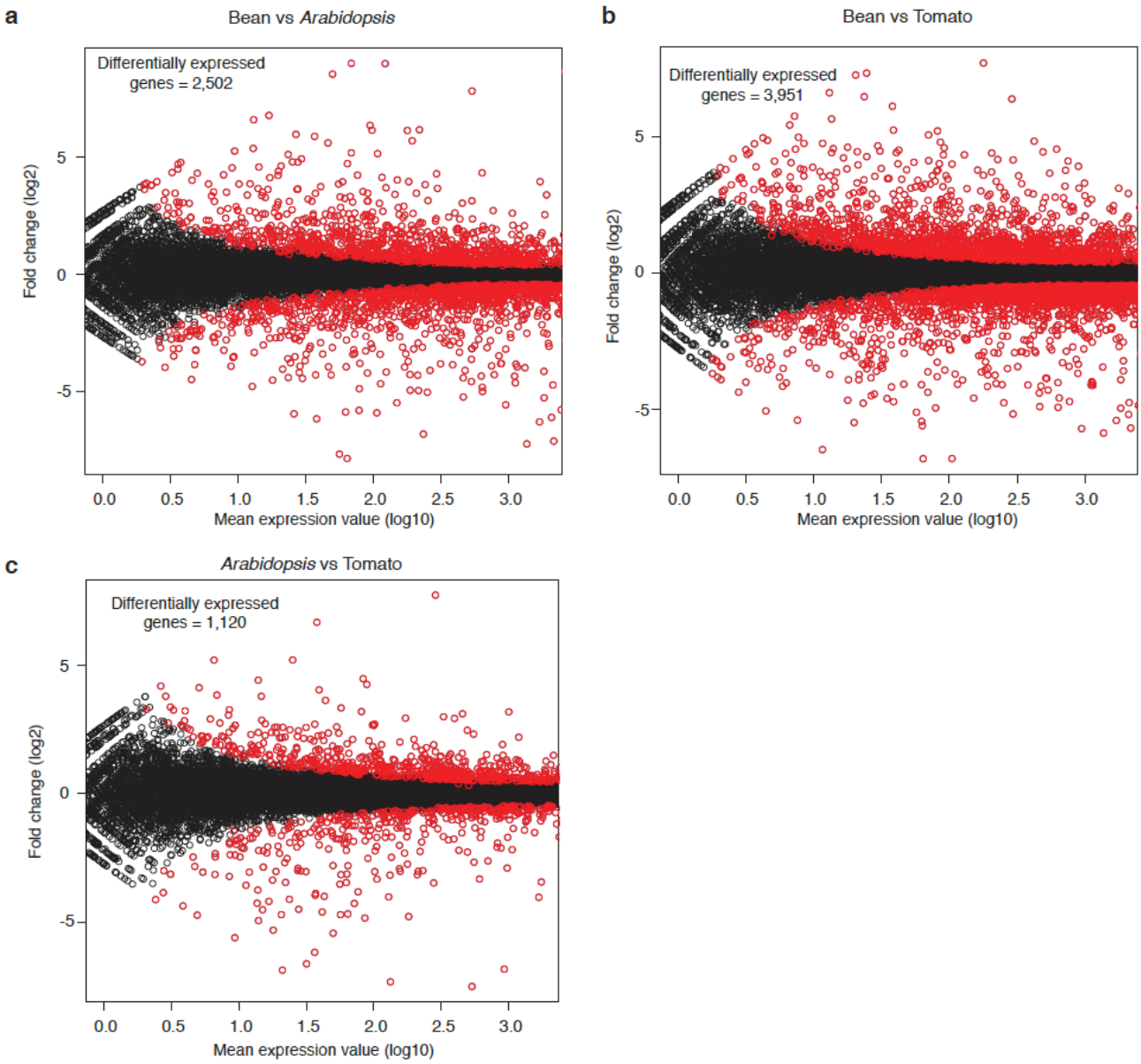


Figure S6.2.2. Genic fold change by mean expression level analysis for the mite feeding experiment. Each plot highlights mean expression values across conditions ($\log_{10}$ transformed) and the fold change between conditions. Points highlighted in red are genes called as differentially expressed by DEseq[90] with a FDR p-value < 0.05. A) Comparisons of larval gene expression between larvae fed on bean versus *Arabidopsis*. B) Comparisons of larval gene expression between larvae fed on bean versus tomato. C) Comparisons of larval gene expression between larvae fed on *Arabidopsis* versus tomato. Many more genes are differentially expressed in all comparisons to bean (A & B), and at somewhat higher fold changes, while there are fewer differentially expressed genes with lower fold changes in the *Arabidopsis* versus tomato comparison (C).

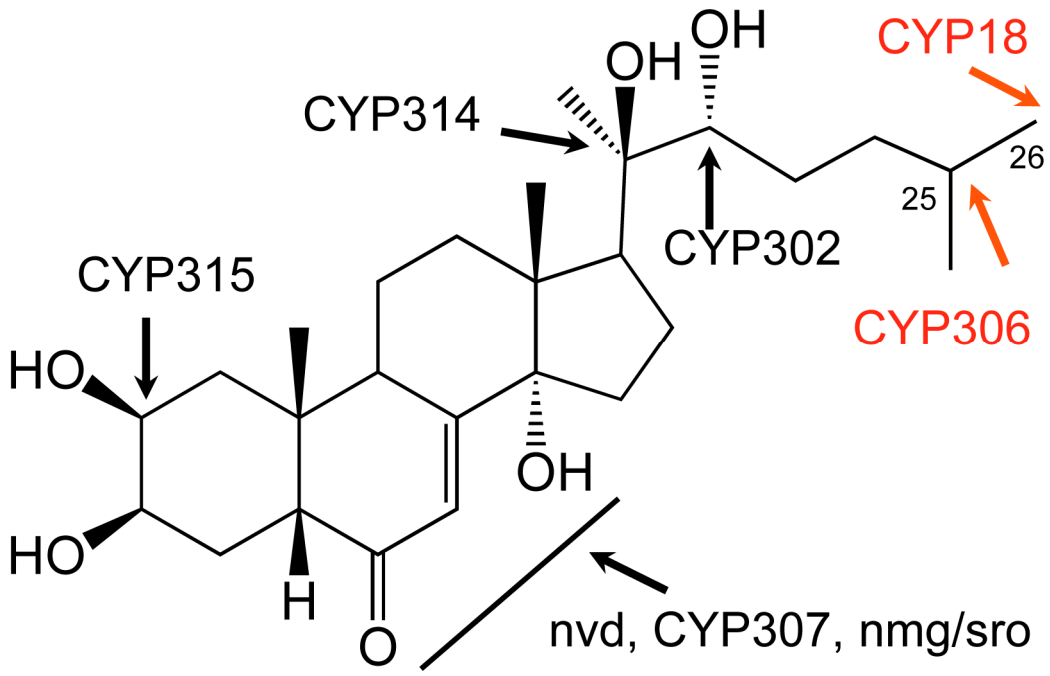
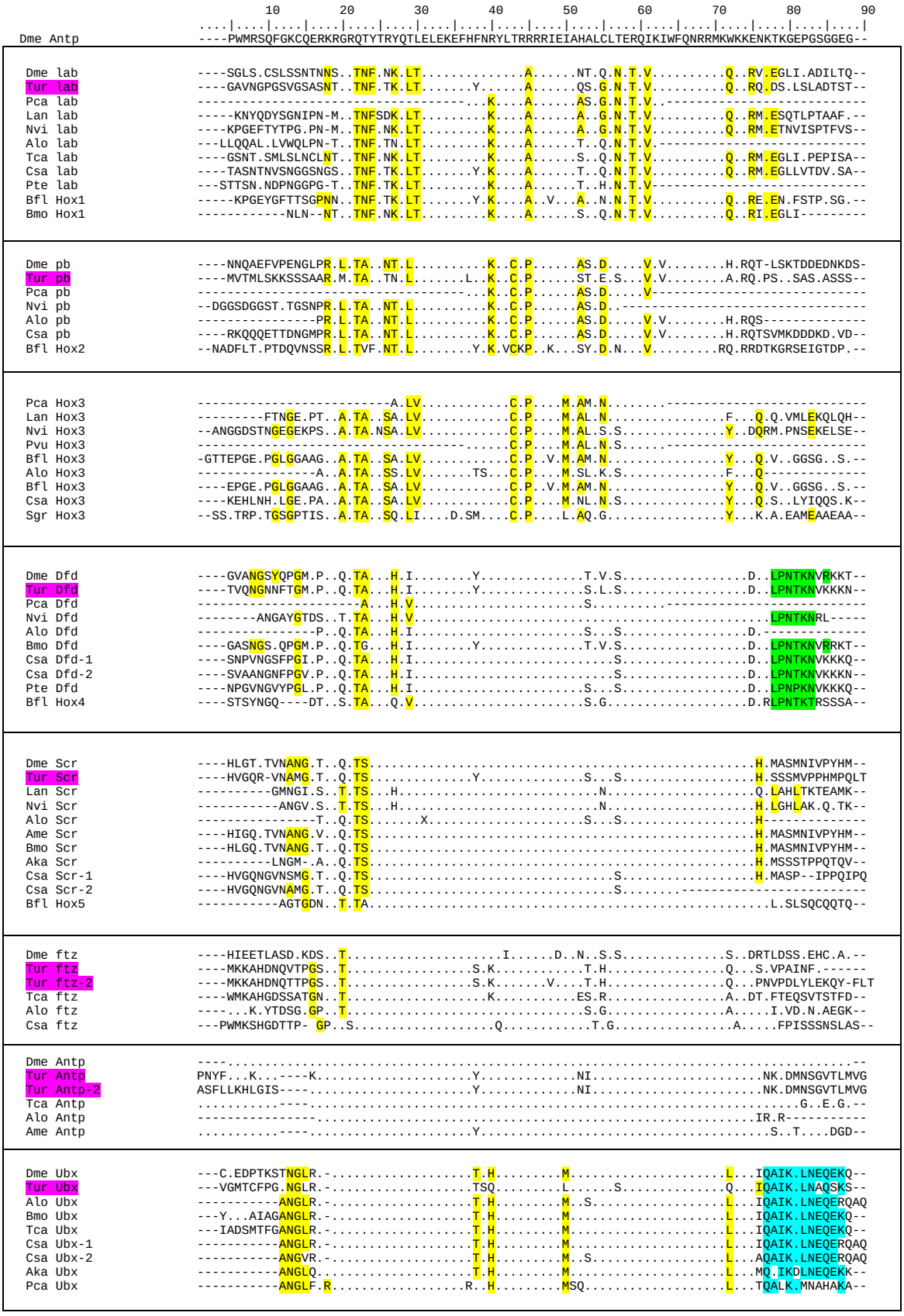


Figure S7.1.1. Structure of ponasterone A, the moulting hormone of *T. urticae*. Arrows point to the sites of metabolism of a sterol precursor by the products of six orthologs to insect ecdysteroid biosynthesis genes. Red arrows indicate sites of action of enzymes lost in *T. urticae*: CYP306, a biosynthetic C25 hydroxylase and CYP18, a C26 hydroxylase/oxidase involved in ecdysteroid degradation.


10 20 30 40 50 60 70 80 90
Dme Antp ----PWMRSQFGKCQERKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKENKTKGEPGSGGEG--
Dme lab ----SGLS.CSLSSNTNNS..TNF.NK.LT.....A.....NT.Q.N.T.V.....Q..RV.EGLI.ADILTQ--
Tur lab ----GAVNGPGSVGSASNT..TNF.TK.LT.....Y.....A.....QS.G.N.T.V.....Q..RQ.DS.LSLADTST--
Pca lab ----------K....A.....AS.G.N.T.V..----------
Lan lab -----KNYQDYSGNIPN-M..TNFSDK.LT.....K....A.....A..G.N.T.V.....Q..RM.ESQTLPTAAF.--
Nvi lab ----KPGEFTYTPG.PN-M..TNF.NK.LT.....K....A.....A..G.N.T.V.....Q..RM.ETNVISPTFVS--
Alo lab ---LLQQAL.LVWQLPN-T..TNF.TN.LT.....K....A.....T..Q.N.T.V----------
Tca lab ----GSNT.SMLSLNCLNT..TNF.NK.LT.....K....A.....S..Q.N.T.V.....Q..RM.EGLI.PEPISA--
Csa lab ----TASNTNVSNGGSNGS..TNF.TK.LT.....Y.K....A.....T..Q.N.T.V.....Q..RM.EGLLVTDV.SA--
Pte lab ---STTSN.NDPNGGPG-T..TNF.TK.LT.....K....A.....T..H.N.T.V----------
Bfl Hox1 -----KPGEYGFTTSGPNN..TNF.TK.LT.....Y.K....A..V...A..N.N.T.V.....Q..RE.EN.FSTP.SG.--
Bmo Hox1 -----------NLN--NT..TNF.NK.LT.....K....A.....S..Q.N.T.V.....Q..RI.EGLI---------
Dme pb ----NNQAEFVPENGLPR.L.TA..NT.L.....K..C.P.....AS.D.....V.V.....H.RQT-LSKTDDEDNKDS-
Tur pb ----MVTMLSKKSSSAAR.M.TA..TN.L.....L..K..C.P.....ST.E.S...V.V.....A.RQ.PS..SAS.ASSS--
Pca pb ----------K..C.P.....AS.D.....V----------
Nvi pb --DGGSDGGST.TGSNPR.L.TA..NT.L.....K..C.P.....AS.D.----------
Alo pb ----------PR.L.TA..NT.L.....K..C.P.....AS.D.....V.V.....H.RQS----------
Csa pb ----RKQQQETTDNGMPR.L.TA..NT.L.....K..C.P.....AS.D.....V.V.....H.RQTSVMKDDDKD.VD--
Bfl Hox2 --NADFLT.PTDQVNSSR.L.TVF.NT.L.....Y.K.VCKP..K...SY.D.N...V.....RQ.RRDTKGRSEIGTDP.--
Pca Hox3 ----------A.LV.....C.P....M.AM.N.....----------
Lan Hox3 ---------FTNGE.PT..A.TA..SA.LV.....C.P....M.AL.N.....F...Q.Q.VMLEKQLQH--
Nvi Hox3 --ANGGDSTNGEGEKPS..A.TA.NSA.LV.....C.P....M.AL.S.S.....Y..DQRM.PNSEKELSE--
Pvu Hox3 ----------C.P....M.AL.N.S.....----------
Bfl Hox3 -GTTEPGE.PGLGGAAG..A.TA..SA.LV.....C.P..V.M.AM.N.....Y...Q.V..GGSG..S.--
Alo Hox3 ----------A..A.TA..SS.LV.....TS...C.P....M.SL.K.S.....F...Q----------
Bfl Hox3 ----EPGE.PGLGGAAG..A.TA..SA.LV.....C.P..V.M.AM.N.....Y...Q.V..GGSG..S.--
Csa Hox3 ----KEHLNH.LGE.PA..A.TA..SA.LV.....C.P....M.NL.N.S.....Y...Q.S..LYIQQS.K--
Sgr Hox3 --SS.TRP.TGSGPTIS..A.TA..SQ.LI....D.SM....C.P....L.AQ.G.....Y...K.A.EAMEAAEAA--
Dme Dfd ----GVANGSYQPGM.P..Q.TA...H.I.....Y.....T.V.S.....D..LPNTKNVRKKT--
Tur Dfd ----TVQNGNNFTGM.P..Q.TA...H.I.....Y.....S.L.S.....D..LPNTKNVKKKN--
Pca Dfd ----------A...H.V.....S.....----------
Nvi Dfd --------ANGAYGTDS..T.TA...H.V.....LPNTKNRL-----
Alo Dfd ----------P..Q.TA...H.I.....S...S.....D.----------
Bmo Dfd ----GASNGS.QPGM.P..Q.TG...H.I.....Y.....T.V.S.....D..LPNTKNVRRKT--
Csa Dfd-1 ----SNPVNGSFPGI.P..Q.TA...H.I.....S.....D..LPNTKNVKKKQ--
Csa Dfd-2 ----SVAANGNFPGV.P..Q.TA...H.I.....S.....D..LPNTKNVKKKN--
Pte Dfd ----NPGVNGVYPGL.P..Q.TA...H.I.....S...S.....D..LPNPKNVKKKQ--
Bfl Hox4 ----STSYNGQ----DT..S.TA...Q.V.....S.G.....D.RLPNTKTRSSSA--
Dme Scr ----HLGT.TVNANG.T..Q.TS.....H.MASMNIVPYHM--
Tur Scr ----HVGQR-VNAMG.T..Q.TS.....Y.....S...S.....H.SSSMVPPHMPQLT
Lan Scr ----------GMNGI.S..T.TS...H.....N.....Q.LAHLTKTEAMK--
Nvi Scr ----------ANGV.S..T.TS...H.....N.....H.LGHLAK.Q.TK--
Alo Scr ----------T..Q.TS.....X.....S...S.....H----------
Ame Scr ----HIGQ.TVNANG.V..Q.TS.....H.MASMNIVPYHM--
Bmo Scr ----HLGQ.TVNANG.T..Q.TS.....H.MASMNIVPYHM--
Aka Scr ----------LNGM-.A..Q.TS.....H.MSSSTPPQTQV--
Csa Scr-1 ----HVGQNGVNSMG.T..Q.TS.....S.....H.MASP--IPPQIPQ
Csa Scr-2 ----HVGQNGVNAMG.T..Q.TS.....S.....----------
Bfl Hox5 ----------AGTGDN..T.TA.....L.SLSQCQQTQ--
Dme ftz ----HIEETLASD.KDS..T.....I.....D..N..S.S.....S..DRTLDSS.EHC.A.--
Tur ftz ----MKKAHDNQVTPGS..T.....S.K.....T.H.....Q...S.VPAINF.-------
Tur ftz-2 ----MKKAHDNQTTPGS..T.....S.K.....V....T.H.....Q...PNVPDLYLEKQY-FLT
Tca ftz ----WMKAHGDSSATGN..T.....K.....ES.R.....A..DT.FTEQSVTSTFD--
Alo ftz ------...K.YTDSG.GP..T.....S.G.....A.....I.VD.N.AEGK--
Csa ftz ---PWMKSHGDTTP- GP..S.....Q.....T.G.....A.....FPISSSNSLAS--
Dme Antp ----.....--
Tur Antp PNYF...K.------K.....Y.....NI.....NK.DMNSGVTLMVG
Tur Antp-2 ASFLLKHLGIS----.....Y.....NI.....NK.DMNSGVTLMVG
Tca Antp .....G..E.G.--
Alo Antp ----------.....IR.R----------
Ame Antp .....Y.....S..T....DGD--
Dme Ubx ---C.EDPTKSTNGLR.-.....T.H.....M.....L...IQAIK.LNEQEKQ--
Tur Ubx ---VGMTCFPG.NGLR.-.....TSQ.....L.....S.....Q...IQAIK.LNAQSKS--
Alo Ubx ----------ANGLR.-.....T.H.....M..S.....L...IQAIK.LNEQERQAQ
Bmo Ubx ---Y...AIAGANGLR.-.....T.H.....M.....L...IQAIK.LNEQEKQ--
Tca Ubx ---IADSMTFGANGLR.-.....T.H.....M.....L...IQAIK.LNEQEKQ--
Csa Ubx-1 ----------ANGLR.-.....T.H.....M.....L...IQAIK.LNEQERQAQ
Csa Ubx-2 ----------ANGVR.-.....T.H.....M..S.....L...AQAIK.LNEQERQAQ
Aka Ubx ----------ANGLQ.....T.H.....M.....L...MQ.IKDLNEQEKK--
Pca Ubx ----------ANGLF.R.....R..H.....MSQ.....L...TQALK.MNAHAKA--

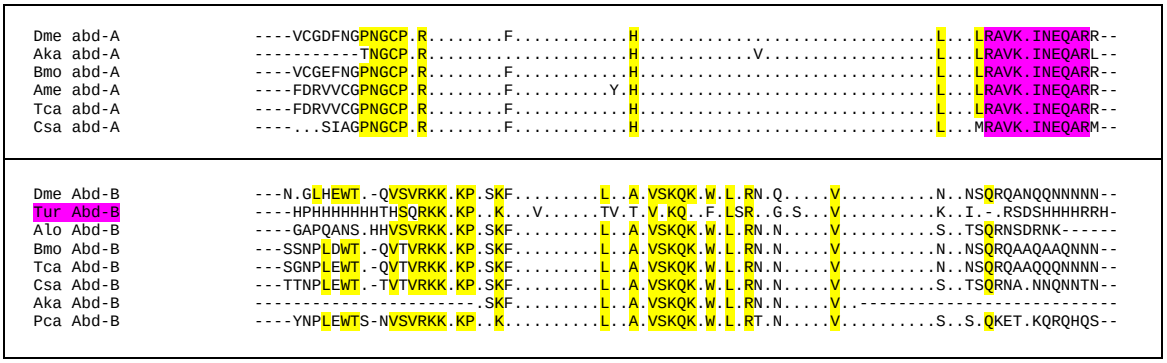


Figure S8.1.1. Alignment of *T. urticae* Hox protein homeodomain and flanking regions.

Dashes indicate identity to *Drosophila melanogaster* Antennapedia amino acid sequence. Yellow highlighting shows conserved residues that differ from the *Drosophila melanogaster* Hox protein Antennapedia within each group of orthologs. Examples of conserved peptide motifs located outside of the homeodomain for a selection of Hox proteins are highlighted as described: Dfd - green, Ubx - blue, and Abd-B - purple. Species names, and abbreviations used in the protein alignment: *Acanthokara kaputensis*, Aka; *Apis mellifera*, Ame; *Archegozetes longisetosus*, Alo; *Artemia franciscana*, Afr; *Bombyx mori*, Bmo; *Branchiostoma floridae*, Bfl; *Cupiennius salei*, Csa; *Drosophila melanogaster*, Dme; *Lingula anatine*, Lan; *Nereis virens*, Nvi; *Parasteatoda tepidariorum*, Pte; *Patella vulgate*, Pvu; *Priapulus caudatus*, Pca; *Tetranychus urticae*, Tur; *Tribolium castaneum*, Tca. (Protein alignment based on de Rosa *et al.*[98] and this study).

Sequence of Tetur01g16320	Domain
MVFKMYLNLLILAITATNYVSTRSMGSMPGMELDVNMPM DMMSNVLGGSAFAGSNADTENEGSEAASNAESTAGANAE ATTYEEPDGEDDGLTYGNDESDADAKATAESAAKA	N-terminal domain
GSDNGSGNNGGNGY---GNNGGSSSATSSSSASGSSTSE GSDNGSGNNGGNGYNNNGNNGGSSSATSSSSASGSSTSE GSDNGSGNNGGNGYNNNGNNGGSSSATSSSSASGSSTSE GSDNGSGNNAGNGYNNNGNNGGSSSATSSSSASGSSTSE GSDNGSGNNGGNGY---GNNGGSSSATSSSSASGSSTSE GSDNGSGNNGGNGYNNNGNNGGSSSATSSSSASGSSTSE GSDN--------GYNNNGNNGGSSSATSSSSASSSSTSE GSDNGSGNNGGNGYNNNVNNGGSSSATSSSSASGSS---	Internal domain 1
NQRDLTMVAVTTEETVITTMVTMEDQAQQHHRPQHQVHQ LQRDLTTVAVTTEETVITTMVTREDQAQQHH-H-HQ-H-	Spacer domain
SGNNGNNGSSSAAASSAAAASGSSAS--NGSDNN GGNNGNNGSSSAAASSAAAASGASAS--NGSDNN GGNNGNNGSSSAAASSAAAASGSSAS--HGSDNN GGNNGNNGSSSAAASSAAASSDASAS--NGSDNN GGNNGNNGSSSAAASSAAAASGASAS--NGSDNN GGNNGNNGSSSAAASSAAAASGSSAS--HGSDNN GGNNGNNGSSSAAASSAAASSGSSAS--NGSDNN GGNNGNNGSSSAAASSAAAASGSSAS--NGSDNN GGHN--R--S----SSAAAASGSSAS--NGSDNN GGTNGNNGSSSAAASSAAAASGSSAS--HGSDNN GGNNGNNGSSSAAASSAAAASGSSAS--NGSDNN GGNNGNNGSSSAAASSAAASSGSSAS--NGSDNN GGNNGNNGSSSAAASSAAAASGSSAS--NGSDNN GGNNGNNGSSSAAASSAAAASGSSAS--NGSDNN GGNNGNNGSSSAAASSAAASSGSSAS--NGSDNN GGNNGNNGSSSAAAASAAAASGSNAKKNNGS-NN	Internal domain 2
SGSNSAATSSNSSGKKVNNSGSSSGSAAGSGSNRGNGQN NGGSKGSNGSAASSATSAAAASGAAGNGNSKKGAKQGNG PGNSAASASAAASSASGKGSKSGKSPAKQGIIPAMMSKI PTLSVSMF	C-terminal domain

Figure S9.1.1. Domain composition of tetur01g16320 protein. Tetur01g16320 protein shows typical modular domain composition of silk proteins. Charged N- and C-terminal domains (residues 1-113, pI 3.70; 998-1112, pI11.59) are also conserved in other silk proteins of *T. urticae*. Two internal domains (domain1: 114-408 and domain2: 483-987), with distinct repetitive motifs (forming β-sheets), are separated by spacer domain containing two almost identical repeats (409-482). Repetitive motifs found in internal domains are shared (although with less conservation) by other silk proteins of *T. urticae.* Note high content of serine (27% of total residue number (TRN)), glycine (17% TRN), alanine (17%TRN) and asparagine (17%TRN) residues.

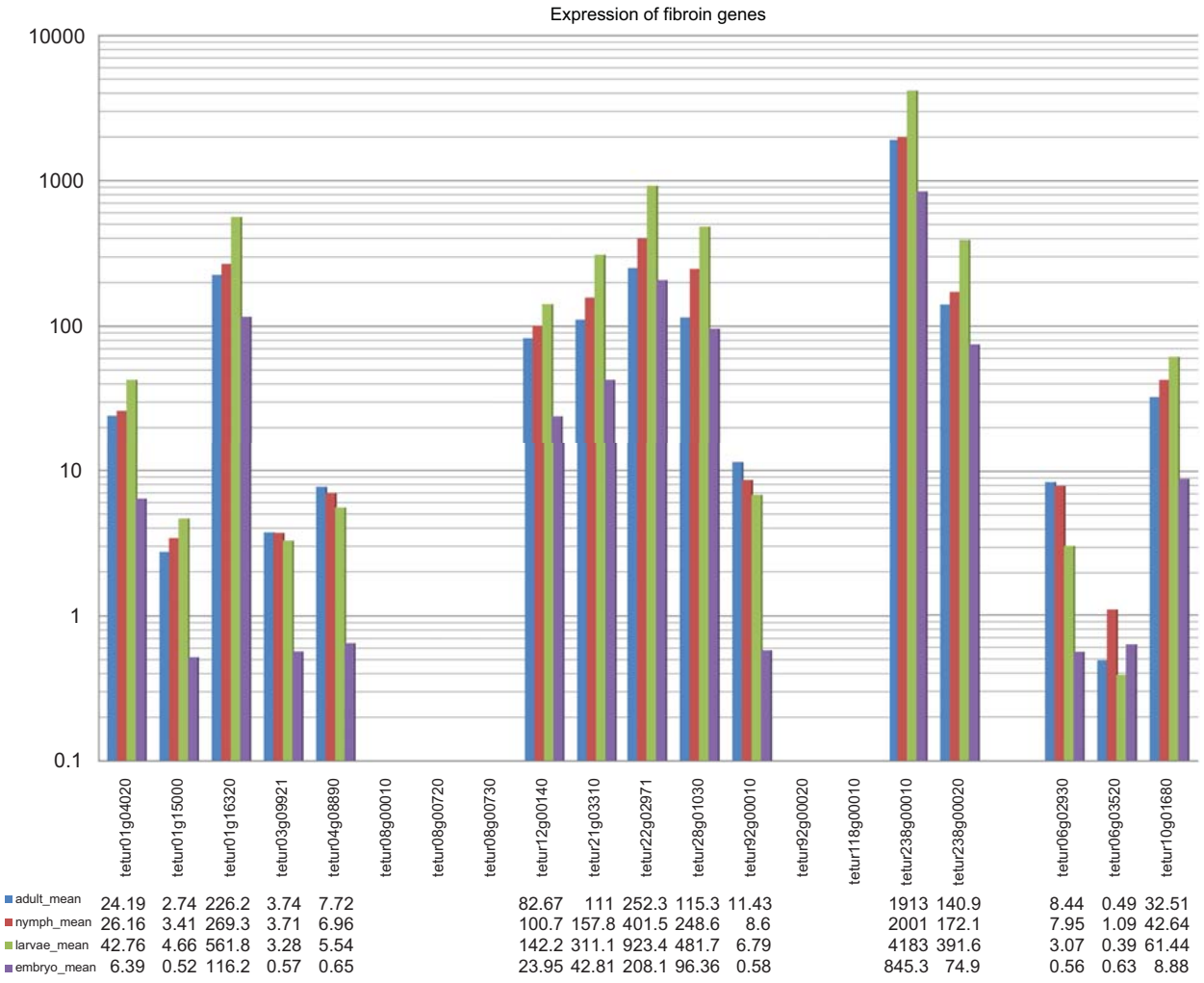


Figure S9.1.2. The expression pattern of putative *T. urticae* fibroin genes in different developmental stages.

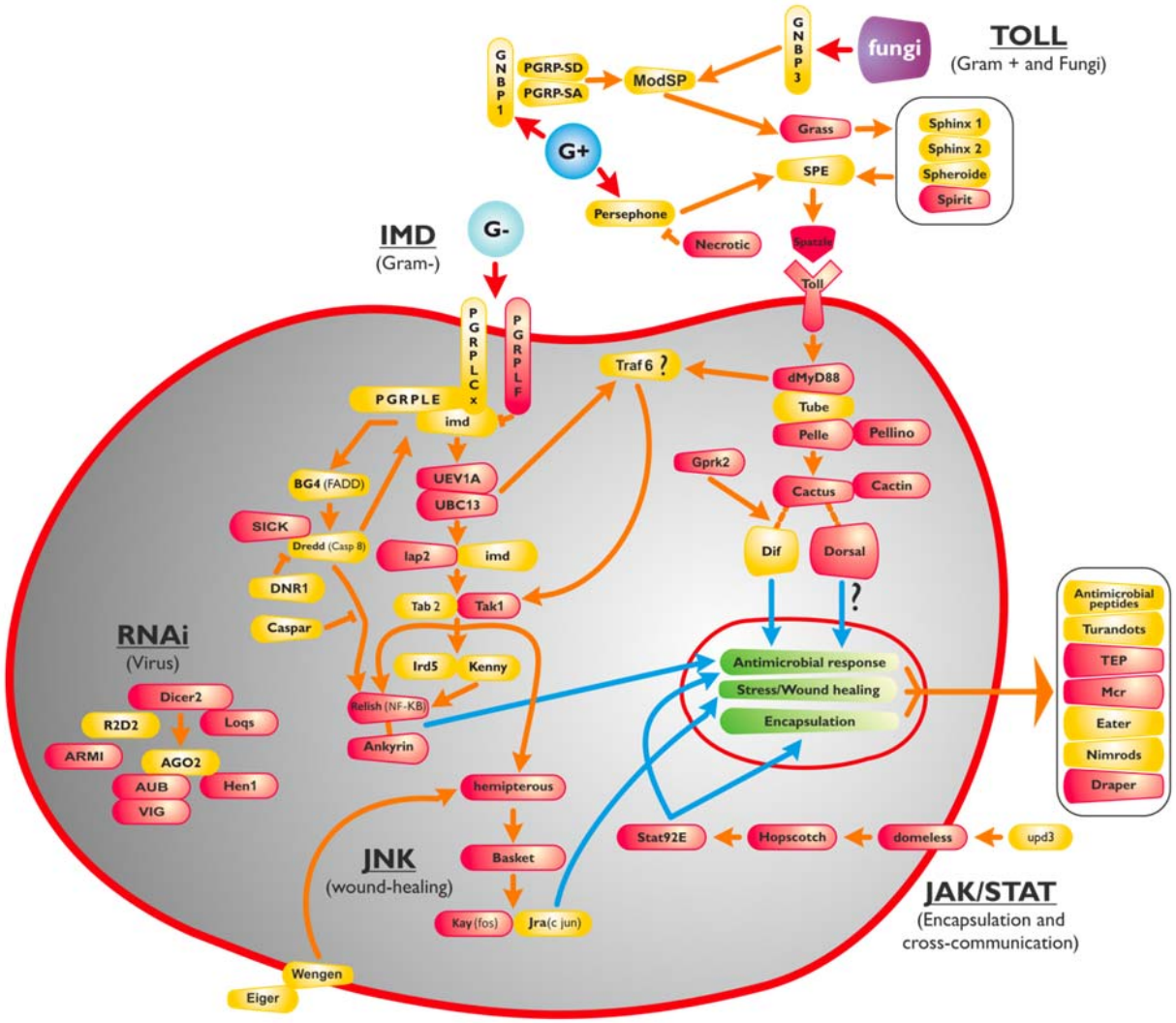


Figure S10.1.1. The four main immunity pathways in *D. melanogaster* and their counterparts in *T. urticae*. The IMD, Toll, JNK and JAK/STAT pathways respond to different challenges and converge into the transcriptional activation of genes involved in phagocytosis, encapsulation and humoral responses. Important genes in the described *D. melanogaster* immune response that were not detected in the *T. urticae* genome are in yellow (conserved genes shown in red). Question marks on the genes *Dorsal* and *Traf6* represent the fragile nature of the evidence for their involvement in these pathways during immune response in *D. melanogaster*.

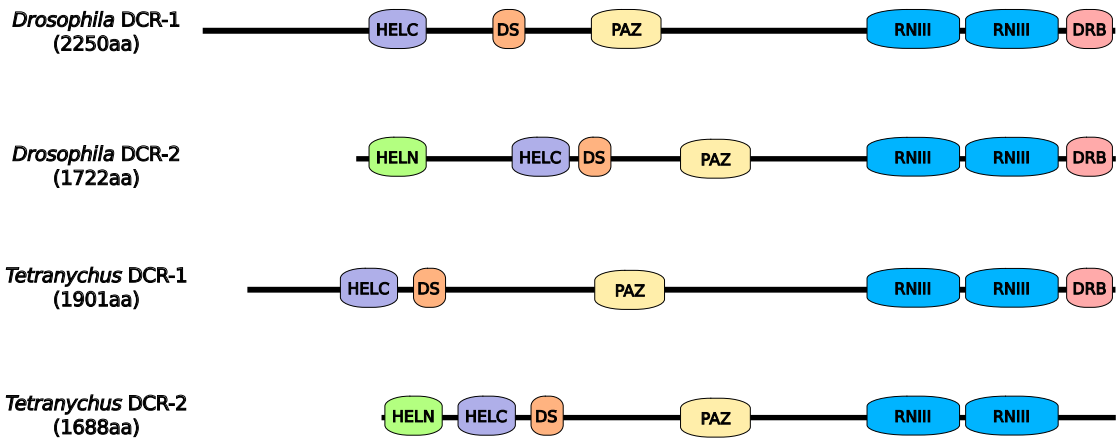


Figure S10.2.1. Comparison of dicer homologs in *Drosophila* and *Tetranychus.* Highlighted domains are PAZ, DExH box (HELN), helicase C (HELC), ribonuclease III (RNIII), and double stranded RNA binding (DRB and DS).

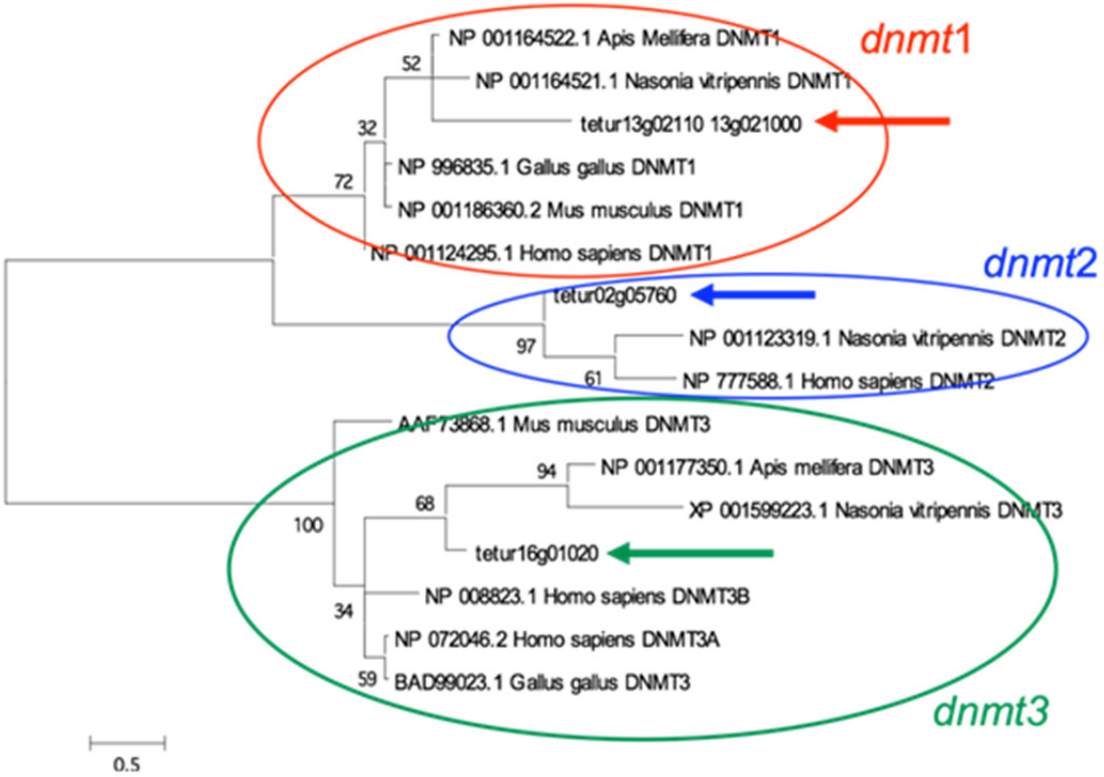


Figure S12.1.1. Phylogenetic relationship among several *dnmt*s and the putative *dnmt*s from *T. urticae* genome. We used the JTT model of protein evolution and the Neighbor-joining algorithm to draw the tree. The bootstrap values are overall low. Nevertheless, the three classes of *dnmts* cluster together and the designated *dnmt* homologs from *T. urticae* are placed in corresponding positions.

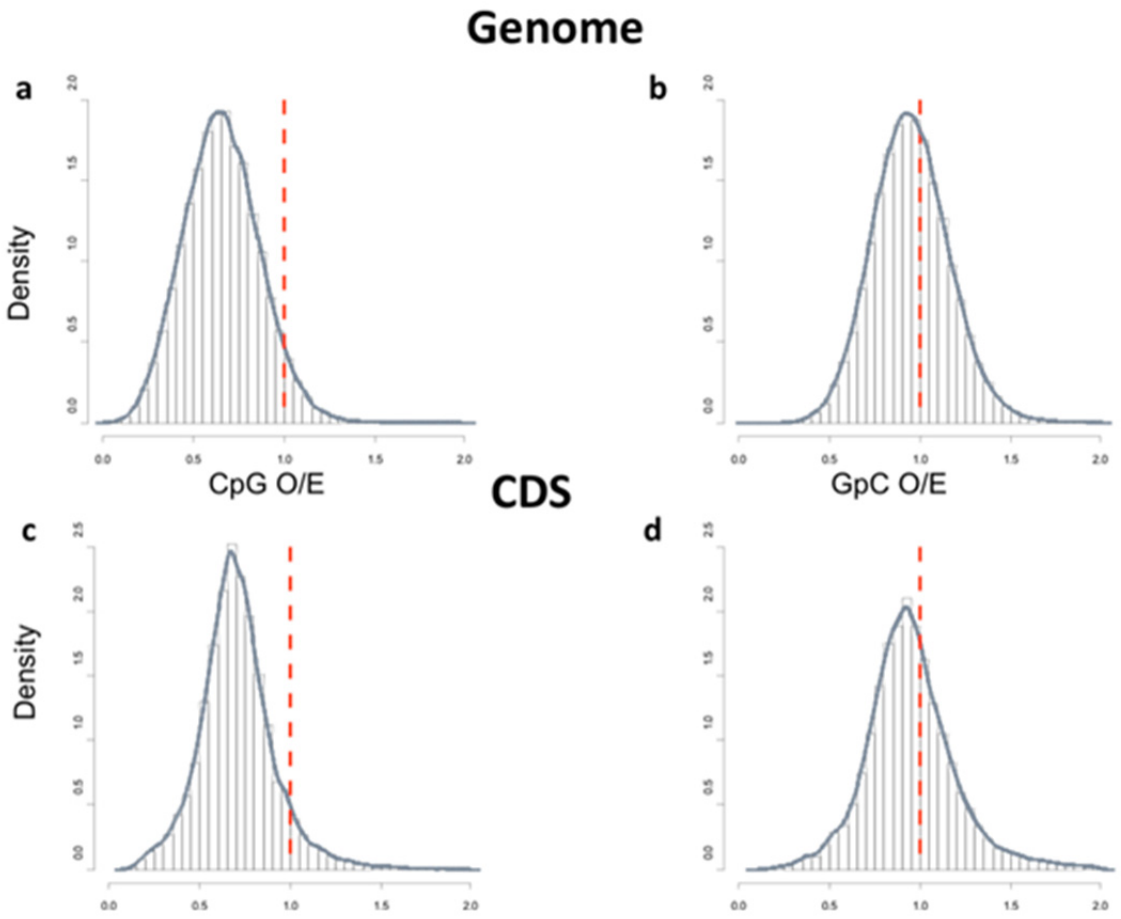


Figure S12.2.1. Distributions of CpG O/E and GpC O/E in *T. urticae* genome. The red dashed line indicates the expected CpG observed/expected ratio (O/E) and GpC O/E if C and G nucleotides associate randomly. A) CpG O/E values from CDs show a moderate level of depletion. B) CpG O/E from random genomic fragments of 1000 bps show a similar pattern to the observed from CDs. C) and D) Distributions of GpC O/Es from CDs and random genomic fragments show that contrary to CpG O/E, GpC are observed as expected under the assumption of random association.

References

1 Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91-96 (2003).

2 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).

3 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621-628 (2008).

4 Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).

5 Kal, A. J. *et al.* Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Mol. Biol. Cell* **10**, 1859-1872 (1999).

6 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Meth.* **57**, 289-300 (1995).

7 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-29 (2000).

8 Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257-258 (2007).

9 Foissac, S. *et al.* Genome annotation in plants and fungi: EuGene as a model platform. *Curr. Bioinform.* **3**, 87-97 (2008).

10 Degroeve, S., Saeys, Y., De Baets, B., Rouze, P. & Van de Peer, Y. SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics* **21**, 1332-1338 (2005).

11 Tweedie, S. *et al.* FlyBase: enhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Res.* **37**, D555-559 (2009).

12 Richards, S. *et al.* The genome of the model beetle and pest *Tribolium castaneum*. *Nature* **452**, 949-955 (2008).

13 Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* **33**, W686-689 (2005).

14 Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100-3108 (2007).

15 Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335-1337 (2009).

16 Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851-1858 (2008).

17 Oliver, J. H. Cytogenetics of mites and ticks. *Annu. Rev. Entomol.* **22**, 407-429 (1977).

18 Maddox, P. S., Oegema, K., Desai, A. & Cheeseman, I. M. "Holo"er than thou: chromosome segregation and kinetochore function in *C. elegans*. *Chromosome Res.* **12**, 641-653 (2004).

19 Ono, T. *et al.* Differential contributions of condensin I and condensin II to mitotic chromosome architecture in vertebrate cells. *Cell* **115**, 109-121 (2003).

20 Gerlich, D., Hirota, T., Koch, B., Peters, J.-M. & Ellenberg, J. Condensin I stabilizes chromosomes mechanically through a dynamic interaction in live cells. *Curr. Biol.* **16**, 333-344 (2006).

21 Phillips, C. M. *et al.* Identification of chromosome sequence motifs that mediate meiotic pairing and synapsis in *C. elegans*. *Nat. Cell Biol.* **11**, 934-942 (2009).

22 Vitkova, M., Kral, J., Traut, W., Zrzavy, J. & Marec, F. The evolutionary origin of insect telomeric repeats, (TTAGG)n. *Chromosome Res.* **13**, 145-156 (2005).

23 Kubo, Y., Okazaki, S., Anzai, T. & Fujiwara, H. Structural and phylogenetic analysis of TRAS, telomeric repeat-specific non-LTR retrotransposon families in Lepidopteran insects. *Mol. Biol. Evol.* **18**, 848-857 (2001).

24 Bolzan, A. D. & Bianchi, M. S. Telomeres, interstitial telomeric repeat sequences, and chromosomal aberrations. *Mutat. Res.* **612**, 189-214 (2006).

25 Mohan, K. N., Rani, B. S., Kulashreshta, P. S. & Kadandale, J. S. Characterization of TTAGG telomeric repeats, their interstitial occurrence and constitutively active telomerase in the mealybug *Planococcus lilacinus* (Homoptera; Coccoidea). *Chromosoma* **120**, 165-175 (2011).

26 Regad, F., Lebas, M. & Lescure, B. Interstitial telomeric repeats within the *Arabidopsis thaliana* genome. *J. Mol. Biol.* **239**, 163-169 (1994).

27 Feschotte, C. Merlin, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences. *Mol. Biol. Evol.* **21**, 1769-1780 (2004).

28 Simonet, T. *et al.* The human TTAGGG repeat factors 1 and 2 bind to a subset of interstitial telomeric sequences and satellite repeats. *Cell Res.* **21**, 1028-1038 (2011).

29 Yang, D. *et al.* Human telomeric proteins occupy selective interstitial sites. *Cell Res.* **21**, 1013-1027 (2011).

30 Kalyanaraman, A. & Aluru, S. Efficient algorithms and software for detection of full-length LTR retrotransposons. *J. Bioinform. Comput. Biol.* **4**, 197-216 (2006).

31 Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**, i351-358 (2005).

32 Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575-1584 (2002).

33 Patel, A. A. & Steitz, J. A. Splicing double: insights from the second spliceosome. *Nat. Rev. Mol. Cell Biol.* **4**, 960-970 (2003).

34 Russell, A. G., Charette, J. M., Spencer, D. F. & Gray, M. W. An early evolutionary origin for the minor spliceosome. *Nature* **443**, 863-866 (2006).

35 Wahl, M. C., Will, C. L. & Luhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**, 701-718 (2009).

36 Lin, C. F., Mount, S. M., Jarmolowski, A. & Makalowski, W. Evolutionary dynamics of U12-type spliceosomal introns. *BMC Evol. Biol.* **10**, 47 (2010).

37 Alioto, T. S. U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res.* **35**, D110-115 (2007).

38 Lau, N. C. *et al.* Abundant primary piRNAs, endo-siRNAs, and microRNAs in a *Drosophila* ovary cell line. *Genome Res.* **19**, 1776-1785 (2009).

39 Malone, C. D. & Hannon, G. J. Small RNAs as guardians of the genome. *Cell* **136**, 656-668 (2009).

40 Kasschau, K. D. *et al.* Genome-wide profiling and analysis of *Arabidopsis* siRNAs. *PLoS Biol.* **5**, e57 (2007).

41 O'Donnell, K. A. & Boeke, J. D. Mighty Piwis defend the germline against genome intruders. *Cell* **129**, 37-44 (2007).

42 Brennecke, J. *et al.* An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* **322**, 1387-1392 (2008).

43 Hofacker, I. L., Priwitzer, B. & Stadler, P. F. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics* **20**, 186-190 (2004).

44 Bonnet, E., Wuyts, J., Rouze, P. & Van de Peer, Y. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* **20**, 2911-2917 (2004).

45 Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **39**, D152-157 (2011).

46 Ruby, J. G., Jan, C. H. & Bartel, D. P. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448**, 83-86 (2007).

47 Rehmsmeier, M., Steffen, P., Hochsmann, M. & Giegerich, R. Fast and effective prediction of microRNA/target duplexes. *RNA* **10**, 1507-1517 (2004).

48 Werren, J. H. *et al.* Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science* **327**, 343-348 (2010).

49 Colbourne, J. K. *et al.* The ecoresponsive genome of *Daphnia pulex*. *Science* **331**, 555-561 (2011).

50 Pruitt, K. D. *et al.* The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* **19**, 1316-1323 (2009).

51 Putnam, N. H. *et al.* Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**, 86-94 (2007).

52 van Dongen, S. M. *Graph Clustering by Flow Simulation* PhD thesis, Universiteit Utrecht, (2000).

53 Felsenstein, J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* **266**, 418-427 (1996).

54 Farris, J. S. Phylogenetic analysis under Dollo's law. *Syst. Zool.* **26**, 77-88 (1977).

55 De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269-1271 (2006).

56 Sanchez-Rodriguez, A., Martens, C., Engelen, K., Van de Peer, Y. & Marchal, K. The potential for pathogenicity was present in the ancestor of the Ascomycete subphylum Pezizomycotina. *BMC Evol. Biol.* **10**, 318 (2010).

57 Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258-261 (2004).

58 Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-3676 (2005).

59 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9440-9445 (2003).

60 Sturn, A., Quackenbush, J. & Trajanoski, Z. Genesis: cluster analysis of microarray data. *Bioinformatics* **18**, 207-208 (2002).

61 Kummerfeld, S. K. & Teichmann, S. A. DBD: a transcription factor prediction database. *Nucleic Acids Res.* **34**, D74-81 (2006).

62 Pfreundt, U. *et al.* FlyTF: improved annotation and enhanced functionality of the *Drosophila* transcription factor database. *Nucleic Acids Res.* **38**, D443-447 (2010).

63 Helle, W. Genetic variability of photoperiodic response in an arrhenotokous mite (*Tetranychus urticae*). *Entomol. Exp. Appl.* **11**, 101-113 (1968).

64 Cook, J. M. & Crozier, R. H. Sex determination and population biology in the Hymenoptera. *Trends Ecol. Evol.* **10**, 281-286 (1995).

65 Whiting, P. W. Multiple alleles in complementary sex determination of *Habrobracon*. *Genetics* **28**, 365-382 (1943).

66 Boudreaux, H. B. in *Proceedings of the 2nd International Congress of Acarology, Sutton Bonington, England, 19-25 July, 1967* (ed G. O. Evans) 485 (Akademiai Kiado, 1969).

67 Helle, W. & Overmeer, W. P. J. Variability in tetranychid mites. *Annu. Rev. Entomol.* **18**, 97-120 (1973).

68 Overmeer, W. P. J., van Zon, A. Q. & Helle, W. Androgenesis in spider mites. *Entomol. Exp. Appl.* **15**, 256-257 (1972).

69 Wilkins, A. S. Variation in the human genome. 12-15 June, 1995. CIBA Foundation symposium 197, London, UK. *Bioessays* **17**, 905-906 (1995).

70 Schutt, C. & Nothiger, R. Structure, function and evolution of sex-determining systems in Dipteran insects. *Development* **127**, 667-677 (2000).

71 Marin, I. & Baker, B. S. The evolutionary dynamics of sex determination. *Science* **281**, 1990-1994 (1998).

72 Raymond, C. S. *et al.* Evidence for evolutionary conservation of sex-determining genes. *Nature* **391**, 691-695 (1998).

73 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).

74 Feyereisen, R. in *Comprehensive Molecular Insect Science* Vol. 4 (eds I. G. Lawrence, I. Kostas, & S. G. Sarjeet) 1-77 (Elsevier, 2005).

75 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797 (2004).

76 Feyereisen, R. Evolution of insect P450. *Biochem. Soc. Trans.* **34**, 1252-1255 (2006).

77 Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104-2105 (2005).

78 Jobb, G., von Haeseler, A. & Strimmer, K. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol. Biol.* **4**, 18 (2004).

79 Hayes, J. D., Flanagan, J. U. & Jowsey, I. R. Glutathione transferases. *Annu. Rev. Pharmacol. Toxicol.* **45**, 51-88 (2005).

80 Oakeshott, J. G. *et al.* Metabolic enzymes associated with xenobiotic and chemosensory responses in *Nasonia vitripennis*. *Insect Mol. Biol.* **19 Suppl 1**, 147-163 (2010).

81 Claudianos, C. *et al.* A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. *Insect Mol. Biol.* **15**, 615-636 (2006).

82 Yu, Q. Y., Lu, C., Li, W. L., Xiang, Z. H. & Zhang, Z. Annotation and expression of carboxylesterases in the silkworm, *Bombyx mori*. *BMC Genomics* **10**, 553 (2009).

83 Sturm, A., Cunningham, P. & Dean, M. The ABC transporter gene family of *Daphnia pulex*. *BMC Genomics* **10**, 170 (2009).

84 Kerr, I. D. Structure and association of ATP-binding cassette transporter nucleotide-binding domains. *Biochim. Biophys. Acta* **1561**, 47-64 (2002).

85 Gasteiger, E. *et al.* ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784-3788 (2003).

86 Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596-1599 (2007).

87 Roth, C. W. *et al.* Identification of the *Anopheles gambiae* ATP-binding cassette transporter superfamily genes. *Mol. Cells* **15**, 150-158 (2003).

88 Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540-552 (2000).

89 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696-704 (2003).

90 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

91 Pascual, N., Belles, X., Delbecque, J. P., Hua, Y. J. & Koolman, J. Quantification of ecdysteroids by immunoassay: comparison of enzyme immunoassay and radioimmunoassay. *Z. Naturforsch. C* **50**, 862-867 (1995).

92 Li, Y. *et al.* Profiling of ecdysteroids in complex biological samples using liquid chromatography/ion trap mass spectrometry. *Rapid Commun. Mass Spectrom.* **20**, 185-192 (2006).

93 Noriega, F. G. *et al.* Comparative genomics of insect juvenile hormone biosynthesis. *Insect Biochem. Mol. Biol.* **36**, 366-374 (2006).

94 Helvig, C., Koener, J. F., Unnithan, G. C. & Feyereisen, R. CYP15A1, the cytochrome P450 that catalyzes epoxidation of methyl farnesoate to juvenile hormone III in cockroach corpora allata. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4024-4029 (2004).

95 King-Jones, K., Horner, M. A., Lam, G. & Thummel, C. S. The DHR96 nuclear receptor regulates xenobiotic responses in *Drosophila*. *Cell Metab.* **4**, 37-48 (2006).

96 Veenstra, J. A. Neurohormones and neuropeptides encoded by the genome of *Lottia gigantea*, with reference to other mollusks and insects. *Gen. Comp. Endocrinol.* **167**, 86-103 (2010).

97 Veenstra, J. A. Neuropeptide evolution: neurohormones and neuropeptides predicted from the genomes of *Capitella teleta* and *Helobdella robusta*. *Gen. Comp. Endocrinol.* **171**, 160-175 (2011).

98 de Rosa, R. *et al.* Hox genes in brachiopods and priapulids and protostome evolution. *Nature* **399**, 772-776 (1999).

99 Katoh, K. & Toh, H. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* **26**, 1899-1900 (2010).

100 Dearden, P. K., Donly, C. & Grbic, M. Expression of pair-rule gene homologues in a chelicerate: early patterning of the two-spotted spider mite *Tetranychus urticae*. *Development* **129**, 5461-5472 (2002).

101 Omenetto, F. G. & Kaplan, D. L. New opportunities for an ancient material. *Science* **329**, 528-531 (2010).

102 Hutter, J. L. & Bechhoefer, J. Calibration of atomic-force microscope tips. *Rev. Sci. Instrum.* **64**, 1868-1873 (1993).

103 Butt, H. J. & Jaschke, M. Calculation of thermal noise in atomic force microscopy. *Nanotechnology* **6**, 1 (1995).

104 Walters, D. A. *et al.* Short cantilevers for atomic force microscopy. *Rev. Sci. Instrum.* **67**, 3583-3590 (1996).

105 Hutter, J. L. Comment on tilt of atomic force microscope cantilevers: effect on spring constant and adhesion measurements. *Langmuir* **21**, 2630-2632 (2005).

106 Heidelberg, A. *et al.* A generalized description of the elastic properties of nanowires. *Nano Lett.* **6**, 1101-1106 (2006).

107 Poinar, G. & Poinar, R. Parasites and pathogens of mites. *Annu. Rev. Entomol.* **43**, 449-469 (1998).

108 Charroux, B. & Royet, J. *Drosophila* immune response: from systemic antimicrobial peptide production in fat body cells to local defense in the intestinal tract. *Fly (Austin)* **4**, 40-47 (2010).

109 McTaggart, S. J., Conlon, C., Colbourne, J. K., Blaxter, M. L. & Little, T. J. The components of the *Daphnia pulex* immune system as revealed by complete genome sequencing. *BMC Genomics* **10**, 175-175 (2009).

110 Gerardo, N. M. *et al.* Immunity and other defenses in pea aphids, *Acyrthosiphon pisum*. *Genome Biol.* **11**, R21 (2010).

111 Iwanaga, S. The molecular basis of innate immunity in the horseshoe crab. *Curr. Opin. Immunol.* **14**, 87-95 (2002).

112 Burmester, T. Origin and evolution of arthropod hemocyanins and related proteins. *J. Comp. Physiol. B, Biochem. Syst. Environ. Physiol.* **172**, 95-107 (2002).

113 Evans, J. D. *et al.* Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Mol. Biol.* **15**, 645-656 (2006).

114 Loker, E. S., Adema, C. M., Zhang, S. M. & Kepler, T. B. Invertebrate immune systems - not homogeneous, not simple, not well understood. *Immunol. Rev.* **198**, 10-24 (2004).

115 Carthew, R. W. & Sontheimer, E. J. Origins and mechanisms of miRNAs and siRNAs. *Cell* **136**, 642-655 (2009).

116 Yigit, E. *et al.* Analysis of the *C. elegans* Argonaute family reveals that distinct Argonautes act sequentially during RNAi. *Cell* **127**, 747-757 (2006).

117 Tomari, Y. & Zamore, P. D. Perspective: machines for RNAi. *Genes Dev.* **19**, 517-529 (2005).

118 Lipardi, C. & Paterson, B. M. Identification of an RNA-dependent RNA polymerase in *Drosophila* involved in RNAi and transposon suppression. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 15645-15650 (2009).

119 Cerutti, H. & Casas-Mollano, J. A. On the origin and functions of RNA-mediated silencing: from protists to man. *Curr. Genet.* **50**, 81-99 (2006).

120 Khila, A. & Grbic, M. Gene silencing in the spider mite *Tetranychus urticae*: dsRNA and siRNA parental silencing of the *Distal-less* gene. *Dev. Genes Evol.* **217**, 241-251 (2007).

121 Panganiban, G., Sebring, A., Nagy, L. & Carroll, S. The development of crustacean limbs and the evolution of arthropods. *Science* **270**, 1363-1366 (1995).

122 Angelini, D. R. & Kaufman, T. C. Insect appendages and comparative ontogenetics. *Dev. Biol.* **286**, 57-77 (2005).

123 Willis, J. H., Iconomidou, V. A., Smith, R. F. & Hamodrakas, S. J. in *Comprehensive Molecular Insect Science* Vol. 4 (eds I. G. Lawrence, I. Kostas, & S. G. Sarjeet) 79-109 (Elsevier, 2005).

124 Cornman, R. S. *et al.* Annotation and analysis of a large cuticular protein family with the R&R Consensus in *Anopheles gambiae*. *BMC Genomics* **9**, 22 (2008).

125 Bestor, T. H. The DNA methyltransferases of mammals. *Hum. Mol. Genet.* **9**, 2395-2402 (2000).

126 Jurkowski, T. P. *et al.* Human DNMT2 methylates tRNA(Asp) molecules using a DNA methyltransferase-like catalytic mechanism. *RNA* **14**, 1663-1670 (2008).

127 Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425 (1987).

128 Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465-476 (2008).

129 Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499-1504 (1980).

130 Fryxell, K. J. & Moon, W. J. CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.* **22**, 650-658 (2005).

131 Elango, N., Kim, S. H., Vigoda, E. & Yi, S. V. Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. *PLoS Comput. Biol.* **4**, e1000015 (2008).

132 Wang, Y. *et al.* Functional CpG methylation system in a social insect. *Science* **314**, 645-647 (2006).

133 Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916-919 (2010).

134 Feyereisen, R. in *Insect Molecular Biology and Biochemistry* (ed L. I. Gilbert) Ch. 8, 236-316 (Elsevier, 2011).
135 Elices, M., Perez-Rigueiro, J., Plaza, G. R. & Guinea, G. V. Finding inspiration in *Argiope trifasciata* spider silk fibers. *JOM* **57**, 60-66 (2005).
136 Virgin, W. P. & Wakeham, W. Cotton quality and fiber properties. Part IV. The relation between single fiber properties and the behavior of bundles, slivers, and yarns. *Text. Res. J.* **26**, 177-191 (1956).
137 Regier, J. C. *et al.* Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* **463**, 1079-1083 (2010).