Chromosome-level genome assembly of *Hestina assimilis* (Lepidoptera: Nymphalidae) provides sights to the saprophagy in brush-footed butterfly

lu zhao

Shaanxi Normal University

Hang Wang

Shaanxi Normal University

Zhicuo Dan

Shaanxi Normal University

Sheng-Quan Xu

Shaanxi Normal University

De-Long Guan (✉ guandelong@snnu.edu.cn)

Shaanxi Normal University

Research Article

Keywords: chromosome-level genome assembly, comparative genomic analysis, *Hestina assimilis*, Nymphalidae, saprophagy

Posted Date: February 18th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1358422/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: While most adult butterflies are nectarivores, some brush-footed species evolved saprophagy. High-quality genomes will provide a backbone for untangling this unique feeding strategy's genetic mechanism and evolutionary history.

Results: We assembled a chromosome-level genome for a typical saprophagous butterfly species, *Hestina assimilis* (Lepidoptera: Nymphalidae). The assembled genome was 423.87 Mb long with the scaffold N50 of 15.84 Mb, and 96.2% scaffolds were anchored onto 30 chromosomes. The BUSCO (version 5.2.2, lineage insecta) completeness was 99.3%. We predicted 16,815 coding genes and found that 46.37% of the genome were repetitive sequences. The populations of *H. assimilis* increased dramatically after the last interglacial period of the Pleistocene. Comparative genomics analysis showed that *H. assimilis* clustered with another saprophagous species, *Vanessa tameamea*, and they diverged from a nectarivorous butterfly species, *Heliconius erato*, approximately 78.4 million years ago, suggesting that the transition to saprophagy was an ancient evolutionary event. The unique and significantly expanded gene families were relevant to feeding strategies and food intake functions, such as digestion, detoxification, chemoreception, immunity, nerves and development. The expansion events were also involved in processes of acid chemical response and alcohol biosynthesis. Furthermore, we detected 23 odorant-binding protein (OBP) genes—two key antennal OBP genes, PBP-A and PBP-B, were absent.

Conclusions: Our results provided insights into the genetic mechanism of feeding habits as well as provides an important resource for future butterfly genome sequencing projects.

Background

Most adult butterflies are flower visitors and nectarivorous with their strong dependence on flower nectar, but a considerable number of butterflies forage for various non-nectar foods, such as rotting fruit, honeydew, tree sap, mud, carrion, and dung [1-3]. These species have significant morphological variation in their proboscis, which appears to be related to their specialization and adaptation with respect to the physical properties of non-nectar foods [4]. Compared with nectar foods, non-nectar foods are significantly different from flower nectar in their chemical properties, including low sugar concentrations and the presence of fermentation products (ethanol and acetic acid) [1, 5]. Saprophagous butterfly could adapt to the compositional characteristics of such rotting foods and showed feeding responsiveness to the low sugars with the fermentative products [1, 6]. Although such feeding habits are intriguing from the ecological and evolutionary aspects of food utilisation in adult butterflies, the saprophagous mechanism at the genetic level is still not clarified.

Hestina assimilis Linnaeus, 1758 (Lepidoptera) is a larger species of Nymphidae with strong adaptability and wide distribution, which occurs in all provinces of China and mostly in forest areas [7]. It flies quickly and has a beautiful appearance. It is marked by 4-5 Red spots on the outer edge of its rear wing, so it has another name Red Ring skirt [8]. It is a typical saprophagous butterfly and the adults could utilize the

diluted liquid of humus such as tree sap, rotting fruits even feces. Due to this trait, *H. assimilis* have increased suitability, with their successful range expansion and growth population [9]. Therefore, *H. assimilis* can be used as ideal model for exploring the mechanisms of saprophagy [10]. However, the absence of genomic information about *H. assimilis* has limited further study into the molecular mechanisms of its eating habits and evolutionary characteristics [11].

With the development of new sequencing technologies, the increasing species' genome has been acquired [12-16]. Deep sequencing data can provide comprehensive information about genomes and gene expression profiles [17]. High-quality genomes of some butterflies have provided new insights into the molecular mechanisms of their ecological and evolutionary characteristics [18-23]. For instance, the analyses of the pigmentary genes based on the genome of *Kallima inachus*, revealed the mechanism underlying protective resemblance and insights into the molecular basis of protective coloration [24]. *Heliconius melpomene* genome reveals promiscuous exchange of mimicry adaptations among species [25]. The *Danaus plexippus* genome enhances our ability to better understand the genetic and molecular basis of long-distance migration [26].

To unravel the genomic signatures behind the saprophagy in butterfly, we first performed a high-quality chromosome-level reference genome assembly of *H. assimilis*. Then, we conducted a genomic comparative approach and performed a phylogenomic reconstruction to identify gene families related to feeding dietary. Finally, we identified the key genes associated with the diet.

Results

Genome sequencing and de novo assembly

We obtained 62.39 Gb (150×) of nanopore long-reads and 49.65 Gb (120×) of Illumina paired-end reads to construct the contig-level genome for this butterfly. We first used nanopore long-reads for de novo assembly. Due to the high error rate of the nanopore reads, the assembly was then polished three times using Illumina short reads. Our result indicate that 95.27% Illumina reads were mapped to the initially assembled genome with a near normal distribution. For completeness assessment, the Nanopore reads were mapped back to the assembly. After above steps, a 423.87 Mb assembly (scaffold:53, N50: 14.32 Mb) was yielded, which is consistent with the genome size estimated based on the k-mer analysis (421.41 Mb) (Fig. S1). The BUSCO completeness value in assembly was 99.1%.

Chromosome-level assembly and syntenic analysis

Combined with 24.7 Gb (50×) of Hi-C fragment libraries sequencing, 96.2% scaffolds from *H. assimilis* were anchored onto 30 chromosomes, ranging from 1.48 Mb to 34.8 Mb in length, and 2 small unplaced scaffolds (Fig. S2). The karyotype of *H. assimilis* (n=30) was suggested. Finally, a 423.87 Mb (scaffold N50: 15.84 Mb, GC content: 35.4%) of haplotype genome assembly without heterozygous sequence was yielded. The BUSCO completeness value in the assembly was 99.3%. We also applied BUSCO analysis to other published Nymphalidae genomes using the same parameters and subsequently compared the

completeness of each genome assembly. The results showed that genome assembly integrity of *H. assimilis* is comparable to, or higher than, the others and is therefore adequate for further analysis (Table. S1, Fig. S3).

Using *D. plexippus* as reference, which is a kind of nectarivorous butterfly, synteny analysis revealed that *H. assimilis* and *D. plexippus* have highly consistent and extensive gene collinearity, which suggests that the close relationship between these two species at the gene level and the similarity of chromosome structure (Fig. 1). Compared with *D. plexippus*, *H. assimilis* has no large-scale chromosome rearrangement. Our results preliminarily excluded the influence of the chromosome rearrangement on the feeding habits of butterflies.

Synteny alignment of the *H. assimilis* and *D. plexippus*. The tagged HA01 to HA30 refer to the chromosomes of the sequenced *H. assimilis* and the black columns refers to the chromosomes of *D. plexippus*. The densities of GC contents, DNA variations in *H. assimilis* are shown in the outer rims in orders.

Gene prediction and functional annotation

Using EDTA, we identified ~ 46.37% of the assembled genome as repetitive sequences (Table S2). Among them, the repetitive sequences were dominated by DNA transposons (25.75%) and long terminal repeats (LTRs, 15.68%). By integrating the methods of ab initio, homology-based and transcriptome-based prediction, a total of 16,815 protein-coding genes were identified and the BUSCO value was 98.0% in the *H. assimilis* genome. In total, 13,747 (81.8%), 15,084 (89.7%) and 9,807 (58.3%) predicted genes were supported by functional annotation from the eggNOG-mapper, Interproscan and UniProt/SwissProt databases (Table S3-5). The average lengths of gene and CDS were 1,447 bp and 8,678 bp, which was comprehensive in the comparison of 14 other lepidoptera species (Fig. S4). Pfam motif results show that the most abundant three classes of motif were RVT_1, Exo_endo_phos_2 and rve (Fig. S5), which were important proteins closely related to transcription.

Demographic history reconstruction

H. assimilis is highly adaptable and widely distributed. This species adapts to the present distribution environment, showing rich genetic polymorphism. PSMC analysis showed that after the last interglacial period of pleistocene, the effective population size of *H. assimilis* increased dramatically and reached to millions (Fig. 2). During this period, the climate changed warm and wet, with the widely distributed tree residues or pollen, which provided an indispensable material basis for saprophagous butterfly.

Comparative genomic analysis

Comparative genomic analyses were performed between *H. assimilis* and 14 other lepidopteran species (*Aricia agestis*, *Danaus plexippus*, *Heliconius erato*, *Leptidea sinapis*, *Manduca sexta*, *Papilio polytes*, *Papilio xuthus*, *Papilio bianor*, *Pararge aegeria*, *Parnassius apollo*, *Pieris rapae*, *Spodoptera frugiperda*,

Vanessa tameamea and *Zerene cesonia*). We found 112 unique gene families in *H. assimilis* (Fig. 3A). Among the common gene families, *H. assimilis* has 1000 expansion and 513 contraction gene families.

By reconstructing the phylogenetic tree, *H. assimilis* and the same saprophagous butterfly *V. tameamea* formed a branch, and diverged approximately 55.69 million years ago (Mya). They diverged from nectarivorous butterfly *H. erato* about 78.4 Mya, suggesting that the change of feeding habits was caused by early evolutionary events (Fig. 3B). In addition, 93 gene families were expanded and 426 were contracted at *H. assimilis* and *V. tameamea*'s common nodes, among which 11 and 8 were significant ($p < 0.01$). The significant expansion gene families were involved in the activation of immune cells and process for oligosaccharide metabolic, but this branch also showed the contraction events were engaged in regulation of gonad development (Table S6).

A Venn diagram of *H. assimilis* and 14 other lepidopteran species gene family. The petals represent the number of unique gene families to each species. **B** Phylogenetic analysis of *H. assimilis*. The number beside each node denotes the estimated divergence time (million years ago) of two connected clades. The arrow indicates the gene family that significantly expanded (+) and contracted (-) at this node. The red pentagram indicates the gene families: immune cells activation (leukocyte GO:0045321, B cell GO:0042113 and lymphocyte GO:0046649), oligosaccharide catabolic process (GO:0009313), mannose metabolic process (GO:0006013) and RNA splicing (GO:0008380), etc., expand at this node.

Positive selection analysis

The values of K_a , K_s and K_a/K_s reveals patterns of functional gene evolution. Compared with *V. tameamea* and *H. erato*, we searched for the rapid evolution of genes associated with dietary habits in *H. assimilis*. Positive selection analysis showed that the rapidly evolving genes could not be found (K_a/K_s ratios < 1), due to too long differentiation time in these three species (Fig. 4). The result also indicated that the genetic background of the three butterfly species had been changed along with differentiation, and most of the gene functions were mainly related to the collateral origin specialized genes (specific genes and gene families) (Fig. S6).

Functional enrichment of target gene family

To understand and trace changes associated with saprophagous dietary, functional analysis was performed on the horizontal transfer gene family, unique gene family and expansion gene family of *H. assimilis*. Among them, 32 gene families were horizontal transfer gene family and the function were mainly associated with chromosome assembly, cellular component biogenesis and other structural functions (Fig. 5A, Table S7). For 112 unique gene families, the significant gene family enrichment functions were related to macromolecule metabolism, response to external stimulus, cell development and differentiation, locomotion and immune cell activation (Fig. 5B, Table S8). Furthermore, 186 gene families across *H. assimilis* exhibited expansions, and the significant gain on gene families was involved in process for organonitrogen compound, carbohydrate and lipid metabolic, the cellular response to external stimulus and detoxification, etc. And we also found expansion events involved in the response to

acid chemical and regulation of alcohol biosynthetic process (Fig. 5C, Table. 1). In *H. assimilis*, most of the enzymes analysed involved in nitrogen compound and lipid metabolic pathways such as trypsin-like serine protease, acyltransferase and enoyl reductase might reflect the diversity of its diet (Table S9).

A GO enrichment of horizontal transfer gene family. **B** GO enrichment of unique gene family. **C** GO enrichment of expansion gene family.

Table. 1 GO enrichment for significant gene families of habit food.

Species and nodes		Function and metabolic pathway	GO	P-value < 0.01
<i>H. assimilis</i> Expansions		+ response to extracellular stimulus	GO:0009991	0.0028
		+ response to acid chemical	GO:0001101	0.0032
		+ response to nutrient	GO:0007584	4.34E-04
		+ response to starvation	GO:0042594	0.0025
		+ fucose metabolic process	GO:0006004	6.64E-08
		+ immune system process	GO:0002376	0.0021
<i>H. assimilis</i> Unique		+ leukocyte activation	GO:0045321	9.13E-04
		+ B cell activation	GO:0042113	8.53E-09
		+ locomotion	GO:0040011	4.52E-04
		+ nitrogen compound metabolic process	GO:0006807	0.0030
		+ lymphocyte activation	GO:0046649	1.47E-06
		+ heterocycle metabolic process	GO:0046483	0.0085
<i>H. assimilis</i> and <i>V. tameamea</i> node Expansions		+ leukocyte activation	GO:0045321	9.13E-04
		+ B cell activation	GO:0042113	0.0010
		+ oligosaccharide catabolic process	GO:0009313	0.0035
		+ mannose metabolic process	GO:0006013	0.0045
		+ lymphocyte activation	GO:0046649	0.0059

Note: “+”: expansion.

Key gene screening

The successful capture of food is attributed to the ability of *H. assimilis* to effectively perceive information in the environment. The results of comparative genomic analysis showed that the OBP gene family was contracted. Here, we have comprehensively screened the genome of *H. assimilis* and determined 23 putative odorant binding proteins (OBPs), including 2 General Odorant-Binding Proteins (GOBPs), 2 Pheromone-Binding Proteins (PBPs) and 19 OBPs. These genes were distributed across 12 scaffolds and some of the *H. assimilis* OBP genes were located within the same cluster. For example, Hass-OBP10, Hass-OBP11, Hass-OBP12, Hass-OBP13 and Hass-OBP14 were located on the region of scaffold 17. Another four OBP genes (Hass-GOBP1, Hass-GOBP2, Hass-PBP-C and Hass-PBP-D) resided in the scaffold 16 (Table S8). Among them, 10 OBPs belonged to minus-C subfamily and others were classified as classic subfamily (Fig. 6A Table S10). We also found that GOBP1, GOBP2, PBP-C and PBP-D of *H. assimilis* defined a unique clade, which was species-specific and related to host location. Interestingly, compared with other species of lepidopteran, PBP-A and PBP-B were absent in *H. assimilis* (Fig. 6B, 6C). We speculate that the change of feeding habits of *H. assimilis* may be related to the loss of OBP gene. Future studies could evaluate the expression levels for OBP genes, and their regulation.

Discussion

Sap-feeding and saprophagous habits are observed in a number of butterfly species, as well as in beetles, hornets and flies [1]. However, little is currently known about related genetic mechanisms. Our study provides unprecedented knowledge of the genomic signatures behind the saprophagy in butterfly. We assembled a high-quality chromosome-level reference genome of *H. assimilis*, which is the representative of saprophagous butterflies in China. The final assembled genome was 423.87 Mb, and approximately 96.2% of the scaffolds were ordered onto 30 chromosomes. The scaffold N50 is 15.84 Mb and BUSCO completeness values is 99.3%, which set a new paradigm for future butterfly genome assemblies, and even a model to investigate gene function and evolution.

Feeding is a complex behavior, including at least food intake itself and foraging or appetite behavior [27]. Surprisingly, we found that feeding habits have little to do with the overall structure of the genome. For instance, compared *H. assimilis* with *D. plexippus*, which is the nectarivorous butterfly of the same family, we could not find the large-scale chromosome rearrangements. The gene interval, intron length and genome size were not much different from other butterflies (Fig. S7).

Then, we found *H. assimilis*'s unique gene family has characteristics associated with the ability to consume humus [1, 2], including gene families which were related to non-sugar (nitrogen compound, heterocycle, and aromatic compound), carbohydrate (oligosaccharide, mannose, hexose and monosaccharide) metabolic process (Table. 1). The result showed that although the sugar profiles of rotten food was significantly lower than that of nectar, saprophagous butterfly species could still use them [1]. On the other hand, we detected the significant expansion in gene families for the dietary

digestion (organonitrogen compound, carbohydrate and lipid metabolism) and signaling processes (response to acid chemical, starvation, nutrient and external stimulus) (Table. 1). These are associated with the strong ability of feeding, suggesting that this species is more adaptive to rotting foods (lower sugar concentrations and fermentative products including alcohols and carboxylic acids). Our analyses also found the expansion of genes involved in immune response (leukocyte, lymphocyte, B cell) and detoxification (cytochrome P450), which could help to explain how butterflies that feed on non-nectar can avoid the potentially adverse effects of their peculiar diet. These differences are likely associated with lineage-specific aspects of physiology, ecology (e.g., niche resources, immune system, interactions), and microbiomes.

We predicted that the changes in key genes may affect eating habits. To support our prediction, we have comprehensively screened the genes which were related to eating habits in *H. assimilis*. We found OBP genes were shown contraction. In this study, we have identified 23 OBP genes for the first time in *H. assimilis* genome. The number was less than *D. plexippus*, which has 32 OBP genes. Compared with *M. sexta*, *B. mori* and *D. plexippus* of lepidopteran, PBP-A and PBP-B were absent in *H. assimilis*. Previous studies have found that PBP transcripts were abundantly expressed in antennae, which appear to contribute to the activation of odorant receptor neurons in the *Antheraea polyphemus* and *B. mori* [28, 29]. Moreover, PBP-A genes were shown to be closely associated with pheromone recognition in *M. sexta* and *Drosophila melanogaster* [30]. Loss of the PBP-A genes was also proposed to be correlated with the olfactory-visual shift in *D. plexippus* [16], while the function of PBP-B gene has not yet been reliably verified. Olfactory is the first step in food acquisition [3]. We consider that the loss PBP genes reduce the sensitivity of sensory neurons to their odor ligands, which allows *H. assimilis* to perceive the odours of rotting foods as attractive odors, even as attractive tastes. The next step is to evaluate the expression level and regulation of this gene.

Conclusions

In conclusion, we provided a high-quality a chromosomal-level genome for *H. assimilis* which has been a typical saprophagous butterfly in China. Comparative genomic analysis provided novel insights into the genomic mechanisms of the feeding habit. Our results could not only serve as the genetic basis for in-depth investigations of butterfly evolution and biological functions but also offers a valuable genetic resource for future butterfly genome sequencing project.

Materials And Methods

Sample collection and sequencing

Ten female adult *H. assimilis* individuals (NCBI txid 378390) were collected from Hu County, Shaanxi Province, China, in June 2021. Genomic DNA was extracted from both individuals using a DNAeasy Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. Using the retrieved genomic DNA, two DNA libraries were constructed for DNA sequencing on Illumina X-ten (Illumina HiSeq

XTen, San Diego, USA) and Nanopore PromethION (Oxford Nanopore Technologies, Oxford, UK) platforms [31]. A short-read Illumina sequencing library was obtained using approximately 1.5 ug of DNA according to the standard sequencing kit protocol (NEBNext Ultra DNA Library Prep Kit for Illumina). A long-read nanopore sequencing library was constructed using approximately 10 ug of DNA and the SQK-LSK109 Ligation Sequencing Kit, following the manufacturer's instructions. The retrieved library had a mean DNA fragment length of approximately 20 kb. Hi-C library was constructed, including cross-linking of formaldehyde, restriction enzyme digestion, ends repair of fragments, DNA cyclization, DNA purification and other steps with Mbol as the restriction enzyme.

Genome assembly and assessment

Before assembly, strict quality control was performed on the raw Illumina and Nanopore sequencing data using Trimmomatic v0.39 [32] and Nanofilt v2.3.0[33], respectively. Low quality reads (Q30<90%) or those that contained more than 5% unknown bases were removed. Environmental microbial contaminants were removed by deleting sequences with hits in the GenBank Env_nt database (<ftp://ftp.ncbi.nlm.nih.gov/blast/db/>). Approximately 62.39 Gb of raw reads were generated on the Nanopore platform. A k-mer based analysis was performed to approximate genome size using genome characteristics estimation [34, 35] using all short-read DNA sequences. This estimated genome size then guided further assembly by assisting with software parameter adjustments.

NextDenovo (<https://github.com/Nextomics/NextDenovo>) was then used for the assembly of the genome sequencing data with the recommended parameters. To improve the assembly accuracy, the generated Nanopore sequence data assembly was polished using Pilon [36] with all Illumina data. For completeness assessment, the Illumina reads and Nanopore reads were mapped back to the assembly using Minimap2 [37], checking the correctness. ALLHiC pipeline [38] was used to conduct chromosome anchoring based on the Hi-C sequences. The parameters were adopted from the software's manu. Synteny analysis was conducted with the reference genome of *Danaus plexippus* (GCA_009731565.1) using MCscanX [39]. The graphical view of the consensus map and other genome basic information, including GC content and DNA variations, was displayed by CIRCOS [40]. Whole genome completeness was assessed using BUSCO v 5.2.2 which benchmarks universal single-copy orthologs [41].

Repetitive elements and gene prediction

Repetitive elements were identified using the EDTA pipeline [42]. The parameters were set as “-sensitive 1 -anno 1”. The divergence time of transposons for *H. assimilis* was estimated using parseRM using a script ParsingRM.pl (<https://github.com/4ureliek/Parsing-RepeatMasker-Outputs>). We used the integrated methods of *ab initio* and homology-based predictions to obtain a high-quality gene set. PASA (PASA, RRID:SCR_014656) [43] and Augustus (Augustus: Gene Prediction, RRID:SCR_008417) [44] were used to search for gene models. Species-specific transcript evidences were downloaded from NCBI SRA database (SRR11149958, SRR11149959, SRR11149960, SRR11149961, SRR11149962, SRR11149963). For the homology-based method, we downloaded gene sets for three related species in the same family: *Vanessa tameamea* (GenBank accession: GCA_002938995), *Heliconius erato* (GenBank accession:

GCA_018249695), and *Danaus plexippus* (GenBank accession: GCA_009731565). These homologous protein sequences were concatenated and imported into GeneWise [45] to search for genes. Pseudogenes were filtered based on correct translation and the presence of mature stop codons. In addition, a concatenated *ab initio* and homology-based gene prediction pipeline identified genes using BRAKER[46]. Finally, all gene prediction results were merged using the Evidence Modeler [47].

Gene functional annotation

The gene set, along with their corresponding protein sequence, was extracted and inserted into functional annotations. Functional annotation was performed by an integrated software EggNog-Mapper (eggNOG, RRID:SCR_002456, Version 2) [48]. The protein sequences of the *H. assimilis* genome were searched for UniProt/SwissProt databases [49]. Interproscan software version 4.8 [50] was used for searching known motifs and domains in the *H. assimilis* genome. The structural domain of interesting proteins was searched in the Pfam 32.0 website. All these annotations can be found in supplementary Table. S3-5.

Phylogenetic reconstruction and gene family analyses

Gene families were identified using the OrthoFinder software[51]. The protein sequences of *H. assimilis*, along with the other 14 species, including *Aricia agestis*, *Danaus plexippus*, *Heliconius erato*, *Leptidea sinapis*, *Manduca sexta*, *Papilio polytes*, *Papilio xuthus*, *Papilio bianor*, *Pararge aegeria*, *Parnassius apollo*, *Pieris rapae*, *Spodoptera frugiperda*, *Vanessa tameamea* and *Zerene cesonia*, were used to search for orthologous gene families. The coding sequences of the single-copy genes in these 15 species were identified, extracted, and aligned using the Multiple Alignment using Fast Fourier Transform program [52] to construct a super-matrix. The r8s software[53] was used to build the phylogenetic tree and estimate the divergence time. The calibration time was queried at www.time-tree.org. According to divergence times and phylogenetic relationships, café [54] was used to determine the expansion and constriction of orthologous gene families. Horizontal transfer, unique and significant expansion ($P < 0.01$) gene families were also used for GO enrichment analysis (supplementary Table. S7-9). The enrichment of GO terms was conducted using KOBAS 3.0[55].

Positive selection analysis

The non-synonymous substitutions (Ka) and synonymous substitutions (Ks) rate values were calculated using the pipeline WGD (<https://github.com/arzwa/wgd>) and the results were further filtered according to the following parameters: Alignment Coverage > 0.6, Alignment Identity > 0.7, both Ka and Ks < 10.

The OBP identification and analysis

We collected the published OBP sequences of insects from GenBank, and used this dataset to search for OBP genes in the *H. assimilis* genome by using tblastn [56]. The resulting genomic regions containing OBP genes were predicted using FGENESH [57]. Each predicted OBP protein was double-checked by the common standard: a conserved cysteine pattern, a predicted size of 14 kDa and a signal peptide located in the N-terminal [3]. For the phylogenetic analyses, we also collected GOBP/PBPs in 10 other

lepidopteran species, including *Danaus plexippus*, *Heliconius Melpomene*, *Bombyx mori*, *Helicoverpa armigera*, *Manduca sexta*, *Grapholita molesta*, *Chilo suppressalis*, *Plutella xylostella*, *Spodoptera littoralis* and *Pieris rapae* were used as queries [16, 58]. All the output sequences were manually checked, and duplicate and redundant candidates were removed. Sequences were aligned using Clustal Omega (<http://www.ebi.ac.uk/tools/msa/clustalo/>). Phylogenetic trees were constructed in MEGA7 software [59] using the neighbor-joining method with 500-fold bootstrap resampling. Tree images were saved as nwk and edited in ITOL (<https://itol.embl.de/>) [60]. All branches were manually collapsed to > 50% bootstrap support.

Abbreviations

OBP: Odorant Binding Protein; GO: Gene ontology; LTR: Long Terminal Repeated; Ka: nonsynonymous substitution rate. Ks: synonymous substitution rate.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All sequence data were deposited in GenBank under BioProject accession no. PRJNA793373.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

L.Z., D-L.G and S-Q.X conceived the study and designed the experiments. H.W. and Z-C.D performed the sequencing experiments. L.Z. and D-L.G analyzed the data. L.Z. wrote the manuscript. D-L.G and S-Q.X revised the manuscript and should be considered as co-corresponding author. All authors read and approved the final manuscript.

Funding

This research was supported by National Animal Collection Resource Center, China, Fundamental Research Funds for the Central Universities (GK201903063, GK202105003). This work was partly supported by the National Natural Science Foundation of China (No. 31872273).

Acknowledgments

Thanks to Professor Yu-Fei Li from Xi'an Jiaotong University for providing the ecological photos of *H. assimilis* and Professor Hua-Teng Huang from Shaanxi Normal University for insightful discussions regarding this research.

References

1. Omura H, Honda K. Feeding responses of adult butterflies, *Nymphalis xanthomelas*, *Kaniska canace* and *Vanessa indica*, to components in tree sap and rotting fruits: synergistic effects of ethanol and acetic acid on sugar responsiveness. *Journal of insect physiology*. 2003; 49(11):1031-1038.
2. Omura H. Identification of feeding attractants in oak sap for adults of two nymphalid butterflies, *Kniska canace* and *Vanessa indica* (vol 25, pg 281, 2000). *Physiological Entomology*. 2001; 26(3):283-283.
3. Zhou YN, Xie S, Chen JN, Wang ZH, Yang P, Zhou SC et al. Expression and functional characterization of odorant-binding protein genes in the endoparasitic wasp *Cotesia vestalis*. *Insect science*. 2021; 28(5):1354-1368.
4. Hill J, Hamer K, Tangah J, Dawood M. Ecology of tropical butterflies in rainforest gaps. *Oecologia*. 2001; 128(2):294-302.
5. Omura H, Honda K, Asaoka K, Inoue TA. Tolerance to fermentation products in sugar reception: gustatory adaptation of adult butterfly proboscis for feeding on rotting foods. *Journal of comparative physiology A, Neuroethology, sensory, neural, and behavioral physiology*. 2008; 194(6):545-555.
6. Kaliszewska ZA, Lohman DJ, Sommer K, Adelson G, Rand DB, Mathew J et al. When caterpillars attack: biogeography and life history evolution of the Miletinae (Lepidoptera: Lycaenidae). *Evolution; international journal of organic evolution*. 2015; 69(3):571-588.
7. Li ZH, Hua C, Wei-Yan YU, Song DJ, Chen H, Shan-Jin XU et al. Bionomics of the butterfly, *Hestina assimilis*. 2008; 45(5):754-757.
8. Au TF, Bonebrake TC. Increased Suitability of Poleward Climate for a Tropical Butterfly (*Euripus nyctelius*) (Lepidoptera: Nymphalidae) Accompanies its Successful Range Expansion. *Journal of insect science (Online)*. 2019; 19(6).
9. Takashi, TANIO, Noboru, Science KJL. Effects of temperature on the initiation of post-diapause development in larvae of *Hestina assimilis assimilis* (Lepidoptera, Nymphalidae) and *Hestina persimilis japonica*. *Lepidoptera Science*. 2019; 70(2):59-63.
10. Zhang S, Zhang Z, Wang H, Kong X. Antennal transcriptome analysis and comparison of olfactory genes in two sympatric defoliators, *Dendrolimus houi* and *Dendrolimus kikuchii* (Lepidoptera: Lasiocampidae). *Insect biochemistry and molecular biology*. 2014; 52:69-81.
11. Wen JP, Wang JP, Wu YP, Cao TW, Li ZQJMDPB. Complete mitochondrial genome of *Hestina assimilis* (Lepidoptera: Nymphalidae). *Mitochondrial DNA Part B*. 2020; 5(2):1269-1271.

12. Gan HM, Grandjean F, Jenkins TL, Austin CM. Absence of evidence is not evidence of absence: Nanopore sequencing and complete assembly of the European lobster (*Homarus gammarus*) mitogenome uncovers the missing nad2 and a new major gene cluster duplication. BMC genomics. 2019; 20(1):335.
13. Tondini F, Jiranek V, Grbin PR, Onetto CA. Genome Sequence of Australian Indigenous Wine Yeast *Torulaspora delbrueckii* COFT1 Using Nanopore Sequencing. Genome announcements. 2018; 6(17).
14. Laszlo AH, Derrington IM, Ross BC, Brinkerhoff H, Adey A, Nova IC et al. Decoding long nanopore sequencing reads of natural DNA. Nature biotechnology. 2014; 32(8):829-833.
15. Pečnerová P, Garcia-Erill G, Liu X, Nursyifa C, Waples RK, Santander CG et al. High genetic diversity and low differentiation reflect the ecological versatility of the African leopard. Current biology : CB. 2021.
16. Vogt RG, Große-Wilde E, Zhou JJ. The Lepidoptera Odorant Binding Protein gene family: Gene gain and loss within the GOBP/PBP complex of moths and butterflies. Insect biochemistry and molecular biology. 2015; 62:142-153.
17. Wang Y, Zhang R, Wang M, Zhang L, Shi CM, Li J et al. The first chromosome-level genome assembly of a green lacewing *Chrysopa pallens* and its implication for biological control. Mol Ecol Resour. 2021.
18. Lu S, Yang J, Dai X, Xie F, He J, Dong Z et al. Chromosomal-level reference genome of Chinese peacock butterfly (*Papilio bianor*) based on third-generation DNA sequencing and Hi-C analysis. GigaScience. 2019; 8(11).
19. Kim SR, Kwak W, Kim H, Caetano-Anolles K, Kim KY, Kim SB et al. Genome sequence of the Japanese oak silk moth, *Antheraea yamamai*: the first draft genome in the family Saturniidae. GigaScience. 2018; 7(1):1-11.
20. Nowell RW, Elsworth B, Oostra V, Zwaan BJ, Wheat CW, Saastamoinen M et al. A high-coverage draft genome of the mycalesine butterfly *Bicyclus anynana*. GigaScience. 2017; 6(7):1-7.
21. Shen J, Cong Q, Kinch LN, Borek D, Otwinowski Z, Grishin NV. Complete genome of *Pieris rapae*, a resilient alien, a cabbage pest, and a source of anti-cancer proteins. F1000Research. 2016; 5:2631.
22. Cong Q, Borek D, Otwinowski Z, Grishin NV. Tiger Swallowtail Genome Reveals Mechanisms for Speciation and Caterpillar Chemical Defense. Cell reports. 2015; 10(6):910-919.
23. Cong Q, Borek D, Otwinowski Z, Grishin NV. Skipper genome sheds light on unique phenotypic traits and phylogeny. BMC genomics. 2015; 16(1):639.
24. Yang J, Wan W, Xie M, Mao J, Dong Z, Lu S et al. Chromosome-level reference genome assembly and gene editing of the dead-leaf butterfly *Kallima inachus*. Mol Ecol Resour. 2020; 20(4):1080-1092.
25. Kanchon, W JR, B AD, D JW. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature. 2012; 487(7405):94-98.
26. Zhan S, Merlin C, Boore JL, Reppert SM. The monarch butterfly genome yields insights into long-distance migration. Cell. 2011; 147(5):1171-1185.

27. Ding W, Zhang X, Zhao X, Jing W, Cao Z, Li J et al. A Chromosome-Level Genome Assembly of the Mandarin Fish (*Siniperca chuatsi*). *Frontiers in genetics*. 2021; 12:671650.
28. Liu Y, Hu Y, Bi J, Kong X, Long G, Zheng Y et al. Odorant-binding proteins involved in sex pheromone and host-plant recognition of the sugarcane borer *Chilo infuscatellus* (Lepidoptera: Crambidae). *Pest management science*. 2020; 76(12):4064-4076.
29. Pophof B. Pheromone-binding proteins contribute to the activation of olfactory receptor neurons in the silkmoths *antheraea polyphemus* and *Bombyx mori*. *Chemical senses*. 2004; 29(2):117-125.
30. Vogt RG, Rogers ME, Franco MD, Sun M. A comparative study of odorant binding protein genes: differential expression of the PBP1-GOBP2 gene cluster in *Manduca sexta* (Lepidoptera) and the organization of OBP genes in *Drosophila melanogaster* (Diptera). *The Journal of experimental biology*. 2002; 205(Pt 6):719-744.
31. De Coster W, De Rijk P, De Roeck A, De Pooter T, D'Hert S, Strazisar M et al. Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome research*. 2019; 29(7):1178-1187.
32. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*. 2014; 30(15):2114-2120.
33. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics (Oxford, England)*. 2018; 34(15):2666-2669.
34. Manekar SC, Sathe SR. A benchmark study of k-mer counting methods for high-throughput sequencing. *GigaScience*. 2018; 7(12).
35. Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Quantitative Biology*. 2013;arXiv:1308.2012.
36. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*. 2014; 9(11):e112963.
37. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)*. 2018; 34(18):3094-3100.
38. Zhang X, Zhang S, Zhao Q, Ming R, Tang H. Assembly of allele-aware, chromosomal-scale autoploid genomes based on Hi-C data. *Nature plants*. 2019; 5(8):833-845.
39. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research*. 2012; 40(7):e49.
40. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D et al. Circos: an information aesthetic for comparative genomics. *Genome research*. 2009; 19(9):1639-1645.
41. Waterhouse RM, Seppey M, Simão FA, Zdobnov EM. Using BUSCO to Assess Insect Genomic Resources. *Methods in molecular biology (Clifton, NJ)*. 2019; 1858:59-74.
42. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome biology*. 2019;

20(1):275.

43. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research*. 2003; 31(19):5654-5666.
44. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research*. 2006; 34(Web Server issue):W435-439.
45. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*. 2012; 7(3):562-578.
46. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-Genome Annotation with BRAKER. *Methods in molecular biology* (Clifton, NJ). 2019; 1962:65-95.
47. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J et al. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome biology*. 2008; 9(1):R7.
48. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular biology and evolution*. 2017; 34(8):2115-2122.
49. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research*. 2000; 28(1):45-48.
50. Mulder N, Apweiler R. InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods in molecular biology* (Clifton, NJ). 2007; 396:59-70.
51. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG*. 2000; 16(6):276-277.
52. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*. 2002; 30(14):3059-3066.
53. Sanderson MJ. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* (Oxford, England). 2003; 19(2):301-302.
54. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* (Oxford, England). 2006; 22(10):1269-1271.
55. Bu D, Luo H, Huo P, Wang Z, Zhang S, He Z et al. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic acids research*. 2021; 49(W1):W317-w325.
56. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*. 1997; 25(17):3389-3402.
57. Solovyev V, Kosarev P, Seledsov I, Vorobyev D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome biology*. 2006; 7 Suppl 1(Suppl 1):S10.11-12.

58. Li MY, Jiang XY, Qi YZ, Huang YJ, Li SG, Liu S. Identification and Expression Profiles of 14 Odorant-Binding Protein Genes From *Pieris rapae* (Lepidoptera: Pieridae). Journal of insect science (Online). 2020; 20(5).
59. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Molecular biology and evolution. 2016; 33(7):1870-1874.
60. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic acids research. 2019; 47(W1):W256-w259.

Figures

Figure 1

Chromosome-level genome assembly of the *H. assimilis*.

Synteny alignment of the *H. assimilis* and *D. plexippus*. The tagged HA01 to HA30 refer to the chromosomes of the sequenced *H. assimilis* and the black columns refers to the chromosomes of *D. plexippus*. The densities of GC contents, DNA variations in *H. assimilis* are shown in the outer rims in orders.

Figure 2

Historical effective population size of *H. assimilis* with bootstrap results indicated with thin lines. The unit of the x-coordinate is time (years).

Figure 3

Comparative genomics analysis of *H. assimilis*.

A Venn diagram of *H. assimilis* and 14 other lepidopteran species gene family. The petals represent the number of unique gene families to each species. **B** Phylogenetic analysis of *H. assimilis*. The number beside each node denotes the estimated divergence time (million years ago) of two connected clades. The arrow indicates the gene family that significantly expanded (+) and contracted (-) at this node. The red pentagram indicates the gene families: immune cells activation (leukocyte GO:0045321, B cell GO:0042113 and lymphocyte GO:0046649), oligosaccharide catabolic process (GO:0009313), mannose metabolic process (GO:0006013) and RNA splicing (GO:0008380), etc., expand at this node.

Figure 4

Distributions Ka/Ks ratio for paralogs in *H. assimilis*.

Figure 5

GO enrichment analysis summarized and visualized as a scatter plot.

A GO enrichment of horizontal transfer gene family. **B** GO enrichment of unique gene family. **C** GO enrichment of expansion gene family.

A

B

C

Figure 6

The odorant-binding proteins (OBPs) identification and analysis.

A Alignment of deduced protein sequences of OBPs identified in *H. assimilis*. Six positionally conserved cysteine residues are highlighted in red box. **B** GOBP/PBP gene clusters in *H. assimilis* and 10 other lepidopteran species, including *D. plexippus*, *H. melpomene*, *B. mori*, *H. armigera*, *M. sexta*, *G. molesta*, *C.*

suppressalis, *P. xylostella*, *S. littoralis* and *P. rapae*. **C** Structures of GOBP/PBP gene clusters from *M. sexta*, *B. mori*, *D. plexippus* and *H. assimilis*.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterial.doc](#)
- [Supplementarytable.xls](#)