www.nature.com/scientificdata

scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the spotted alfalfa aphid *Therioaphis trifolii*

Tianyu Huang[1,2,3], Yang Liu[1], Kang He[4], Frédéric Francis[3], Bing Wang[1] ✉ & Guirong Wang[1,2] ✉

The spotted alfalfa aphid (SAA, *Therioaphis trifolii*) (Hemiptera: Aphididae) is a destructive pest of cultivated alfalfa (*Medicago sativa* L.) that leads to large financial losses in the livestock industry around the world. Here, we present a chromosome-scale genome assembly of *T. trifolii*, the first genome assembly for the aphid subfamily Calaphidinae. Using PacBio long-read sequencing, Illumina sequencing, and Hi-C scaffolding techniques, a 541.26 Mb genome was generated, with 90.01% of the assembly anchored into eight scaffolds, and the contig and scaffold N50 are 2.54 Mb and 44.77 Mb, respectively. BUSCO assessment showed a completeness score of 96.6%. A total of 13,684 protein-coding genes were predicted. The high-quality genome assembly of *T. trifolii* not only provides a genomic resource for the more complete analysis of aphid evolution, but also provides insights into the ecological adaptation and insecticide resistance of *T. trifolii*.

Background & Summary

Alfalfa (*Medicago sativa* L.), also called lucerne, is one of the world's most important cultivated fodder plants. It is cultivated in at least 80 countries, and because it is an abundant and stable source of nutrients, it has become the backbone of the global livestock industry[1–3]. The spotted alfalfa aphid (SAA), *Therioaphis trifolii*, is one of the most serious insect pests of legumes, mainly causing the wide-scale destruction of alfalfa crops[4]. *T. trifolii* was first recorded in New Mexico in the United States of America[5], and it also occurs in many regions of Australia, China, Europe, India, the Middle East, and the Mediterranean[5–7]. SAA damages its host plants by extracting nutrients from the leaves and phloem, and also by transmitting plant-pathogenic viruses, such as alfalfa mosaic virus and bean yellow mosaic virus[8], thereby severely restricting the growth of plants and causing devastating losses in alfalfa production[9,10].

The intensive use of chemical insecticides is the primary means of controlling aphids on many crops; however, this approach has become more challenging because aphids possess a great capacity to overcome multiple insecticides through the evolution of resistance[11–13]. Detoxifying enzymes contribute considerably to the development of insecticide resistance in aphids. For example, the peach potato aphid *Myzus persicae* is able to generate resistance to sulfoxaflor via overexpression of many detoxification-related enzymes, including UDP-glucuronosyltransferase (UGT) and cytochrome P450 (CYP) enzymes[14], and *Aphis gossypii* overcomes sulfoxaflor through the up-regulation of ATP-binding cassette (ABC) transporter expression[15]. One practical and sustainable strategy to reduce insecticide applications is the cultivation of aphid-resistant plants[16]. Many studies have mined for specific genes that can generate durable genetic resistance to *T. trifolii* in plants[17–19]; however, the molecular mechanisms by which *T. trifolii* responds to aphid-resistant alfalfa plants still remain unclear. Studies on other aphids have revealed that many digestive proteases may be involved in overcoming the defenses of aphid-resistant plants. For example, significant changes in the expression levels of various digestion-related genes, such as serine proteases (SPs) and carboxypeptidases (CPs), have been detected in *Aphis glycines* after

[1]State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing, 100193, China. [2]Guangdong Laboratory of Lingnan Modern Agriculture, Shenzhen; Genome Analysis Laboratory of the Ministry of Agriculture; Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, 518120, China. [3]Functional and Evolutionary Entomology, Gembloux Agro-Bio Tech, University of Liège, Gembloux, 5030, Belgium. [4]Ministry of Agriculture Key Lab of Molecular Biology of Crop Pathogens and Insects, Institute of Insect Sciences, Zhejiang University, Hangzhou, China. ✉e-mail: wangbing02@caas.cn; wangguirong@caas.cn

feeding on resistant soybean[20]. The availability of a high-quality genome sequence will be considerably beneficial for gaining an improved understanding of the molecular mechanisms underlying SAA resistance to pesticides and aphid-resistant alfalfa.

Taking advantage of the feasibility of inexpensive sequencing, researchers have sequenced the genomes of many aphids[21–26], but the number of available genomes is still limited compared with the number of recorded aphids (more than 5000 species)[27]. In addition, most of the sequenced aphids belong to the subfamily Aphidinae, one large group consisting of various important pests, and only a few efforts have focused on other subfamilies[28,29]. The lack of genome sequences for other subfamilies has greatly limited our understanding of the genomic diversity and evolution of aphids. Calaphidinae is the second largest subfamily within the family Aphididae[30]; it consists of nearly 400 valid species, some of which are notorious pests damaging a distinctive range of host plants[31,32]. However, despite its importance, no reference genome is yet available for this group.

Here, we present a high-quality chromosome-level genome assembly of *T. trifolii*, generated using a combination of PacBio, Illumina, and chromatin conformation capture (Hi-C) techniques. Phylogenetic analysis was performed to determine the relationship of SAA with other members of the superfamily Aphidoidea. Moreover, annotation and comparative analyses of digestion- and detoxification-related gene families and genome synteny analyses were carried out between *T. trifolii* and other representative aphid species. Our study provides the first genome assembly for a Calaphidinae aphid, which will facilitate studies on the genome evolution of aphids and also significantly benefit efforts to control this important alfalfa pest.

Methods

Sample preparation and genomic sequencing. The *T. trifolii* colony was collected from the alfalfa fields at the Langfang Experimental Station of the Chinese Academy of Agricultural Sciences and reared on alfalfa (*Medicago sativa*) in natural light in a greenhouse maintained at 20 ± 2°C and relative humidity of 75%. We aimed to create a colony consisting entirely of asexual females, so we carefully selected a single female from the original population to establish a new colony. From this colony, we selected one offspring to generate the next colony, and we repeated this process until we obtained the fifth aphid colony, which comprised solely and steadily asexual females. This pure parthenogenetic colony was used as the sample for all the genome sequencing experiments.

For PacBio sequencing, DNA was extracted from about 200 individuals, consisting of wingless parthenogenetic female adults and nymphs. Two single-end 20-kb libraries were constructed with the PacBio SMRT (Single-Molecule Real-Time) sequencing (Pacific Biosciences). Raw reads were generated from one cell sequenced on the PacBio Sequel II platform. After quality control filtering, 118.55 Gb (~220 × coverage) of SMRT PacBio sequences were obtained, with a mean read length of 14.40 kb (N50 = 21.04 kb). For Illumina sequencing, about 200 wingless parthenogenetic female adults and nymphs were used for DNA extraction, and the library (400-bp inserts) was constructed using standard Illumina protocols and sequenced on the Illumina HiSeq X Ten platform, generating 33.73 Gb of data with 150 bp paired-end reads. To further assemble the contigs into chromosomes, we generated a Hi-C library using protocols described in a previous study[33]. Fresh tissues from about 150 individual samples (including adults and nymphs) were crosslinked with paraformaldehyde to obtain the interacting DNA segments. The cross-linked sample was digested with DpnII, and biotinylated nucleotides were used to label the ends of the restriction fragments. The library was quantified and sequenced on the Illumina Novaseq/MGI-2000 platform, and ~49.21 Gb of data with 150 bp paired-end sequencing raw reads were generated.

RNA sequencing. Total RNA was extracted from 100 adult parthenogenetic female adults using TRIzol reagent (Invitrogen, Carlsbad, CA, USA)[34] and dissolved in RNase-free water. The integrity of the RNA was assessed by 1% agarose gel electrophoresis. RNA purity and concentration were assessed using a Nanodrop ND-2000 spectrophotometer (ThermoFisher, USA). The qualified RNA was used for constructing cDNA libraries. Raw sequencing data were generated using an Illumina NovaSeq 6000 platform (Illumina, San Diego, CA, USA) with the 200 bp paired-end strategy. A total of 93,816,908 clean reads were generated with a Q30 rate exceeding 90%.

Genome assembly. The quality control of raw Illumina reads was carried out using FASTP v0.20.0[35]. Clean reads were used to construct a 17-mer frequency distribution map using JELLYFISH v2.3.0[36]. The genome size of *T. trifolii* was estimated to be 542.4 Mb based on k-mer analysis.

For contig assembly, we first used FALCON v1.8.7 (reads_cutoff: 1k,seed_cutoff: 33k)[37] for the error correction of PacBio reads. The corrected reads were assembled into the preliminary genome assembly using SMARTDENOVO v1.0 with parameters -J 3000 and -k 19[38]. To correct errors generated during the assembly process, PacBio reads were mapped to the genome using BLASR v5.1[39], and ARROW v2.2.2 was used for one round of genome polishing with default parameters. Illumina reads were also mapped to the assembly using BWA v0.7.12[40], and then four iterations of contig polishing were carried out using NEXTPOLISH v1.0.5 with default parameters[41]. A contig-level assembly with a total length of 541.26 Mb was generated, which is comparable to the estimated genome size, and the contig N50 length was 2.54 Mb (Table 1).

Hi-C scaffolding. Low-quality raw reads (quality score < 20 and shorter than 30 bp) and adaptors were removed using FASTP v0.20.0, then the clean reads were mapped to the contig assembly using BOWTIE2 v2.3.2 (-end-to-end–very-sensitive -L 30)[42]. HI-C PRO v2.8.1[43] with default parameters was used to identify valid interaction paired reads and to filter out reads with multiple hits and singleton reads. LACHESIS[44] was used to cluster, order, and orient the contigs with parameters CLUSTER MIN RE SITES = 100; CLUSTER MAX LINK DENSITY = 2.5; CLUSTER NONINFORMATIVE RATIO = 1.4; ORDER MIN N RES IN TRUNK = 60; ORDER MIN N RES IN SHREDS = 60.

Features	Statistics
Estimated genome size (bp)	542,395,090
Assembly size (bp)	541,263,359
Contigs N50 (bp)	2,544,558
Scaffolds number	575
Scaffolds N50 (bp)	44,770,504
BUSCO genes	C: 96.6% [S: 93.5%, D: 3.1%], F: 0.7%
Number of protein-coding genes	13,684

Table 1. Major indicators of the *Therioaphis trifolii* genome.

As a result, Hi-C data were combined with the contig-level assembly to generate a chromosome-level assembly comprising eight large scaffolds (Fig. 1a), which corresponds to the previously reported haploid chromosome number for this species[45]. Around 90.07% of the contigs were anchored onto chromosomes, resulting in a scaffold N50 length of 44.77 Mb (Table 1). The longest chromosome was 149.16 Mb while the shortest was 37.54 Mb (Fig. 1b).

Repeat annotation. TANDEM REPEAT FINDER v4.07b (parameters: 2 7 7 80 10 50 500 -f -d -h -r)[46] was used to identify all tandem repeat elements. Transposable elements (TEs) were identified using a combination of two methods. First, a de novo repeat library was generated using REPEATMODELER v1.0.11 and MITE-hunter[47] with default parameters. This library was searched against the Repbase[48] to classify repeat families using REPEATMASKER v1.331, and then merged with Repbase to generate the final repeat sequence library. Next, REPEATMASKER v1.331 was used to predict TEs based on the final TE library. The result showed that repeat sequences make up 36.86% of the genome, most of which are TEs (33.31%) (Table 2).

Protein coding gene prediction and functional annotation. Gene model prediction from the TE soft-masked *T. trifolii* genome was performed using multiple approaches, namely transcriptome-based prediction, ab initio prediction, and homology-based gene prediction. For transcriptome-based analysis, clean reads were aligned to the genome assembly using STAR v2.7.3a[49] with the default parameters. Next, STRINGTIE v1.3.4d[50] was used to obtain transcript locations, and open reading frames of the transcripts were predicted using PASA v2.3.3[51]. For de novo gene model prediction, the transcript set generated by PASA was utilized by GENEMARK-ST v5.1[52] for self-training. The training set was applied to AUGUSTUS v3.3.1[53] for gene model prediction. For the homology-based gene modeling process, protein sets from three aphids with high-quality genome assemblies (consisting of *A. pisum*, *R. maidis* and *S. miscanthi*) were aligned to the genome assembly via GEMOMA v1.6.1[54]. Finally, we combined the results from the three gene prediction approaches to create a consensus gene model set using EVIDENCEMODELER v1.1.1 (–segmentSize 1000000–overlapSize 100000)[51]. As a result, 13,684 protein-coding gene models were generated, with an average gene length of 15 kb, average coding sequence length of 1.5 kb, and average exon number of 7.1.

For gene functional annotation, protein sequences encoded by the predicted gene models were aligned to the non-redundant (nr), SWISS-PROT, Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000), and eukaryotic orthologous groups (KOG) databases (Galperin *et al.*, 2015) using BLASTP v2.7.1 with a cutoff of 1e-5. We also used INTERPROSCAN v5.32-71.0[55] to obtain gene ontology (GO) annotations for the proteins.

Phylogenetic and comparative genomic analyses. The longest predicted protein sequences of 12 aphid genomes, namely *Aphis glycines*[56], *Acyrthosiphon pisum*[22], *Cinara cedri*[28], *Diuraphis noxia*[57], *Eriosoma lanigerum*[29], *Myzus cerasi*[58], *Myzus perisicae*[22], *Pentalonia nigronervosa*[59], *Rhopalosiphum maidis*[24], *Rhopalosiphum padi*[58], *Sitobion miscanthi*[23], and *T. trifolii*, and the greenhouse whitefly *Trialeurodes vaporariorum*[60], which was used as an outgroup, were utilized for identifying orthologous groups among aphids using ORTHOFINDER v2.4.0[61]. A total of 2758 single-copy orthogroups were identified and used to generate a concatenated alignment for inferring phylogenetic relationships. The species tree of the 12 aphids was also inferred using ORTHOFINDER[62] and rooted by STRIDE[63]. Divergence times among aphids were calculated by R8S[64] based on divergence information extracted from TimeTree (http://www.timetree.org/): *A. pisum* vs *M. persicae* 42.5–48.0 million years ago (mya) (Fig. 2). We also used CAFE v4.2.1[65] to analyze the expansion and contraction of gene families in all 12 tested aphid lineages. The results from the phylogenetic tree with divergence times were used as inputs (Fig. 2).

Analysis of detoxification- and digestion- related genes. The amino acid sequences of each aphid were searched against the NR and SWISS-PROT databases using DIAMOND v0.9.21[66] with an e-value cut-off of 1e-5, and INTERPROSCAN was used to predict functional domains in each sequence. Genes encoding the best scoring protein hits of digestion-related and detoxification-related enzymes were annotated according to the best hit (Fig. 3).

Synteny analysis. The synteny analyses were carried out between the chromosome-level genome assemblies of *T. trifolii*, *A. pisum* (JIC1 v1), and *E. lanigerum*. To obtain syntenic blocks, we uploaded the official gene sets to the ORTHOVENN2 server[67]. The 1:1 single-copy ortholog pairs from each comparison (*T. trifolii* vs *A. pisum* and *T. trifolii* vs *E. lanigerum*) were identified using the parameters e-value = 1e-5 and inflation

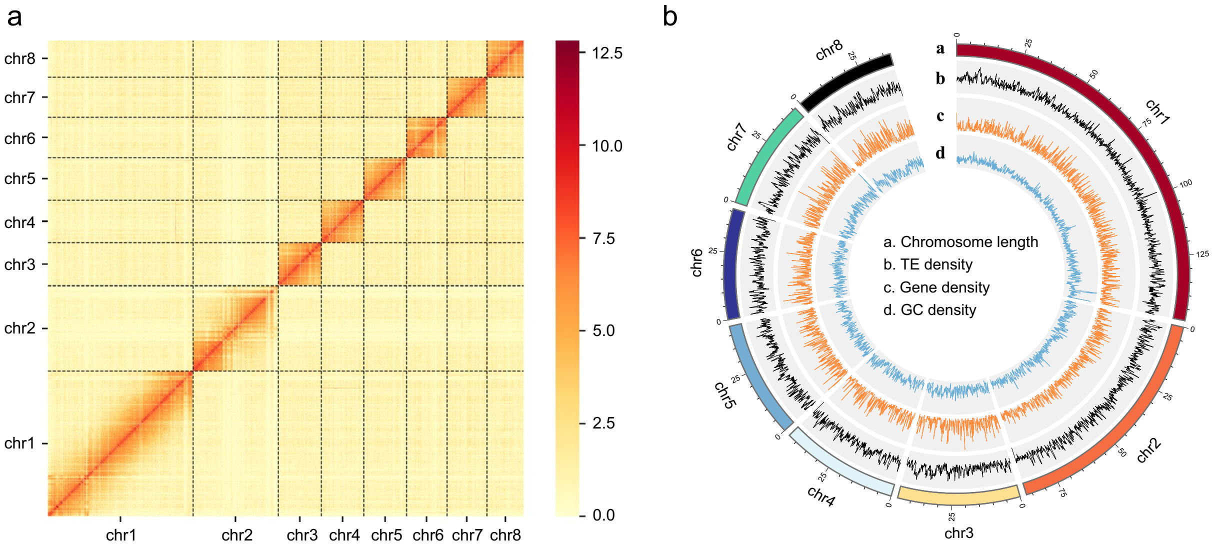


Fig. 1 Heatmap of genome-wide Hi-C data and circular representation of the chromosomes of *Therioaphis trifolii*. (**a**) The heatmap of chromosome interactions in *T. trifolii*. The frequency of Hi-C interaction links is represented by colours, which ranges from yellow (low) to red (high). (**b**) Circos plot of distribution of the genomic elements in *T. trifolii*. The tracks indicate **a**) length of the chromosome, **b**) distribution of transposable element (TE) density ranges from 11 to 502, **c**) gene density ranges from 0 to 17, and **d**) GC density ranges from 22 to 62. The densities of TEs, genes, and GC were calculated in 100 kb windows.

Repeat types	Number of elements	Length occupied (bp)	Percentage of sequence
SINE	5,290	506,439	0.09%
LINE	135,921	27,613,234	5.10%
LTR	73,841	14,343,500	2.65%
MITE	53,954	13,738,880	2.54%
DNA	716,470	121,720,828	22.49%
RC	18,203	2,394,844	0.44%
Unknown	45,373	7,137,015	1.32%
Total base masked	1,236,028	199,531,663	36.86%

Table 2. Statistics of the transposable elements in *Therioaphis trifolii* genome.

value = 1.5. These gene pairs were selected for genome synteny analyses using MCSCANX v1.1[68] with default parameters. SYNVISIO (https://synvisio.github.io) was used to visualize genome synteny (Fig. 4).

Data Records

The genome sequencing, RNA sequencing reads data has been updated to the National Center for Biotechnology Information (NCBI) as a BioProject no. PRJNA804007. Pacbio, Hi-C, Illumina and transcriptome sequencing reads have been deposited in the Sequence Read Archive (SRA) databases with the accession number of SRP359015[69]. Genome assembly has been deposited at the NCBI under the accession number of JALBXZ000000000[70], and can be download from National Genomic Data Center (NGDC) under accession number GWHBQDZ00000000.1. The annotated detoxification and digestion related genes among aphids have been uploaded to the NGDC under accession number OMIX002672. Gene sequences predicted from the genome assembly are also publicly available in NGDC, under the accession number OMIX002673. The statistics of RNA-Seq data have been deposited to the NGDC under accession number OMIX003518. All data in NGDC were related to the BioProject PRJCA014018. The TE library, sequences used for orthology analysis, the results of ORTHOFINDER and ORTHOVENN2 are available at Zenodo[71].

Technical Validation

The accuracy and completeness of the contig assembly were validated using four methods. First, clean Illumina reads were mapped to the contigs assembled by BWA v0.7.12, and the total mapped reads and mapping rate were calculated using SAMTOOLS v1.4[72], resulting in a mapping rate of 99.40%. Second, clean reads from 4 whole-body transcriptomes were mapped onto the genome assembly, more than 97% of the RNA-Seq reads can be aligned to the coding regions of the genome assembly. Third, Benchmarking Universal Single-Copy Orthologs (BUSCO) v4.0.5[73] was employed to assess the completeness of the genome assembly based on the insecta_odb10 database (-l insecta_odb10 -m genome), the BUSCO analysis indicated that 97.3% of gene

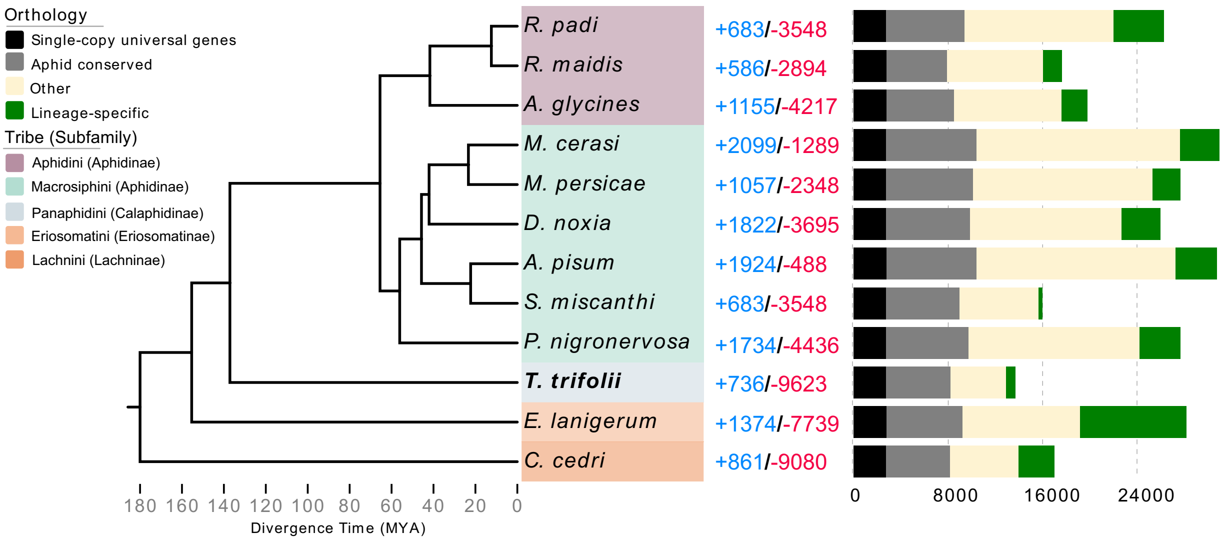
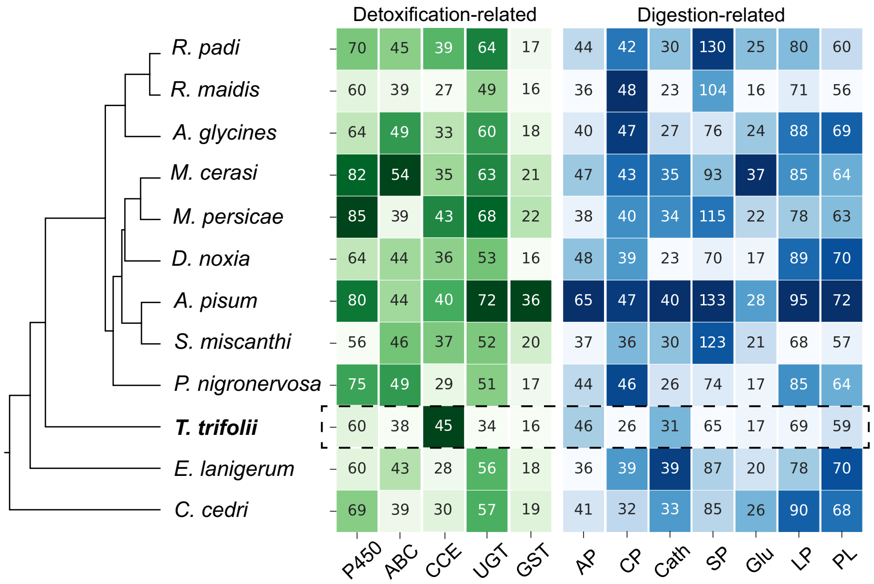


Fig. 2 Phylogeny and orthology analyses between *Therioaphis trifolii* and other aphid species. The phylogenetic tree was constructed based on 2,758 single-copy orthogroups obtained from the genomes of all tested aphids. The greenhouse whitefly *Trialeurodes vaporariorum* (not shown) was selected as the outgroup. Aphid species are colored according to their tribe. Numbers of expanded (blue) and contracted (red) gene families are presented alongside the species and nodes. The bar chart shows the comparison of orthologs between 12 aphids. 'Single-copy universal' refers to a single-copy ortholog that is present in all the aphids. 'Aphid conserved' refers to genes that can be detected in at least 11 aphid genomes. 'Lineage-specific' refers to genes that do not have an ortholog in any other aphid. 'Other' indicates orthologs are found in some of the aphids (e.g., in 1 to 10 aphids).



	Detoxification-related					Digestion-related						
	P450	ABC	CCE	UGT	GST	AP	CP	Cath	SP	Glu	LP	PL
R. padi	70	45	39	64	17	44	42	30	130	25	80	60
R. maidis	60	39	27	49	16	36	48	23	104	16	71	56
A. glycines	64	49	33	60	18	40	47	27	76	24	88	69
M. cerasi	82	54	35	63	21	47	43	35	93	37	85	64
M. persicae	85	39	43	68	22	38	40	34	115	22	78	63
D. noxia	64	44	36	53	16	48	39	23	70	17	89	70
A. pisum	80	44	40	72	36	65	47	40	133	28	95	72
S. miscanthi	56	46	37	52	20	37	36	30	123	21	68	57
P. nigronervosa	75	49	29	51	17	44	46	26	74	17	85	64
T. trifolii	60	38	45	34	16	46	26	31	65	17	69	59
E. lanigerum	60	43	28	56	18	36	39	39	87	20	78	70
C. cedri	69	39	30	57	19	41	32	33	85	26	90	68

Fig. 3 Detoxification- and digestion-related gene families in *Therioaphis trifolii*. The maximum likelihood phylogeny, based on a concatenated alignment of 2,758 single-copy orthogroups, illustrates the phylogenomic relationship between *T. trifolii* and other aphids. Numbers in the heatmaps indicate the size of the corresponding gene family in aphid species. Detoxification-related and digestion-related gene families are labeled with a green and blue shade, respectively.The darker shade of color indicates a higher number of identified genes.

orthologs were identified in *T. trifolii*, including complete and fragment scores of 96.6% and 0.7%, respectively. Finally, CEGMA v2[74] with default parameters was used to validate the integrity of the core genes in the assembly, 242 core eukaryotic genes were assembled, among which 94.76% were complete.

To ensure the completeness of the annotated gene set, four validation methods were employed. Firstly, the annotation was subjected to BUSCO analysis using the insecta_odb10 database (-l insecta_odb10 -m prot). The results indicated that 95.98% of the conserved single copy ortholog genes, including 95.39% of complete genes

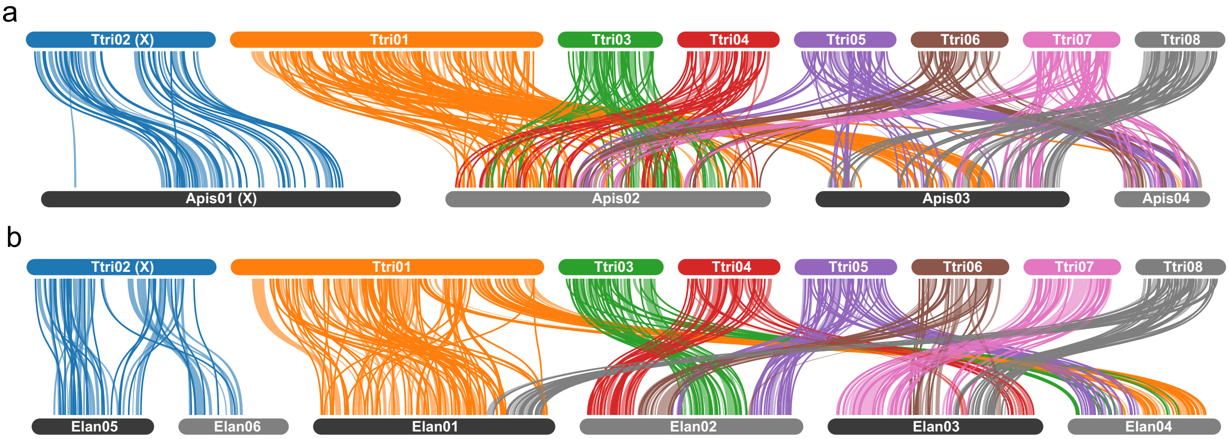


Fig. 4 Genome synteny between (**a**) *Therioaphis trifolii* (Calaphidinae) and *Acythosiphon pisum* (Aphidinae) and (**b**) *T. trifolii* and *Eriosoma lanigerum* (Eriosomatinae). Links indicate the edges of syntenic blocks of gene pairs identified by synteny analyses and are shown in the same color as that of the chromosome ID of *T. trifolii*. Ttri indicates *T. trifolii*, Apis indicates *A. pisum*, and Elan indicates *E. lanigerum*.

and 0.59% of fragmented genes, were present in the annotated protein set. Secondly, gene expression analysis was conducted using RNA-Seq reads from four whole-body transcriptomes. The analysis revealed that 11,422 (83.47%) annotated genes were expressed in at least one transcriptomic sample, and 72.91%~74.32% of the RNA-seq reads could be assigned onto the coding region of the genome assembly. Thirdly, to assess the completeness of annotated gene structures, the length ratio of predicted proteins to their best hit in the proteomes of three high-quality aphid genome assemblies (*A. pisum*, *M. persicae*, *and R. maidis*) was analyzed. Predicted proteins with a ratio of 0.9–1.1 were considered high confidence predictions. The results showed that a large number of high confidence predictions were obtained, including 8,850 (64.7%), 8,876 (64.9%), and 8,611 (62.9%) from the *T. trifolii* versus *A. pisum*, *T. trifolii* versus *M. persicae*, and *T. trifolii* versus *R. maidis* comparisons, respectively. Finally, the predicted gene models were compared against several protein databases (nr, SWISS-PROT, GO, KOG, and KEGG). The results showed that 12,995 (94.96%) of the predicted gene models had significant homology to proteins in at least one of these databases.

Code availability

All software and pipelines used for data processing were executed according to the manuals and protocols of the bioinformatics software cited above, and the parameters are clearly described in the Methods section. If no detailed parameters are mentioned for a software, the default parameters were used. The version of the software has been described in Methods.

Received: 9 January 2023; Accepted: 24 April 2023;
Published online: 12 May 2023

References

1. Radović, J., Sokolović, D. & Marković, J. Alfalfa-most important perennial forage legume in animal husbandry. *Biotechnol. Anim. Husb.* **25**, 465–475 (2009).
2. Frank, D. *et al.* Impact of brassica and lucerne finishing feeds and intramuscular fat on lamb eating quality and flavor. A cross-cultural study using Chinese and non-Chinese Australian consumers. *J. Agric. Food Chem.* **64**, 6856–6868 (2016).
3. Bai, Z. *et al.* China's livestock transition: Driving forces, impacts, and consequences. *Sci. Adv.* **4**, eaar8534 (2018).
4. Blackman, R. L. & Eastop, V. F. *Aphids on the world's crops* (John Wiley and Sons, Chichester, 2000).
5. Dickson, R., Laird, E. & Pesho, G. The spotted alfalfa aphid (yellow clover aphid on alfalfa). *Hilgardia* **24**, 93–118 (1955).
6. Lake, A. Spotted alfalfa aphid survival and reproduction on annual medics with various levels of aphid resistance. *Aust. J. Agric. Res.* **40**, 117–123 (1989).
7. Wang, L. *et al.* Forage yield, water use efficiency, and soil fertility response to alfalfa growing age in the semiarid Loess Plateau of China. *Agric. Water Manag.* **243**, (2020).
8. Jones, R. A. C. Occurrence of virus infection in seed stocks and 3-year-old pastures of lucerne (Medicago sativa). *Aust. J. Agric. Res.* **55**, 757–764 (2004).
9. He, C. G. & Zhang, X. G. Field evaluation of lucerne (*Medicago sativa* L.) for resistance to aphids in northern China. *Aust. J. Agric. Res.* **57**, 471–475 (2006).
10. Irwin, J. A. G., Lloyd, D. L. & Lowe, K. F. Lucerne biology and genetic improvement - an analysis of past activities and future goals in Australia. *Aust. J. Agric. Res.* **52**, 699–712 (2001).
11. Bass, C. *et al.* The evolution of insecticide resistance in the peach potato aphid, *Myzus persicae*. *Insect Biochem. Mol. Biol.* **51**, 41–51 (2014).
12. Lokeshwari, D., Krishna Kumar, N. K. & Manjunatha, H. Multiple mutations on the second acetylcholinesterase gene associated with dimethoate resistance in the melon aphid, *Aphis gossypii* (Hemiptera: Aphididae). *J. Econ. Entomol.* **109**, 887–897 (2016).
13. Chen, A., Zhang, H., Shan, T., Shi, X. & Gao, X. The overexpression of three cytochrome P450 genes CYP6CY14, CYP6CY22 and CYP6UN1 contributed to metabolic resistance to dinotefuran in melon/cotton aphid, *Aphis gossypii* Glover. *Pestic. Biochem. Physiol.* **167**, 104601 (2020).
14. Pym, A. *et al.* Overexpression of UDP-glucuronosyltransferase and cytochrome P450 enzymes confers resistance to sulfoxaflor in field populations of the aphid, *Myzus persicae*. *Insect Biochem. Mol. Biol.* **143**, 103743 (2022).

15. Wang, L. *et al*. Overexpression of ATP-binding cassette transporters associated with sulfoxaflor resistance in *Aphis gossypii* glover. *Pest Manag. Sci.* **77**, 4064–4072 (2021).
16. Smith, C. M. & Chuang, W. P. Plant resistance to aphid feeding: behavioral, physiological, genetic and molecular cues regulate aphid host selection and feeding. *Pest Manag. Sci.* **70**, 528–540 (2014).
17. Kamphuis, L. G. *et al*. Characterization and genetic dissection of resistance to spotted alfalfa aphid (*Therioaphis trifolii*) in *Medicago truncatula*. *J. Exp. Bot.* **64**, 5157–5172 (2013).
18. Jacques, S. *et al*. A functional genomics approach to dissect spotted alfalfa aphid resistance in *Medicago truncatula*. *Sci. Rep.* **10**, 22159 (2020).
19. Zhao, H. *et al*. Inhibitory effects of plant trypsin inhibitors Msti-94 and Msti-16 on *Therioaphis trifolii* (Monell) (Homoptera: Aphididae) in alfalfa. *Insects* **10**, 154 (2019).
20. Bansal, R., Mian, M. A., Mittapalli, O. & Michel, A. P. RNA-Seq reveals a xenobiotic stress response in the soybean aphid, *Aphis glycines*, when fed aphid-resistant soybean. *BMC Genomics* **15**, 972 (2014).
21. Li, Y., Park, H., Smith, T. E. & Moran, N. A. Gene family evolution in the pea aphid based on chromosome-level genome assembly. *Mol. Biol. Evol.* **36**, 2143–2156 (2019).
22. Mathers, T. C. *et al*. Chromosome-scale genome assemblies of aphids reveal extensively rearranged autosomes and long-term conservation of the X chromosome. *Mol. Biol. Evol.* **38**, 856–875 (2021).
23. Jiang, X. *et al*. A chromosome-level draft genome of the grain aphid *Sitobion miscanthi*. *Gigascience* **8**, giz101 (2019).
24. Chen, W. *et al*. Genome sequence of the corn leaf aphid (*Rhopalosiphum maidis* Fitch). *Gigascience* **8**, giz033 (2019).
25. Wenger, J. A. *et al*. Whole genome sequence of the soybean aphid, *Aphis glycines*. *Insect Biochem. Mol. Biol.* **123**, 102917 (2020).
26. International Aphid Genomics, C. Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol.* **8**, e1000313 (2010).
27. Emden, H. F. V. & Harrington, R. *Aphids as crop pests*. (CAB International, 2017).
28. Julca, I. *et al*. Phylogenomics identifies an ancestral burst of gene duplications predating the diversification of aphidomorpha. *Mol. Biol. Evol.* **37**, 730–756 (2020).
29. Biello, R. *et al*. A chromosome-level genome assembly of the woolly apple aphid, *Eriosoma lanigerum* Hausmann (Hemiptera: Aphididae). *Mol. Ecol. Resour.* **21**, 316–326 (2021).
30. Favret, C. *Aphid species file* http://Aphid.SpeciesFile.org (2013).
31. Nebreda, M. *et al*. Activity of aphids associated with lettuce and broccoli in Spain and their efficiency as vectors of Lettuce mosaic virus. *Virus Res.* **100**, 83–88 (2004).
32. Herbert, J. J., Mizell, R. F. 3rd & McAuslane, H. J. Host preference of the crapemyrtle aphid (Hemiptera: Aphididae) and host suitability of crapemyrtle cultivars. *Environ. Entomol.* **38**, 1155–1160 (2009).
33. Liu, Y. *et al*. *Apolygus lucorum* genome provides insights into omnivorousness and mesophyll feeding. *Mol. Ecol. Resour.* **21**, 287–300 (2020).
34. Huang, T. *et al*. Identification and functional characterization of ApisOr23 in pea aphid *Acyrthosiphon pisum*. *J. Integr. Agric.* **21**, 1414–1423 (2022).
35. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
36. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
37. Chin, C. S. *et al*. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
38. Liu, H., Wu, S., Li, A. & Ruan, J. SMARTdenovo: a *de novo* assembler using long noisy reads. *Gigabyte* **2021**, gigabyte15 (2021).
39. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
41. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
42. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
43. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
44. Burton, J. N. *et al*. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol* **31**, 1119–1125 (2013).
45. Sunnucks, P. *et al*. Biological and genetic characterization of morphologically similar Therioaphis trifolii (Hemiptera: Aphididae) with different host utilization. *Bull. Entomol. Res.* **87**, 425–436 (1997).
46. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
47. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199 (2010).
48. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
49. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
50. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
51. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
52. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, e78 (2015).
53. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
54. Keilwagen, J. *et al*. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, e89 (2016).
55. Zdobnov, E. M. & Apweiler, R. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
56. Mathers, T. C. Improved Genome Assembly and Annotation of the Soybean Aphid (*Aphis glycines* Matsumura). *G3 (Bethesda)* **10**, 899–906 (2020).
57. Nicholson, S. J. *et al*. The genome of *Diuraphis noxia*, a global aphid pest of small grains. *BMC Genomics* **16**, 429 (2015).
58. Thorpe, P., Escudero-Martinez, C. M., Cock, P. J. A., Eves-van den Akker, S. & Bos, J. I. B. Shared transcriptional control and disparate gain and loss of aphid parasitism genes. *Genome Biol. Evol.* **10**, 2716–2733 (2018).
59. Mathers, T. C., Mugford, S. T., Hogenhout, S. A. & Tripathi, L. Genome sequence of the banana Aphid, *Pentalonia nigronervosa* Coquerel (Hemiptera: Aphididae) and its symbionts. *G3 (Bethesda)* **10**, 4315–4321 (2020).
60. Xie, W., He, C., Fei, Z. & Zhang, Y. Chromosome-level genome assembly of the greenhouse whitefly (*Trialeurodes vaporariorum* Westwood). *Mol. Ecol. Resour.* **20**, 995–1006 (2020).
61. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
62. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
63. Emms, D. M. & Kelly, S. STRIDE: species tree root inference from gene duplication events. *Mol. Biol. Evol.* **34**, 3267–3278 (2017).
64. Sanderson, M. J. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* **19**, 101–109 (2002).

65. 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).
66. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
67. Xu, L. *et al*. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* **47**, w52–w58 (2019).
68. Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
69. *NCBI Sequence Read Archive* https://identifiers.org/insdc.sra:SRP359015 (2022).
70. Huang, T. *Therioaphis trifolii* isolate LF-2019, whole genome shotgun sequencing project. *Genbank* https://identifiers.org/nucleotide:JALBXZ000000000 (2022).
71. Huang, T. *et al*. Supplymentary data for chromosome-level genome assembly of the spotted alfalfa aphid. *Therioaphis trifolii*. *Zenodo* https://doi.org/10.5281/zenodo.7700460 (2023).
72. Li, H. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
73. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
74. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).

Acknowledgements

This study was supported by a grant from Shenzhen Science and Technology Program (Grant No. KQTD20180411143628272) and the National Natural Science Foundation of China (31872039), Major special projects for green pest control (110202201017(LS-01)), and the Central Public-interest Scientific Institution Basal Research Fund (CAASZDRW202108) and the Agricultural Science and Technology Innovation Program.

Author contributions

B.W., G.W. and F.F. conceived this study. T.H. prepared DNA and RNA for sequencing. T.H. performed the experiments and analyzed the data. T.H., K.H., Y.L., B.W., F.F. and G.W. wrote the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.W. or G.W.

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

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