BMC Genomics

RESEARCH

Open Access

Gene novelty and gene family expansion in the early evolution of Lepidoptera

Asia E. Hoile[1], Peter W. H. Holland[1*] and Peter O. Mulhair[1*]

Abstract

Background Almost 10% of all known animal species belong to Lepidoptera: moths and butterflies. To understand how this incredible diversity evolved we assess the role of gene gain in driving early lepidopteran evolution. Here, we compared the complete genomes of 115 insect species, including 99 Lepidoptera, to search for novel genes coincident with the emergence of Lepidoptera.

Results We find 217 orthogroups or gene families which emerged on the branch leading to Lepidoptera; of these 177 likely arose by gene duplication followed by extensive sequence divergence, 2 are candidates for origin by horizontal gene transfer, and 38 have no known homology outside of Lepidoptera and possibly arose via de novo gene genesis. We focus on two new gene families that are conserved across all lepidopteran species and underwent extensive duplication, suggesting important roles in lepidopteran biology. One encodes a family of sugar and ion transporter molecules, potentially involved in the evolution of diverse feeding behaviours in early Lepidoptera. The second encodes a family of unusual propeller-shaped proteins that likely originated by horizontal gene transfer from *Spiroplasma* bacteria; we name these the Lepidoptera *propellin* genes.

Conclusion We provide the first insights into the role of genetic novelty in the early evolution of Lepidoptera. This gives new insight into the rate of gene gain during the evolution of the order as well as providing context on the likely mechanisms of origin. We describe examples of new genes which were retained and duplicated further in all lepidopteran species, suggesting their importance in Lepidoptera evolution.

Keywords Insect evolution, HGT, Gene duplication, Genome evolution

Background

Diversification and adaptation depend on genetic change but associating genomic drivers underpinning phenotypic change is challenging. Many studies have approached this problem by starting with phenotypic polymorphisms within a species or differences between closely related species and then using genomic and experimental approaches to identify underlying causative mutations. Several of these studies have uncovered sequence changes in non-coding DNA affecting the expression of conserved genes [1–4]. Other studies have identified coding sequence changes causing amino acid substitutions, or loss of function, as causative mutations that were subsequently fixed under selection [5–7]. It is clear, however, that changes in existing genes, whether they affect gene expression or protein sequence, cannot explain all adaptive evolution. Perhaps the best evidence lies in comparative genomics: when genome sequences are compared ample evidence is uncovered for the role of gene number variation, gene duplications, and gene novelty in driving evolution and adaptation [8–13].

Gene novelty is a multi-faceted concept [14–17]. We define novel genes as protein-coding loci that are

*Correspondence:
Peter W. H. Holland
peter.holland@biology.ox.ac.uk
Peter O. Mulhair
peter.mulhair@biology.ox.ac.uk
[1] Department of Biology, University of Oxford, Mansfield Road, Oxford OX1 3SZ, UK

lineage-specific (i.e. taxonomically restricted genes), without close homologues in other taxa [18]. This is a pragmatic definition rather than a mechanistic one since we cannot always determine the mechanism by which a novel gene arose. The mode of origin of taxonomically restricted genes might be gene duplication followed by extensive sequence divergence [19–24], fusion of distinct loci or a transposable element into a pre-existing locus [25–30], horizontal gene transfer [31–33] or de novo origin from non-coding DNA [17, 34–36]. Whatever the mode of origin, novel genes likely reflect novel biology as they will encode proteins with potentially distinct activity or function not present in the outgroup taxa. Examples in arthropods include horizontally acquired genes from bacteria underpinning adaptations to phytophagy [37] or male courtship behaviour in moths and butterflies [32], and divergent gene duplicates recruited for limb patterning in water striders [38].

Here we investigate the origin of novel genes in the early evolution of the insect order Lepidoptera. Lepidoptera are a holometabolous order of insects consisting of the moths and butterflies and comprise nearly 160,000 described species or 8–10% of known animal species on the planet [39]. The oldest members of the Lepidoptera crown group are estimated to have appeared in the Late Carboniferous (~300 mya) and were likely pollen feeders, with the evolution of a tube-like proboscis and nectar feeding occurring later in the Middle Triassic (~240 Ma). Today the Lepidoptera inhabit almost all terrestrial ecosystems, displaying a large variety of ecological adaptations relating to feeding, defence, and survival [39–41]. Larvae of the earliest lineages were likely endophagous, feeding internally in the tissue of nonvascular land plants, with adults possessing mandibulate chewing mouthparts (as seen in extant members of the family Micropterigidae) suitable for pollen feeding [42, 43]. A period of diversification early in the evolution of Lepidoptera coincided with the development of the tube-like proboscis, used by adults to feed on nectar, and the expansion of angiosperms. The remarkable diversity present in Lepidoptera today can be attributed to continued co-evolution with diverse angiosperm lineages, major transitions in morphology and habitat, and the emergence of diverse feeding behaviours [41].

To assess whether novel genes arose in the early evolution of Lepidoptera, and whether any of these underwent further gene family expansion, we require complete genome sequences from a dense sampling of Lepidoptera and related insect orders. Previous studies have constructed deep-level phylogenies of Lepidoptera using a large density of species but relatively few loci [39], while other studies have studied specific gene families in depth [44–46]. Large genomic datasets have only recently become available through sequencing consortia such as the Darwin Tree of Life Project [47] affiliated to the Earth Biogenome Project [48]. Here, we avail of this data by analysing 115 high quality insect genomes and identify 217 novel genes that arose on the stem lineage of Lepidoptera and 541 novel genes that arose on the stem lineage of the Ditrysia, a major clade encompassing most of lepidopteran diversity [49]. We infer the likely modes of origin for these novel genes. We then focus attention on two gene families gained on the ancestral lepidopteran branch that were subsequently retained across all species, suggestive of recruitment to important roles in lepidopteran biology. One is a gene family encoding divergent sugar transporter proteins; the other is a likely horizontal gene transfer from bacteria.

Materials and methods

Gene family construction and discovery of novel genes

Proteome data from 99 species of Lepidoptera and 16 other arthropod species (Supplementary Table S1) were obtained from Ensembl Rapid Release (rapid.ensembl.org; accessed February 2023); taxon sampling was based on obtaining robust phylogenetic coverage across Lepidoptera while also preferentially selecting species with proteome predictions based on the Ensembl genebuild annotation pipeline (i.e. annotation which incorporated RNA sequence data). Primary transcripts were obtained from the predicted proteome data and Orthofinder v2.3.14 was run with default parameters to determine orthogroups within the dataset [50]. To relate these to a species tree, amino acid sequences from 25 single copy orthologues present in all species, as obtained from the Orthofinder output, were aligned using MAFFT v7.505 [51], trimmed using trimAl v1.4.rev15 build [52], and concatenated with PhyKIT [53]. This concatenated alignment was used to generate a species tree using IQ-TREE version 2.0-rc1 with 1000 bootstrap iterations, the given model LG+G4 and option -nt AUTO which automatically determines the best number of cores given the current data and computer capacity [54]. Orthogroups gained at nodes of interest (i.e. the branch leading to Lepidoptera and the branch leading to Ditrysia) were extracted using Orthoparser (github.com/PeterMulhair/ortho_parser). To test further whether orthogroups inferred by the analysis to be specific to Lepidoptera were actually present in outgroups but missing from predicted proteomes, Trichoptera genomes annotated by the alternative Augustus-Gaius pipeline (BRAKER) [55] were analysed. This was carried out using a BLASTp search of the orthogroups against the trichopteran BRAKER proteomes to find any potential missing homologues (using an e-value cutoff of 1e-5 and filtering hits above 25% sequence identity match along with query and subject

coverage of 60% to remove hits due to partial homology). Downstream of these steps, genes within orthogroups were analysed by exploring gene copy number, conducting synteny analyses, and generating expression matrices using publicly available RNAseq data. Figures including phylogenetic trees and heatmaps were generated in R using ggtree v3.6.2 [56], ggplot2 v3.4.4 [57], and Pheatmap v1.0.12. Protein models were predicted using AlphaFold (ColabFold v1.5.5: AlphaFold2) [58] and imported into Chimera v1.18 [59]. Molecular graphics and analyses of protein models were performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311. Chromosome plots showing gene positions were created using RIdeogram v0.2.2 [60].

Phylogenetic analysis of gene families

Phylogenetic trees of the two gene families of interest were built by aligning deduced protein sequences using MAFFT v7.505 followed by trimming using trimAl v1.4.rev15 build and tree building using maximum likelihood in IQ-TREE version 2.0-rc1. Trees were visualised using ggtree v3.6.2 in Rstudio. In the sugar transporter orthogroup analyses, PfamScan (command line tool pfam_scan.pl) was used to search each orthogroup against the Pfam-A.hmm database with cutoff –cut_ga and an e-value threshold of 1e-3 [61] to annotate functional domains in each gene. This was used to detect additional gene families labelled as belonging to sugar transporters (possessing Pfam domain Sugar_tr; PF00083), followed by phylogenetic analysis including *Drosophila* and other arthropod SLC sequences to infer the class of SLC each orthogroup belonged to [62]. In the propeller protein analyses, putative HGT was investigated using a BLASTp search (e-value threshold of 1e-3) [63] against the BLAST nr database with all lepidopteran sequences removed (Supplementary Table S3). The source of the HGT was then inferred by building a gene tree from the BLAST hits. Additional orthogroups in our datasets possessing the *propellin* gene were discovered by running a BLASTp search of the initial orthogroup (OG0000175) against all orthogroups in our dataset, retaining only those with percent identity equal to or above 25% and query and subject equal to or above 60%. This uncovered 8 additional homologous orthogroups, each of which contained only lepidopteran species. To further test the likely mode of origin of each of the 9 orthogroups, we carried out sequence similarity searches against the non-redundant protein sequence database (nr) and the core nucleotide database (core_nt) using a set of 10 representative species from each of the orthogroups (Supplementary Table S4). In one of the orthogroups (OG0008135), two of the species had hits against genes/proteins belonging to other insects. To test whether these BLAST hits represented true homologs, or the result of spurious homology, we aligned both insect and *Spiroplasma* proteins to a *Manduca sexta* propellin protein. This was carried out using the RCSB pairwise structure alignment tool [64].

Gene expression quantification

RNA-seq data for *Bombyx mori* were obtained from NCBI datasets PRJDB8614 and PRJNA675719 [65, 66], for *Danaus plexippus* from PRJNA663267 [67], and for *Papilio machaon* from PRJNA270386 [68]. RNA reads were trimmed using Trimmomatic v0.39 [69], and mapped to the reference genome using STAR 2.7.10b [70]. Stringtie v2.2.1 was used to quantify expression in each of the species datasets [71] and expression matrices were generated in RStudio using Pheatmap. Where multiple samples were available for a given tissue of lifestage, these were averaged to give one value.

Gene synteny analysis

Synteny analyses were used to test orthology of genes within and beyond Lepidoptera. For genes of interest, the gene ID, chromosome number, and location were determined from the genome annotation and gene track browser on Ensembl Rapid Release [72]. Two conserved 'marker genes' either side of the gene of interest were chosen and BLASTp searches (using Ensembl default parameters) conducted against the genomes of four Lepidoptera (*Danaus plexippus*, *Papilio machaon*, *Tinea trinotell*a and *Micropterix aruncella*) and eight outgroups (*Limnephilus lunatus*, *Limnephilus marmoratus*, *Limnephilus rhombicus*, *Glyphotaelius pellucidus*, *Bibio marci*, *Drosophila melanogaster*, *Adalia bipunctata* and *Vespula vulgaris*). These data were used to compare chromosomal organisation and gene neighbourhoods surrounding the genes of interest, revealing if individual genes within lepidopteran orthology groups were 1:1 homologues between species and also whether highly divergent orthologues were present in outgroups.

Results

Novel genes emerging at the base of Lepidoptera

To build a framework for comparative analyses, a phylogenetic tree was built from 25 single copy genes from 115 species, comprising 99 Lepidoptera species representing 24 families, and 16 outgroup taxa (Fig. 1, Supplementary Table S1). The tree is broadly consistent with previously hypothesized evolutionary relationships, including placing the Micropterigidae family (*Micropterix aruncella* and *Neomicropterix facetella* in our dataset) sister to the rest of the lepidopteran lineages, the presence of

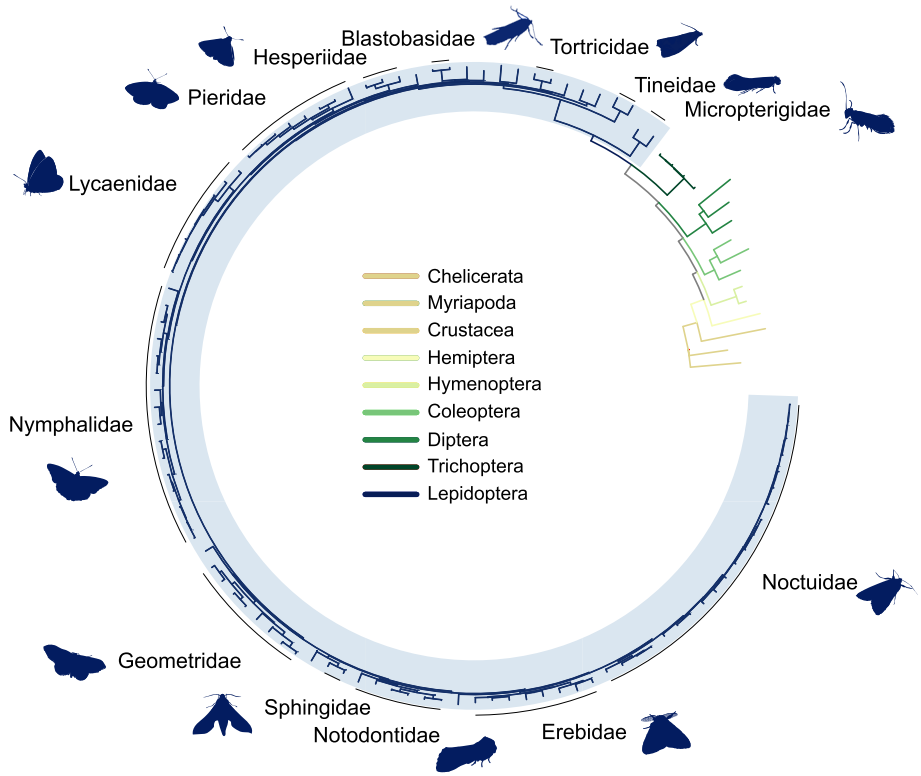

Fig. 1 Molecular phylogenetic tree of the 99 lepidopteran species from 24 families and 16 outgroup species inferred from 25 single-copy orthologues. Branches are coloured by insect order; species belonging to the named lepidopteran families are labelled with black lines on the outside of the tree

the large, established groups of Ditrysia, Apoditrysia, and Macroheterocera [39], and recovering monophyletic groups for all taxonomic families in the dataset [39, 49] (Fig. 1).

To identify novel genes or novel gene families that emerged early in lepidopteran evolution, we first constructed homologous gene groups ('orthogroups') using OrthoFinder [50]. Novel gene families here are defined as orthogroups present in a clade but missing from all outgroup taxa i.e. taxonomically restricted genes. We filtered the complete set of orthogroups to only retain those present in greater than two species. To place each of these orthogroups onto the species tree, we took the parsimonious assumption that the common ancestor of all species present in each orthogroup represented the node of origin (Fig. 2A). We identified 217 putative novel gene families originating on the branch leading to Lepidoptera (Fig. 2A).

To assess the mode of origin for each gene family we applied Pfam annotations to search for protein domains (indicative of duplication and divergence from pre-existing genes) as well as carrying out sequence similarity searches against metazoan (excluding Lepidoptera; further suggestive of duplication) and non-metazoan sequences (suggestive of HGT) from the nr protein database. We deduce that the majority of novel gene families (177 orthogroups) which originated along the lepidopteran branch likely arose via duplication followed by extensive sequence divergence (Fig. 2B). Putative HGTs accounted for only two orthogroups, as indicated by presence in Lepidoptera and non-metazoan proteomes but absent from animals other than Lepidoptera. We suggest that 38 orthogroups are potential orphan genes, candidates for origin by de novo gene genesis, although further analysis and additional data would be needed to test this hypothesis. We also detected 541 putative novel orthogroups on the branch leading to Ditrysia (representing all species outside of Micropterigidae in our dataset) (Fig. 2B). Of these, 398 likely arose from duplication, 13 via HGT, and 130 genes potentially originated de novo.

We hypothesized that novel genes of particular importance to lepidopteran biology would be present in most species of Lepidoptera analysed, with little or no gene loss after gene emergence. Furthermore, some genes of functional importance may have undergone duplication and divergence since their emergence [73]. We therefore calculated gene copy number for every orthogroup

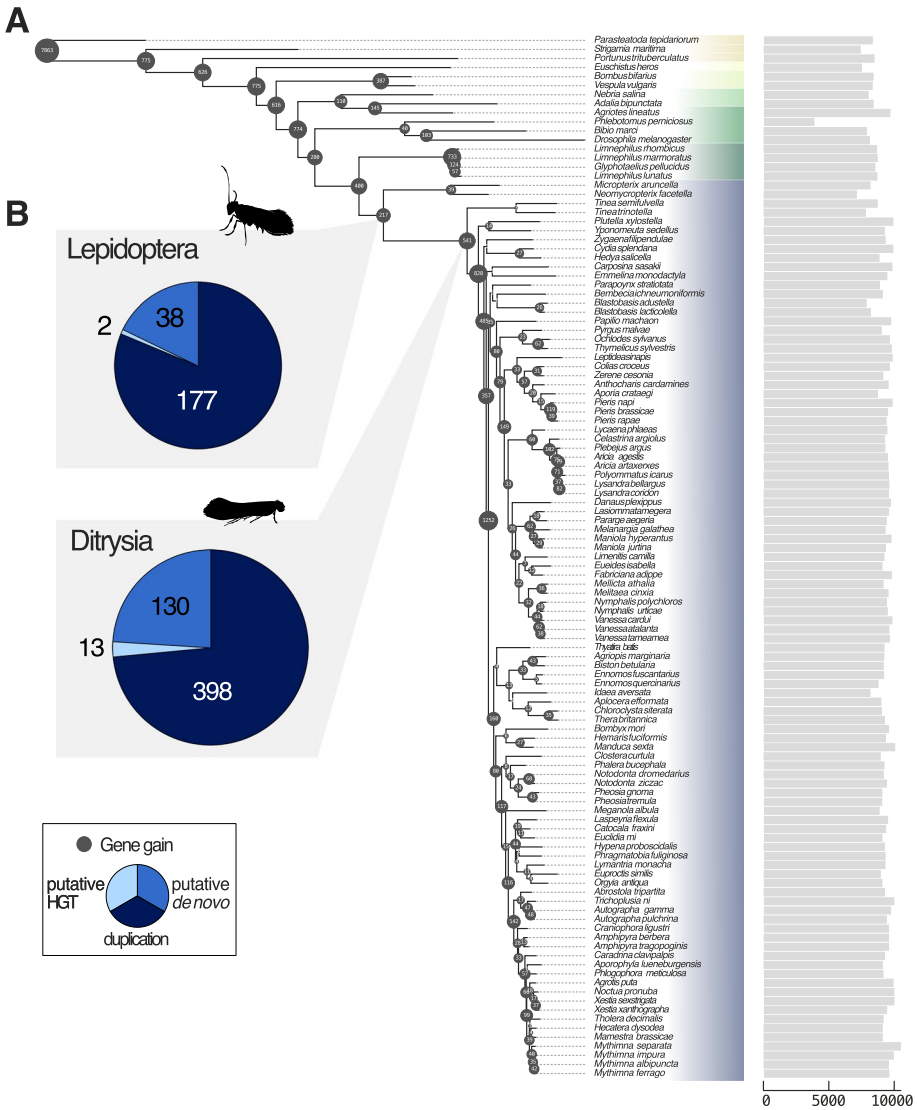

Fig. 2 **A** Species tree showing numbers of orthogroups gained at each phylogenetic node. Insect orders are separated by colours. Bar chart to the right of the tree displays the total number of orthogroups identified in each species. **B** Pie charts show the number of orthogroups originating at the Lepidoptera and Ditrysia nodes and proportions of the putative modes of new gene origin

originating at the Lepidoptera node and plotted these as a heatmap against a phylogenetic tree (Fig. 3A). Approximately half of the 217 orthogroups showed a scattered phylogenetic distribution (present in a low number of species within Lepidoptera); these may represent genes that are frequently lost, or which underwent extensive sequence divergence within Lepidoptera complicating orthology assignment (right-hand columns in Fig. 3A). 87 orthogroups are present in 75% or more of the lepidopteran species in this dataset, with sporadic gene loss and occasional gene duplication on some internal branches (left-hand columns in Fig. 3A). To identify orthogroups with higher rates of duplication patterns, we first determined that the data does not follow a normal distribution (positive, non-symmetric, right skew) and is non-parametric (Anderson–Darling test, $p < 0.05$), and that at least one orthogroup has a gene copy distribution across species which differs from the mean number of gene copies per orthogroup per species (Kruskal–Wallis rank test, $p < 0.05$; mean number of gene copies = 1.1645). We found that two orthogroups deviate significantly from the mean number of gene copies within a given orthogroup gained at the lepidopteran node: OG0000164 and OG0000175 (Dunn test, $p < 0.0001$; Fig. 3B). These two orthogroups have the highest variation in copy number, implying they have undergone extensive gene

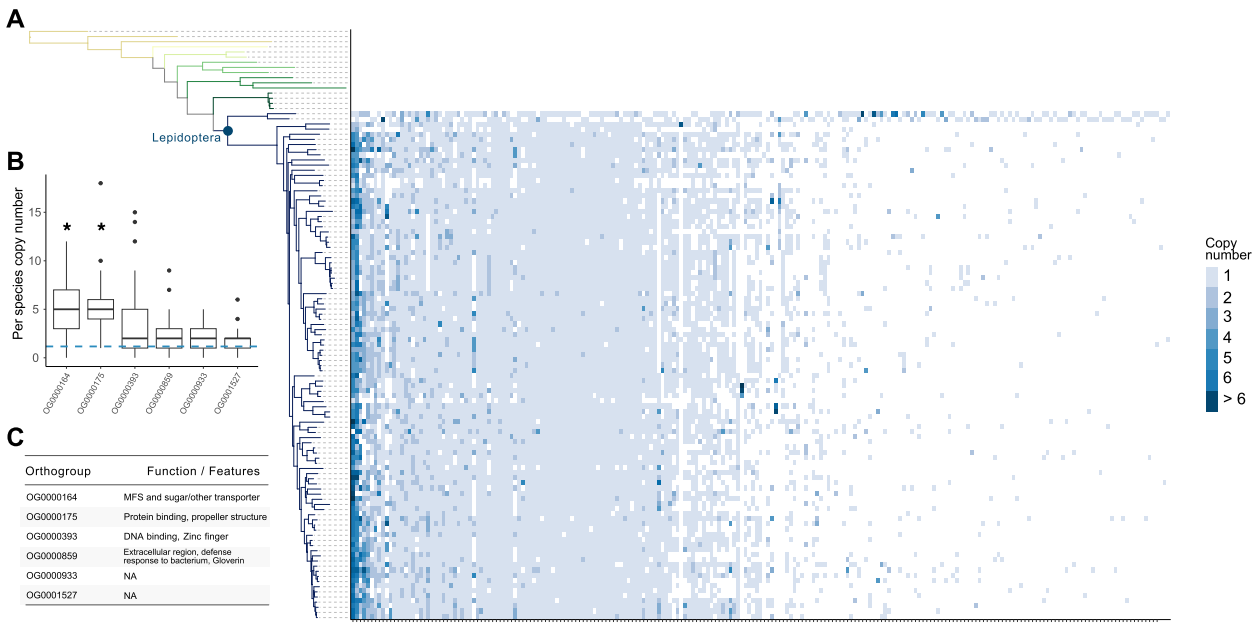

Fig. 3 Copy number of genes gained on the ancestral node of Lepidoptera. **A** Heatmap (right) showing gene copy number for each orthogroup originating at the Lepidoptera node mapped to the species tree (left). Lepidoptera node is labelled with a blue circle. Orthogroups on the right-hand side of the figure have genes present in few species and may include spurious homologies. **B** Boxplots showing copy number variation per species in the top 6 orthogroups present in all or most lepidopteran species. Blue broken line signifies the mean copy number per species for all orthogroups. Orthogroups OG0000164 and OG0000175 have a mean copy number significantly different from the mean copy number of lepidopteran orthogroups, as signified by an asterisk ($p < 0.05$). **C** Table showing functions and features from six orthogroups deviating above the average copy number per orthogroup

duplication within Lepidoptera, and they are also present in every lepidopteran species analysed. Sequence homology from BLASTp searches and domain annotation from Pfam revealed that these proteins have a putative sugar transporter domain (OG0000164; MFS and Sugar/other transporter, PF00083.27, GO:0016020|GO :0022857|GO:0055085) and a 6-bladed beta propeller 3D structure (OG0000175; GO:0005515) (Fig. 3C). To determine whether there were any functions enriched in the full set of 217 orthogroups gained on the lepidopteran node, we analysed the functional domains of each to determine whether there were any categories which were significantly overrepresented. Although no functional categories were found to be enriched within this dataset, approximately 9% of the orthogroups (19 out of 217) were found to contain a zinc finger domain (Supplementary Table S2). We also discover that the Gloverin gene family (OG0000859) emerged on the branch leading to Lepidoptera (Fig. 3C). The *gloverin* gene has previously been described as a lepidopteran novelty, and we confirm its emergence coincident with the evolution of Lepidoptera, where it has been retained in 86 of the 99 lepidopteran species in our dataset including *Micropterix aruncella* (Fig. 3A). Gloverin, first purified from *Hyalophora gloveri* [74], is a glycine rich protein with no detectable homology outside of Lepidoptera. It functions as an antimicrobial peptide against a range of bacteria, with greater specificity to Gram-negative bacteria, and appears to be commonly and widely expressed across a range of life stages and tissues, with significant increases in expression observed following exposure to bacteria [75, 76].

Gene expansion of lepidopteran sugar and solute transporters

The orthogroup originating on the node leading to the Lepidoptera with the highest mean copy number is a sugar transporter gene family (OG0000164) (Fig. 3). Across the species analysed, the copy number for this lepidopteran-specific orthogroup ranged from one gene (*Micropterix aruncella*) to twelve genes (*Manduca sexta*). As the sugar transporter protein superfamily is large and diverse in animals [62], and to understand the significance of this Lepidoptera-specific orthogroup, we extended our analysis to include all orthogroups containing a sugar transporter domain. We found 99 orthogroups with genes possessing a sugar transporter domain present across all species in our dataset (Fig. 4A), nine of which are annotated as emerging on the lepidopteran or ditrysian node; gene copy number for all nine orthogoups

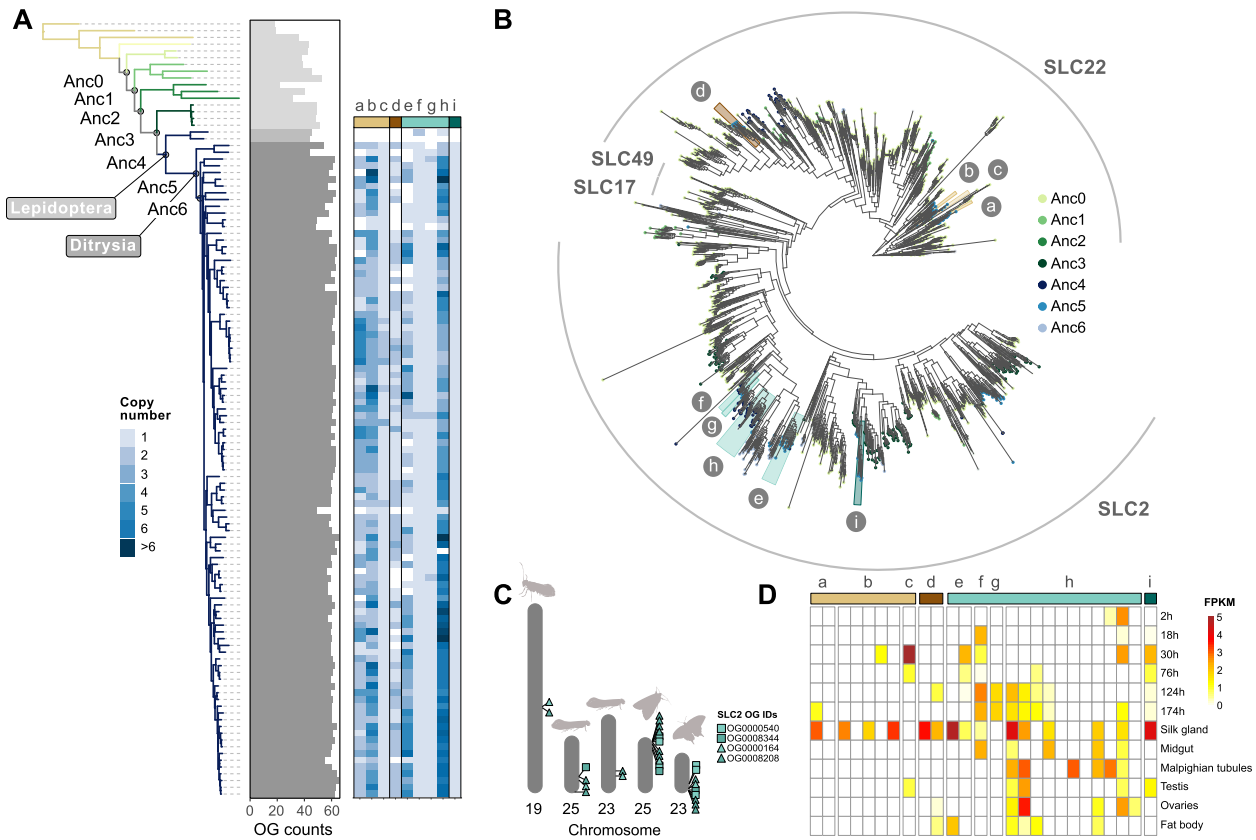

Fig. 4 Origins and evolution of lepidopteran and ditrysian-specific sugar transporter genes. Orthogroups are identified as follows: a—OG0000700, b—OG0000319, c—OG0007512, d—OG0001801, e—OG0000540, f—OG0008208, g—OG0008344, h—OG0000164, i—OG000840 (**A**) Species tree on the left is coloured by taxonomic group, with the Lepidoptera and Ditrysia nodes labelled. The numbered ancestral nodes correlate to the node of origin for the orthogroups shown in the gene tree (part B). The grey bar chart (middle) shows that Lepidoptera (darker grey bars) have a higher total number of sugar transporter orthogroups compared to outgroup species. Copy number of lepidopteran and ditrysian-specific orthogroups varies greatly (heatmap, right): SLC22 transporters are below the light and dark brown bars (labelled a-d); SLC2 transporters are below the light and dark blue/green bars (labelled with letters e-i). **B** Phylogenetic tree built using a representative sample of outgroup sugar transporters, combined with sugar transporters identified in the Lepidoptera. SLC2 transporters are highlighted in light and dark blue/green along with letters e-i, while SLC22 transporters are highlighted in light and dark brown with letters a-d. Tip colours represent the node of origin (as shown in the species tree in part A) for each orthogroup. **C** The four closely related SLC2 transporters were mapped to a selection of lepidopteran chromosomes (left to right: Micropterix aruncella, Tinea trinotella, Tinea semifulvella, Papilio machaon and Autographa gamma). Sugar transporters of lepidopteran origin are represented by a triangle, while those of ditrysian origin are represented by a square. All four transporter orthogroups group in close physical proximity, on the same chromosome. **D** Heatmap of expression patterns of nine lepidopteran/ditrysian originating sugar transporters in Bombyx mori tissues

in each species shows varying rates of copy number expansion and gene loss (Fig. 4A). Four of these orthogroups were single copy in all or most species, while five have undergone extensive gene duplication since their lepidopteran origins (Fig. 4A and B).

Next, we wanted to determine whether these nine lepidopteran sugar transporter-containing orthogroups had a single evolutionary origin or whether they had evolved independently from separate ancestral sugar transporter genes. To resolve this, a phylogenetic tree of transporter proteins from representative lepidopteran and outgroup species was constructed using all sugar transporter orthogroups (Fig. 4B). The tree topology suggests multiple origins of the lepidopteran- and ditrysian-specific sugar transporter genes, although not each of the nine orthogroups had independent origins. Notably, two lepidopteran-originating orthogroups (OG0000164 and OG0008208), and two ditrysian-originating orthogroups (OG0008344 and OG0000540) group close to each other in the phylogenetic tree (e–h). Another orthogroup (i; OG0008400) is located outside this clade, however each of these orthogroups are present in a larger clade consisting of members of the solute carrier 2 (SLC2) family (Fig. 4B). In some taxa SLC2 genes have been shown to

encode proteins that facilitate transport of small sugars across cell membranes [62].

The remaining four sugar transporter orthogroups (a-d) are all ditrysian-specific, three of which form a single monophyletic group. The fourth orthogroup is located in a more phylogenetically distinct group, however, all orthogroups belong to the SLC22 protein subfamily (Fig. 4B). The SLC22 proteins are membrane transporters known to regulate metabolic functions, transporting a broader range of small molecules than SLC2 [62]. In all instances, the most closely related orthogroups in the gene tree contain both outgroups and lepidopteran species (Fig. 4B). This implies that the lepidopteran- and ditrysian-specific transporter orthogroups originated from more ancient gene families that were present across all or most insects including Lepidoptera. These ancestral genes duplicated and underwent extensive amino acid substitutions specifically in the lineage leading to Lepidoptera or Ditrysia.

The four SLC2-like orthogroups [e–h] which group closely together in the gene tree (Fig. 4B) are also co-located in the genome, found consistently in close association with one another across diverse lepidopteran species (Fig. 4C). This suggests that these sugar transporter genes originated from a single ancestral duplication event at the base of the Lepidoptera and subsequently underwent tandem duplication in the ancestral lepidopteran and again in the branch leading to Ditrysia. In contrast, for the SLC22-like orthogroups gained on the ditrysian branch (a-d), we do not see the same close linkage and instead they are scattered on separate chromosomes (Supplementary Figure S1). If these originated from a single ancestral gene, as suggested by branching patterns in the gene tree, they dispersed around the genome after duplication.

To investigate possible functions of these lepidopteran-species transporter gene families, we assessed their patterns of expression across multiple time points and tissues from *Bombyx mori* [65, 66]. All genes from the nine lepidopteran and ditrysian-specific gene families are expressed in at least one tissue or at one developmental time point (Fig. 4D). The silk gland was the most frequent site of expression across the nine orthogroups, but some of the genes have wide and distributed expression (Fig. 4D).

Lepidoptera propellin genes arose through horizontal gene transfer

The second gene family which emerged at the base of Lepidoptera and is maintained in significantly high copy number across all butterfly and moth species analysed is orthogroup OG0000175 (Fig. 3). The genes in this family were previously undescribed in insects. Below we show they encode proteins with a beta-propeller structure; we therefore name this the *propellin* gene family. Intriguingly, this group of genes is conserved across Lepidoptera yet does not have detectable sequence identity to any orthogroups in the arthropod outgroups included in our initial analysis. Furthermore, iterative BLAST searching revealed that OG000175 is not the only set of *propellin* genes in Lepidoptera; the genes are split into eight distinct orthogroups, including the original group OG000175 with the highest copy number. All eight *propellin* orthogroups are specific to Lepidoptera (Supplementary Figure S2). Combining all *propellin* orthogroups together, we find Lepidoptera genomes have an average of 11 gene copies, ranging from 3 copies in *Neofacetella micropterix* (Micropterigidae) to 25 copies in *Phragmatobia fuliginosa* (Erebidae) (Supplementary Table S5).

To assess the likely origin of the *propellin* gene in Lepidoptera, we carried out a BLASTp search using all proteins in orthogroup OG0000175 against the NCBI nr protein database excluding Lepidoptera sequences. This revealed significant sequence similarity matches to proteins from bacterial species. The most frequent bacterial genus in the set of matches was *Spiroplasma*, with additional matches in *Macrococcus* and *Escherichia* (Supplementary Table S3). Using iterative rounds of BLAST searching, we found very few matches outside bacteria; we identified a potentially related unnamed gene in the genome of the plant *Picea sitchensis* (spruce; ABK22491.1) and a fungus gnat *Bradysia coprophila* (30% identity to a *Spiroplasma* homologue of Lepidoptera *propellin* genes over 16% query cover). To confirm the likely bacterial origins of the different *propellin* orthogroups, we also carried out BLASTp and tBLASTn searches against the nr and core nt databases, respectively, for each of the *propellin* orthogroups. We searched protein sequences from 10 representative species in each of the 8 orthogroups using both methods and found that bacterial, specifically *Spiroplasma*, sequences represented the majority of sequence similarity hits (Supplementary Table S4). While two genes from one orthogroup showed sequence similarity to other insect proteins (E3 ubiquitin ligases), we deduce that these hits are likely a result of spurious homology, with low query coverage (34–40%) and sequence similarity likely resulting from convergent amino acid residues in repetitive regions. In addition to this, all other hits from other species in the same orthogroup showed sequence similarity to *Spiroplasma* and other bacterial proteins. These *Spiroplasma* proteins were deduced to have similar tertiary structures to *propellin* (i.e. 6-bladed beta propeller; RMSD value of 3.73); in contrast, the spurious insect protein hits possessed multiple alpha helices and no structural similarity (RMSD value of 5.57).

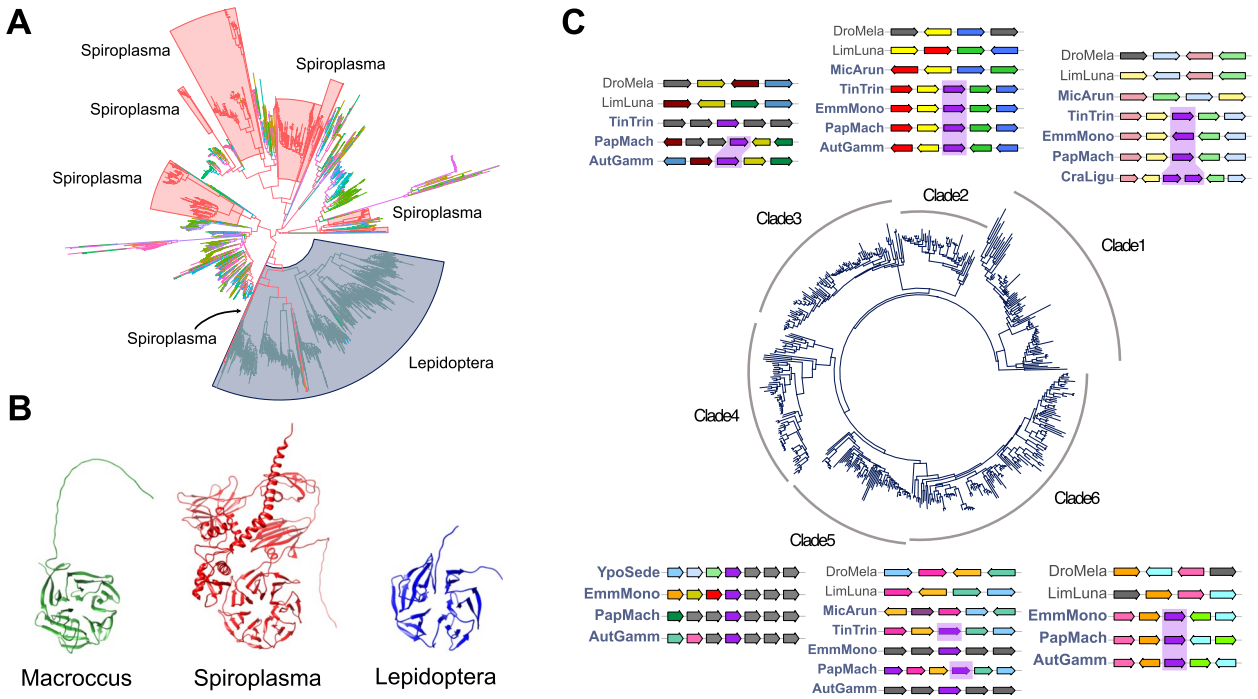

Fig. 5 Lepidoptera-specific genes encoding proteins with sequence identity and structural similarity to bacterial 6-bladed propeller proteins. **A** Gene tree of propellin and putative bacterial homologs. The Lepidoptera clade (blue) and *Spiroplasma* clades (red) are labelled with coloured boxes and text. All other branches represent a range of bacterial species which are shown in Supplementary Figure S3. Molecular phylogenetic analysis indicates the propellin genes of Lepidoptera are monophyletic, whose most closely branching lineages are *Spiroplasma* genes, and sister group to a clade dominated by *Spiroplasma* genes (highlighted in red). **B** AlphaFold predictions suggest lepidopteran propellin proteins form 6-bladed propeller structures similar to bacterial homologues; examples shown from *Macrococcus* (green), *Spiroplasma* (red) and the lepidopteran *M. sexta* (blue). Additional protein structure predictions in Supplementary Figure S4. **C** Molecular phylogenetic analysis indicates that the largest orthogroup of lepidopteran propellin genes divides into 6 clades, each gene (purple) located at a different chromosomal location, most of which show conserved synteny between lepidopteran species (synteny indicated by shaded purple regions). The Micropterigidae species *M. aruncella* only has a gene in clades 1. Marker genes are shown by various colours

Next, we constructed a phylogenetic tree of all *propellin* copies and their putatively homologous sequences. This shows that all lepidopteran *propellin* sequences are closely related in the gene tree (Fig. 5A, Supplementary Figure S3). The most closely related branches to the Lepidoptera clade are *Spiroplasma* sequences which, along with a larger sister clade dominated by *Spiroplasma* sequences, suggests there has been a putative horizontal gene transfer from bacteria to Lepidoptera (Fig. 5A). Based on the gene tree topology, we cannot exclude the possibility of multiple horizontal transfer events into Lepidoptera. Although there are clear sequence similarity matches between Lepidoptera and *Spiroplasma*, the level of primary sequence identity is low. The highest percentage identity found between a lepidopteran protein and a *Spiroplasma* protein had only 35% identity over a sequence alignment of 134aa. This represents 45% coverage of a lepidopteran *propellin* protein (ENSAGMG00005008917.1) and 18% of a *Spiroplasma* protein (WP_164028422.1). To further assess legitimacy of the homologous relationships between these genes with low sequence identity, we predicted 3D structures of the deduced proteins from *Spiroplasma*, *Macrococcus*, and Lepidoptera (using six genes from *Manduca sexta* as representative of Lepidoptera) with AlphaFold [58] (Fig. 5B; Supplementary Figure S4). We find clear similarity in predicted protein structure with all lepidopteran and bacterial sequences having a 6-bladed propeller structure (Fig. 5B). Further support for homology between lepidopteran *propellin* proteins and bacterial proteins was found when we aligned representative protein structures, which showed an RMSD value of 3.83 and TM-score of 0.69 (Supplementary Figure S5). Each structured propeller region within a *propellin* protein is approximately 221aa long consisting of blades of 30aa in length. There is variability outside of the beta-propeller domain including regions of varying length and structure, most notably in the additional domains in the *Spiroplasma* protein model (Fig. 5B). To reflect this protein structure, we name the Lepidoptera genes the *propellin* gene family.

To further investigate the evolution of *propellin* genes, we focussed attention on the high-copy number *propellin* orthogroup (OG0000175; Supplementary Figure S2). Phylogenetic analysis divides this orthogroup into six clades within Lepidoptera, which we refer to as gene subfamilies (Fig. 5C). The early diverging lineage of Lepidoptera, represented by the family Micropterigidae, has a *propellin* gene in clade 1 in the gene tree (Fig. 5C). Next, we examined the local gene synteny for these six subfamilies across representative lepidopteran species. Genes from each subfamily, excluding clade four, were found in a microsyntenic cluster of genes ('marker genes') which are homologous across most or all species (Fig. 5C). This confirms that each of these 6 *propellin* subfamilies are one-to-one orthologues across Lepidoptera. Importantly, many of the marker genes also exist in microsyntenic blocks in the arthropod outgroups, consistent with these being the genomic sites where the Lepidoptera-specific *propellin* gene was integrated (Fig. 5C). Since the six *propellin* subfamilies are at distinct chromosomal locations, yet form a monophyletic group in molecular phylogenetic analysis, we propose that this orthogroup emerged through a single HGT event from *Spiroplasma* or another bacterial source to Lepidoptera, followed by duplication and transposition around the genome. These duplications generated not only the six subfamilies analysed in detail, but also likely the additional *propellin* genes referred to above. We note that genes in subfamily 1 are intronless (or have one intron), while the remaining subfamilies and additional *propellin* orthogroups have between 0 and 8 introns, with the median count being 1 intron. This could reflect transposition via an RNA intermediate or could be a legacy of the gene's bacterial origin (Supplementary Table S5).

For a first insight into the possible functional role of the lepidopteran *propellin* genes, we analysed the expression of all copies of *propellin* using transcriptomic data sets from three species: *Bombyx mori*, *Danaus plexippus*, and *Papilio machaon* (Supplementary Figure S6). While we find evidence for expression of all gene copies in each species, the patterns are complex and variable within and between species. In *Danaus plexippus* for example, while most *propellin* copies show some expression in larval or pupal stages (8 of the 11 genes), levels of expression are highest in the adult life stage, with particularly high expression found in the thorax, compared to the head or abdomen (Supplementary Figure S6). In *Papilio machaon* most copies are restricted to one or two life stages, while others are strongly expressed throughout the life cycle of the butterfly. Expression in *Bombyx mori* shows wider coverage across life stages and tissue types, with most gene copies expressed in early developmental and adult life stages. While expression is common across most adult tissue types in *Bombyx mori*, there is little, or no expression found in the midgut or silk glands (Supplementary Figure S6). While there is little correlation in expression between homologous copies of *propellin* across all three species, we note that transcriptomic datasets available are not comprehensive. However, such pervasive expression across life stages and tissues in multiple species provides support to the fact that these genes are functional across a wide range of lepidopteran species.

We noted above that there were some sequence similarity matches outside bacteria and Lepidoptera. The putatively homologous gene from Diptera is an uncharacterised locus (LOC119081672, encoding putative protein XP_037046651) on an unplaced scaffold in the genome assembly of a fungus gnat *Bradysia coprophila* [77]. We find this gene is present in two species of *Bradysia.* It is unlikely that the fungus gnat scaffold is a contaminating sequence since it is present in two species, and because it is adjacent in the genomes to recognisable insect genes (Supplementary Figure S6). Analysis of the unplaced scaffold reveals clearly dipteran genes immediately 3' (LOC119081668) and relatively close 5' (LOC119081673 and LOC119081675) to the gene of interest. Intriguingly, a locus immediately 5' (LOC119081585) has high similarity to springtail (*Collembola*) tyrosine kinases, and the next neighbouring gene (LOC119081673) is *Bradysia*-specific (Supplementary Figure S6). We therefore suggest the Diptera gene LOC119081672 arose by an independent HGT from *Spiroplasma* in the *Bradysia* fungus gnat genus, which has likely also acquired other genes by HGT. We have not deduced the origin of the loci with a sequence match in *Picea sitchensis* (spruce).

Discussion

In this study we identified 217 'novel' genes arising on the evolutionary lineage leading Lepidoptera, after it had diverged from outgroups including the closest related order Trichoptera (caddisflies). We caution, however, against this as a quantitative measure of genomic novelty. First, we are using a pragmatic definition of novelty that includes de novo genes, horizontally transferred genes, and gene duplication followed by sequence divergence; altering parameters relating to sequence divergence could increase or decrease the gene count [78]. To improve inference of new genes in early lepidopteran evolution, we employed a phylogenetically informed approach to construct gene families, minimising the effects of bias resulting from rapid sequence divergence [50]. Second, novelty at the Lepidoptera node could be 'undercounted' if some genome annotations are incomplete, particularly those of early diverging lepidopteran taxa. Third, there are factors that could spuriously 'overcount' novelty. For example, in our study around half the

novel orthogroups were found sporadically in a small number of distantly related Lepidoptera species. This could indicate repeated gene loss following the origin of the novel gene but could also include 'noise' as a result of some proteins being grouped incorrectly due to spurious sequence identity. Secondary loss of genes from caddisfly genomes could theoretically cause overcounting of genes on the Lepidoptera node, but we have minimized this risk through use of four caddisfly genomes. We also noted a small degree of overcounting (< 2%) emerging from alternative genome annotation methods [79], but we accounted for this (see Methods). Specifically, the initial input data consisted of proteomes predicted from the Ensembl Genebuild annotation which incorporates RNA sequence data and filters poorly supported potential coding transcript proteins. A second method of genome annotation, the BRAKER method, is potentially less stringent and found some genes that had been missed by Genebuild. The difference amounted to just two orthogroups. The same caveats apply to counts of novel genes at other similarly deep phylogenetic nodes. Despite this caveat, we find it interesting that even more apparent gene novelty (541 gene families) dates to the node leading to Ditrysia. These genes require further analysis, but the observation suggests that the evolution of new biological traits continued during the early evolutionary radiation of moths. More important than an absolute number of novel genes, the analysis gives us a first look into the relative importance of different modes of gene origin during the emergence of Lepidoptera. We find the majority of novel gene families gained on the ancestral lepidopteran branch arose via gene duplication and divergence (~ 82%) while around ~ 18% genes had no sequence matches or any recognisable domains. These are putative candidates for genes arising de novo from non-coding genes. Just two genes (< 1%) are candidates for having arisen via HGT (including the *propellin* gene), with hits to bacterial or fungal species.

One of the genes that likely arose via HGT was highlighted in our analysis as a novel gene that underwent extensive gene duplication in Lepidoptera to generate a large gene family. This gene family, which we name the *propellin* genes, is potentially functional as evidenced by the extensive retention through evolution and conserved domain structure. Currently, however, its precise role in lepidopteran biology is unclear. Phylogenetic analyses suggest that the progenitor of the *propellin* gene family was transferred to an insect from *Spiroplasma* bacteria, some time on the Lepidoptera stem lineage. *Spiroplasma* is a well-known intracellular symbiont in arthropods. Furthermore, *Spiroplasma* is known to colonise reproductive tissues, which in turn impacts upon the host's reproduction, and indeed this genus is one of two bacterial symbionts in Lepidoptera for which maternal transmission has been demonstrated [80]. In some cases, transmission is enhanced by manipulation of host physiology, such as male-killing which increases the number of female offspring as observed in *Danaus chrysippus* [81]. Clearly, persistent association with reproductive tissues gives opportunity for horizontal gene transfer, as the symbiont DNA is in close physical proximity to the DNA of the host germline. This has been seen in the relationship between a mealybug and two endosymbiont species *Tremblaya* and *Moranella* [82]. Interestingly, we also found a putatively homologous gene in two species of Diptera (genus *Bradysia*), possibly reflecting an independent HGT event. This is consistent with previous findings that some types of gene are more prone to HGT than others, perhaps those encoding proteins with few interaction partners [83]. The evolutionary retention of the likely HGT-derived *propellin* gene, plus its extensive gene duplication in Lepidoptera, suggest this gene family likely evolved to perform functions that are important for the biology of moths and butterflies. We do not know the biological role, or roles, of *propellin* genes in Lepidoptera, but note that their bacterial homologues have diverse functions including ligand-binding proteins, signalling proteins, lysases, structural proteins, isomerases and hydrolases [84]. It is worth noting that the *Spiroplasma* genes which group closest to the *propellin* genes in the gene tree are annotated as hypothetical proteins without known function, suggesting more work is needed to understand the functional context of this gene.

The only other novel lepidopteran gene to show such widespread retention and extensive gene duplication encodes a family of SLC2-like sugar transporter proteins. In other animals, members of the SLC2 sugar transporter superfamily encode glucose-uptake proteins, ribose transport proteins, and several putative membrane proteins probably involved in sugar transport [62, 85, 86]. The functional link to sugars is particularly intriguing since the ecological association between Lepidoptera and sugar-feeding changed markedly in early lepidopteran evolution. Specifically, adult moths in the basal family Micropterigidae primitively lack a proboscis and are pollen feeders, whereas adult moths and butterflies in the Ditrysia use a proboscis to access sugar-rich nectar in flowers. Our wider comparative survey of sugar transporter gene families picks up potentially interesting co-evolution between this ecological shift and the sugar transporter genes. We find that although OG0000164 (and one other sugar transporter gene family) are present in pollen-feeding Micropterigidae, it is not until the evolution of the nectar-feeding Ditrysia that we see extensive gene duplication, widespread gene retention and the emergence of additional SLC-like sugar transporter

gene families [62]. We suggest, therefore, that novel sugar transporter gene families emerged at the base of Lepidoptera, but it was only later in lepidopteran evolution that massive gene duplication and functional divergence of sugar-transporter genes took place, in association with nectar feeding. The causal link between these genetic changes and the evolution of novel feeding behaviour in the early evolution of Lepidoptera warrants further study.

Conclusion

We have demonstrated the emergence of 217 novel gene families (orthogroups) on the node leading to Lepidoptera and 541 novel gene families emerging on the node leading to the Ditrysia. Two orthogroups have significantly higher gene copy per species across Lepidoptera indicative of extensive gene duplication following their origins. One likely originated by horizontal gene transfer from the endosymbiont bacterium *Spiroplasma* and then duplicated to generate a diverse group of '*propellin*' genes encoding a 6-bladed propeller domain. The other encodes a large set of sugar transporter proteins and is part of a diverse set of sugar and solute transporter genes that duplicated and diverged extensively in early lepidopteran evolution.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11338-x.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements
We thank the taxonomic experts who collected the insects used in the study and all members of the DToL project at the Sanger Institute for sequencing and assembling the genomes. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Authors' contributions
P.W.H.H. and P.O.M. conceived the study and oversaw the research. A.E.H. and P.O.M. designed analyses and carried out the bioinformatic research presented. A.E.H., P.W.H.H. and P.O.M. interpreted all results. A.E.H. wrote the initial draft of the manuscript, and P.W.H.H. and P.O.M. edited versions. All authors read and approved the final manuscript.

Funding
AEH was supported by the Oxford Interdisciplinary DTP and funding from the Biotechnology and Biological Sciences Research Council (UKRI-BBSRC) [grant number BB/T008784/1]; POM and PWHH acknowledge funding from Wellcome Trust Darwin Tree of Life Awards (grant agreements 218328, 226458).

Data availability
Genome data associated with this study is listed in Supplementary Table S1 along with accession numbers. Data and code generated in this study can be found on figshare; figshare.com/s/32d1b9055257dad1892f.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 6 December 2024 Accepted: 10 February 2025
Published online: 19 February 2025

References

1. Carroll SB, Gates J, Keys DN, Paddock SW, Panganiban GE, Selegue JE, et al. Pattern formation and eyespot determination in butterfly wings. Science. 1994;265:109–14.
2. Wucherpfennig JI, Howes TR, Au JN, Au EH, Roberts Kingman GA, Brady SD, et al. Evolution of stickleback spines through independent cis-regulatory changes at HOXDB. Nat Ecol Evol. 2022;6:1537–52.
3. Tian S, Asano Y, Banerjee TD, Wee JLQ, Lamb A, Wang Y, et al. A micro-RNA is the effector gene of a classic evolutionary hotspot locus. bioRxivorg. 2024;:2024.02.09.579741.
4. Livraghi L, Hanly JJ, Evans E, Wright CJ, Loh LS, Mazo-Vargas A, et al. A long noncoding RNA at the cortex locus controls adaptive coloration in butterflies. Proc Natl Acad Sci U S A. 2024;121: e2403326121.
5. Hoekstra HE, Coyne JA. The locus of evolution: evo devo and the genetics of adaptation: The locus of evolution. Evolution. 2007;61:995–1016.
6. Ota KG, Kuraku S, Kuratani S. Hagfish embryology with reference to the evolution of the neural crest. Nature. 2007;446:672–5.
7. Dutrow EV, Serpell JA, Ostrander EA. Domestic dog lineages reveal genetic drivers of behavioral diversification. Cell. 2022;185:4737–55.e18.
8. Richter DJ, Fozouni P, Eisen MB, King N. Gene family innovation, conservation and loss on the animal stem lineage. eLife. 2018;7:e34226.
9. Paps J, Holland PWH. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. Nat Commun. 2018;9:1730.
10. Thomas GWC, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K, et al. Gene content evolution in the arthropods. Genome Biol. 2020;21:15.
11. Fernández R, Gabaldón T. Gene gain and loss across the metazoan tree of life. Nat Ecol Evol. 2020;4:524–33.
12. Guijarro-Clarke C, Holland PWH, Paps J. Widespread patterns of gene loss in the evolution of the animal kingdom. Nat Ecol Evol. 2020;4:519–23.
13. Cicconardi F, Milanetti E, Pinheiro de Castro EC, Mazo-Vargas A, Van Belleghem SM, Ruggieri AA, et al. Evolutionary dynamics of genome size and content during the adaptive radiation of Heliconiini butterflies. Nat Commun. 2023;14:5620.
14. Long M, Betrán E, Thornton K, Wang W. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 2003;4:865–75.
15. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010;20:1313–26.
16. Haggerty LS, Jachiet P-A, Hanage WP, Fitzpatrick DA, Lopez P, O'Connell MJ, et al. A pluralistic account of homology: adapting the models to the data. Mol Biol Evol. 2014;31:501–16.
17. Van Oss SB, Carvunis A-R. De novo gene birth. PLoS Genet. 2019;15: e1008160.
18. Rödelsperger C, Prabh N, Sommer RJ. New gene origin and deep taxon phylogenomics: Opportunities and challenges. Trends Genet. 2019;35:914–22.
19. Ohno S. Evolution by gene duplication. 1970th ed. Berlin, Germany: Springer; 2013.
20. Rastogi S, Liberles DA. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol Biol. 2005;5:28.
21. Conrad B, Antonarakis SE. Gene duplication: a drive for phenotypic diversity and cause of human disease. Annu Rev Genomics Hum Genet. 2007;8:17–35.

22. Sémon M, Wolfe KH. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis. Proc Natl Acad Sci U S A. 2008;105:8333–8.
23. Holland PWH, Marlétaz F, Maeso I, Dunwell TL, Paps J. New genes from old: asymmetric divergence of gene duplicates and the evolution of development. Philos Trans R Soc Lond B Biol Sci. 2017;372:20150480.
24. DuBose JG, de Roode JC. The link between gene duplication and divergent patterns of gene expression across a complex life cycle. Evol Lett. 2024;8:726–34.
25. McClintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 1950;36:344–55.
26. Long M, Langley CH. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science. 1993;260:91–5.
27. Leonard G, Richards TA. Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. Proc Natl Acad Sci U S A. 2012;109:21402–7.
28. Bornberg-Bauer E, Albà MM. Dynamics and adaptive benefits of modular protein evolution. Curr Opin Struct Biol. 2013;23:459–66.
29. Cosby RL, Judd J, Zhang R, Zhong A, Garry N, Pritham EJ, et al. Recurrent evolution of vertebrate transcription factors by transposase capture. Science. 2021;371:eabc6405.
30. Mulhair PO, Moran RJ, Pathmanathan JS, Sussfeld D, Creevey CJ, Siu-Ting K, et al. Bursts of novel composite gene families at major nodes in animal evolution. bioRxiv. 2023;:2023.07.10.548381.
31. Husnik F, McCutcheon JP. Functional horizontal gene transfer from bacteria to eukaryotes. Nat Rev Microbiol. 2018;16:67–79.
32. Li Y, Liu Z, Liu C, Shi Z, Pang L, Chen C, et al. HGT is widespread in insects and contributes to male courtship in lepidopterans. Cell. 2022;185:2975–87.e10.
33. Keeling PJ. Horizontal gene transfer in eukaryotes: aligning theory with data. Nat Rev Genet. 2024;25:416–30.
34. Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. Proc Natl Acad Sci U S A. 2006;103:9935–9.
35. McLysaght A, Hurst LD. Open questions in the study of de novo genes: what, how and why. Nat Rev Genet. 2016;17:567–78.
36. Zhao L, Svetec N, Begun DJ. De Novo genes Annu Rev Genet. 2024. https://doi.org/10.1146/annurev-genet-111523-102413.
37. Wybouw N, Pauchet Y, Heckel DG, Van Leeuwen T. Horizontal gene transfer contributes to the evolution of arthropod herbivory. Genome Biol Evol. 2016;8:1785–801.
38. Santos ME, Le Bouquin A, Crumière AJJ, Khila A. Taxon-restricted genes at the origin of a novel trait allowing access to a new environment. Science. 2017;358:386–90.
39. Kawahara AY, Plotkin D, Espeland M, Meusemann K, Toussaint EFA, Donath A, et al. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. Proc Natl Acad Sci U S A. 2019;116:22657–63.
40. Mitter C, Davis DR, Cummings MP. Phylogeny and evolution of Lepidoptera. Annu Rev Entomol. 2017;62:265–83.
41. Kawahara AY, Storer C, Carvalho APS, Plotkin DM, Condamine FL, Braga MP, et al. A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. Nat Ecol Evol. 2023;7:903–13.
42. Krenn HW. Feeding mechanisms of adult Lepidoptera: structure, function, and evolution of the mouthparts. Annu Rev Entomol. 2010;55:307–27.
43. Bazinet AL, Mitter KT, Davis DR, Van Nieukerken EJ, Cummings MP, Mitter C. Phylotranscriptomics resolves ancient divergences in the Lepidoptera: Ancient divergences in Lepidoptera. Syst Entomol. 2017;42:305–16.
44. Macias-Muñoz A, Rangel Olguin AG, Briscoe AD. Evolution of phototransduction genes in Lepidoptera. Genome Biol Evol. 2019;11:2107–24.
45. Mulhair PO, Crowley L, Boyes DH, Harper A, Lewis OT, Darwin Tree of Life Consortium, et al. Diversity, duplication, and genomic organization of homeobox genes in Lepidoptera. Genome Res. 2023;33:32–44.
46. Mulhair PO, Crowley L, Boyes DH, Lewis OT, Holland PWH. Opsin gene duplication in Lepidoptera: Retrotransposition, sex linkage, and gene expression. Mol Biol Evol. 2023;40:msad241.
47. Darwin Tree of Life Project Consortium. Sequence locally, think globally: The Darwin Tree of Life Project. Proc Natl Acad Sci U S A. 2022;119: e2115642118.
48. Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, et al. The Earth BioGenome Project 2020: Starting the clock. Proc Natl Acad Sci U S A. 2022;119: e2115635118.
49. Rota J, Twort V, Chiocchio A, Peña C, Wheat CW, Kaila L, et al. The unresolved phylogenomic tree of butterflies and moths (Lepidoptera): Assessing the potential causes and consequences. Syst Entomol. 2022;47:531–50.
50. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238.
51. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
52. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.
53. Steenwyk JL, Buida TJ, Labella AL, Li Y, Shen X-X, Rokas A. PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. Bioinformatics. 2021;37:2325–31.
54. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–4.
55. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 2024;34:769–77.
56. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. Ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8:28–36.
57. Wickham H. Ggplot2: Elegant graphics for data analysis. 2nd ed. Cham, Switzerland: Springer International Publishing; 2016.
58. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.
59. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem. 2004;25:1605–12.
60. Hao Z, Lv D, Ge Y, Shi J, Weijers D, Yu G, et al. RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. PeerJ Comput Sci. 2020;6: e251.
61. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42 Database issue:D222–30.
62. Denecke SM, Driva O, Luong HNB, Ioannidis P, Linka M, Nauen R, et al. The identification and evolutionary trends of the solute carrier superfamily in arthropods. Genome Biol Evol. 2020;12:1429–39.
63. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
64. Bittrich S, Segura J, Duarte JM, Burley SK, Rose Y. RCSB Protein Data Bank: exploring protein 3D similarities via comprehensive structural alignments. Bioinformatics. 2024;40:btae370.
65. Xu G-F, Gong C-C, Lyu H, Deng H-M, Zheng S-C. Dynamic transcriptome analysis of Bombyx mori embryonic development. Insect Sci. 2022;29:344–62.
66. Yokoi K, Tsubota T, Jouraku A, Sezutsu H, Bono H. Reference Transcriptome Data in Silkworm Bombyx mori. Insects. 2021;12:519.
67. Ranz JM, González PM, Clifton BD, Nazario-Yepiz NO, Hernández-Cervantes PL, Palma-Martínez MJ, et al. A de novo transcriptional atlas in Danaus plexippus reveals variability in dosage compensation across tissues. Commun Biol. 2021;4:791.
68. Li X, Fan D, Zhang W, Liu G, Zhang L, Zhao L, et al. Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies. Nat Commun. 2015;6:8212.
69. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
70. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
71. Shumate A, Wong B, Pertea G, Pertea M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. PLoS Comput Biol. 2022;18: e1009730.
72. Harrison PW, Amode MR, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, et al. Ensembl 2024. Nucleic Acids Res. 2024;52:D891–9.

73. Copley SD. Evolution of new enzymes by gene duplication and divergence. FEBS J. 2020;287:1262–83.
74. Axén A, Carlsson A, Engström A, Bennich H. Gloverin, an antibacterial protein from the immune hemolymph of Hyalophora pupae. Eur J Biochem. 1997;247:614–9.
75. Hwang J, Kim Y. RNA interference of an antimicrobial peptide, gloverin, of the beet armyworm, Spodoptera exigua, enhances susceptibility to Bacillus thuringiensis. J Invertebr Pathol. 2011;108:194–200.
76. Sparks ME, Blackburn MB, Kuhar D, Gundersen-Rindal DE. Transcriptome of the Lymantria dispar (gypsy moth) larval midgut in response to infection by Bacillus thuringiensis. PLoS ONE. 2013;8: e61190.
77. Urban JM, Foulk MS, Bliss JE, Coleman CM, Lu N, Mazloom R, et al. High contiguity de novo genome assembly and DNA modification analyses for the fungus fly, Sciara coprophila, using single-molecule sequencing. BMC Genomics. 2021;22:643.
78. Weisman CM, Murray AW, Eddy SR. Many, but not all, lineage-specific genes can be explained by homology detection failure. PLoS Biol. 2020;18: e3000862.
79. Weisman CM, Murray AW, Eddy SR. Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. Curr Biol. 2022;32:2632–9.e2.
80. Duplouy A, Hornett EA. Uncovering the hidden players in Lepidoptera biology: the heritable microbial endosymbionts. PeerJ. 2018;6: e4629.
81. Jiggins FM, Hurst GD, Jiggins CD, d Schulenburg JH v, Majerus ME. The butterfly Danaus chrysippus is infected by a male-killing Spiroplasma bacterium. Parasitology. 2000;120(Pt 5):439–46.
82. Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, et al. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. Cell. 2013;153:1567–78.
83. Cohen O, Gophna U, Pupko T. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol Biol Evol. 2011;28:1481–9.
84. Chen CK-M, Chan N-L, Wang AH-J. The many blades of the β-propeller proteins: conserved but versatile. Trends Biochem Sci. 2011;36:553–61.
85. Fiegler H, Bassias J, Jankovic I, Brückner R. Identification of a gene in Staphylococcus xylosus encoding a novel glucose uptake protein. J Bacteriol. 1999;181:4929–36.
86. Ioannidis P, Buer B, Ilias A, Kaforou S, Aivaliotis M, Orfanoudaki G, et al. A spatiotemporal atlas of the lepidopteran pest Helicoverpa armigera midgut provides insights into nutrient processing and pH regulation. BMC Genomics. 2022;23:75.

Publisher's Note