

Fall webworm genomes yield insights into rapid adaptation of invasive species

Ningning Wu^{1,2,6}, Sufang Zhang^{3,6}, Xiaowei Li^{1,6}, Yanghui Cao¹, Xiaojing Liu¹, Qinghua Wang³, Qun Liu¹, Huihui Liu^{1,3}, Xiao Hu¹, Xuguo J. Zhou⁴, Anthony A. James⁵, Zhen Zhang^{3*}, Yongping Huang^{1*} and Shuai Zhan^{1*}

Invasive species cause considerable ecological and economic damage. Despite decades of broad impacts of invasives on diversity and agriculture, the genetic adaptations and near-term evolution of invading populations are poorly understood. The fall webworm, *Hyphantria cunea*, a highly successful invasive species that originated in North America, spread throughout the Northern Hemisphere during the past 80 years. Here, we use whole-genome sequencing of invasive populations and transcriptome profiling to probe the underlying genetic bases for the rapid adaptation of this species to new environments and host plants. We find substantial reductions in genomic diversity consistent with founder effects. Genes and pathways associated with carbohydrate metabolism and gustatory receptors are substantially expanded in the webworm genome and show strong signatures of functional polymorphisms in the invasive population. We also find that silk-yielding-associated genes maintained a relatively low level of functional diversity, and identify candidate genes that may regulate the development of silk glands in fall webworms. These data suggest that the fall webworm's ability to colonize novel hosts, mediated by plasticity in their gustatory capabilities along with an increased ability to utilize novel nutrition sources and substrates, has facilitated the rapid and successful adaptation of the species throughout its range.

In addition to economic impacts and severe negative consequences for biodiversity, invasive species offer excellent models to study how rapid adaptation occurs¹. They represent natural ecological and evolutionary experiments unfolding in a recent historical time frame, involving classic events in population genetics, such as bottlenecks, inbreeding depression, and variations and mutations in response to intense selection². The fall webworm, *Hyphantria cunea*, is native to North America and was introduced inadvertently into Europe and Asia in 1940 and 1945, respectively³ (Supplementary Fig. 1). Since then, they have spread widely and colonized tens of Northern Hemisphere countries³ (Supplementary Fig. 1). As is typical of many invasive species, the fall webworm has become a major destructive pest to forests, fruit trees and even field crops in invaded regions⁴, while maintaining minor economic importance in its native range.

The robust adaptability of the fall webworm probably results from its high polyphagy and reproductive capability. Fall webworms feed on more than 600 species of trees and shrubs⁵ and adult females may lay as many as 900 eggs each⁶. As the name denotes, the fall webworm also is a remarkable web-maker, and newly emerged larvae construct massive silk web nests to increase their survival by facilitating feeding and escaping natural enemies (Supplementary Fig. 2). However, the molecular bases underlying these attributes and whether they are involved in adaptations of invasive forms are unknown. The introduction in 1979 of the webworm in China and its subsequent spread are well documented⁷. Since the first reports, it has expanded its range rapidly, causing unprecedented economic

losses⁴. Thus, the Chinese fall webworm is a representative population for studying invasion. We use genome and transcriptome analyses to probe the genetic bases of the rapid adaptation of the invasive forms in China.

Results and discussion

Assembly, annotation and gene editing for the fall webworm genome. We carried out a classic shotgun sequencing approach by combining high coverage of long reads and stepwise-increased long-insert libraries (Supplementary Table 1). A total of 148 gigabases (Gb) of sequencing data (289× coverage) was used to generate the draft assembly, with an assembled size of 513.8 Mb and a scaffold N50 size of 1.1 Mb (Supplementary Table 2). The fall webworm genome has a high degree of both completeness and congruity when compared with published genomes of other lepidopteran species (Supplementary Table 3). The consensus gene set was generated by integrating ab initio predictions, lepidopteran homology and transcriptome data from different *H. cunea* samples (Supplementary Tables 4 and 5; Methods). A total of 15,799 protein-encoding genes comprise the core set, with 99.3% of these expressed in at least 1 tissue or developmental stage (Supplementary Table 6). Quality controls support the high quality of our gene set in comparison with other lepidopteran species (Supplementary Table 7).

We developed a gene manipulation system in the fall webworm to probe gene function in vivo to complement and verify aspects of the sequencing data. We implemented a CRISPR–cas9 genome-editing approach by focusing on genes with easily visible

¹CAS Key Laboratory of Insect Developmental and Evolutionary Biology, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China. ²University of Chinese Academy of Sciences, Shanghai, China. ³Key Laboratory of Forest Protection, Research Institute of Forest Ecology, Environment and Protection, Chinese Academy of Forestry, State Forestry Administration, Beijing, China. ⁴Department of Entomology, University of Kentucky, Lexington, KY, USA. ⁵Department of Molecular Biology and Biochemistry, University of California, Irvine, Irvine, CA, USA. ⁶These authors contributed equally: Ningning Wu, Sufang Zhang, Xiaowei Li. *e-mail: szhan@sibs.ac.cn; yphuang@sibs.ac.cn; zhangzhen@caf.ac.cn

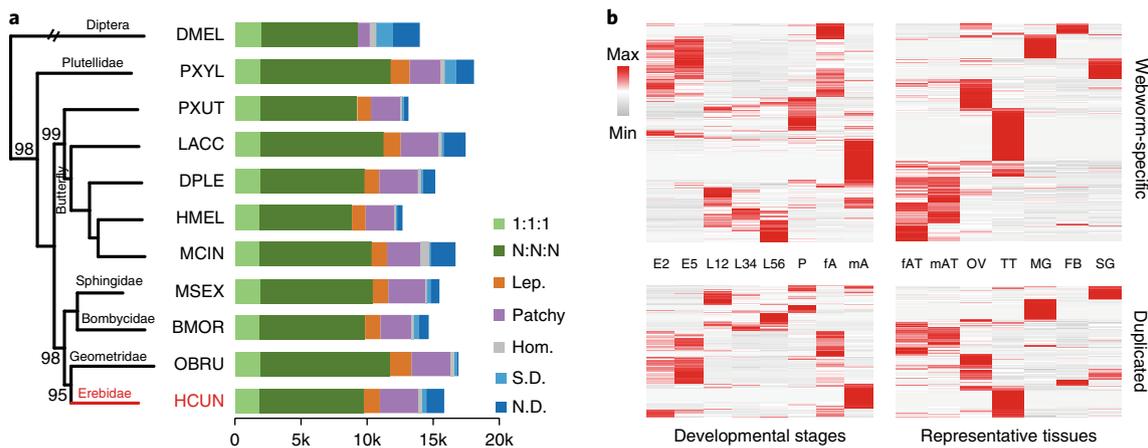


Fig. 1 | Genome evolution of the fall webworm. a, Lepidopteran orthology across ten sequenced species as follows: PXYL, *Plutella xylostella*⁴⁵; PXUT, *Papilio xuthus*⁷²; LACC, *Lerema accius*⁷³; DPLE, *Danaus plexippus*⁴²; HMEL, *Heliconius melpomene*⁶⁶; MCIN, *Melitaea cinxia*⁷⁴; MSEX, *Manduca sexta*⁷⁵; BMOR, *Bombyx mori*⁷⁴; OBRU, *Operophtera brumata*⁷⁶; HCUN, *Hyphantria cunea*. The maximum-likelihood phylogeny was calculated on the basis of the concatenated alignments of 558 exactly single-copy proteins in each species (Supplementary Table 16). The tree was rooted using the Diptera species *D. melanogaster* (DMEL)⁴³. Bootstrap values based on 100 replicates are equal to 100 for most nodes except those labelled with a number. The gene set of each species was subdivided to represent different types of orthology cluster as indicated. '1:1:1' represents universal single-copy gene families across all examined species, but absence and/or duplication in at most one genome is tolerated; 'N:N:N' indicates other universal genes; 'Lep.' indicates common gene families unique to Lepidoptera; 'S.D.' means species-specific duplication; 'N.D.' represents species-specific genes. See detailed information in Supplementary Table 17. **b**, Expression profiles of fall webworm-specific genes and -specific duplicated genes. Representative tissues were dissected from the third day of fifth instar larvae, including the ovary (OV), the testis (TT), the midgut (MG), the fatbody (FB) and the silk gland (SG). Antennae were dissected from female (fAT) and male adults (mAT), respectively. Whole bodies were used to represent every developmental stage as follows: E2, the second day of eggs; E5, the fifth day of eggs; L12, a pool of the second days of first and second instar larvae; L34, a pool of the second days of third and fourth instar larvae; L56, a pool of the second days of fifth and sixth instar larvae; P, pupae; fA, female adults; mA, male adults. Scaled expression per gene is presented.

morphological phenotypes. For example, the gene encoding the yellow/major royal jelly protein is conserved in insects and generally affects coloration^{8,9}. We successfully introduced a variety of mutations around the targeted sites and observed phenotypes in different lines (Supplementary Fig. 3). Successful mutagenesis of *yellow* demonstrates the availability of gene editing in *H. cunea* and affords a set of tools to do functional research in the webworm.

Specific genes and gene family expansions in the fall webworm. Comparative genome evolution was initiated by identifying orthologous genes across 10 lepidopteran species, using *Drosophila melanogaster* as an outgroup, and a phylogenetic tree was constructed using 558 single-copy conserved genes (Fig. 1). The fall webworm genome represents the first sequenced Erebidae genome, which was placed as the sister group of Geometridae (Fig. 1a). Overall, we found that the composition of lepidopteran proteome is relatively common; that is, approximately 10,000 genes (~66% of each set) are generally conserved across most lepidopteran species (Fig. 1a).

Novel genes and duplications are thought to be primary forces in species-specific adaptation and divergence by allowing the differentiation of gene functions^{10,11}. Thus, they may provide insights into the understanding of fall webworm-specific attributes. We annotated 1,648 species-specific genes and 225 orthologous groups with specific duplications in the fall webworm genome (Fig. 1a). Transcriptomes of representative developmental stages and tissues were used to assess their possible functions in biological processes (Supplementary Table 5). Compared with other tissues, we found that a substantial proportion of webworm-specific and -duplicated genes are highly expressed in the antennae (Fig. 1b). Given that the antennae are the primary olfactory sensors of insects^{4,12}, the biased expression pattern supports the hypothesis that these play a central role in locating hosts and/or mating. We also found specific genes that are expressed highly in ovary or testis samples and exclusive to

each other (Fig. 1b). Correspondingly, temporal profiles revealed that the expression of specific genes was more likely to occur in adult, exclusively, or egg stages, consistently (Fig. 1b). Further functional studies may help explore the relationship between the enriched pattern in reproduction-related tissues or stages and the great reproductive capacity in the fall webworm.

Classification by domains provides an alternative way to predict expansions of gene families. We found 142 InterPro domains¹³ over-represented in the webworm genome, and unexpectedly found that a substantial fraction of these over-represented domains are related to carbohydrate metabolism (Supplementary Table 8). Other over-represented domains highlighted gene regulation and DNA repair. Interestingly, 'gustatory receptor' (GR) is the only chemoreception-related domain across the whole list in the fall webworm (Supplementary Table 8). Gene families of expansion may enable the potential of regulation plasticity in related functions.

Population signatures of invasive fall webworms in China.

We sought to characterize evolutionary signatures of the invasive population, as well as the adaptive mechanism that may have facilitated the recent invasion of fall webworms. To do this, we sequenced 20 diploid genomes of the invasive population in China (Supplementary Table 9). The pilot population genomics study included samples of five distinct sites along the stepwise expansion routes of fall webworms in China (Supplementary Fig. 1), and found that the geographic relationship of these sites was well supported by both the population structure and the inferred phylogenetic relationship across these samples (Supplementary Fig. 4).

We first applied *k*-mers analyses to compare genome characteristics between invasive fall webworms and other native lepidopteran species in China (Fig. 2a). All local species represented a bimodal distribution of the *k*-mer coverage, as well as evident heterozygous peaks at half the coverage, indicating a typical signature of

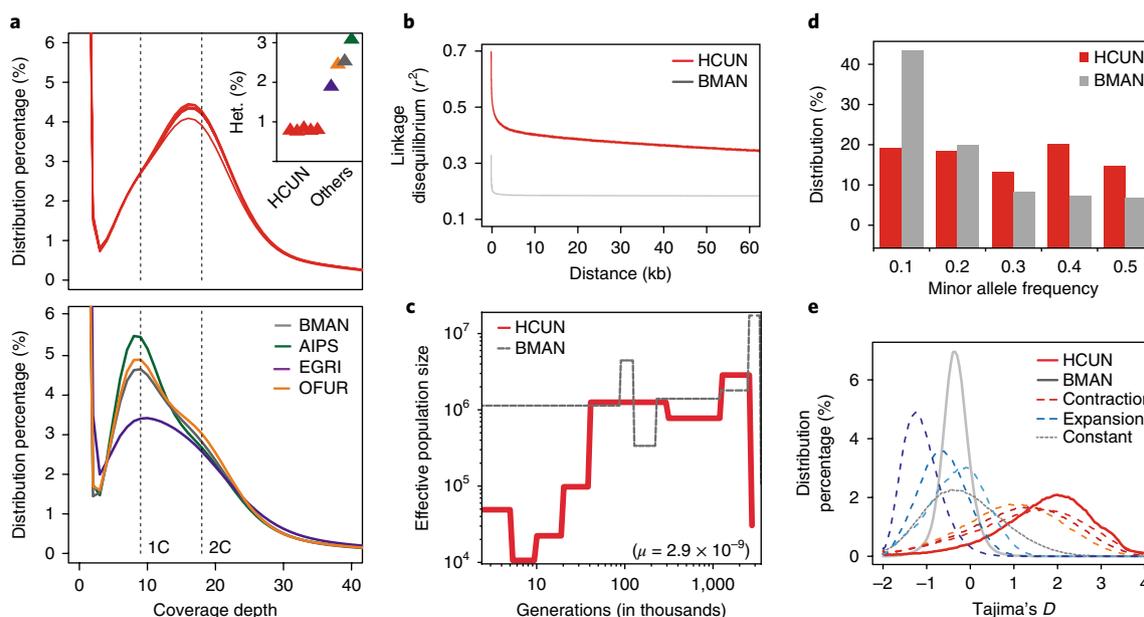


Fig. 2 | Genetic signatures of recent invasion in the fall webworm genomes. **a**, Distributions of k -mer (17-base oligonucleotide) frequency spectrum in invasive fall webworms in China (a male individual from each sampling site is presented; HCUN). In comparison, we selected four representative Chinese native lepidopteran species (AIPS, the black cutworm, *Agrotis ipsilon*; OFUR, the Asian corn borer, *Ostrinia furnacalis*; BMAN, the wild silkworm, *Bombyx mandarina*; EGRI, the tea Geometrid, *Ectropis grisescens*). To exclude the potential impacts of sex chromosomes, male samples (ZZ) were used in the analyses, except for EGRI (Supplementary Table 18). Each sample was normalized to $\sim 18\times$ sequencing coverage for cross-species comparisons. Inset: triangles indicate the genome-wide heterozygosity of fall webworms (HCUN, in red) and other involved species (colours correspond to below). **b**, Decay of linkage disequilibrium, indicated as r^2 , in five diploid individuals from distinct sampling sites (Supplementary Tables 9 and 18). **c**, Inferred population history by the SMC++ algorithm¹⁴. Per-generation mutation rate was assumed as 2.9×10^{-9} , which was estimated in *Heliconius* (Lepidoptera)⁶⁶. See an alternative estimation using MSMC¹⁵ in Supplementary Fig. 5. **d**, Distributions of MAF across all genomic SNPs. See the analysis based on the entire MAF spectrum in Supplementary Fig. 6. **e**, Distribution of Tajima's D^{17} across all 5-kb genomic windows. The dashed lines indicate distributions of simulated data sets under different evolutionary scenarios, which were simulated by ms⁶⁹ (grey, constant; light blue, instantaneously growing; blue, exponentially growing; dark blue, instantaneous expansion followed by bottleneck; golden red, exponential decline; red, severe bottleneck followed by exponential growth; dark red, instantaneously shrinking; see Methods for detailed information; see Supplementary Fig. 15 for the distribution free of demographic effects).

appreciable heterozygosity in diploid genomes (Fig. 2a). In contrast, the k -mer distributions of fall webworms presented minor heterozygous peaks (Fig. 2a). Consistently, the genome-wide heterozygosity rates of invaded fall webworms were estimated in a range between 0.75% and 0.83%, which is much lower than that of local species (from 1.89% to 3.08%). The extent of linkage disequilibrium (LD) can also provide insights into the demographic history of a population. The level of LD in the invasive population of fall webworms decayed to its half-maximum until 60kb, while the wild silkworm (*Bombyx mandarina*) native to China rapidly dropped to a much lower background level within 3 kb (Fig. 2b). The observations on global patterns of diversity and LD both suggest the scenario that this invasive population experienced bottlenecks in population size during introduction to China. We further inferred the population size history of invasive fall webworms in China from distributions of coalescent times. Two independent approaches, SMC++¹⁴ and multiple sequentially Markovian coalescence (MSMC)¹⁵, both revealed a concordant demography that invasive fall webworms have experienced dramatic declines in population size recently. The time frame of tens of thousands of generations indicates that the decline might have started before the invasion of China (Fig. 2c and Supplementary Fig. 5).

When a population undergoes a bottleneck, allelic diversity is generally reduced faster than heterozygosity¹⁶. Indeed, we found a relatively low frequency of rare alleles in the invasive population of fall webworms compared with the local species (Fig. 2d and Supplementary Fig. 6). We also observed an overall positive distribution of Tajima's D^{17} in the invasive population of fall webworms

(Fig. 2e), which was in line with the excess pattern of high-frequency polymorphisms, the consequence of a population contraction. These multiple signatures support a consistent bottleneck scenario in the demography of invasive fall webworms in China. We note that the outbreak of fall webworms in China is so recent that it does not leave a signature of population expansion in the genome.

Functional polymorphisms are enriched in carbohydrate metabolism-related pathways. We next sought to explore how fall webworms became successful invaders. Selection and adaptive genetic changes may play important roles in the evolution of invasive species¹⁸. We note that the distribution of Tajima's D in invasive fall webworms even deviates in a more positive direction from the simulated scenario of a severe bottleneck, for example, reduction in population size to a particularly low fraction ($1/10,000$) within a very short period ($0.4N_0$ generations) (Fig. 2e), which led us to speculate that potential selection favouring polymorphisms may jointly shape the observed distribution in addition to the demographic effect.

We investigated the evolutionary forces on different classes of variants, since they are interspersed throughout the genome and equally affected by demographic events. Genetic diversity, as measured by the value of heterozygosity (H_e) with Hardy–Weinberg equilibrium (HWE) expectation, showed similar levels across protein-coding regions (CDS), introns, untranslated regions and intergenic sequences (Supplementary Fig. 7a). However, we found that the observed frequency of heterozygotes (H_o) is significantly elevated in CDS compared to other classes of sites ($P < 2.6 \times 10^{-49}$, Z test)

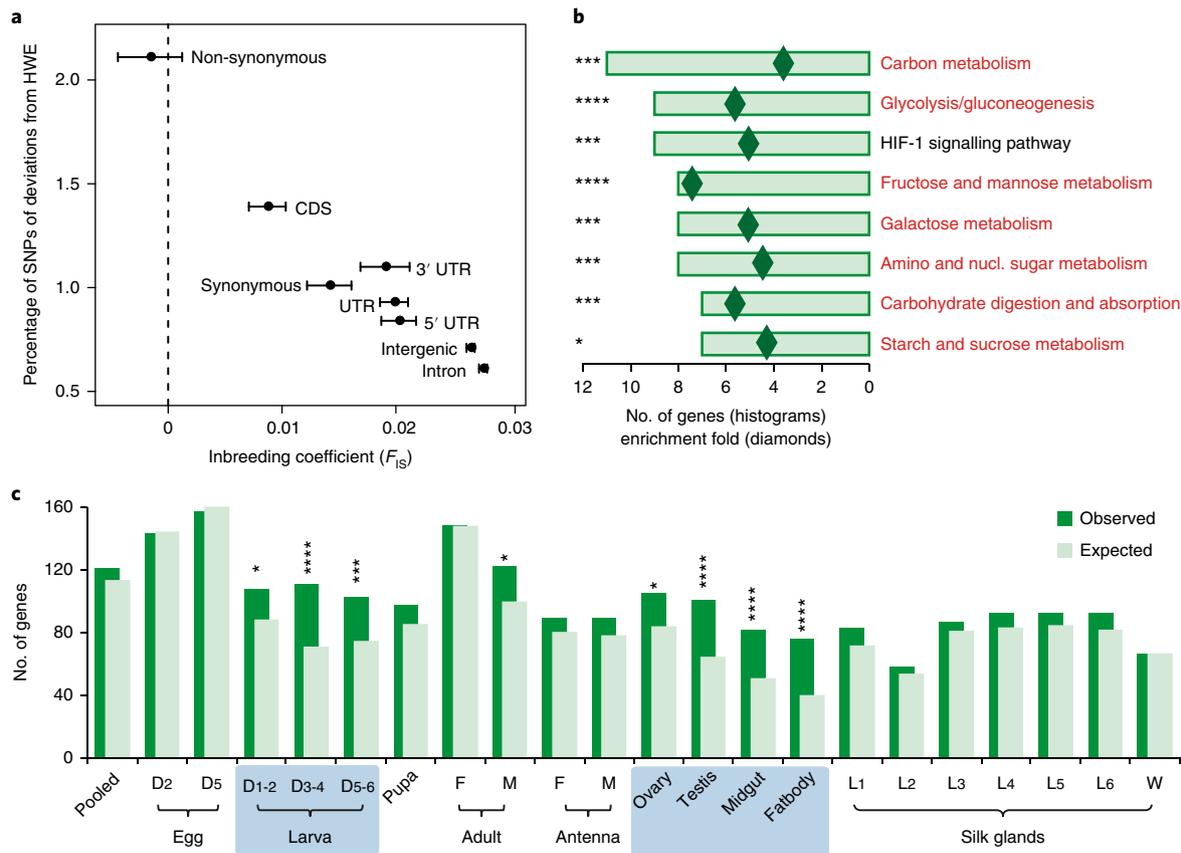


Fig. 3 | Genetic diversity associated with gene functions. **a**, Statistics of SNPs of significant deviation from HWE (chi-squared test, $P < 0.01$) and of higher observed heterozygosity than expected. The error bars indicate \pm s.e.m. $\times 3$. We note that variants located within splice sites (GU-AG), another type of variations that cause functional polymorphisms (alternative splicing), are of the most negative F_{IS} (mean = -0.0268 ; s.e.m. = 0.0104) and of the highest proportion of SNPs deviated from HWE (2.54%), although the data were not plotted due to the extremely small sample size for statistics. **b**, All significantly enriched pathways (with multiple testing) corresponding to the gene set of greater frequency of non-synonymous polymorphisms ($\pi_N > \pi_S$). Pathways in red indicate those related to carbohydrate metabolism. Gene counts and enrichment fold are presented by histograms and diamonds, respectively. Hypergeometric test (FDR-adjusted): * $P < 0.05$, *** $P < 0.005$, **** $P < 0.001$. Details of significance are listed in Supplementary Table 10. **c**, Enrichment analysis of expression in a full list of carbohydrate metabolism-related genes including all non-synonymous variations. Larvae and related tissues are shown in blue shadows. Hypergeometric test: * $P < 0.05$, *** $P < 0.005$, **** $P < 0.001$.

(Supplementary Fig. 7a). We further classified CDS polymorphisms into non-synonymous (amino-acid-changing) and synonymous (silent) sites. As expected, the genetic diversity (H_e) of non-synonymous sites is lower than synonymous sites (Supplementary Fig. 7a). However, we found an unusually high level of H_e , relative to H_e , in non-synonymous sites (hence, a slightly negative inbreeding coefficient F_{IS}), while H_e is significantly lower than expected in synonymous and other non-coding sites (Fig. 3a and Supplementary Fig. 7a). Consistently, non-synonymous sites contained the highest proportion of heterozygote-excess sites of significant deviation from HWE (Fig. 3a). Remarkably, the heterozygotes excess in non-synonymous sites was getting more evident along the increase of minor allele frequency (Supplementary Fig. 7b,c). We speculate that the striking observations of biased functional heterozygotes are probably due to the selection that maintains multiple functional alleles of advantageous adaptation.

We next analysed effects of selection on local variations by comparing the diversities of synonymous (π_S) and non-synonymous polymorphisms (π_N). Approximately half of the fall webworm genes showed values of π_N/π_S lower than 0.1, due to a much higher fraction of non-synonymous sites being generally deleterious than synonymous mutations. Although demographic effects, such as relaxed

selection on nearly neutral mutations in a contracted population, may lead to a limited elevation of π_N/π_S , the further elevated π_N/π_S is more likely owing to selection that favours diversity (Supplementary Fig. 8), for example, diversifying selection or balancing selection^{19,20}. We isolated a total of 771 genes of greater π_N than π_S ($\pi_N/\pi_S > 1$) as the potential repertoire of functional diversity. This gene set was significantly enriched in 11 biological pathways (Supplementary Table 10). We were surprised to find that a majority of these enriched pathways are related to carbohydrate metabolism (Fig. 3b). In insects, nutrients are essential for meeting energetic needs associated with fuelling growth in larvae and reproduction in adults²¹. The fall webworm is supposed to accumulate more nutrients in the sub-adult stages that would support high reproductive performance in the ensuing adults. Correspondingly, we further identified a full set of 455 carbohydrate metabolism-related genes of potential function diversity and found that they were significantly more likely to be expressed in the larval stages, the only feeding stage of most lepidopterans, and associated tissues (Fig. 3c). Carbohydrates and proteins are particularly important dietary nutrients. Since the variation of these two nutrients in plants is much greater than in animals²², the capability of response to nutritional imbalance is thus more important for herbivores, especially for polyphagous species,

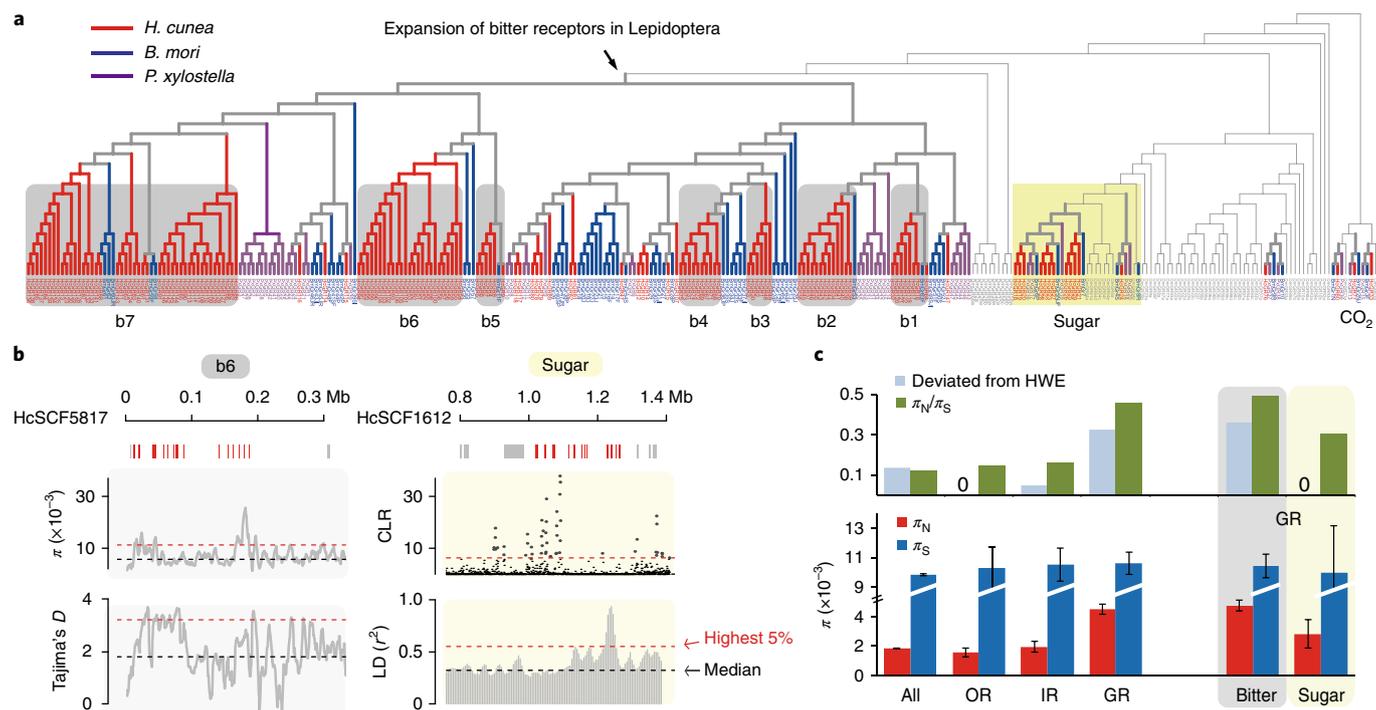


Fig. 4 | GR genes experienced substantial expansion and potential selection. **a**, Maximum-likelihood phylogenetic analysis of GR genes in the fall webworm (in red) and three other species in comparison. The domesticated silkworm, *B. mori* (in blue), and the diamondback moth, *P. xylostella* (in purple), are the representative monophagous and oligophagous species in Lepidoptera, respectively. *D. melanogaster* (in grey) was also included in the analysis to annotate the subgroups of GRs. Grey shadows (b1–b7) indicate seven clusters of bitter GRs with massive expansion in the fall webworm, while the yellow shadow indicates the cluster of sugar GRs. **b**, Population genetic statistics in 5-kb sliding windows across the b6 and sugar cluster, respectively. The red and black dashed lines indicate the upper 5% cutoffs and median values, respectively, across the fall webworm genome. **c**, Distribution of heterozygote-excess polymorphisms and ratios of π_N/π_S in different categories of genes. Upper blue, percentage of SNPs of higher observed heterozygosity than expected and of significant deviation from HWE ($P < 0.01$, chi-squared test); upper green, the ratios of π_N/π_S ; lower, mean levels of π_N and π_S are plotted in red and blue histograms, respectively, with s.e.m. indicated by error bars. OR, olfactory receptor; IR, ionotropic receptor; GR, gustatory receptor.

including the fall webworm. Supporting evidence for this is that the fall webworms in China present a great ability to consume mulberry leaves, which are fairly high in protein content and used exclusively for rearing domesticated silkworms in Asia. Thus, we supposed an adaptive scenario that maintaining multiple functional polymorphisms in carbohydrate metabolism-related genes is likely to be selectively advantageous in invasive fall webworms, probably owing to the improved flexibility and plasticity of utilizing nutrients from a wide range of novel hosts.

GRs are massively expanded and under divergent selections.

Equally important as utilizing hosts is gene families probably involved in perception of hosts and energy. Invasive species and populations require interaction with novel ecological spaces and host niches. During interactions with plants, insect herbivores evolved both host-recognition capabilities and resistance to the accompanying defensive compounds, via chemoreception and detoxification, respectively. We annotated the full set of chemosensory and detoxification-related genes in the available lepidopteran genomes (Supplementary Table 11). Overall, the fall webworm genome encodes a similar repertoire of these components to other moths. A notable exception is the GR family in which we identified a total of 147 components, representing a massive gene expansion in the fall webworm genome (Supplementary Tables 8 and 11). Phylogenetic analysis reveals greater expansions of GRs in putative bitter receptors (Fig. 4a). The genomic arrangements of GR genes shows that a majority of them clustered on the same scaffold (Fig. 4b), suggesting that the expansions originate from duplication events after speciation. Combined with the recent discoveries in

polyphagous Noctuidae^{23,24}, we speculate that the expansion of bitter GRs occurred in parallel in polyphagous species and represents a general adaptation mechanism for polyphagy in Lepidoptera.

In addition to the association between gene expansion and polyphagy, we subsequently examined whether the evolution of GRs play roles in the invasive adaptation. We found evidence of selection along the genomic locations of a subset of GRs and diverse signatures within different clusters (Fig. 4b). A subset of sugar receptor genes may be subject to directional selection, based on substantial signatures of both the composite likelihood ratio (CLR) test and the near-complete LD (Fig. 4b). In contrast, a number of bitter GRs presented evidence of selection favouring diversity, such as the top level of D in the Tajima test and prominent peaks of π (Fig. 4b). The increased diversity in bitter GRs was further supported by the relatively high proportion of heterozygote-excess variants as well as the greater distribution of π_N/π_S than other classes of genes (Fig. 4c). We hypothesize that variations that occurred in sugar receptors are probably deleterious for perception of sugars, which serve as some of the most easily metabolized forms of energy for fall webworms. Instead, selection favouring polymorphisms on bitter receptors may beneficially allow invasive populations to detect novel hosts as intake of alternative nutrients. It is worth noting that we found only relatively neutral signatures across olfactory receptor genes, and that the number of olfactory receptors in the fall webworm is comparable to other lepidopterans (Supplementary Fig. 9 and Supplementary Table 11). We thus hypothesize that fall webworms apply taste but not odour, which leads to less selectivity and permits an expanded host range to explore novel ecosystems.

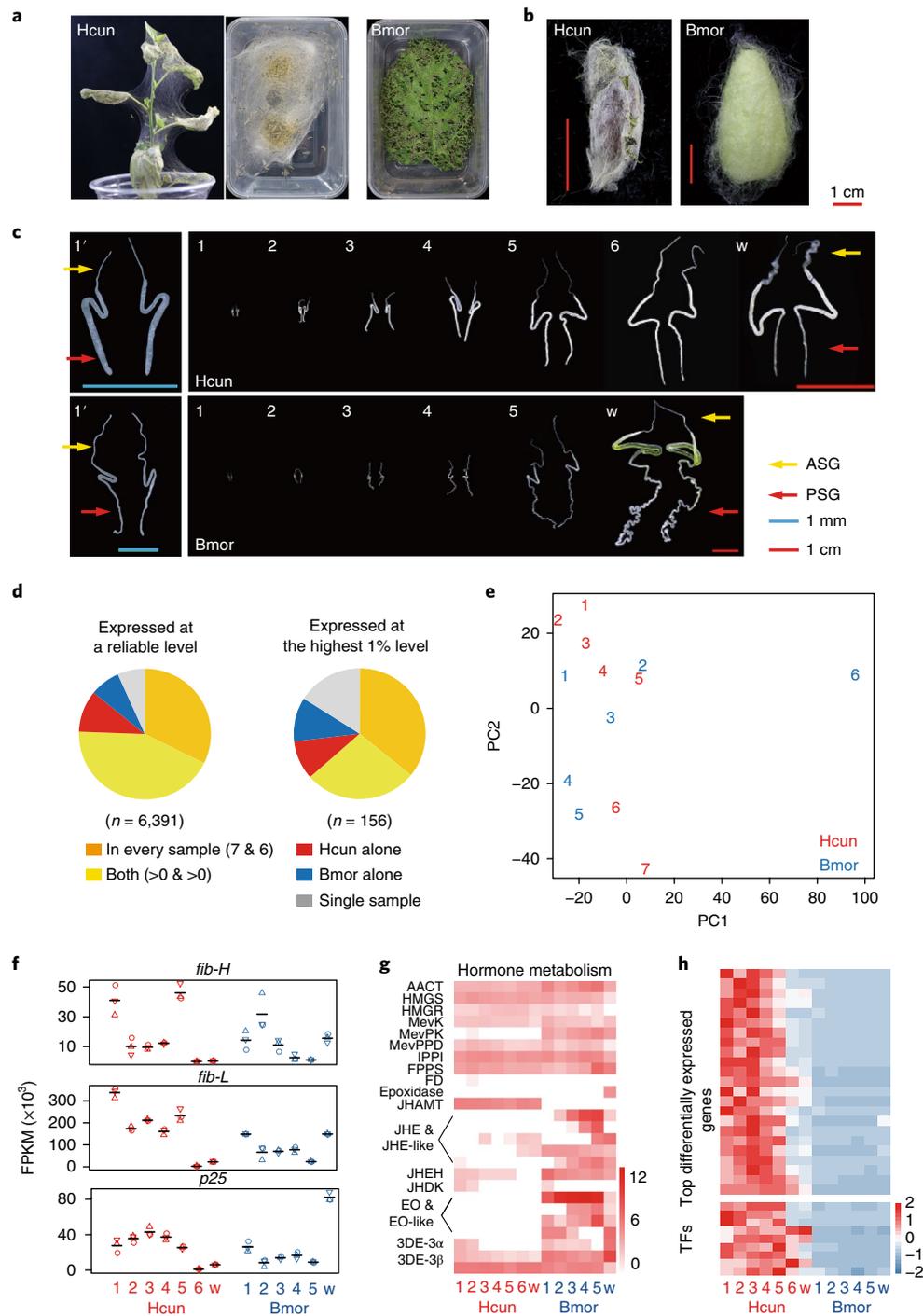


Fig. 5 | Divergent regulation of silk gland development between the domesticated silkworm and the fall webworm. **a**, Fall webworms (Hcun) secrete a massive amount of silk to build webs on both host plants and artificial diet at the early stages of larvae. **b**, In contrast, silkworms (Bmor) produce a large amount of silk at a late developmental stage of larvae and spin cocoons of better size and quality. **c**, Anatomical comparison of the silk glands between the silkworm and the webworm. 1–6, the instar of a larval stage; 1', the first instar in enlarged view; w, the wandering stage; ASG, anterior silk gland; PSG, posterior silk gland. **d**, Distributions of expressed genes at different levels. Every sample, genes are expressed in the silk glands of all seven instars of fall webworm larvae as well as all six instars of silkworm larvae; both, genes are expressed in at least one instar of fall webworm larvae as well as one of silkworm larvae; Hcun alone, genes are expressed only in the fall webworm; Bmor alone, genes are expressed only in the silkworm; single sample, genes are expressed only in one instar of fall webworm larvae or silkworm larvae. Genes were shown inclined to commonly expressed in the silk glands of each larval instar in both webworms and silkworms. **e**, Principal component analysis of the examined silk gland samples based on their overall expression profiles. The first two eigenvectors are plotted. See the same analysis on independent biological replicates in Supplementary Fig. 17. **f**, Expression patterns of three core genes involved in silk production: the heavy-chain fibroin (*fib-H*), light-chain fibroin (*fib-L*) and P25 (*p25*). The horizontal lines indicate mean values of three biological replicates. **g**, Expression profiles of genes related to JH and ecdysone metabolism. **h**, Expression profiles of genes differentially expressed in silk glands of fall webworms. TFs, transcription factors. See details in Supplementary Tables 14 and 15.

On the other hand, we did not find evident gene-copy variations or specific evolutionary signatures for well-acknowledged detoxification-related gene families, such as P450 and ABC, in fall webworm genomes (Supplementary Table 11). Instead, we identified a cluster of organic cation transporter (OCT) genes in the strongest selective sweep across the whole genome (Supplementary Fig. 10 and Supplementary Table 12). These genes showed selective signatures of the highest CLR, extremely negative values of Tajima's D and near-complete loss of diversity (Supplementary Fig. 10). OCT genes generally serve as uptake transporters of both endogenous substrates and a variety of xenobiotics, such as pesticides and other environmental toxins. Evidence of positive selection in OCT genes therefore suggests a potential role in the environmental adaptation of invasive fall webworms.

Divergent development of the silk glands between the fall webworm and the silkworm. Another aspect of fall webworm biology is their conspicuous webs. We found that genes expressed in silk glands preserve the lowest proportions of functional polymorphisms compared with any other tissues or developmental stages (Supplementary Fig. 11 and Supplementary Table 13), indicative of the role of web creation in housekeeping processes for fall webworms. The biology of silk production has been well characterized in the domesticated silkworm *B. mori*, in which silk utilization is mostly restricted to the construction of cocoons in which the insects pupate. In contrast, the fall webworm provides a model in which massive silk production occurs immediately after the larvae hatch (Fig. 5a) and relatively poor production of cocoons for pupae (Fig. 5b). Given that the yield of silk depends closely on the development of the silk glands, its comparison between the silkworm and the webworm offers an invaluable opportunity to characterize the genetics and evolution of silk production in Lepidoptera.

The silk glands differentiate into anterior, middle and posterior parts²⁵. We dissected the entire silk glands of each larva instar for both silkworms and fall webworms, and found evident anatomical differences in differentiation and composition during development (Fig. 5c). The posterior silk gland of webworms is substantially longer and thicker, in proportion, than that of silkworms in early instars (Fig. 5c). Along with the development of larvae, the volume of the posterior silk gland increases in the silkworm but gradually degenerates in the fall webworm (Fig. 5c). By sequencing the transcriptomes of the silk glands, we generated a real-time profile to explore the difference at the molecular level (Supplementary Table 5). Although a majority of genes, especially of highly expressed genes, were found commonly expressed in both species and across all instars (Fig. 5d), the dynamic changes in expression suggest distinctive patterns across these developmental stages. Principal component analysis of all expressed genes reveals three features of potential interest (Fig. 5e). First, the wandering stages, in which larvae stop feeding and are ready for cocooning, of both webworms and silkworms stood out as outliers and exclusively to each other. Second, developmental stages of both webworms and silkworms were placed continuously along the negative slope between the first two principal components. Third, the first two instars of webworms were placed relatively uniquely along the negative slope, suggesting an advanced developmental process of silk glands in the webworm. The posterior silk gland mainly synthesizes fibroins, the central fibre proteins, which are composed of six sets of heavy-chain (*fib-H*) and light-chain (*fib-L*) proteins and a fibrohexamerin protein (*P25*)²⁶. We identified all of these three genes and preliminarily confirmed the relatedness with the development of the silk glands by mutagenesis of *fib-H* or *fib-L* in the webworm (Supplementary Fig. 12). Consistent with the anatomical difference, the expression profiles of these three crucial genes presented a diverse pattern along the larval development between webworms and silkworms, leading to larval and pupal silk, respectively (Fig. 5f).

We then characterized the underlying mechanisms that result in the diversified development patterns of the silk glands. Hormones, such as juvenile hormone (JH) and moulting hormone (ecdysone), play important roles in the regulation of the development of the silk glands²⁷. It is puzzling that while the enzymes in the initial steps of JH biosynthesis are expressed, the expression of those in subsequent steps was almost absent (Fig. 5g). Notably, the enzyme that mediates the last step leading to JH, JHAMT, was specifically expressed in all larval stages of webworms. Genes related to JH degradation and ecdysone metabolism also showed an opposite pattern between the fall webworm and the silkworm; a majority of them showed deficiency of expression in the webworm.

Transcription factors are other main regulators of silk genes that attracted attention in the silkworm^{28,29}. We identified a total of 330 transcription factor genes, including several known *fib-H* regulators in *B. mori*, and found that they are of overall similar expression patterns between the fall webworm and the silkworm (Supplementary Table 14). Remarkably, we identified nine transcription factors that are expressed substantially and specifically in the webworm as potential key regulators of silk production at early developmental stages (Fig. 5h and Supplementary Table 14). We also identified 23 genes of consistently biased expression in webworms in comparison to silkworms (Supplementary Table 15).

Conclusion

We have leveraged extensive sequencing and comparisons to characterize the genetic basis of biological attributes in the fall webworm and examine their roles in the context of population invasion. It is inspiring that these insights reciprocally complement each other by identifying energy supply and consumption as significant in the fall webworm invasion. Moreover, our studies provide high-quality genomic resources along with an efficient genome editing system for the fall webworm. These basic advances uncover the potential of the webworm as a novel model organism not only for invasive adaptations but also for important biological phenomena in Lepidoptera. Additional population genomic studies, especially for native populations and other stepwise-spread populations, should be forthcoming.

Methods

De novo sequencing and assembly. To minimize the hindrance caused by heterozygosity, the line of *H. cunea* (the fall webworm) used for de novo sequencing has been maintained in the laboratory with inbreeding for more than 60 generations, starting from a wild individual that was caught in 2005 in Tianjin, China (39°N, 117°E). The line was kept at 25°C, in a 16:8 light/dark photoperiod and at a relative humidity of 35–40% in plastic cages and animals were fed an artificial diet. All live fall webworms for this study were processed at the Research Institute of Forest Ecology (CAF) in Beijing, China.

Genomic DNA was isolated from individual pupa using a standard protocol as follows: RNase treatment, proteinase K digestion, phenol and chloroform extraction, and isopropanol precipitation. We used Illumina sequencing platforms to generate deep coverage of sequencing reads. The MiSeq platform was employed to generate paired-end reads of long read length (300 base pairs (bp) in each end), while the HiSeq platform was employed to generate regular paired-end reads and mate-pair reads (150 bp in each end). The MiSeq library was designed to apply an average size of 500 bp, which enabled the paired ends to overlap each other. This library was used to build initial contigs. Libraries of stepwise increased inserts, ranging from 500 bp to 13 kb, were applied to assemble scaffolds. Genomic DNA used for most libraries was prepared from a single individual, except those two libraries of long inserts (8 kb and 13 kb), which required a great amount of DNA. All libraries were constructed following the Illumina standard protocols. Library construction and sequencing were performed by the Chinese National Human Genome Center at Shanghai.

A total of 148.2 Gb of sequencing data, which equals a 288.5-fold change of the genome, was generated to perform de novo assembly of the *H. cunea* reference genome. Details of sequencing data are presented in Supplementary Table 1. On the basis of the clean reads with removal of adaptors, we additionally used seqtk v1.0 (<https://github.com/lh3/seqtk>) to trim bases of low sequencing quality. MiSeq read pairs were first merged to a long read by FLASH v1.2.11³⁰. These merged long reads were subsequently utilized to build contigs using Newbler v2.5.3 (454 Life Sciences) with '-large -force' and other default parameters. Before scaffolding,

initial contigs were processed by Rabbit v2.6.18 (<https://github.com/gigascience/rabbit-genome-assembler>), under default settings, to remove potential redundancy. We used the paired information of long-insert libraries, step-by-step from 500-bp to 13-kb insert size, to join the initial contigs into scaffolds by ScaffoldMatch v0.9³¹. Remaining gaps within these scaffolds were iteratively filled with paired-end reads of 500-bp and 800-bp inserts using GapCloser v1.12 available in SOAPdenovo³². The resulting v1 assembly had a final scaffold N50 length of 1.1 Mb (spanning 513.8 Mb) and contig N50 length of 52.6 Kb (spanning 511.3 Mb).

We first assessed the accuracy of the genome assembly using REAPR, which aims to score every base for accuracy and to automatically pinpoint mis-assemblies³³. As suggested, the mate-pair library of the largest inserts (13 kb) was used to call potential scaffolding errors and a small insert library (500 bp) was used to call base-by-base errors. As a result, REAPR reported a corrected N50 length of 966 kb after breaking the original assembly at breakpoints and a ratio of error-free bases of 93.2% in our assembly. Compared with the presented cases³³, the assembly approach used in this study is at a relatively low level. We further evaluated the completeness of coverage of our assembly using two common pipelines for quantitative measures for the assessment of genome assembly: CEGMA v2.4 (Core Eukaryotic Genes Mapping Approach³⁴) and BUSCO v3 (Benchmarking Universal Single-Copy Orthologs³⁵). Both of these pipelines revealed a high quality of the *H. cunea* reference assembly, compared with other sequenced lepidopteran genomes (Supplementary Table 2). We also aligned the assembled fall webworm transcripts (see below) onto the scaffolds using BLASTN. As a result, 99.5% of the webworm transcripts could be aligned back to the assembly, suggesting a nearly complete representation of genes in the v1 assembly.

Transcriptome sequencing and analysis. Webworms from critical stages of development and representative tissues from the third day of sixth instar larva were used to ensure a good representation of transcripts. Total RNA was independently extracted from each sample using Trizol and subsequently stored at -80°C for further use. Construction of the cDNA library and paired-end RNA-seq (Illumina) were carried out by Berry Genomics Co. Ltd. Statistics of transcriptome sequencing data are listed in Supplementary Table 5.

Raw RNA-seq reads were first filtered of low-quality bases using seqtk v1.0 (<https://github.com/lh3/seqtk>). We used Trinity v2.3.2³⁶ to de novo assemble transcripts based on the RNA-seq reads of a pool library, which was constructed by pooling equal amounts of RNA from all preparations. To help predict genes, the pool library was also mapped to the genome according to the recommended pipeline of TopHat (v2.0.12)³⁷ and Cufflinks (v2.2.1)³⁸. The independent sequenced samples were mapped to the reference using Bowtie v2.2.9³⁹ and determined expression profiles as fragments per kilobase of transcript per million mapped reads (FPKM) by RSEM v1.3.0⁴⁰. On the basis of the overall distribution, we determined the genes with FPKM higher than 12.2 as reliably expressed genes.

Genome annotation. We identified repetitive sequences and transposable elements using RepeatMasker v4.0.5 (<http://www.repeatmasker.org>) against a de novo repeat library that was built by RepeatModeler v1.0.7 (<http://www.repeatmasker.org>), as well as the arthropod set of Repbase v1.40⁴¹. Non-interspersed repeat sequences were also identified by RepeatMasker with the ‘-noimt’ option.

Official gene models were determined as described previously⁴². Briefly, we first generated independent gene sets. For homology-based sets, we used NCBI RefSeq proteins, *D. melanogaster* proteins (FlyBase⁴³) and several lepidopteran gene sets, including *Bombyx mori*⁴⁴, *Plutella xylostella*⁴⁵ and *Danaus plexippus*⁴², for homology search by TBLASTN with $E < 10^{-5}$. The high-scoring pairs were then processed by genblastA v1.0.1⁴⁶ and predicted gene structures by GeneWise v2.2.0⁴⁷. For ab initio predictions, we applied three different predictors: AUGUSTUS v2.5.5⁴⁸, Genscan v1.0⁴⁹ and SNAP v20060728⁵⁰ (Supplementary Table 1G). We utilized the de novo assembled transcripts, as described above, to train these predictors. Then, all of the above individual gene sets, as well as the transcript-based set, were integrated into a consensus gene set using GLEAN⁵¹ and Maker v2.26⁵², respectively. We evaluated these two consensus sets using 20 cloned fall webworm genes that were downloaded from NCBI. As GLEAN presented better quality than any other set, our official gene set (OGS1.0) was based on the non-redundant GLEAN models, with additional removal of genes that were not supported by either homology or the transcriptome. Details of all gene sets are presented in Supplementary Table 4. Quality controls of OGS1.0 are listed in Supplementary Table 7.

Gene annotation. For each gene in OGS1.0, we searched the homology by querying the *Bombyx*, *Drosophila*, NCBI RefSeq and UniProt protein sets. We also carried out a local run of InterProScan (IPR) search⁵³ with all implemented methods to identify domains, GeneOntology terms and KEGG KO terms for the gene set. All above databases were updated to June 2016 for annotation. Function enrichment analysis (KEGG) was calculated on the basis of a hypergeometric test with multiple testing (false discovery rate (FDR)-corrected).

For chemosensory receptors that are difficult to identify from automated predictions, we directly identified them in the genome assembly using a classic pipeline as described previously⁵⁴. Briefly, candidate genomic loci were identified using TBLASTN searches with the corresponding sets of *B. mori* and *D. melanogaster* as queries; GeneWise v2.2.0 was then used to predict gene

structures based on the loci with significant hits ($E < 10^{-5}$). This process was iteratively performed until no additional gene was found. Abnormal structures were manually curated on the basis of multiple alignments of selected protein sequences using ClustalX. The well-aligned regions were analysed for maximum-likelihood phylogenetic analysis using MEGA 6⁵⁵ with the JTT model and 100 replicates of bootstrapping.

Orthology and phylogenomics. We froze the orthology analysis by the end of 2016; thus, most available gene sets of lepidopteran moths and representative butterflies were employed (Supplementary Table 16). For each set, proteins of very short length (<30 amino acids) were filtered out, and alternative transcripts of a single gene were compared to keep the longest isoform. Orthology analysis was performed following the pipeline of OrthoMCL⁵⁶. In brief, all-against-all protein alignments were performed using BLASTP with $E < 10^{-5}$; high-scoring pairs were processed by orthomclSoftware-v2.0.9 and calculated orthologues, inparalogues and co-orthologues by MCL v14-137⁵⁶ following the suggested parameter values. Orthology grouping data are listed in Supplementary Table 17. The orthologue groups that consisted of only webworm proteins were defined as webworm-specific orthologues, while the groups in which only the webworm set had multiple-copy genes were defined as webworm-specific expansions.

To construct the species tree, we utilized single-copy universal genes across all examined species. Multiple alignments were independently performed for each group using Muscle v3.8.31⁵⁷; the conserved blocks of these alignments were subsequently isolated using Gblocks v0.91b⁵⁸. Conserved blocks of 558 employed genes were concatenated to a super gene, consisting of 96,530 amino acids, for each of the 11 gene sets. The phylogeny was calculated using PhyML v3⁵⁹ with the JTT model for 100 replicates of bootstrapping.

Genome sequencing of the fall webworm population and other species of lepidopteran moths. To investigate unique features of invasive populations, we additionally sequenced wild-caught fall webworms from China to generate population genomic data. Locations of sampling, as shown in Supplementary Table 9, are based on the actual expansion range of fall webworms in China. We sequenced a total of 20 individuals, with 4 individuals (2 females and 2 males) from each sampling site. The caught individuals were immediately subject to DNA preparation without further breeding in the laboratory. Genomic DNA was isolated from individual thoraces using the protocol as described above. All standard PE libraries were constructed and sequenced on Illumina HiSeq platforms by Berry Genomics Co. Ltd. Each sample was sequenced independently for at least 25 \times coverage (Supplementary Table 9).

To perform comparison with local moth populations in China, we further sequenced the whole genomes of three representative species of lepidopteran pests (Supplementary Table 18). These species included the black cutworm (*Agrotis ipsilon*), the Asian corn borer (*Ostrinia furnacalis*) and the tea Geometrid (*Ectropis griseascens*). Importantly, all of these pests were supposed to be distributed in China for a long time. Approximately 30 \times sequencing data were generated by a wild-caught male individual for each species, except for *E. griseascens*, following the same approach of resequencing webworms. We also utilized the raw sequencing data of the wild silkworm⁶⁰, *B. mandarina*, which was also supposed to originate in China. Since this species is not supposed to be a pest and was rarely reported with population explosion in the history, we assumed it as a population under nearly neutral evolution.

k-mer analysis. To exclude the impact of heterozygous sex chromosomes, we used only sequencing data of male samples (ZZ in lepidopteran males), except for *E. griseascens*. One male individual from each sampling site was analysed to represent the pattern of *H. cunea* (Supplementary Table 9). Since male wild-caught samples of *E. griseascens* were unavailable for us when the project was in process, we note that the heterozygosity in *E. griseascens* is partially overestimated due to the existence of unpaired sex chromosomes and also partially underestimated due to a short period of laboratory crossing.

We first trimmed reads with adaptors and bases of low quality as described above. A further run of filtering was performed to remove PCR duplicated pairs. For comparison across species, we normalized the sequencing data to 18 \times for each sample. k-mers were counted by jellyfish v1⁶¹ with 17-base oligonucleotide. Heterozygosity and other characteristics were determined by GenomeScope v1.0⁶².

Population genomics. Reads were first trimmed with bases of low quality using seqtk (<https://github.com/lh3/seqtk>). Clean reads were aligned to the reference assembly using BWA-MEM v0.7.10 with default parameters. Single-nucleotide polymorphism (SNP) calling was performed according to the GATK best practice⁶³. Briefly, a series of subsequent processes were carried out before SNP calling, including sorting reads by SAMTOOLS v0.1.19⁶⁴, removing PCR duplicates by Picard v1.82 (<http://broadinstitute.github.io/picard/>) and realignment of InDels by GATK v3.1.160. We applied the UnifiedGenotyper module of GATK v3.1.160 to call SNPs across all sequenced samples simultaneously. SNPs were further filtered with the following cutoffs: SNP quality > 50 and r.m.s. mapping quality > 40. Genotyping on heterozygous alleles is sensitive to sequencing coverage; that is, incomplete coverage might cause false negative calling of heterozygous SNPs while

excess coverage, usually led by improper mapping on duplicated sequences, might generate false positive heterozygotes. Since problematic heterozygotes can affect numerous signatures, we used only SNPs of sequencing coverage within a normal region in subsequent population genomics ($19\times\text{--}47.5\times$ coverage, mean = 23.8 \times ; see Supplementary Fig. 13).

Genetic relationship across sequenced individuals was inferred on the basis of both the principal component analysis and the phylogenetic tree. Principal component analysis was carried out on synonymous sites and performed with the package EIGENSOFT v7.2.1⁶⁵. Individual phylogeny was based on the matrix of genetic distance as described previously⁶⁰. Briefly, pairwise genetic distances were calculated across all samples using a simple model at each locus: 1, identical genotypes; 0, distinctive homozygous genotypes; 0.5, heterozygous genotypes. Neighbour-joining phylogeny was built on the basis of the resulting distance matrix using MEGA 6⁵⁵.

To perform the demographic analysis, we used two independent approaches based on distributions of coalescent times to estimate the historical population sizes: the MSMC¹⁵ and SMC++¹⁴. The calculation of coalescence in MSMC depends on inferred haplotypes; we thus used BEAGLE v4.1 with default settings to phase the input data in VCF format generated by SAMTOOLS mpileup⁶⁴. We also limited the analysis in uniquely mappable genomic regions that were generated using the SNPable toolkit (<http://lh3lh3.users.sourceforge.net/snpable.shtml>) with default settings. We used MSMC v2.0.0, under the default settings, to estimate the historical effective population sizes of a population on two individuals and all scaffolds with length over 100 kb. All four diploid fall webworms (that is, eight haplotypes) were analysed together for each sampling site. All analyses of five sampling sites resulted in a consistent demographical pattern for invasive fall webworms in China (Supplementary Fig. 5). SMC++ was designed to analyse a large number of samples together while requiring only unphased genomes. We utilized SMC++ to analyse all 20 diploid sequenced fall webworms together. As suggested, input files for SMC++ were transformed from the same VCF file as MSMC by merging all individuals for each genomic scaffold. SMC++ v1.14.0.dev0 was utilized to fit the demographic model with the inputs of all genomic regions under the optimization algorithm of 'Powell'. Estimated rate parameters were scaled to real generations and population sizes using the mutation rate per base pair per generation as 2.9×10^{-9} , which was estimated for *Heliconius*⁶⁶.

Population statistics were calculated in a 5-kb sliding window (500-bp steps) across the whole genome. Tajima's D ¹⁷ and nucleotide diversity (π) were calculated by custom Perl scripts. The CLR was determined by SweepFinder2⁵⁷ with parameters '-sg 500'. LD was initially estimated as r^2 across each scaffold, using Haploview v4.2⁶⁸ with parameters '-maxdistance 100 -dprime -minGeno 0.6 -minMAF 0.1 -hwcutoff 0.001'. We estimated that the genome-wide LD in the fall webworm decays to half at ~ 50 kb; local LD was subsequently calculated in each 50-kb genomic window along the whole genome. Haplotype blocks were called using Haploview v4.2⁶⁸ with parameters '-n -q -memory 16384 -maxdistance 50 -minGeno 0.6 -minMAF 0.03 -hwcutoff 0.001 -blockoutput GAB'. As calling haplotype blocks is an extremely time-consuming process, we only called blocks around each gene by extending 50 kb to both upstream and downstream. Importantly, the above population statistics were shown to be insensitive to sequencing coverage (Supplementary Fig. 14).

Further analyses were focused on local variations by categories. Coding regions (CDS), untranslated regions and introns were determined on the basis of the annotation of OGS 1.0. Observed frequency of heterozygotes (H_o) and expected heterozygosity under HWE (H_e) were calculated independently for each SNP of normal sequencing coverage. A Z test was applied to compare the genome-wide difference between different types. Comparisons were also performed based on the entire minor allele frequency (MAF) spectrum. Non-synonymous diversity per non-synonymous site (π_N) and synonymous diversity per synonymous site (π_S) were independently calculated for each gene of OGS 1.0. Genes of greater π_N than π_S from at least three effective loci were used to identify genes under potential selection. Function enrichment analysis was based on KO terms and was further mapped to the KEGG pathway using the OMICS SHARE cloud platform (<http://www.omicsshare.com/tools/Home/Soft/pathwaygsea>). Hypergeometric and FDR multiple tests were utilized to identify significantly enriched pathways. A full set of carbohydrate metabolism-related genes was further sorted out based on the KO annotations (Class B). Genes of at least three effective loci were considered as functional diversity and processed for enrichment analysis (hypergeometric test) of expression in tissues or developmental stages.

As a control, we also performed the same population analyses using the resequencing data of the wild silkworm, *B. mandarina* (SRP119041)⁶⁰. Controls of neutral population were simulated using ms⁶⁹ under the same sample size and average SNPs per window for a comparable number of data sets with the webworm genome, under the following scenarios: constant, -eN 0.6 1; instantaneously growing, -eN 0.6 0.1; exponentially growing, -G 6; instantaneously shrinking, -eN 0.1 100; exponential decline, -G -6 -eN 0.1 100; instantaneous expansion followed by bottleneck, -eN 0.6 0.1 -eN 1.2 0.5; severe bottleneck followed by exponential growth, -G 6 -eN 0.09 0.9 -eN 0.1 100. We note that we have simulated bottlenecking scenarios at different levels, that is, different fractions in size changes and/or periods, and finally presented patterns of the most positively deviated Tajima's D ¹⁷ in Fig. 2e. The population structure analysis showed partial population

isolation between Shenyang and other samples. Population subdivision may lead a positive Tajima's D . We calculated the distribution of Tajima's D by excluding Shenyang samples, and found that the resulting pattern is marginally more negative than the pattern based on all samples and quite distinct from any other pattern we simulated for comparison (Supplementary Fig. 15).

CRISPR-cas9-mediated genome editing. To establish a genetic manipulation system for functional studies, we implemented the CRISPR-cas9 system in the fall webworm and tested the availability on selected morphological genes, such as *yellow* and *fibroin* genes. We designed two single-guide RNA (sgRNA) target sites according to the feature of 5'-(20N)NGG-3' and located in exons for efficient knockout. These sites were also checked with uniqueness in the genome via BLASTN. Cas9 messenger RNAs were synthesized using the mMACHINE mMACHINE kit (Ambion). sgRNA templates were synthesized based on oligonucleotides that encode the T7 polymerase binding site, and subsequently annealed to common oligonucleotides that encode the remainder of the sgRNA sequence. The reaction conditions were as described previously⁷⁰. sgRNAs were synthesized in vitro using the MEGAscript Kit (Ambion).

Fall webworm eggs were injected within 6 h after oviposition following a standard approach described for *B. mori*⁷⁰. The Cas9-coding mRNA and the sgRNA were mixed at final concentrations of approximately 300 ng μl^{-1} . The injected eggs were incubated at 25 °C for ~ 10 days until hatching. Genotyping assays of mutants were conducted by PCR amplifications and subsequent Sanger sequencing. The CRISPR-cas9 system in the fall webworm was shown to have relatively high efficiency. For example, we observed approximately 13.5% (54/400) and 16.5% (66/400) hatching ratios for *fib-L* and *fib-H*, respectively, as well as mutation rates of 13.6% (3/22) and 8.3% (3/36).

Anatomical and transcriptome analyses of silk glands. Fall webworms were reared as described above. Domesticated silkworms (P50) were reared under the condition of 25 °C and a photoperiod of 12:12 light/dark with fresh mulberry leaves. Larvae were dissected in PBS, under a Motic SMZ168-BLED microscope. We sampled the silk glands on the second day of every instar and the first day of the wandering stage. Note that the fall webworm experiences six instars in the larva stage, while the silkworm experiences five. Approximately 400, 80, 30, 30, 10, 5 and 5 individuals were dissected, respectively, for each examined stage. Three biological replicates of pooled samples were prepared (Supplementary Fig. 16). Images of silk glands were captured with either a fluorescence microscope (Nikon AZ100), for the silk glands at early stages, or a Nikon EOS 70D camera for the silk glands of large size.

The silk glands were collected into Trizol and kept frozen at -80 °C immediately for further use. RNA was extracted as described. Construction of the cDNA library and paired-end RNA-seq (Illumina) were carried out by GENEWIZ Suzhou, China. Quality control and mapping of reads were performed as described above. The reference genomic data of the silkworm were downloaded from SilkDB 2.0⁷¹. Orthologue pair between the webworm and the silkworm was determined by the reciprocal BLASTP ($E < 10^{-5}$). Principal component analysis based on expression profiles was performed by the built-in R function 'prcomp'.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All sequence data from the fall webworm genome project have been deposited in GenBank under the accession code PKRV00000000. Transcriptome data have been deposited in SRA under SRR6433091. Genomic sequencing data have been deposited in SRA under SRR6432897, SRR8103939 and SRR8103940.

Received: 19 February 2018; Accepted: 6 November 2018;
Published online: 10 December 2018

References

- Sakai, A. K. et al. The population biology of invasive species. *Annu. Rev. Ecol. Syst.* **32**, 305–332 (2001).
- Bock, D. G. et al. What we still don't know about invasion genetics. *Mol. Ecol.* **24**, 2277–2297 (2015).
- Sullivan, G. T. & Ozman-Sullivan, S. K. Tachinid (Diptera) parasitoids of *Hyphantria cunea* (Lepidoptera: Arctiidae) in its native North America and in Europe and Asia - A literature review. *Entomol. Fenn.* **23**, 181–192 (2012).
- Zhang, L. W. et al. Analysis of the antennal transcriptome and insights into olfactory genes in *Hyphantria cunea* (Drury). *PLoS ONE* **11**, e0164729 (2016).
- Zhu, G. et al. Chemical investigations of volatile kairomones produced by *Hyphantria cunea* (Drury), a host of the parasitoid *Chouioia cunea* Yang. *Bull. Entomol. Res.* **107**, 234–240 (2017).
- Luo, J. et al. Characterization and polymorphism analysis of phosphoglucose isomerase gene in the fall webworm (*Hyphantria cunea*). *Bull. Entomol. Res.* **102**, 477–488 (2012).

7. Cao, L. J., Wei, S. J., Hoffmann, A. A., Wen, J. B. & Chen, M. Rapid genetic structuring of populations of the invasive fall webworm in relation to spatial expansion and control campaigns. *Divers. Distrib.* **22**, 1276–1287 (2016).
8. Drapeau, M. D., Albert, S., Kucharski, R., Prusko, C. & Maleszka, R. Evolution of the Yellow/Major Royal Jelly protein family and the emergence of social behavior in honey bees. *Genome Res.* **16**, 1385–1394 (2006).
9. Xia, A. H. et al. Identification and analysis of YELLOW protein family genes in the silkworm, *Bombyx mori*. *BMC Genomics* **7**, 195 (2006).
10. Ting, C. T. et al. Gene duplication and speciation in *Drosophila*: evidence from the Odysseus locus. *Proc. Natl Acad. Sci. USA* **101**, 12232–12235 (2004).
11. Wang, S. et al. Comparative genomics reveals adaptive evolution of Asian tapeworm in switching to a new intermediate host. *Nat. Commun.* **7**, 12845 (2016).
12. Tanaka, K. et al. Highly selective tuning of a silkworm olfactory receptor to a key mulberry leaf volatile. *Curr. Biol.* **19**, 881–890 (2009).
13. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
14. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2017).
15. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
16. Nei, M., Maruyama, T. & Chakraborty, R. The bottleneck effect and genetic variability in populations. *Evolution* **29**, 1–10 (1975).
17. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
18. Zayed, A. & Whitfield, C. W. A genome-wide signature of positive selection in ancient and recent invasive expansions of the honey bee *Apis mellifera*. *Proc. Natl Acad. Sci. USA* **105**, 3421–3426 (2008).
19. Hiwatashi, T. et al. An explicit signature of balancing selection for color-vision variation in new world monkeys. *Mol. Biol. Evol.* **27**, 453–464 (2010).
20. Moncla, L. H. et al. Selective bottlenecks shape evolutionary pathways taken during mammalian adaptation of a 1918-like avian influenza virus. *Cell Host Microbe* **19**, 169–180 (2016).
21. Badisco, L., Van Wielendael, P. & Vanden Broeck, J. Eat to reproduce: a key role for the insulin signaling pathway in adult insects. *Front. Physiol.* **4**, 202 (2013).
22. Le Gall, M. & Behmer, S. T. Effects of protein and carbohydrate on an insect herbivore: the vista from a fitness landscape. *Integr. Comp. Biol.* **54**, 942–954 (2014).
23. Cheng, T. et al. Genomic adaptation to polyphagy and insecticides in a major East Asian noctuid pest. *Nat. Ecol. Evol.* **1**, 1747–1756 (2017).
24. Xu, W., Papanicolaou, A., Zhang, H. J. & Anderson, A. Expansion of a bitter taste receptor family in a polyphagous insect herbivore. *Sci. Rep.* **6**, 23666 (2016).
25. Xia, Q., Li, S. & Feng, Q. Advances in silkworm studies accelerated by the genome sequencing of *Bombyx mori*. *Annu. Rev. Entomol.* **59**, 513–536 (2014).
26. Inoue, S. et al. Silk fibroin of *Bombyx mori* is secreted, assembling a high molecular mass elementary unit consisting of H-chain, L-chain, and P25, with a 6:6:1 molar ratio. *J. Biol. Chem.* **275**, 40517–40528 (2000).
27. Sehna, F. & Akai, H. Insect silk glands: their types, developmental and function, and effects of environmental factors and morphogenetic hormones on them. *Int. J. Insect Morphol. Embryol.* **19**, 79–132 (1990).
28. Ma, L. et al. Ras1(CA) overexpression in the posterior silk gland improves silk yield. *Cell Res.* **21**, 934–943 (2011).
29. Xia, Q. et al. Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science* **326**, 433–436 (2009).
30. Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
31. Mandric, I. & Zelikovsky, A. ScaffoldMatch: scaffolding algorithm based on maximum weight matching. *Bioinformatics* **31**, 2632–2638 (2015).
32. Li, R. et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
33. Hunt, M. et al. REAPR: a universal tool for genome assembly evaluation. *Genome Biol.* **14**, R47 (2013).
34. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
35. Waterhouse, R. M. et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2017).
36. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
37. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
38. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
39. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
40. Liu, P., Sanalkumar, R., Bresnick, E. H., Keles, S. & Dewey, C. N. Integrative analysis with ChIP-seq advances the limits of transcript quantification from RNA-seq. *Genome Res.* **26**, 1124–1133 (2016).
41. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
42. Zhan, S., Merlin, C., Boore, J. L. & Reppert, S. M. The monarch butterfly genome yields insights into long-distance migration. *Cell* **147**, 1171–1185 (2011).
43. FlyBase Consortium FlyBase: a *Drosophila* database. *Nucleic Acids Res.* **26**, 85–88 (1998).
44. International Silkworm Genome Consortium The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochem. Mol. Biol.* **38**, 1036–1045 (2008).
45. You, M. et al. A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.* **45**, 220–225 (2013).
46. She, R., Chu, J. S., Wang, K., Pei, J. & Chen, N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res.* **19**, 143–149 (2009).
47. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
48. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
49. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
50. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
51. Elisk, C. G. et al. Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).
52. Cantarel, B. L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
53. Hunter, S. et al. InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).
54. Zhou, X. et al. Phylogenetic and transcriptomic analysis of chemosensory receptors in a pair of divergent ant species reveals sex-specific signatures of odor coding. *PLoS Genet.* **8**, e1002930 (2012).
55. Tamura, K., Stecher, G., Peterson, D., Filipowski, A. & Kumar, S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013).
56. Li, L., Stoeckert, C. J. Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
57. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
58. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577 (2007).
59. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
60. Xiang, H. et al. The evolutionary road from wild moth to domestic silkworm. *Nat. Ecol. Evol.* **2**, 1268–1279 (2018).
61. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
62. Vurture, G. W. et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
63. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 1–33 (2013).
64. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
65. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
66. Keightley, P. D. et al. Estimation of the spontaneous mutation rate in *Heliconius melpomene*. *Mol. Biol. Evol.* **32**, 239–243 (2015).
67. DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I. & Nielsen, R. SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics* **32**, 1895–1897 (2016).
68. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
69. Hudson, R. R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).
70. Wang, Y. et al. The CRISPR/Cas system mediates efficient genome engineering in *Bombyx mori*. *Cell Res.* **23**, 1414–1416 (2013).
71. Duan, J. et al. SilkDB v2.0: a platform for silkworm (*Bombyx mori*) genome biology. *Nucleic Acids Res.* **38**, D453–D456 (2010).
72. Nishikawa, H. et al. A genetic mechanism for female-limited Batesian mimicry in Papilio butterfly. *Nat. Genet.* **47**, 405–409 (2015).
73. Cong, Q., Borek, D., Otwinowski, Z. & Grishin, N. V. Skipper genome sheds light on unique phenotypic traits and phylogeny. *BMC Genomics* **16**, 639 (2015).

74. Ahola, V. et al. The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nat. Commun.* **5**, 4737 (2014).
75. Kanost, M. R. et al. Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth, *Manduca sexta*. *Insect Biochem. Mol. Biol.* **76**, 118–147 (2016).
76. Derks, M. F. et al. The genome of winter moth (*Operophtera brumata*) provides a genomic perspective on sexual dimorphism and phenology. *Genome Biol. Evol.* **7**, 2321–2332 (2015).

Acknowledgements

We thank S. Wang and Y. Zhu for kind suggestions in initiating the sequencing. We also acknowledge H. Li and X. Huang for comments on the population genetics analyses. This study was supported by the National Science Foundation of China (31672370, 31522053, 31420103918, 91631103 and 31501877), the National Key Research and Development Program of China (2015CB755703), the Chinese Academy of Sciences (XDB27040205, 173176001000162007 and QYZDB-SSW-SMC029), the Special Fund for Forest Scientific Research in the Public Welfare (201504302), the Fundamental Research Funds for the Central Non-profit Research Institution of CAF (CAFYBB2017QB003) and the Shanghai Yangfan Program (16YF1413700).

Author contributions

Y.H., S. Zhan and Z.Z. initiated the project. Y.H. coordinated the project. S. Zhan designed and directed the project. Z.Z., S. Zhang and Q.W. collected samples. Y.C., N.W., S. Zhang and H.L. prepared DNA and RNA. S. Zhan led analyses. N.W., S. Zhan and X.H. performed analyses. N.W., X. Liu and X. Li dissected the silk glands. N.W., X. Li and Q.L. performed the CRISPR–cas9 experiments. S. Zhan, Y.H., Z.Z., A.A.J. and X.J.Z.

interpreted the data. S. Zhan and N.W. wrote the manuscript. A.A.J., X.J.Z. and Y.H. improved the manuscript. N.W., S. Zhang and X. Li contributed equally to the work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41559-018-0746-5>.

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

Correspondence and requests for materials should be addressed to Z.Z. or Y.H. or S.Z.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s), under exclusive licence to Springer Nature Limited 2018

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated
- Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequence data from the fall webworm genome project have been deposited in GenBank under an accession code PKRV00000000. Transcriptome data has been deposited in SRA under SRR6433091. Genomic sequencing data of other lepidopteran species has been deposited in SRA under SRR6432897, SRR6432898, and SRR6432899.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For de novo sequencing, all libraries with short-inserts were derived from ONE single individual to limit the heterozygous level which may cause problems to assembling. For RNA sequencing, we pooled a varying quantity of samples, from 3 to ~300, to represent each developmental stage or tissue, respectively, depending on the actual size of tissues at that stage; e.g., some tissues such as the silk glands at the 1st instar larvae or fat bodies are quite small, thus hundreds of individuals were dissected to collect enough RNA; we sampled from at least three individuals, with independent preparation of RNA, for some large tissues such as the silk glands at the last instar larvae or the whole body of pupae. For pilot resequencing, we sampled 20 individuals at five distinct locations (2 males and 2 females for each location). For CRISPR/Cas9 experiments, we injected at least 400 eggs for genetic manipulation of each gene. The injected quantity was based on our preliminary experiences on the actual hatching ratio and mutation ratio of injection with fall webworms.
Data exclusions	No data was excluded.
Replication	For the CRISPR/Cas9 experiments, we observed mutagenesis-induced phenotypes for each gene from independently injected individuals and only reported phenotypes that were presented by multiple individuals for each gene. For profiling silkglands between webworms and silkworms, we sequenced three biological replicates of transcriptome samples for each time point.
Randomization	For phenotypic comparisons between the fall webworm and the silkworm, laboratory-maintained wildtype lines of the fall webworm and the silkworm were reared under a standard condition independently. Individuals of abnormal phenotypes or developmental sizes were eliminated from the colony. At the corresponding stage, individuals were randomly selected from the colony for dissection and imaging, without any further filtering steps.
Blinding	Not applicable in this study.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

Methods

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Not applicable in this study.
Wild animals	Related information has been listed in Supplementary Tables 9 and 15.
Field-collected samples	Not applicable in this study.