scientific data

OPEN

DATA DESCRIPTOR

Genome assembly of the grassland caterpillar *Gynaephora qinghaiensis*

Youpeng Lai[1,5], Shan Xiao[2,5], Minggang Qin[1], Xinhai Ye[3], Fang Wang[4] & Qi Fang[4] ✉

The grassland caterpillars are the most damaging insect pests to the alpine meadow of the Qinghai-Tibetan Plateau in China. In this study, we present a genome assembly of one grassland caterpillar *Gynaephora qinghaiensis* by using Oxford Nanopore long-read and BGI short-read sequencing. The genome assembly of 861.04 Mb in size consists of 107 contigs, with a contig N50 size of 18.65 Mb. The BUSCO analysis revealed the presence of 99.56% (99.27% complete and 0.29% fragmented) BUSCO genes in the assembly. 580.2 Mb (67.4% of genome) of repetitive sequences and 16,618 protein-coding genes were predicted in *G. qinghaiensis* genome. Phylogenomic analysis indicated that *G. qinghaiensis* and the rusty tussock moth *Orgyia antiqua* diverged approximately 18.3 million years ago. Moreover, gene family evolution analysis suggested that 130 gene families significantly expanded and 43 contracted in the *G. qinghaiensis* genome. The availability of the reference genome could provide genetic resources to uncover adaptive evolutionary mechanisms of grassland caterpillars to high-altitude environments and contributes to the development of integrated pest management strategies.

Background & Summary

Grassland caterpillars (Lepidoptera: Lymantriinae: *Gynaephora*) are a small group with 15 recorded species worldwide, distributing in the arctic tundra and high-altitude areas of the northern hemisphere[1]. In China, all eight nominated *Gynaephora* species inhabit Qinghai-Tibetan Plateau (QTP) at altitudes of 2900 to 5000 m above sea level (masl), accounting for most parts of the area[2]. The grassland caterpillars are the most damaging insect pests to the alpine meadow of the QTP. They not only devour forage vegetation, leading to serious feed shortages and grassland degradation, also cause mouth mucous membrane canker in domestic and wild animals[1].

The grassland caterpillars are well adapted to the harsh high-altitude environments in the QTP. Morphologically, the larvae are covered with dense black body hair, which can help them to resist high UV radiation and regulate body temperature[3]. Physiologically, female adults do not develop their wings and antennae in comparison to males, showing significant sexual dimorphism. The optimized allocation of energy to reproduction rather than metamorphosis contributes to increasing fitness[4]. Genetically, several genes associated with response to hypoxia, energy metabolism and DNA repair, are positively selected in *G. menyuanensis and G. alpherakii* compared to no-QTP insects[5]. And sequence variations and expression pattern changes of mitochondrial genes are also associated with adaptation to different high-elevation environments[6]. The QTP *Gynaephora* species were proposed to be derived from a common ancestor and the genetic differentiation and speciation occurred in association with QTP uplift and climate changes[2,7]. The grassland caterpillars distributed at divergent and specific elevations have high levels of genetic diversity, which are promising models for studying adaptive evolution in high-altitude insects[8]. However, limited genetic information is available for the grassland caterpillars as of now[5,7,9].

Here, we present a genome assembly of one grassland caterpillar *G. qinghaiensis*, which is widely distributed in Tibet, Qinghai Province, Sichuan Province, and Gansu Province of China from 3000 to 4000 masl[2]. Whole genome is sequenced with Oxford Nanopore long-read and BGI short-read sequencing technologies. Genome assembly is 861.04 Mb in size, with both high completeness (BUSCO score: 99.56%) and continuity (contig N50: 18.65 Mb). The availability of grassland caterpillars reference genome could provide genetic resources to uncover

[1]Key Laboratory of Agricultural Integrated Pest Management of Qinghai Province, Qinghai University, Xining, 810016, China. [2]Ningbo Academy of Agricultural Science, Ningbo, 315040, China. [3]College of Advanced Agriculture Science, Zhejiang A&F University, Hangzhou, China. [4]State Key Laboratory of Rice Biology, Ministry of Agricultural and Rural Affairs Key Laboratory of Molecular Biology of Crop Pathogens and Insects, Institute of Insect Sciences, Zhejiang University, Hangzhou, 310058, China. [5]These authors contributed equally: Youpeng Lai, Shan Xiao. ✉e-mail: fangqi@zju.edu.cn

Characteristics	Features
Genome assembly	
Genome size	861.04 Mb
GC content	35.23%
Number of Contigs	107
Contig N50	18.65 Mb
BUSCOs	
Complete	99.27%
Complete and single copy	98.03%
Complete and duplicated	1.24%
Fragmented	0.29%
Missing	0.44%

Table 1. Statistics of *Gynaephora qinghaiensis* genome assembly.

Type	Number of elements	Length (bp)	Percent (%)
SINEs	14,457	2,252,336	0.26
LINEs	1,173,819	192,607,078	22.37
LTRs	395,039	70,651,345	8.21
DNAs	175,929	45,779,223	5.32
Rolling-circles	478,722	85,657,418	9.95
Unclassified	1,021,157	165,906,701	19.27
Small RNA	953	138,190	0.02
Satellites	1,133	300,604	0.03
Simple repeats	163,846	16,162,256	1.88
Low complexity	15,412	743,971	0.09
Total	3,440,467	580,199,122	67.4

Table 2. Annotation of repeat elements in the *Gynaephora qinghaiensis* genome. Note: SINEs, short interspersed repetitive DNA sequences; LINEs, long interspersed nuclear elements; LTRs, long terminal repeat elements; DNAs, DNA transposons.

Database	Number	Percent (%)
Swiss-Prot	9,125	54.91
Nr	15,601	93.88
KEGG	5,751	34.61
GO	9,482	57.06
pfamA	11,118	66.90
Total annotated	15,649	94.17

Table 3. Functional annotation of *Gynaephora qinghaiensis* proteins.

adaptive evolutionary mechanisms of the *Gynaephora* species to high-altitude environments and contributes to the development of integrated pest management strategies.

Methods

Genome sequencing and assembly. The grassland caterpillar *G. qinghaiensis* were collected in Yushu County (33°45′N/95°48′E), Qinghai province, China. Total genomic DNA from a male adult was extracted for genome sequencing using a Qiagen DNA purification kit (Qiagen, USA). The genome was sequenced using a combination of short and long-read strategies. Short-read sequencing library was constructed using a Truseq Nano DNA HT Sample Preparation Kit (Illumina, USA), and sequenced on MGISEQ-2000 platform (BGI, China). After quality-filtering using fastp v0.20.0[10], 28.17 Gb clean short reads were obtained. For long-read sequencing, library was generated using the DNA Ligation Sequencing Kit (SQK-LSK109) (Oxford Nanopore Technologies, England) and sequenced on PromethION platform (Oxford Nanopore Technologies, England). A total of 129.06 Gb Nanopore reads were attained after quality-filtering using Guppy v3.2.2 + 9fe0a78[11], which covered 149.9 folds of *G. qinghaiensis* genome.

Genome size and heterozygosity rate of *G. qinghaiensis* were inferred by kmerFreq[12] using BGI short reads, based on 17-mer frequency distribution. The estimated genome size was 844.04 Mb and the heterozygosity was

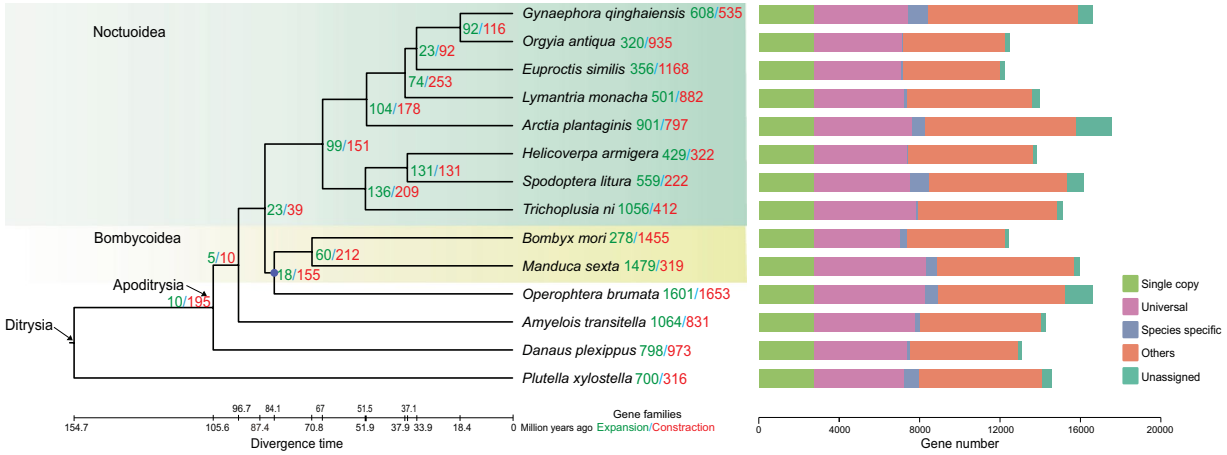

Fig. 1 Comparative genomics analysis of the *Gynaephora qinghaiensis* genome. To the left is the maximum likelihood phylogenetic tree displaying gene family expansion and contraction in *G. qinghaiensis* compared with other 13 lepidopteran species. The green and red numbers indicate the numbers of expanded and contracted gene families, respectively. Branch lengths represent divergence times. All nodes received 100% bootstrap support, except for the nodes in blue, whose bootstrap support value is 90%. To the right is the gene counts for different types of orthologous groups in the genomes. "Single copy" indicates universal one-to-one orthologs present in all species; "Universal" indicates other universal genes; "Species specific" indicates species-specific genes with more than one copy. "Unassigned genes" indicates species-specific genes with only one copy in the genome; "Others" indicates remaining genes.

1.3%. For *de novo* genome assembly, genome was initially assembled with NextDenovo v2.3.1[13] using Nanopore long reads. The preliminary assembly was then polished with NextPolish v1.3.0[14] using both long reads (three iterations) and short reads (four iterations). Redundant contigs was eliminated using Purge Haplotigs[15]. The final assembly is 861.04 Mb in size, encompassing 107 contigs, with a contig N50 of 18.65 Mb, and the GC content of 35.23% (Table 1).

Transcriptome sampling and sequencing. For transcriptome sequencing, eggs, first instar larvae, late instar larvae, female pupae, female adults of *G. qinghaiensis* were collected separately with three replicates. Total RNA was extracted using TRIzol® Reagent (Thermo Fisher, Shanghai, China) in accordance with the manufacturer's protocol. RNA-Seq libraries were prepared using TruSeq RNA Sample Prep Kit (Illumina, USA) and sequenced on Illumina NovaSeq 6000 platform (Illumina, USA). The reads with adaptors and low-quality reads were filtered using fastp v0.20.0[10].

Genome prediction and functional annotation. The repeat elements in the genome were identified with a combination of *de novo* and homology approaches. *De novo* predictions for long terminal repeat element (LTR) and non-LTR repeat sequences were performed using LTR_Finder[16] and RepeatModeler v2.0.1[17], separately. The results from two softwares combined with RepBase library v20170127[18] and Dfam v3.1[19] database were used as the library for RepeatMasker v4.0.7[20] to identify and classify repeat elements in genome. The repeat annotation result showed that repetitive elements sequences account for 67.4% of the genome sequence (Table 2). Among them, long interspersed nuclear elements (LINE) (22.37%), Rolling-circles (9.95%) and LTRs (8.21%) represented the three most abundant repeat types.

Protein-coding gene prediction was performed using a combination of ab initio, homology, and transcriptome-based approaches. For ab initio prediction, Braker2 v2.1.2[21] were used. For homology-based predictions, we downloaded the protein sequences of nine closely related species including *Arctia plantaginis* (RefSeq assembly accession: GCA_902825455.1), *Bombyx mori* (GCF_014905235.1), *Helicoverpa armigera* (GCF_023701775.1), *Manduca sexta* (GCF_014839805.1), *Spodoptera litura* (GCF_002706865.2) and *Trichoplusia ni* (GCF_003590095.1) from the NCBI database, *Euproctis similis* (GCA_905147225.2), *Lymantria monacha* (GCA_905163515.2) and *Orgyia antiqua* (GCA_916999025.1) from the Darwin Tree of Life Data Portal database (https://portal.darwintreeoflife.org/). The protein sequences were matched with the *G. qinghaiensis* genome using tblastn v2.13.0[22] with an E-value cutoff of 1e-5, and the matched proteins were then mapped against the homologous genomic sequences using Exonerate v2.2.0[23] and GenomeThreader v1.7.1[24]. For transcriptome-based prediction, RNA-Seq data were aligned with genome assembly using Hisat v2.2.1[25] and gene structures were predicted using Stringtie v2.1.7[26]. The non-redundant reference gene set was generated by integrating genes predicted from three methods using EvidenceModeler v1.1.1[27]. Eventually, 16,618 protein-coding genes were predicted (Table 3). To unravel the functions of protein-coding genes in *G. qinghaiensis*, we annotated the official gene set against the NCBI non-redundant protein database (Nr, https://ftp.ncbi.nlm.nih.gov/blast/db/) and SwissProt databases (https://www.uniprot.org/) using blastp v2.13.0[22] with an E-value cutoff of 1e−5. In addition, Protein domains were predicted using HMMER v3.3.2[28]. Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) were predicted using BlastKOALA[29]

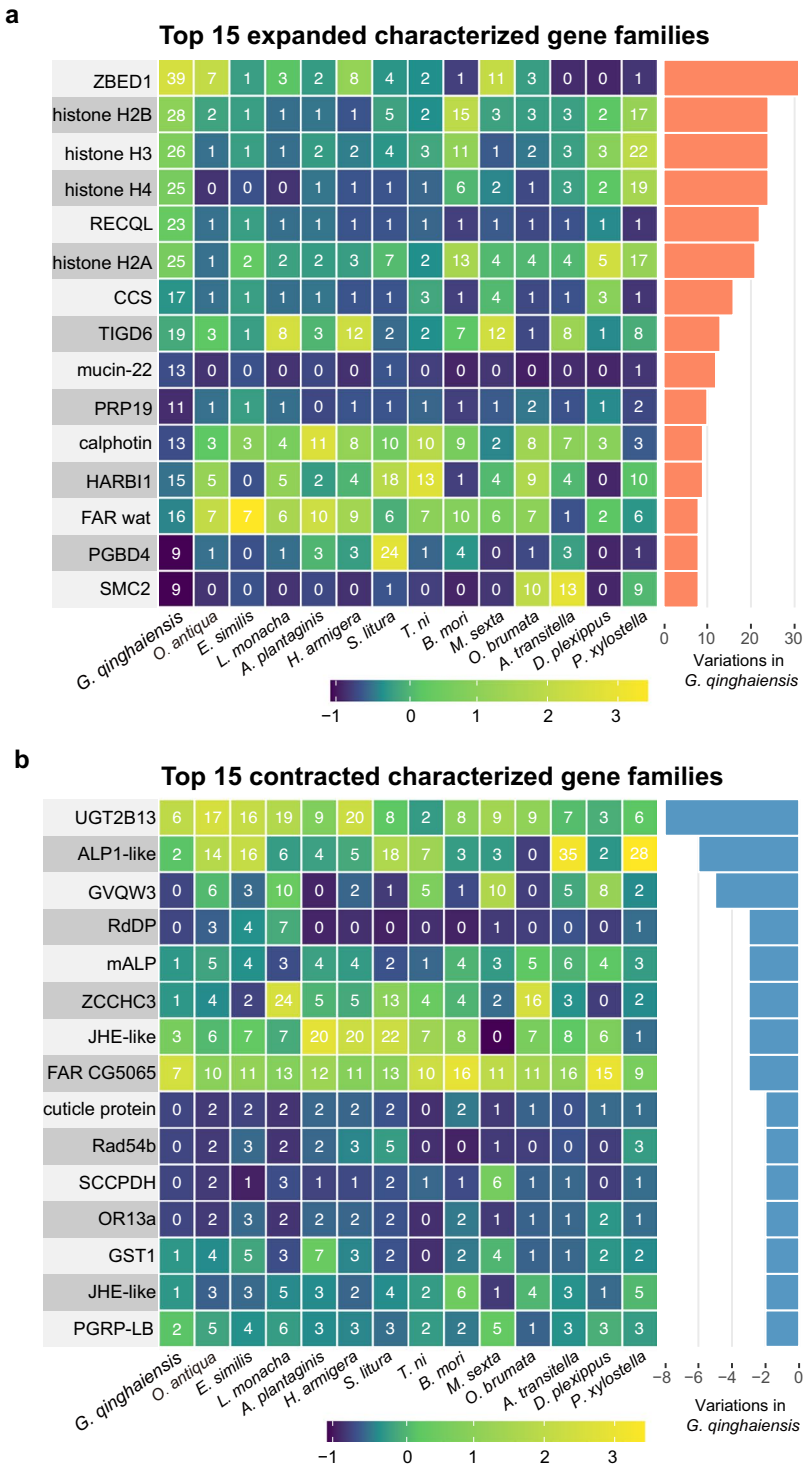

Fig. 2 Gene family evolution in the grassland caterpillar *Gynaephora qinghaiensis*. 15 most significantly **(a)** expanded and **(b)** contracted characterized gene families of *G. qinghaiensis* are illustrated using heatmap plots. The bars on the right depict the variations of gene families in *G. qinghaiensis*. ZBED1, zinc finger BED domain-containing protein 1; RECQ1, ATP-dependent DNA helicase Q1; CCS, copper chaperone for superoxide dismutase; TIGD6, tigger transposable element-derived protein 6; PRP19, pre-mRNA-processing factor 19; HARBI1, putative nuclease HARBI1; FAR wat, fatty acyl-CoA reductase wat; PGBD4, piggyBac transposable element-derived protein 4; SMC2, structural maintenance of chromosomes protein 2; UGT2B13, UDP-glucuronosyltransferase 2B13; ALP1-like, protein ALP1-like; GVQW3, protein GVQW3; RdDP, RNA-directed DNA polymerase; mALP, membrane-bound alkaline phosphatase-like; ZCCHC3, zinc finger CCHC domain-containing protein 3; JHE-like, juvenile hormone esterase-like; FAR CG5065, fatty acyl-CoA reductase CG5065; Rad54b, fibrinogen silencer-binding protein-like; SCCPDH, saccharopine dehydrogenase-like oxidoreductase; OR13a, odorant receptor 13a; GST1, glutathione S-transferase 1; PGRP-LB, peptidoglycan-recognition protein LB.

(https://www.kegg.jp/blastkoala/) and PANNZER2[30] (http://ekhidna2.biocenter.helsinki.fi/sanspanz/), respectively. Eventually, 15,649 genes (94.17%) were successfully annotated by at least one public biological function database.

Phylogenetic analysis. For comparison analysis between different genomes, the longest protein of each gene locus was retained and the orthologous genes across *G. qinghaiensis* and 13 other lepidopteran insects were identified with Orthofinder v2.5.4[31]. The 13 lepidopteran insects including *Amyelois transitella* (GCF_001186105.1), *Danaus plexippus* (GCF_009731565.1), *Operophtera brumata* (http://v2.insect-genome.com/Organism/590), *Plutella xylostella* (GCF_905116875.1) and nine former mentioned species, whose protein sets were used for homology-based prediction. Finally, 15,090 orthogroups (OGs) were identified, among which 8,807 OGs were present in all 14 lepidopteran insects and 2,741 were single copy OGs. The 2,741 single-copy OGs were used for phylogenetic tree construction. All protein sequences were aligned with MAFFT v7.123b[32] and poorly aligned regions were removed by trimAl v1.4.rev22[33]. Alignments were concatenated into a supergene sequence and the phylogenetic tree was constructed using the maximum likelihood (ML) method by IQ-TREE v2.1.4-beta[34] software with 1000 ultrafast bootstrap replicates. The best-fit substitution model Q.insect + F + R6 was determined by ModelFinder[35] implemented in IQ-TREE and the diamondback moth *P. xylostella* was used as an outgroup[36]. Phylogeny of 14 lepidopteran insects shows that *G. qinghaiensis* is a sister taxon to *O. antiqua* (rusty tussock moth), forming a lineage together with three other Erebidae insects: the yellow-tail moth *E. similis*, the black-arched tussock moth *L. monacha*, and the wood tiger moth *A. plantaginis* (Fig. 1). Divergence times between species were estimated using r8s v1.81[37] software. Four calibration points were applied according to a former study: 67–88.6 million years ago (Mya) for Noctuoidea, 70.1–89.9 Mya for Bombycoidea, 105.6–132.1 Mya for Apoditrysia and 154.7 Mya for Ditrysia[36]. *G. qinghaiensis* and *O. antiqua* were estimated to be diverged approximately 18.3 Mya.

Gene family expansion and contraction analysis. Gene family expansion and contraction of *G. qinghaiensis* were analyzed using CAFÉ v5.0[38] software. The gene family results inferred from OrthoFinder and estimated divergence time by r8s were used as inputs. CAFÉ results suggested that 608 and 535 gene families were expanded and contracted in *G. qinghaiensis*, respectively (Fig. 1). And 173 (130 expanded and 43 contracted families) gene families were rapidly evolved (P-value < 0.01). The significantly expanded gene families were associated with various molecular functions and biological processes, including, zinc finger BED domain-containing protein 1 (ZBED1), core histones (histone H2A, H2B, H3 and H4), fatty acyl-CoA reductase (FAR), copper chaperone for superoxide dismutase (CCS), (Fig. 2, Table S1[39]). The contracted gene families include UDP-glucuronosyltransferase 2B13 (UGT2B13), membrane-bound alkaline phosphatase-like (mALP), zinc finger CCHC domain-containing protein 3 (ZCCHC3), juvenile hormone esterase-like (JHE-like) and so on. Variations of characterized gene families were illustrated with the Superheat[40] R package.

Data Records

Oxford Nanopore long-read (SRA accessions: SRR24032119[41]) and BGI short-read (SRR24032120[42]) sequencing data for *G. qinghaiensi* genome are available as NCBI BioProject PRJNA950575. Illumina transcriptomic data for eggs (SRR31034734[43], SRR31034740[44], SRR31034741[45]), first instar larvae (SRR31034731[46], SRR31034732[47], SRR31034733[48]), late instar larvae (SRR31034728[49], SRR31034729[50], SRR31034730[51]), female pupae (SRR31034727[52], SRR31034738[53], SRR31034739[54]), female adults (SRR31034735[55], SRR31034736[56], SRR31034737[57]) with three replicates are available as NCBI BioProject PRJNA1174259. The genome assembly has been submitted to NCBI under accession number GCA_042920415.1[58]. Gene CDS[59], protein[60], and genome annotation[61] files are deposited in the Figshare database.

Technical Validation

DNA and RNA qualities were assessed by 0.7% gel electrophoresis and Agilent 2100 Bioanalyzer (Agilent, USA), respectively. Only high-quality samples were used for library preparation and sequencing.

BUSCO[62] v5.5.0 was used to evaluate the genome assembly completeness with the insecta_odb10 database. 99.56% (99.27% complete, 0.29% fragmented) BUSCO genes were detected in the assembly, with a low duplication rate (1.24%), indicating high completeness of the reference genome (Table 1).

Code availability

All software were executed with default parameters except for those which were listed below.

fastp: -n 0 -f 5 -F 5 -t 5 -T 5.

Guppy: –c dna_r9.4.1_450bps_fast.cfg.

NextDenovo: read_cutoff: 1k, seed_cutoff: 36660.

Purge Haplotigs: -l 0 -m 110 -h 200 -a 60.

minimap2: -x map-ont.

exonerate: -model protein2genome -percent 50 -score 100 -minintron 20 -maxintron 20000.

GenomeThreader: -species drosophila -intermediate

HMMER: -E 1e-5.

BlastKOALA: Eukaryotes for taxonomy group and family_eukaryotes for KEGG GENES database file to be searched.

IQ-TREE: -m Q.insect + F + R6 -B 1000 -T AUTO.

The codes and scripts for the complete workflow are privided at GitHub (https://github.com/xiaoshan40/Genome-annotation).

Received: 5 November 2024; Accepted: 10 January 2025;
Published online: 27 January 2025

References

1. Zhang, Q. L. & Yuan, M. L. Research status and prospect of grassland caterpillars (Lepidoptera: Lymantriidae). *Pratacultural Science* **30**, 638–646 (2013).
2. Yuan, M. L., Zhang, Q. L., Wang, Z. F., Guo, Z. L. & Bao, G. S. Molecular phylogeny of grassland caterpillars (Lepidoptera: Lymantriinae: *Gynaephora*) endemic to the Qinghai-Tibetan Plateau. *PLoS One* **10**, e0127257 (2015).
3. Zhao, J. R. *et al*. Differential gene expression patterns between the head and thorax of *Gynaephora aureata* are associated with high-altitude adaptation. *Front Genet* **14**, 1137618 (2023).
4. Berman, T. S., Laviad-Shitrit, S., Lalzar, M., Halpern, M. & Inbar, M. Cascading effects on bacterial communities: cattle grazing causes a shift in the microbiome of a herbivorous caterpillar. *ISME J* **12**, 1952–1963 (2018).
5. Zhang, Q. L. *et al*. Comparative transcriptomic analysis of Tibetan *Gynaephora* to explore the genetic basis of insect adaptation to divergent altitude environments. *Sci Rep* **7**, 16972 (2017).
6. Zhang, Q. L. *et al*. Gene sequence variations and expression patterns of mitochondrial genes are associated with the adaptive evolution of two *Gynaephora* species (Lepidoptera: Lymantriinae) living in different high-elevation environments. *Gene* **610**, 148–155 (2017).
7. Yuan, M. L. *et al*. Mitochondrial phylogeny, divergence history and high-altitude adaptation of grassland caterpillars (Lepidoptera: Lymantriinae: *Gynaephora*) inhabiting the Tibetan Plateau. *Mol Phylogenet Evol* **122**, 116–124 (2018).
8. Wang, H. *et al*. Genetic diversity and population structure of *Gynaephora qinghaiensis* in Yushu prefecture, Qinghai province based on the mitochondrial *COI* gene. *Biochem Genet* **59**, 1396–1412 (2021).
9. Wang, H. Z. *et al*. Analysis of the *Gynaephora qinghaiensis* pupae immune transcriptome in response to parasitization by *Thektogaster* sp. *Arch Insect Biochem Physiol* **100**, e21553 (2019).
10. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
11. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol* **20**, 129 (2019).
12. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. *arXiv preprint arXiv:1308.2012* (2013).
13. Hu, J. *et al*. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol* **25**, 107 (2024).
14. 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).
15. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
16. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob DNA* **10**, 48 (2019).
17. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457 (2020).
18. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11 (2015).
19. Hubley, R. *et al*. The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**, D81–89 (2016).
20. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, 4 10 11–14 10 14 (2009).
21. Bruna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform* **3**, lqaa108 (2021).
22. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
23. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
24. Gremme, G., Brendel, V., Sparks, M. E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology* **47**, 965–978 (2005).
25. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).
26. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
27. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
28. Potter, S. C. *et al*. HMMER web server: 2018 update. *Nucleic Acids Res* **46**, W200–W204 (2018).
29. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol* **428**, 726–731 (2016).
30. Toronen, P., Medlar, A. & Holm, L. PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res* **46**, W84–W88 (2018).
31. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019).
32. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772–780 (2013).
33. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
34. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268–274 (2015).
35. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587–589 (2017).
36. Kawahara, A. Y. *et al*. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc Natl Acad Sci USA* **116**, 22657–22663 (2019).
37. Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301–302 (2003).
38. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2021).
39. Xiao, S. Rapidly evolving gene families in *Gynaephora qinghaiensis*. *Figshare* https://doi.org/10.6084/m9.figshare.27285153 (2024).
40. Barter, R. L. & Yu, B. Superheat: an R package for creating beautiful and extendable heatmaps for visualizing complex data. *J Comput Graph Stat* **27**, 910–922 (2018).
41. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR24032119 (2024).
42. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR24032120 (2024).
43. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31034734 (2024).
44. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31034740 (2024).
45. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31034741 (2024).
46. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31034731 (2024).
47. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31034732 (2024).
48. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31034733 (2024).

49. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31034728 (2024).
50. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31034729 (2024).
51. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31034730 (2024).
52. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31034727 (2024).
53. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31034738 (2024).
54. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31034739 (2024).
55. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31034735 (2024).
56. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31034736 (2024).
57. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31034737 (2024).
58. *NCBI Assembly* https://identifiers.org/insdc.gca:GCA_042920415.1 (2024).
59. Xiao, S. GynQin.OGS.cds.fa. *figshare* https://doi.org/10.6084/m9.figshare.27292971 (2024).
60. Xiao, S. GynQin.OGS.pep.fa. *figshare* https://doi.org/10.6084/m9.figshare.27292983 (2024).
61. Xiao, S. GynQin.OGS.gff. *figshare* https://doi.org/10.6084/m9.figshare.27292989 (2024).
62. Simao, 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).

Acknowledgements

We are grateful to Professor Gongyin Ye from Zhejiang University for help in collecting grassland caterpillar samples and promoting the study. This work was supported by the National Key Research and Development Program of China (no. 2022YFD1401102 to Q.F.), and the Key Research and Development Program of Qinghai Province (no. 2023-NK-152 to Y.L.).

Author contributions

Q.F. and Y.L. conceived of this project. S.X. participated in the data analysis. S.X., Y.L. and M.Q. collected the samples. S.X. and Y.L. wrote the manuscript. Q.F., X.Y. and F.W. revised the manuscript. All authors have read, revised, and approved the final manuscript for submission.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Q.F.