Received: 29 August 2022 | Accepted: 17 April 2023

DOI: 10.1111/imb.12844

ORIGINAL ARTICLE

Insect Molecular Biology

Genome sequencing provides insights into the evolution of gene families encoding plant cell wall-degrading enzymes in longhorned beetles

Na Ra Shin [1,2] | Yu Okamura [1,2] | Roy Kirsch [1,2] | Yannick Pauchet [1,2]

[1]Department of Entomology, Max Planck Institute for Chemical Ecology, Jena, Germany

[2]Department of Insect Symbiosis, Max Planck Institute for Chemical Ecology, Jena, Germany

Correspondence
Yannick Pauchet, Entomology, Max Planck Institute for Chemical Ecology, Hans-Knoell-Str. 8, Jena 07745, Germany.
Email: ypauchet@ice.mpg.de

Present address
Yu Okamura, Department of Biological Sciences, Graduate School of Science, University of Tokyo, Tokyo, Japan.

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: PA2808/4-1; Max-Planck-Gesellschaft

Abstract

With more than 36,000 species, the longhorned beetles (family Cerambycidae) are a mega-diverse lineage of mostly xylophagous insects, all of which are represented by the sole sequenced genome of the Asian longhorned beetle (*Anoplophora glabripennis*; Lamiinae). Their successful radiation has been linked to their ability to degrade plant cell wall components using a range of so-called plant cell wall-degrading enzymes (PCWDEs). Our previous analysis of larval gut transcriptomes demonstrated that cerambycid beetles horizontally acquired genes encoding PCWDEs from various microbial donors; these genes evolved through multiple duplication events to form gene families. To gain further insights into the evolution of these gene families during the Cerambycidae radiation, we assembled draft genomes for four beetle species belonging to three subfamilies using long-read nanopore sequencing. All the PCWDE-encoding genes we annotated from the corresponding larval gut transcriptomes were present in these draft genomes. We confirmed that the newly discovered horizontally acquired glycoside hydrolase family 7 (GH7), subfamily 26 of GH43 (GH43_26), and GH53 (all of which are absent from the *A. glabripennis* genome) were indeed encoded by these beetles' genome. Most of the PCWDE-encoding genes of bacterial origin gained introns after their transfer into the beetle genome. Altogether, we show that draft genome assemblies generated from nanopore long-reads offer meaningful information to the study of the evolution of gene families in insects. We anticipate that our data will support studies aiming to better understand the biology of the Cerambycidae and other beetles in general.

KEYWORDS
Cerambycidae draft genome, horizontal gene transfer, nanopore long reads, plant cell wall-degrading enzymes

INTRODUCTION

The Phytophaga clade of beetles represents a mega-diverse assemblage of mostly phytophagous insects encompassing two superfamilies, the Chrysomeloidea (leaf beetles and longhorned beetles) and the Curculionoidea (weevils and bark beetles) (Marvaldi et al., 2009). Specifically, the longhorned beetles (Cerambycidae) form a major radiation of insects (containing an estimated 36,300 extant species), for which the immature life stage is mostly xylophagous (Allison et al., 2004; Haack, 2017; Monné et al., 2017). These larvae bore deep into woody tissues of dead, heavily decaying, or even healthy trees. Consequently, the bulk of their food is rich in plant cell wall polysaccharides such as

cellulose, xylan, and pectin. Cerambycid larvae digest these complex polysaccharides using a set of plant cell wall-degrading enzymes (PCWDEs) encoded by the beetle's genome, which have been acquired from various microbial donors through several independent horizontal gene transfer (HGT) events (McKenna et al., 2016; McKenna et al., 2019; Pauchet et al., 2014; Shin et al., 2021).

Previously, we generated transcriptomes that targeted the larval gut of more than two dozen species representing most subfamilies of Cerambycidae that we used to identify, analyse the distribution of, and clarify the evolution of genes encoding PCWDEs in this diverse lineage of beetles (Shin et al., 2021). We identified 10 enzyme families and subfamilies—all glycoside hydrolases (GHs)—potentially implicated in digesting polysaccharides in the plant cell wall. These enzyme families, which are composed of multiple paralogs, result from several gene-duplication events. Apart from our recent study of the GH5_2 family (Shin et al., 2022), few large-scale functional analyses of cerambycid-derived PCWDEs exist. We showed that paralogs of this enzyme family evolved their substrate specificity in such a way that, together, they are responsible for the breakdown of cellulose, as well as hemicellulose polysaccharides, such as xyloglucan, xylan, and mannans.

Transcriptome sequencing is a powerful tool for gene discovery but provides no information about the structure and organisation in the genome of genes of interest. This knowledge would benefit our case particularly, given that the PCWDE-encoding genes we identified in species of Cerambycidae were horizontally acquired from microbial donors. The sole sequenced genome for species of Cerambycidae available to date, that of the Asian longhorned beetle *Anoplophora glabripennis* (Lamiinae), has shown that several families of PCWDEs, including GH5_2, GH28, GH45, and GH48, were encoded by the beetle's genome and has provided insights into their evolution (McKenna et al., 2016). However, we discovered several new families of PCWDEs, such as GH5_8, GH43_26, and GH53, all of which seem to be restricted to species of the subfamily Cerambycinae and are consequently absent from the genome of *A. glabripennis* (Shin et al., 2021).

To gain further insights into the evolution of gene families encoding PCWDEs, especially newly discovered ones, in Cerambycidae, we extended our previous work by attempting to draft genomes from beetle species of Cerambycidae outside Lamiinae. When we dissected specimens for our larval gut transcriptome sequencing project, we kept small pieces of tissue (rest body) for each species; initially, we planned to use these for DNA barcoding in order to confirm species identity (Wilson, 2012). We investigated whether draft genomes generated from this starting material would be sufficiently contiguous to analyse the structure and organisation of PCWDE-encoding genes. Here, we present the draft genomes of four species of Cerambycidae—two Cerambycinae, one Lepturinae, and one Lamiinae—generated by combining the isolation of high-molecular-weight DNA with long-read nanopore sequencing. We characterised these new genomes by determining their Benchmarking Universal Single-Copy Orthologs (BUSCO) scores, the number of predicted genes, and by generating a species tree based on single-copy orthologous genes. We then assessed the degree of synteny between these genomes. Finally, we screened these genomes for the presence of carbohydrate-active enzymes and identified families of putative PCWDEs. We confirmed that all previously discovered PCWDE-encoding genes based on our larval gut transcriptomes were present in the corresponding draft genomes. Most of these genes harbour introns and were flanked by either known insect genes or insect-like transposable elements, thus providing further evidence supporting their acquisition from microbial donors through horizontal gene transfer (HGT). We identified several extra genes that were not present in our transcriptome datasets, most of which originated from species-specific gene duplication events. Taken together, we conclude that draft genomes generated from a limited amount of starting material and sequenced using long-read technology represent a useful resource for studies focusing on the evolution of gene families.

RESULTS

Genome sequencing, assembly, and annotation

We successfully isolated high-molecular-weight DNA (HMW DNA) out of larval rest body tissue we saved from four cerambycid species (Table S1). We obtained a read N50 after sequencing ranging from as small as 11.03 kb for *Rhamnusium bicolor* (Lepturinae) to as long as 37.2 kb for *Aromia moschata* (Cerambycinae) (Table 1). We generated draft genome assemblies using a pipeline that combined high-accuracy base calling, assembly with Flye, and several rounds of polishing and cleaning (Figure S1). Genome coverage ranged from as little as 15× for *A. moschata* up to 44× for *Exocentrus adspersus* (Lamiinae) (Table 2). The estimated genome size varied greatly, from 262 MB for the lamiine *E. adspersus* to 1.2 GB for the lepturine *R. bicolor* (Table 2). Estimating the BUSCO scores to assess the quality of these draft genomes (Tables 2, S2), we recovered 82.2% complete BUSCO genes from the draft assembly of the *Molorchus minor* (Cerambycinae) genome, and up to 96.7% of the *E. adspersus* genome. Finally, the number of predicted genes ranged from 17,422 in the *E. adspersus* genome to 22,887 in the *A. moschata* genome (Table 2). Raw sequencing data, as well as the assembly and the results of the gene annotation were all deposited in NCBI Genbank (Table 3).

We identified the mitochondrial genome of all four cerambycid species as single circular contigs in the corresponding draft genome assemblies (Figure S2), with coverage varying between 756× in *A. moschata* and 4014× in *E. adspersus* (Table S3). These genomes did not vary much in size, ranging from 15,109 bp in *A. moschata* to 16,140 bp in *E. adspersus* (Figure S2 and Table S3). To improve gene prediction, we took advantage of the RNA-SEQ data we had generated previously (Shin et al., 2021), using them to polish these mitogenomes. Gene order and orientation, which was conserved between the four mitogenomes and other Cerambycidae ones (Nie et al., 2020), comprised 13 protein-coding genes, 22 transfer RNA genes, and 2 ribosomal RNA genes (Figure S2). The size variation observed between these four mitogenomes is mostly due to the control region (non-coding region) (Figure S2). We deposited these annotated mitogenomes in Genbank (Table 3).

TABLE 1 Information on the reads obtained after Nanopore sequencing.

	Aromia moschata[a]	*Molorchus minor*	*Exocentrus adspersus*	*Rhamnusium bicolor*
# raw reads (k)	202.42/354.94	1240	846.92	3430
generated data size (Gb)	4.14/5.46	6.57	11.2	21.74
Read N50 (kb)	37.2/29.6	15.66	26.64	11.03

[a]For *A. moschata*, the numbers on the left correspond to the sequencing of HMW DNA obtained after BluePippin isolation, whereas those on the right correspond to the sequencing of HMW DNA isolated using the Short Read Eliminator kit (Circulomics).

TABLE 2 Summary of the assembly statistics of the four Cerambycidae draft genomes.

	Aromia moschata	*Molorchus minor*	*Exocentrus adspersus*	*Rhamnusium bicolor*
Estimated genome size (Mb)	796.05	468.78	262.13	1220
Total length (bp)	687,741,793	365,739,002	257,537,449	1,169,059,409
# contigs and scaffolds	1816	7052	1083	6054
N50 (contig/scaffold) (bp)	1,105,269	140,248	1,296,204	641,621
Mean contig/scaffold length (bp)	378712.4	51863.2	237,800	193105.3
Longest contig/scaffold (bp)	12,512,031	1,513,889	8,848,599	7,506,537
GC content (%)	35.09	34.58	36.85	32.6
BUSCO (%)	87.5	82.2	96.7	93.7
Coverage	15×	18×	44×	20×
TEs and repeats content (%)	60.48	45.83	34.21	63.43
# predicted genes	22,887	19,533	17,422	19,992

Note: Quality information was generated from Seqkit, and proportion information was assessed manually using annotation from RepeatMasker.

TABLE 3 Accession numbers from NCBI corresponding to the four new genome datasets.

Species name	SRA	Genome assembly & annotation	Mitogenome
Aromia moschata	SRR19837461	JAPWTK000000000	OP096448
Molorchus minor	SRR19837458	JAPWTJ000000000	OP346982
Exocentrus adspersus	SRR19837459	JANEYG000000000	OP096449
Rhamnusium bicolor	SRR19837460	JANEYF000000000	OP346983

Abbreviations: SRA, short read archive.

Analysis of the repeatome

To complement the current and subsequent analyses, in addition to our four newly generated genomes, we included the genome of the model Cerambycidae *Anoplophora glabripennis* (Lamiinae; GCA_000390285.2) and the recently available genome from *Rutpela maculata* (Lepturinae; GCA_936432065.2) generated in the context of the Darwin Tree of Life project. Altogether, our Cerambycidae dataset contains six species: two Lamiinae (*A. glabripennis* and *E. adspersus*), two Lepturinae (*R. bicolor* and *R. maculata*), and two Cerambycinae (*M. minor* and *A. moschata*).

Genome size vary greatly between the species of Cerambycidae we investigated, regardless of the subfamily to which they belong (Table 2). Previous studies in several groups of insects, including grasshoppers (Shah et al., 2020) and Lepidoptera (Talla et al., 2017), have shown that genome size usually correlates positively with the amount of transposable elements (TEs) and simple repeats present in these genomes. Although 34.21% of the relatively small genome of *E. adspersus* was occupied by TEs and repeats, these make up to 63.43% of the large genome of *R. bicolor* (Table 2, Figure 1a). We identified both class I (retrotransposons) and class II (DNA transposons) TEs in all four genome assemblies, as well as in the *A. glabripennis* and *R. maculata* genomes (Table S4). Across the six Cerambycidae genomes analysed, we recovered a statistically significant positive relationship between the proportion of TE content and genome size (Pearson's correlation test: $r = 0.9$, $p = 0.015$) (Figure 1b). Altogether, our data indicate that the repeatome strongly contributes to genome size in Cerambycidae.

Synteny between Cerambycidae genomes

Keeping in mind that the draft genomes generated were still relatively fragmented, we assessed the degree of synteny between the

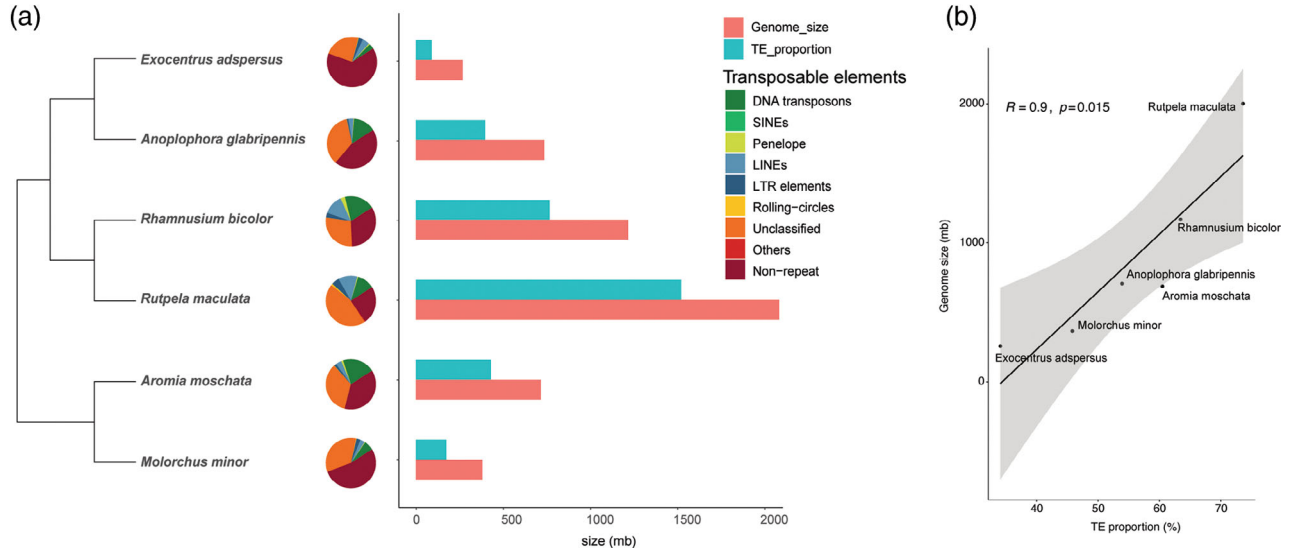

FIGURE 1 Analyses of the repeatome of six Cerambycidae draft genome assemblies. We performed repeat masker to assess the content of transposable elements (TEs) and simple repeats present in these assemblies. (a) Pie charts indicate the proportion of various categories of TEs identified from these genomes, as well as the proportion of those assemblies free of TEs. These TE categories are colour coded and presented on the figure. The bar graph puts in relation the estimated size of each Cerambycidae genomes and the corresponding amount covered by TEs. (b) Linear model of the relationship between the proportion of TEs present on each genome (X-axis) and the corresponding predicted genome size (Y-axis). The grey area indicates the 95% confidence interval (Pearson's correlation test: $r = 0.9$, $p = 0.015$).

genomes of species belonging to different subfamilies. We compared pairwise the two most contiguous draft assemblies we generated, namely those of *E. adspersus* (Lamiinae) and *A. moschata* (Cerambycinae), to the chromosome-scale assembly of *R. maculata* (Lepturinae) (Figure 2). To reduce the complexity of the analysis, we used only contigs that were larger than the N50 contig size of the corresponding assemblies. The genomes of *E. adspersus* to *R. maculata* share a high level of synteny (Figure 2a). Individual contigs from the *E. adspersus* assembly matched large sections of individual chromosomes of *R. maculata*, further illustrating the size difference between the genomes of the two species. Interestingly, our analyses also revealed that sections of individual *E. adspersus* contigs match to two or more *R. maculata* chromosomes (Figure 2b). We obtained similar results from the synteny analyses comparing the genomes of *A. moschata* and *R. maculata* (Figure 2c,d). Altogether, we conclude that the genomes of longhorned beetles belonging to different subfamilies share large syntenic regions, but chromosome rearrangements may have occurred after these subfamilies split.

Species phylogeny

Using BUSCO genes recovered from our four newly generated genomes as well as from publicly available beetle genomes (Table S2), we generated a species tree based on a maximum likelihood (ML) phylogenetic analysis (Figure 3; Supplementary dataset 1). In total, 299 BUSCO genes were held in common in the 34 beetle genomes we analysed and the three lepidopteran genomes included as an outgroup. In general, relationships between this set of beetle species correlate well with recent analyses of their phylogeny (McKenna et al., 2019; Zhang et al., 2018). The six Cerambycidae species, which were recovered as monophyletic, clustered subfamily-wise, and sister to species of Chrysomelidae, formed the Chrysomeloidea. However, *Callosobruchus maculatus* (Chrysomelidae, Bruchinae), which did not cluster together with other species of leaf beetles but was recovered as sister clade to the rest of the Chrysomelidae + Cerambycidae, is the only taxon that is not placed according to current phylogenetic hypotheses (McKenna et al., 2019; Nie et al., 2020; Song et al., 2018; Zhang et al., 2018). Species of weevils (Curculionoidea) were recovered as monophyletic and formed a clade sister to Chrysomelidae + Cerambycidae, confirming the now accepted monophyly of the Phytophaga clade (McKenna et al., 2019, Zhang et al., 2018).

Organisation of genes encoding plant cell wall-degrading enzymes (PCWDEs) in genomes of Cerambycidae

We performed a dbCAN analysis (Zhang et al., 2018) on the predicted gene sets to identify and annotate specific carbohydrate-active enzymes (CAZymes) encoded by the genome of these cerambycid species (Table 4). The enzyme classes with the largest expansions are the glycoside hydrolases (GHs), followed by the glycosyltransferases (GTs). Within GHs, the most abundant families consistently annotated in Cerambycidae genomes are GH1 β-glucosidases, GH18 chitinases, and GH28 polygalacturonases (Supplementary dataset 2).

We then extracted specific CAZymes, namely, those potentially implicated in plant cell wall degradation, manually curated them, and

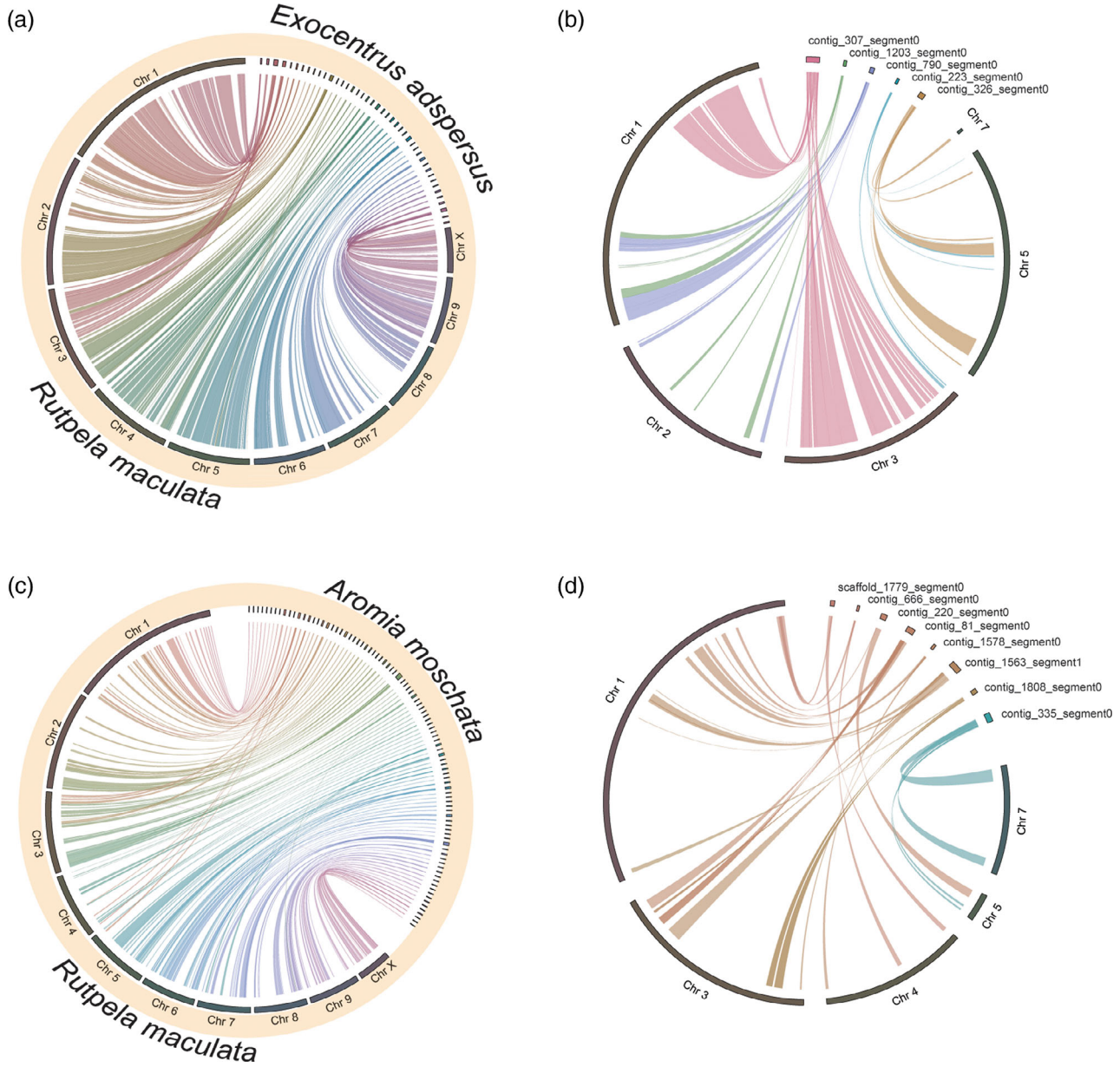

FIGURE 2 Degree of synteny between Cerambycidae genomes from species of different subfamilies. We compared the draft genomes of the Lamiinae *Exocentrus adspersus* (a and b) and of the Cerambycinae *Aromia moschata* (c and d) to the chromosome-scale one of the Lepturinae *Rutpela maculata*. For both genomes generated, we used only those contigs/scaffolds that were bigger than the N50 contig/scaffold size, which corresponded to 43 and 87 contigs for *E. adspersus* and *A. moschata*, respectively. (a) and (c) are Circos maps of the pairwise comparisons of these contigs with the 10 pseudo-chromosomes from *R. maculata*. (b) and (d) are Circos maps of selected contigs to illustrate possible chromosome rearrangements between genomes of species belonging to different subfamilies.

compared the number of predicted genes present in these genomes to the number we annotated from the corresponding larval gut transcriptomes (Shin et al., 2021). First, the enzyme families predicted to play a role in the digestion of PCW polysaccharides are all GHs. Second, the genes encoding all members of the PCWDE families we previously annotated from the larval midgut transcriptome were recovered from the corresponding draft genomes (Table 5), but we were unable to identify any new PCWDE families. However, new genes in already identified PCWDE families were annotated (Table 5). Importantly, most of the genes encoding PCWDE families harbour introns, including those corresponding to the recently discovered families GH7, GH53, and GH43_26 (Figures 4, S3–S9). The only exceptions are genes encoding GH5_2 enzymes, which harbour no intron, confirming previous findings from the genome of *A. glabripennis* (McKenna et al., 2016).

We observed that, in Cerambycidae-derived genes encoding GH45, GH48, and GH28 proteins, the position and phase of many introns were conserved when compared between cerambycid species or to genes derived from species of Chrysomelidae or Curculionidae (Figures 5, S3–S5). Intron gain seems to be frequent for genes

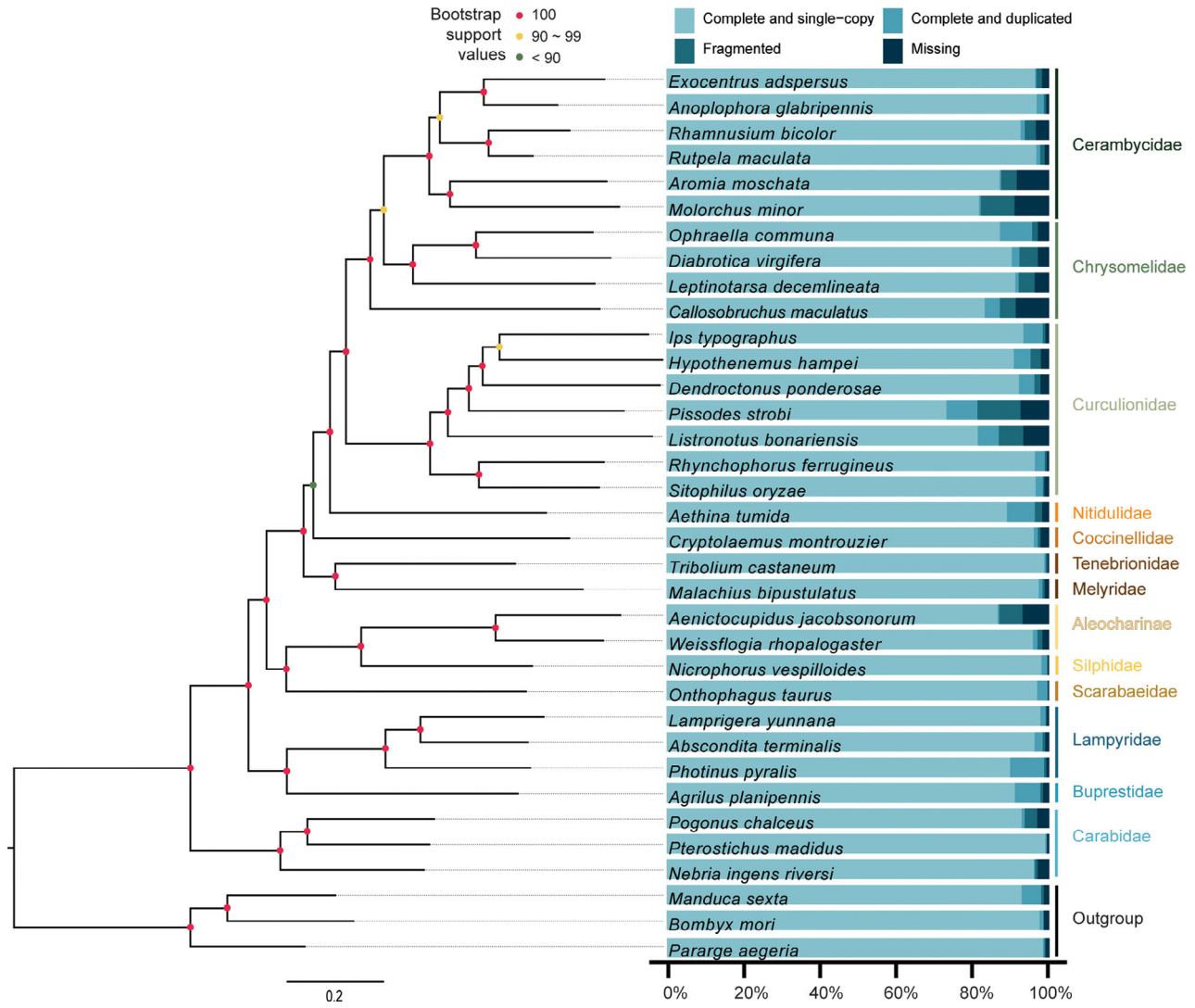

FIGURE 3 Species tree based on single-copy orthologous genes. In addition to the four newly generated Cerambycidae genomes, we retrieved draft genome assemblies available on public databases for 28 beetle species spanning several lineages representative of the order Coleoptera. We used draft genome assemblies of three species of Lepidoptera as an outgroup. First, the quality of all these genome assemblies was assessed using BUSCO (right side of the figure), and all BUSCO genes that were found to be in common between the 35 assemblies were retrieved for a total of 299 genes. A codon-based alignment was performed for each of these genes using MAFFT, these single alignments were then concatenated, and gaps were removed. The resulting sequence alignment was used to perform a maximum likelihood (ML) phylogenetic analysis using IQ-TREE. The best model fit for this dataset was determined to be GTR + F + R6. Node support was assessed by running 1000 ultrafast bootstrap replicates. The resulting support values for each nodes are marked with dots of different colours; red: 100%; yellow: 90%–99%; orange: below 90%. Relevant taxonomic information is indicated on the far right of the figure.

encoding GH45 and GH48 (Figures 4 and 5), whereas the typical four-exon structure previously described for GH28-encoding genes harboured by leaf beetles and weevil genomes (Kirsch et al., 2014) was found to be conserved for almost all the corresponding genes found in Cerambycidae genomes (Figures 4 and 5). We pointed out in our transcriptome survey (Shin et al., 2021) that Lamiinae-derived GH28s and what we called the "ancestral-type" GH28s (distributed in leaf beetles, weevils, and all Cerambycidae, except species of Lamiinae) were acquired through independent HGT events. This fact is also reflected in their gene structure (Figure 4). The genes encoding Lamiinae-derived GH28s harbour a single intron located in the predicted 5'-UTR but no intron in the open reading frame (ORF), and genes encoding "ancestral-type" GH28s are composed of four exons and three introns in their ORF.

Previously, we identified genes encoding xylanases belonging to GH10 in the transcriptomes of three species of cerambycid beetles (Shin et al., 2021). We now have the intron-exon structure of the corresponding genes for two of these species (Figure 4) and could compare it to the structure of the *gh10* gene encoded by the genome of the coffee berry borer, *Hypothenemus hampei* (Curculionidae; Scolytinae), the sole beetle genome available to date that harbours such a gene (Padilla-Hurtado et al., 2012; Vega et al., 2015). Interestingly, the

TABLE 4 Distribution of carbohydrate active enzymes (CAZymes) in the six analysed Cerambycidae genomes.

CAZy	*Aromia moschata*	*Molorchus minor*	*Rhamnusium bicolor*	*Exocentrus adspersus*	*Anoplophora glabripennis*
AA	44	28	35	53	59
CE	32	32	16	48	115
GH	169	171	138	171	310
GT	159	120	119	154	311
PL	2	2	2	3	5

Abbreviations: AA, auxiliary activities; CE, carbohydrate esterases; GH, glycoside hydrolases; GT, glycosyltransferases; PL, polysaccharide lyases.

TABLE 5 Comparison of the number of PCWDE-encoding genes per family identified in Cerambycidae transcriptomes and in the corresponding draft genome assemblies.

	Aromia moschata		*Molorchus minor*		*Rhamnusium bicolor*		*Exocentrus adspersus*	
	T	G	T	G	T	G	T	G
GH9	1	1	-	1	-	1	1	1
GH45	2	4	7	9	1	1	1	2
GH48	2	3	2	4	2	6	-	-
GH28	8	9	6	7	3	3	7	11
GH5_2	-	-	-	-	4	7	5	8
GH5_8	2	2	-	-	-	-	-	-
GH43_26	6	12	-	-	-	-	-	-
GH53	2	2	-	-	-	-	-	-
GH10	1	1	-	-	-	-	1	1
GH7	-	-	-	-	-	-	1	2

Note: T, transcriptome generated from RNA-SEQ data as in our previous study (Shin et al., 2021). G, corresponding draft genome assemblies.

gh10 genes present in the genomes of *A. moschata* and *E. adspersus*, and in the fragmented draft genome of the Cerambycinae *Xylotrechus colonus* (previously investigated in (Shin et al., 2021)), share two introns in a conserved position and phase with the *gh10* genes of *H. hampei* (Figures 5, S6); this overlap suggests that these genes were inherited by these three species from a common ancestor, likely the ancestral Phytophaga.

Coincidently, the coffee berry borer is also the only other species of Phytophaga—other than species of Cerambycinae—to express GH5_8 mannanases (Aguilera-Galvez et al., 2013; Vega et al., 2015). We previously investigated the structure of GH5_8-encoding genes in the two highly fragmented draft genomes of *Phymatodes lengi* and *X. colonus* (Shin et al., 2021). Their structure comprises two exons and a single intron with a conserved position and phase between *gh5_8* genes of these two species. We confirmed these findings with the two extra *gh5_8* genes encoded by the genome of the Cerambycinae *A. moschata* (Figure 4). In contrast, neither of the *gh5_8* genes encoded by the *H. hampei* genome harbour any introns (Aguilera-Galvez et al., 2013, Vega et al., 2015) (Figures 5, S7).

We further analysed the genomic environment of regions harbouring PCWDE-encoding genes in these genomes and assessed the degree of microsynteny between them. These genomic regions can be sorted into three categories: (i) a highly microsyntenic genomic region common to all genomes investigated that harbours a PCWDE-encoding gene (like GH9) (Figure 6a); (ii) a highly conserved microsyntenic region common to all genomes investigated, but PCWDE-encoding genes are present only in the genomes of species belonging to the same subfamily (such as GH45 in Cerambycinae) (Figure 6b); (iii) a highly conserved microsyntenic region common to all genomes investigated, but a PCWDE-encoding gene is found only in the genome of a single species (Figure 6c). Interestingly, we also observed cases where true orthologs (based on phylogenetic analyses and functional data), such as GH5_2-encoding genes (Shin et al., 2022), are located in very different genomic regions, further supporting the hypothesis that genome rearrangements occurred after the corresponding species split.

DISCUSSION

Pros and cons of using draft genomes generated from nanopore long-read only to study gene family evolution

We started this project with limited amount of material, which allowed us to extract only a few hundred nanograms of high-molecular-weight DNA. As our goal was to analyse how PCWDE-encoding genes were organised in genomes of Cerambycidae, we wanted draft

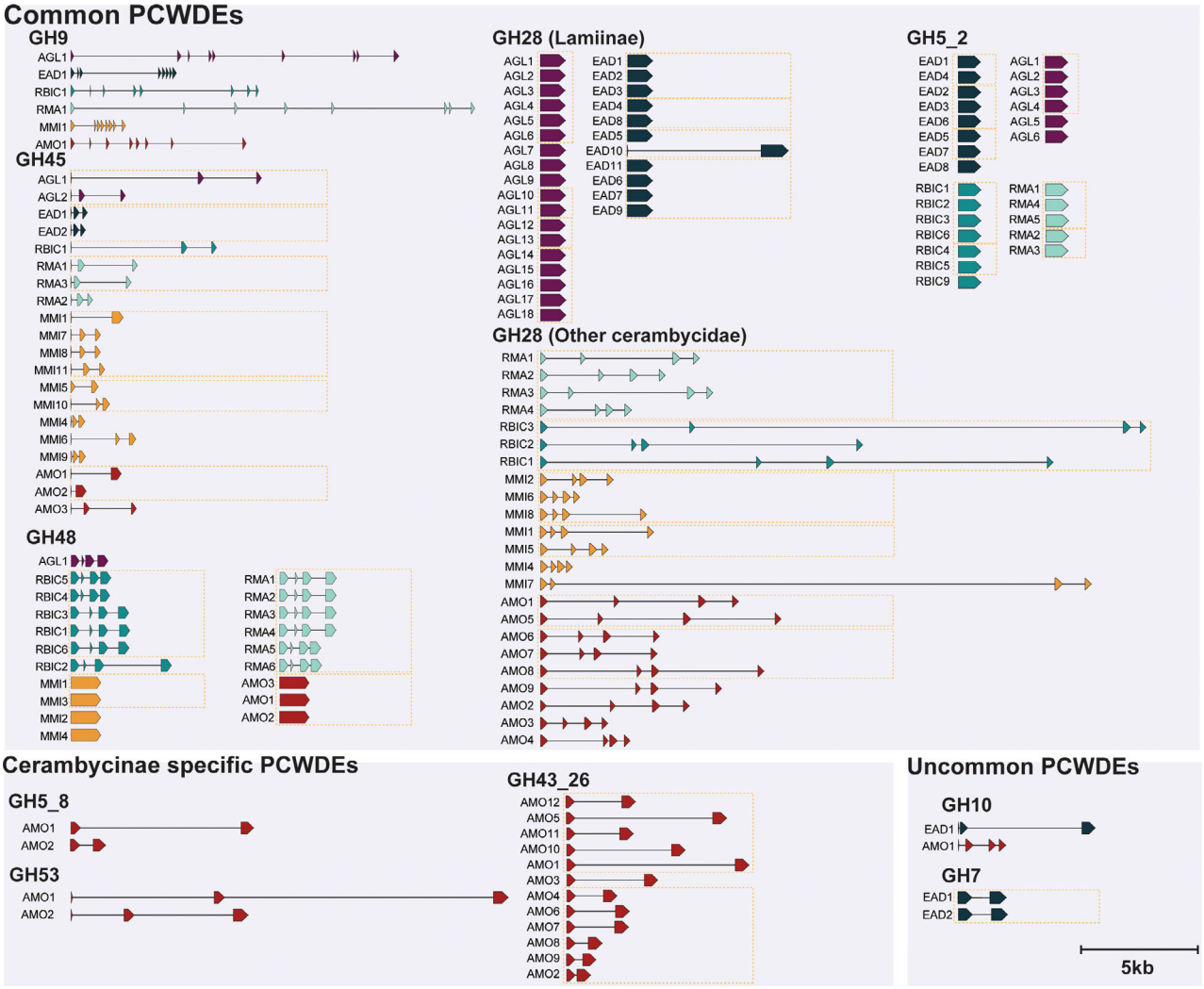

FIGURE 4 Summary of the gene structure of all the PCWDE-encoding genes identified from six Cerambycidae genomes. Exons are indicated by blocks, whereas introns are represented by thin lines. All the gene structures found in a given species have the same colour; purple: *Anoplophora glabripennis* (AGL); petrol: *Exocentrus adspersus* (EAD); blue: *Rhamnusium bicolor* (RBIC); light blue: *Rutpela maculata* (RMA); orange: *Molorchus minor* (MMI); red: *Aromia moschata* (AMO). Genes located in tandem on the same contig are indicated by yellow dashed boxes. Each gene structure is drawn to scale.

assemblies to be as contiguous as possible. Thus, we opted for long-read nanopore sequencing, keeping in mind that this technology tends to be highly error-prone (Rang et al., 2018; Wick et al., 2019). To get the most out of our limited starting material, we successfully enriched our genomic DNA in high-molecular-weight fragments, reaching up to 37 kb read N50 for *A. moschata*. After sequencing of these fragments, we built a bioinformatics pipeline combining high-accuracy base calling using an optimised version of Oxford Nanopore's Guppy base caller before assembly, to several rounds of polishing of the obtained assembly with Racon and Medaka, and finally the removal of contaminating sequences using Kraken.

BUSCO scores, although below 90% of complete single-copy BUSCO genes for two of our draft genomes, were sufficient to obtain a reliable species tree according to current hypotheses about the phylogeny of beetles (McKenna et al., 2019; Zhang et al., 2018). The misplacement of *Callosobruchus maculatus* in our analysis could be due to either the poor representation of species of Chrysomelidae in our dataset, or to the overall quality of the corresponding draft genome. In addition, the fact that all the PCWDE-encoding genes we previously identified in the gut transcriptomes of the corresponding species (Shin et al., 2021) were present in these draft genomes, provided extra evidence that we could use these assemblies for subsequent analyses. On the one hand, we recognise that the suitability of our assemblies for analyses dealing with chromosome synteny and whole genome comparisons is limited. However, we anticipate that chromosome-scale genome assemblies for other species of Cerambycidae, in addition to the recently generated assembly for *Rutpela maculata* and for other species of the Phytophaga clade (Crowley et al., 2021; Van Dam et al., 2021), will become available in the near future and will allow for such analyses. On the other hand, the draft assemblies we generated were contiguous enough and enabled us to (i) identify tandem gene duplications among families of genes

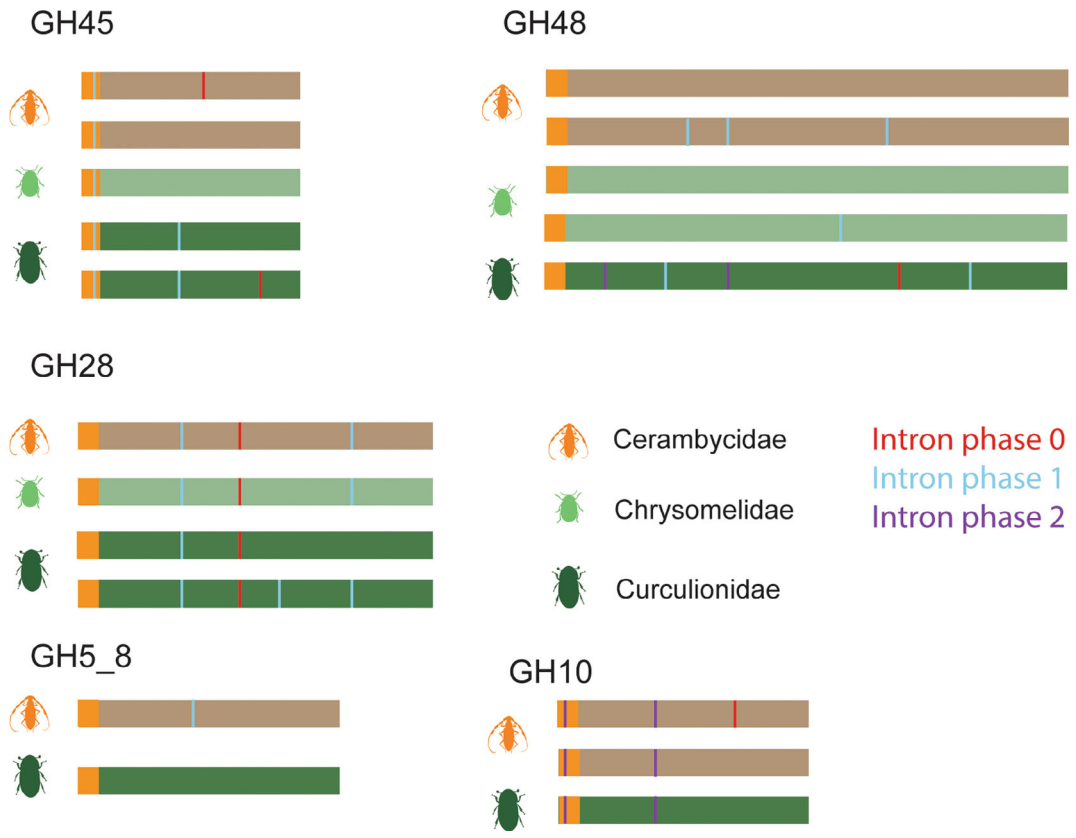

FIGURE 5 Conservation of intron position and phase between PCWDE-encoding genes of different species across the Phytophaga clade. The primary structure of each enzyme family is presented. The sequence corresponding to the predicted signal peptide is marked as a yellow box. Intron position is indicated by a line of different colour according to the phase; red: phase 0; light blue: phase 1; purple: phase 2.

encoding PCWDEs, (ii) determine the intron-exon structure of these genes, and (iii) analyse the genomic regions surrounding PCWDE-encoding genes and compare these between draft assemblies of different species (microsynteny).

In addition to the sequence of the insect's own genome and the associated mitogenome, recent insect genome sequencing projects have also described associated bacteria such as *Wolbachia* and other endosymbionts (Faulk, 2023; Gagalova et al., 2022). These projects typically used a whole insect as starting material for genomic DNA (gDNA) extraction, and in this case, the chance to isolate gDNA of associated bacteria is high. Our Kraken analysis did not reveal any significant traces of associated bacteria or other microorganisms in the four genomes we generated. However, the presence of yeast symbionts in larvae of extended lineages of longhorned beetles has long been known (Grunwald et al., 2010). Future genome projects using species of Cerambycidae should consider this fact.

High variation among size of Cerambycidae genomes

The only published genome of a species of Cerambycidae, that of the Asian longhorned beetle *A. glabripennis*, has an estimated size of about 980 Mb. Our study showed that genomes of Cerambycidae vary greatly in size: the draft genome of *E. adspersus* was the smallest, estimated at about 260 Mb; the genome of *R. maculata* was the biggest, close to 2 Gb. This observation was not unexpected, as such a variation has been observed in other lineages of beetles (Hanrahan & Johnston, 2011; Petitpierre et al., 1993). Our data indicate that TEs and simple repeats likely play an important role in this phenomenon, as exemplified by the positive correlation we observed between the size of the genomes and the part of it covered by the repeatome. Such a correlation explains the dramatic size variation between genomes of the myriapods (So et al., 2022). In some cases, TEs have been shown to be important evolutionary tools that contribute to processes such as chromosome rearrangements and the creation of new regulatory regions and gene mutations. In addition, TEs were implicated in the evolution of gene families by enabling duplications, inversions, and translocations, thus playing an important role in the adaptation of organisms to new environments (Bire & Rouleux-Bonnin, 2012).

A specific look at the evolution of PCWDEs in Cerambycidae

In draft genomes, unlike transcriptomes, HGT events can be better claimed by simply investigating the genomic environment of a gene of interest. From our analysis of larval gut transcriptomes of more than 20 species of Cerambycidae, especially species of the subfamily

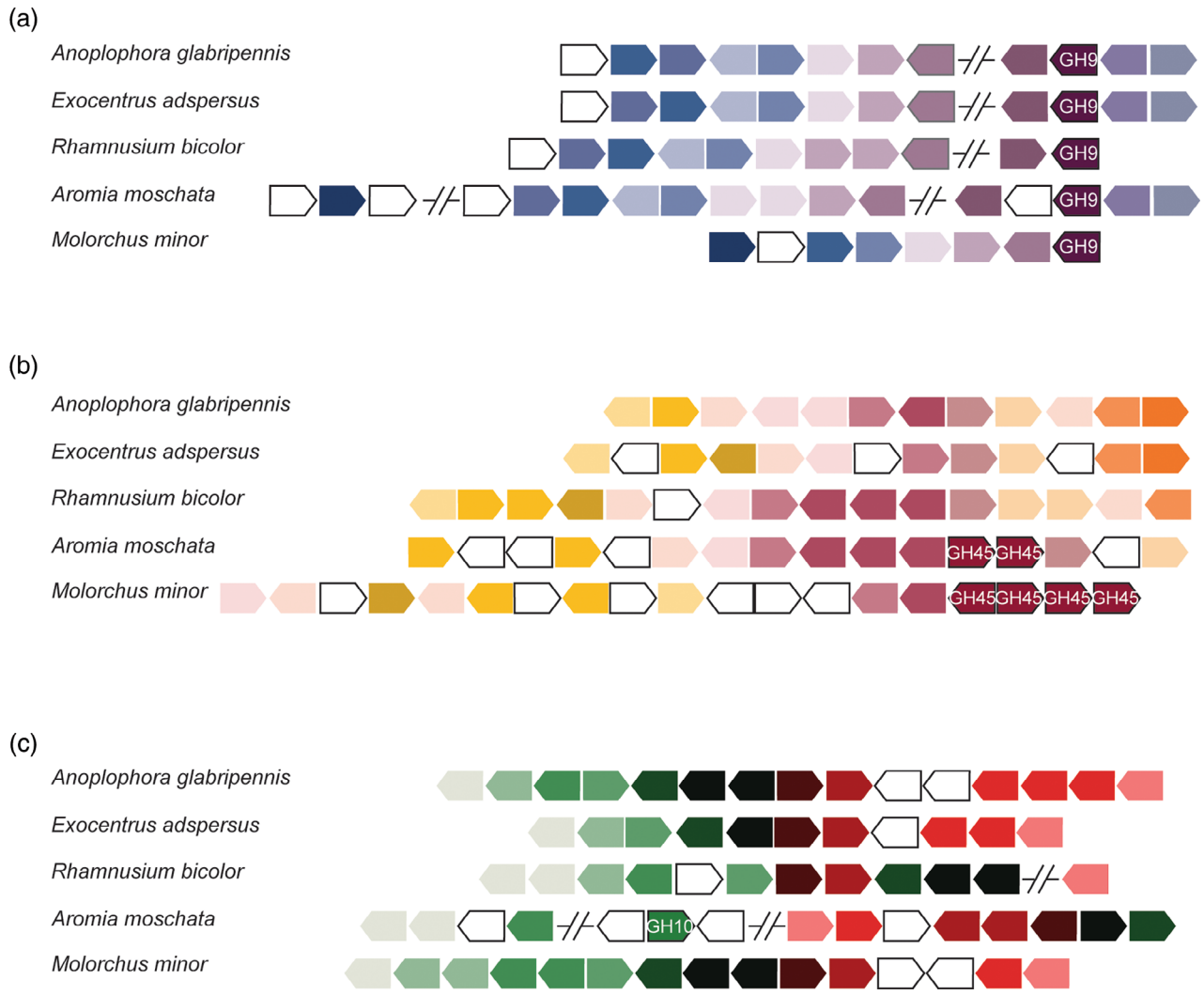

FIGURE 6 Micro-synteny analyses in genome regions harbouring PCWDE-encoding genes. (a) Genomic regions containing the *gh9* gene are presented for the four new draft genomes, and compared with the *A. glabripennis* genome. (b) Corresponding syntenic regions between the five genomes are presented. This region contains GH45-encoding genes only in the two species of Cerambycinae *M. minor* (MMI) and *A. moschata* (AMO). (c) Corresponding syntenic regions between five Cerambycidae genomes are presented. A GH10-encoding gene is present only in this region for the genome of *A. moschata* (AMO). Individual genes are represented by arrows indicating their orientation. Arrows harbouring the same colour correspond to orthologous genes.

Cerambycinae (McKenna et al., 2019; Shin et al., 2021), we identified new families of PCWDEs, such as GH5_8, GH43_26, GH53, and GH7, none of which were present in previous transcriptome analyses (Pauchet et al., 2014; Scully et al., 2013) or in the reference genome of *A. glabripennis* (McKenna et al., 2016). Using phylogenetic analyses, we hypothesized that these genes were acquired from various microbial donors through several independent HGT events (Shin et al., 2021). Here, we confirmed their integration into the Cerambycidae genomes in regions containing known insect-derived genes and showed that the genes acquired introns after their original integration.

The genomes of Cerambycidae encode PCWDEs that are also found in species of other families of Phytophaga, such as Chrysomelidae and Curculionidae. By analysing the genes encoding GH45 cellulases in beetles of these families, we demonstrated that the position and phase of the first intron are conserved in leaf beetle and weevil species (Busch et al., 2019). We observed a similar situation for genes encoding GH28 polygalacturonases, for which the position and phase of the three introns are conserved between species of leaf beetles and weevils (Kirsch et al., 2014). Here, we confirmed that for both enzyme families, intron position and phase are also conserved in the corresponding genes annotated from the Cerambycidae genomes. As intron conservation points out to a common origin, these results consolidate our view that genes encoding GH28, GH45, and GH48 proteins originated from independent HGT events in the most recent common ancestor of the Phytophaga clade (Busch et al., 2019; Kirsch et al., 2014; McKenna et al., 2019; Shin et al., 2021). In addition, given that most of the putative donor organisms of these HGT events were bacteria, it is reasonable to speculate that these genes gained introns shortly (in evolutionary time) after their integration into the insect's genome. Finally, intron gains have often been correlated with

increased gene expression levels due to the improved stability of the corresponding pre-mRNA (Le Hir et al., 2003). Thus, analysing the gene structure of HGT-derived genes provides insights into the origins of these events and the fate of these genes after acquisition.

We observed the conservation of the position and phase of several introns between genes encoding GH10 xylanases originating from the two Cerambycidae *A. moschata* and *E. adspersus*, and from the coffee berry borer, *H. hampei* (Curculionidae; Scolytinae) (Padilla-Hurtado et al., 2012; Vega et al., 2015). In addition, phylogenetic analysis revealed that the GH10 sequences of Cerambycidae and of *H. hampei* cluster together next to several clades of bacterial counterparts (Shin et al., 2021). These facts suggest that these three species inherited these genes from a common ancestor, likely ancestral Phytophaga. The current known distribution of *gh10* genes in species of Phytophaga is quite scarce, exemplified by the fact that we identified *gh10* genes in the transcriptome of only three Cerambycidae species out of the 23 we surveyed (Shin et al., 2021). This distribution pattern is better explained by massive gene loss in several whole lineages of Phytophaga beetles than by independent acquisitions through several HGT events. On the other hand, whereas the structure of Cerambycidae-derived genes encoding GH5_8 mannanases comprises two exons and a single intron with a conserved position and phase, neither of the *gh5_8* genes encoded by the *H. hampei* genome harbour any introns (Aguilera-Galvez et al., 2013; Vega et al., 2015). Moreover, a phylogenetic analysis including Cerambycinae- and *H. hampei*-derived genes together with microbial *gh5_8* genes showed that the beetle-derived sequences were not recovered in a monophyletic clade (Shin et al., 2021). These results suggest that the presence of these genes in Cerambycinae and in the coffee berry borer may have been the result of two independent HGT events, one that occurred at the beginning of the radiation of the subfamily Cerambycinae and a species- or genus-specific one in the coffee berry borer. However, an intron gain in the common ancestor of the Cerambycinae would be equally possible and cannot be excluded at this stage.

CONCLUSION

By adding four new draft genomes, we triple the number of Cerambycidae genomes available to date. These draft genomes cover three of the eight currently recognised subfamilies of Cerambycidae. A single one of these draft genomes is represented by a chromosome-scale assembly, the genome of *Rutpela maculata* derived from the Darwin Tree of Life initiative. Clearly, much remains to be done to improve species coverage and quality of genome assemblies for this megadiverse lineage of mostly xylophagous insects. Nonetheless, we hope that the generated genome assemblies will become a valuable resource for the scientific community worldwide, as these beetles are used to address fundamental questions in development, evolution, ecology, life history, genetics, physiology, and systematics. Although these genome assemblies are fragmented, they still provide a more comprehensive picture of the gene content of a species than transcriptomes, which are often generated from single tissues. Finally, they allow us to draw conclusions about the evolutionary history of HGT-derived genes, including their origin (common introns in distantly related species indicate a common origin), how they have adapted to the recipient's genomic environment, and their fate.

EXPERIMENTAL PROCEDURES

Isolation of high molecular weight genomic DNA

The tissues we used here for genomic DNA isolation were collected from the same specimens described in our previous work reporting on the analysis of larval gut transcriptomes of species of Cerambycidae (Shin et al., 2021). Besides preserving the gut of these specimens, we also kept a small piece of restbody (larval tissue without gut) for each of them. Initially, we planned to use these tissue samples for DNA barcoding for species identification. Insect tissue was mixed with 500 μL buffer CT in a 2-ml microcentrifuge tube, squeezed with a pestle, before being homogenised using a Sunlab Rotator "Digi Mixer" (NeoLab, Heidelberg, Germany). Genomic DNA (gDNA) was further isolated using the Nanobind Big DNA kit (Circulomics, Baltimore, MD, USA) following the manufacturer's instructions. To test methods for the isolation of high-molecular-weight (HMW) DNA molecules, we use the gDNA isolated from tissue of *A. moschata*. An aliquot was used to isolate HMW DNA on a BluePippin (Sage Science, Beverly, MA, USA) using a high-pass filtering threshold of 275 bp – 50 kb. Alternatively, we selectively precipitated HMW fragments from another gDNA aliquot using the Short Read Eliminator kit XS (Circulomics), again following the manufacturer's instructions. Based on a better yield and on the read N50 after sequencing, we decided to use the Short Read Eliminator kit XS to isolate HMW DNA from tissue of the remaining three species. Isolated HMW DNA was further cleaned up using AMPure XP beads (Beckman Coulter, Krefeld, Germany) following the manufacturer's instructions. Final DNA purity and concentrations were measured using Nanodrop (Thermo Fisher, Waltham, MA, USA) and Qubit (Thermo Fisher). The amount of HMW DNA isolated from tissue of the four species is summarised in Table S1.

Library preparation and sequencing

End-DNA repair before library preparation was performed using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs, Ipswich, MA, USA) following the manufacturer's instructions. Sequencing adapters were then ligated using the Ligation Sequencing Kit (Nanopore Technologies, Oxford, UK). Ligation reactions were incubated at room temperature for 10 min followed by a clean-up step with the AMPure XP beads (Beckman Coulter). Sequencing was performed on a MinION platform (Nanopore Technologies). Priming of the MinION flow cell for sequencing was performed with the Flow Cell Priming Kit (Nanopore Technologies). Flow cells were loaded with 100–200 ng of the libraries three times during a run of 72 h total,

with washing steps in between using the Flow Cell Wash Kit EXP-WSH003 (Nanopore Technologies).

High-accuracy base calling, assembly, and polishing

After sequencing, high-accuracy base calling of the raw reads was performed with an optimised version of Guppy v6.0.1 (Nanopore Technologies) (dna_r9.4.1_450bps_hac.cfg model). We de novo assembled the resulting reads in Flye v2.7.1 (Kolmogorov et al., 2019) with setting minimum overlap as 10 kb and with the "--meta" option. Then, we performed four rounds of polishing with Racon v1.3.3 (Vaser et al., 2017) starting from the Flye assembly with option (−m 8 -x − 6 -g − 8 -w 500). After each polishing round, reads were re-aligned to the resulting assembly with minimap2 v2.17 (Li, 2018). A final round of polishing was performed using Medaka v1.2.0 (https://github.com/nanoporetech/medaka) with the r941_min_high_g344 model using the MinION raw reads. After polishing, we merged haplotype redundancies using Purge haplotigs v1.0.4 (Roach et al., 2018) and collapsed duplicated haplotigs using Haplomerger2 v2.01 (Huang et al., 2017). When present, contigs corresponding to potential contaminants of microbial origin were removed using Kraken2 v2.0.8 (Wood et al., 2019). Genome coverage was calculated by mapping the reads to the assembled genome with minimap2 and subsequently using the genomeCoverageBed package in bedtools v2.29.0 (Quinlan, 2014). Genome size was estimated with backmap.pl v0.5 (Schell et al., 2017) based on the coverage information generated from minimap2 and using default settings. GC content was assessed with the seqkit tool (Shen et al., 2016) using the fx2tab option. BUSCO v5.2.1 (Simão et al., 2015) scores were calculated for each genome using the insect_odb10 database. Corresponding accessions numbers from Genbank can be found in Table S4.

Curation of mitochondrial genomes

Even before this project began, the mitogenomes of *Molorchus minor* (MN442323.1) and *Rhamnusium bicolor* (MN473084.1) were available in Genbank. As a benchmark experiment, we first mapped all the reads generated for the *M. minor* genome on the sequence of the publicly available *M. minor* mitogenome using minimap2. We performed the same procedure for *R. bicolor*. For the de novo assembly of the mitogenomes of *Aromia moschata* and *Exocentrus adspersus*, we mapped the corresponding reads generated for genome sequencing onto the publicly available *M. minor* mitogenome using minimap2. Reads that were mapped to the corresponding backbones were recovered for each species and assembled using Flye. During the polishing step, we performed two rounds of Racon. The resulting contig corresponding to the mitogenome of each species was inspected manually to check whether it was circular. Given that reads corresponding to the mitogenomes were found in the RNA-SEQ data of the corresponding species (Shin et al., 2021), we took this opportunity to perform an extra polishing step and mapped the RNA-SEQ Illumina short-reads of the corresponding species on the obtained mitogenomes. Automatic gene annotation was performed with MITOS2 (Donath et al., 2019). Each ORF was manually checked for the presence of indels or frameshifts.

Repeat identification and gene annotation

We employed the Maker2 v3.01.03 (Holt & Yandell, 2011) automated pipeline for both repeat identification and ab initio gene prediction. First, de novo repeat identification was performed using RepeatModeler v2.0.1. Second, the output generated by RepeatModeler was implemented in RepeatMasker v4.1.0 (Tempel, 2012), which classified repeats according to the open source Dfam repeat library and a custom Coleoptera-specific repeat database. The latter database was produced by performing a RepeatModeler analysis on the genomes assemblies of the three coleopteran model species *Leptinotarsa decemlineata* (GCF_000500325.1), *Anoplophora glabripennis* (GCA_000390285.2) and *Tribolium castaneum* (GCA_000002335.3). The outputs produced by RepeatModeler for these three genomes were then combined to generate a custom Coleoptera-specific repeat database. Then, we chose protein datasets derived from (i) the three model beetle species mentioned above, and (ii) RNA-SEQ data generated from larval gut corresponding to the four cerambycid species we assembled the genome for (Shin et al., 2021) to train both Augustus (Nachtweide & Stanke, 2019) and SNAP (Korf, 2004) before performing gene prediction. This step was performed three times to increase the accuracy of the gene prediction.

Synteny analyses

We analysed the degree of synteny between the chromosome-scale assembly of *Rutpela maculata* (Lepturinae) and the two most contiguous assemblies (of *Exocentrus adspersus* (Lamiinae) and *Aromia moschata* (Cerambycinae)). For the latter two species, we used only scaffolds/contigs that were larger than those of the N50 scaffolds/contigs belonging to the corresponding genome assemblies. This constraint corresponded to 43 scaffold/contigs for *E. adspersus* and 87 for *A. moschata*. We performed pairwise alignments between the 10 pseudo-chromosomes of *R. maculata* and (i) the 43 scaffolds/contigs of *E. adpsersus*, and (ii) the 87 scaffolds/contigs of *A. moschata* using PROmer (PROtein MUMmer v3.0.7) (Kurtz et al., 2004) with default parameters (breaklen: 60; maxgap: 30; minmatch: 6; minimum match clustering: 20). This script uses the raw DNA sequences, and performs all matching and alignment routines on the six-frame amino acid translation of the DNA input sequences. Then, we used Circos (Krzywinski et al., 2009) to visualise the results.

We used SynNet (Zhao et al., 2017), which combines DIAMOND and McscanX, to detect syntenic blocks between the draft genomes included in our analyses. We did not include *R. maculata* here because the annotated gene set was not available at the time we performed these analyses. SynNet uses annotation sets as input and automatically performs pairwise inter- and intra-species comparisons, trims the

outputs for synteny detection, and treats outputs containing all synteny blocks to a final network file. In parallel, we used PROmer to perform pairwise comparisons between draft genomes of Cerambycidae. Then, we combined the results of SynNet and PROmer, and manually collected syntenic contigs of interest (containing PCWDE-encoding genes) from the corresponding assemblies. Microsynteny between regions of Cerambycidae genomes was visualised using the R package genoPlotR (Guy et al., 2010).

Species phylogeny

Based on BUSCO scores (Waterhouse et al., 2018), we extracted single-copy as well as duplicated orthologous genes common to the genomes of the 35 beetle species we used, and a codon-based alignment of each individual BUSCO gene was performed with MAFFT v7.388 and PAL2NAL v14 as described earlier (Shin et al., 2021). These individual sequence alignments were trimmed with trimAL v1.4.rev22 with 5% gap cut-off, and badly aligned regions were removed using CIAlign v1.0.14. The resulting trimmed alignments of each individual BUSCO gene were concatenated to generate a super alignment of 299 genes for subsequent phylogenetic analysis. An ML analysis was performed in IQ-TREE v2.1.3 (Nguyen et al., 2015). Statistical measures of nodal support were estimated using ultrafast bootstrap (Hoang et al., 2018) analysis implemented in the IQ-TREE software with 1000 replicates.

Identification of CAZyme and curation of PCWDE-encoding genes

The identification of carbohydrate-active enzymes (CAZymes) was performed using dbCAN v2.0.11 (Huang et al., 2018) through a webserver (https://bcb.unl.edu/dbCAN2/index.php). Genes encoding putative PCWDEs were inspected manually; their intron-exon structure was determined by comparing the sequence of the cDNA to the corresponding sequence of the gene on the genome assembly using splign (Kapustin et al., 2008). During this analysis, we detected additional genes that were absent in the corresponding transcriptomes of each gene family. For these new genes, we relied on the most similar cDNA sequence to determine their intron-exon structure using genewise (Birney et al., 2004) first and then splign.

AUTHOR CONTRIBUTIONS

Na Ra Shin: Conceptualization; investigation; writing – original draft; formal analysis; visualization; methodology; validation; data curation. **Yu Okamura:** Writing – review and editing; methodology; conceptualization; validation. **Roy Kirsch:** Conceptualization; investigation; writing – review and editing; validation. **Yannick Pauchet:** Conceptualization; investigation; funding acquisition; writing – original draft; supervision.

ACKNOWLEDGEMENTS

We are grateful to Bianca Wurlitzer and Henriette Ringys-Beckstein for technical support. This work was supported by the Max Planck Society. Yannick Pauchet acknowledges support from the Deutsche Forschungsgemeinschaft (DFG; PA2808/4-1). We thank Emily Wheeler of Harwich, MA, for editorial support. Open access funding enabled and organised by Projekt DEAL. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest.

DATA AVAILABILITY STATEMENT

- Newly generated genome sequences (raw reads, assembly, annotated gene set, mitogenomes) have been submitted to Genbank. Accession numbers are summarised in Table 3.
- Tables S1–S4 and Figures S1–S9 have been submitted with the present paper and are available online.
- The following supporting data are freely accessible on EDMOND - the Open Research Data Repository of the Max Planck Society with DOI: https://doi.org/10.17617/3.XYUPD3.
  - Supplementary dataset 1 contains the codon-based nucleotide alignment (FASTA file), the IQ-TREE log file and the resulting tree file (Newick file) used to generate Figure 3.
  - Supplementary dataset 2 contains the dbCAN output of the identification of carbohydrate-active enzymes present in the annotated gene set of the four genomes.

ORCID

Na Ra Shin https://orcid.org/0000-0001-8435-3361
Yannick Pauchet https://orcid.org/0000-0002-2918-8946

REFERENCES

Aguilera-Galvez, C., Vasquez-Ospina, J.J., Gutierrez-Sanchez, P. & Acuna-Zornosa, R. (2013) Cloning and biochemical characterization of an endo-1,4-beta-mannanase from the coffee berry borer *Hypothenemus hampei*. *BMC Research Notes*, 6, 333.

Allison, J.D., Borden, J.H. & Seybold, S.J. (2004) A review of the chemical ecology of the Cerambycidae (coleoptera). *Chemoecology*, 14, 123–150.

Bire, S. & Rouleux-Bonnin, F. (2012) Transposable elements as tools for reshaping the genome: it is a huge world after all! *Methods in Molecular Biology*, 859, 1–28.

Birney, E., Clamp, M. & Durbin, R. (2004) GeneWise and Genomewise. *Genome Research*, 14, 988–995.

Busch, A., Danchin, E.G.J. & Pauchet, Y. (2019) Functional diversification of horizontally acquired glycoside hydrolase family 45 (GH45) proteins in Phytophaga beetles. *BMC Evolutionary Biology*, 19, 100.

Crowley, L.M. & University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium. (2021) The genome sequence of the hazel leaf-roller, *Apoderus coryli* (Linnaeus, 1758). *Wellcome Open Research*, 6, 315.

Donath, A., Jühling, F., Al-Arab, M., Bernhart, S.H., Reinhardt, F., Stadler, P.F. et al. (2019) Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. *Nucleic Acids Research*, 47, 10543–10552.

Faulk, C. (2023) *De novo* sequencing, diploid assembly, and annotation of the black carpenter ant, *Camponotus pennsylvanicus*, and its

symbionts by one person for $1000, using nanopore sequencing. *Nucleic Acids Research*, 51, 17–28.

Gagalova, K.K., Whitehill, J.G.A., Culibrk, L., Lin, D., Lévesque-Tremblay, V., Keeling, C.I. et al. (2022) The genome of the forest insect pest *Pissodes strobi* reveals genome expansion and evidence of a Wolbachia endosymbiont. *G3 Genes Genomes. Genetics*, 12, jkac038.

Grunwald, S., Pilhofer, M. & Holl, W. (2010) Microbial associations in gut systems of wood- and bark-inhabiting longhorned beetles [coleoptera: Cerambycidae]. *Systematic and Applied Microbiology*, 33, 25–34.

Guy, L., Kultima, J.R. & Andersson, S.G. (2010) genoPlotR: comparative gene and genome visualization in R. *Bioinformatics*, 26, 2334–2335.

Haack, R.A. (2017) Feeding biology of cerambycids. In: Wang, Q. (Ed.) *Cerambycidae of the world: biology and Pest management*. Boca Raton, FL: CRC Press.

Hanrahan, S.J. & Johnston, J.S. (2011) New genome size estimates of 134 species of arthropods. *Chromosome Research*, 19, 809–823.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q. & Vinh, L.S. (2018) UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, 35, 518–522.

Holt, C. & Yandell, M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, 12, 491.

Huang, S., Kang, M. & Xu, A. (2017) HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics*, 33, 2577–2579.

Huang, L., Zhang, H., Wu, P., Entwistle, S., Li, X., Yohe, T. et al. (2018) dbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation. *Nucleic Acids Research*, 46, D516–D521.

Kapustin, Y., Souvorov, A., Tatusova, T. & Lipman, D. (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. *Biology Direct*, 3, 20.

Kirsch, R., Gramzow, L., Theißen, G., Siegfried, B.D., ffrench-Constant, R.H., Heckel, D.G. et al. (2014) Horizontal gene transfer and functional diversification of plant cell wall degrading polygalacturonases: key events in the evolution of herbivory in beetles. *Insect Biochemistry and Molecular Biology*, 52, 33–50.

Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P.A. (2019) Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37, 540–546.

Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, 5, 59.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D. et al. (2009) Circos: an information aesthetic for comparative genomics. *Genome Research*, 19, 1639–1645.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. et al. (2004) Versatile and open software for comparing large genomes. *Genome Biology*, 5, R12.

Le Hir, H., Nott, A. & Moore, M.J. (2003) How introns influence and enhance eukaryotic gene expression. *Trends in Biochemical Sciences*, 28, 215–220.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34, 3094–3100.

Marvaldi, A.E., Duckett, C.N., Kjer, K.M. & Gillespie, J.J. (2009) Structural alignment of 18S and 28S rDNA sequences provides insights into phylogeny of Phytophaga (coleoptera: Curculionoidea and Chrysomeloidea). *Zoologica Scripta*, 38, 63–77.

McKenna, D.D., Scully, E.D., Pauchet, Y., Hoover, K., Kirsch, R., Geib, S.M. et al. (2016) Genome of the Asian longhorned beetle (*Anoplophora glabripennis*), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface. *Genome Biology*, 17, 227.

McKenna, D.D., Shin, S., Ahrens, D., Balke, M., Beza-Beza, C., Clarke, D.J. et al. (2019) The evolution and genomic basis of beetle diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 116, 24729–24737.

Monné, M.L., Monné, M.A. & Wang, Q. (2017) General morphology, classification, and biology of Cerambycidae. In: Wang, Q. (Ed.) *Cerambycidae of the world: biology and Pest management*. Boca Raton, FL: CRC Press.

Nachtweide, S. & Stanke, M. (2019) Multi-genome annotation with AUGUSTUS. *Methods in Molecular Biology*, 1962, 139–160.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32, 268–274.

Nie, R.E., Andújar, C., Gómez-Rodríguez, C., Bai, M., Xue, H.J., Tang, M. et al. (2020a) The phylogeny of leaf beetles (Chrysomelidae) inferred from mitochondrial genomes. *Systematic Entomology*, 45, 188–204.

Nie, R.E., Vogler, A.P., Yang, X.K. & Lin, M. (2020b) Higher-level phylogeny of longhorn beetles (coleoptera: Chrysomeloidea) inferred from mitochondrial genomes. *Systematic Entomology*, 46, 56–70.

Padilla-Hurtado, B., Flórez-Ramos, C., Aguilera-Gálvez, C., Medina-Olaya, J., Ramírez-Sanjuan, A., Rubio-Gómez, J. et al. (2012) Cloning and expression of an endo-1,4-beta-xylanase from the coffee berry borer, *Hypothenemus hampei*. *BMC Research Notes*, 5, 23.

Pauchet, Y., Kirsch, R., Giraud, S., Vogel, H. & Heckel, D.G. (2014) Identification and characterization of plant cell wall degrading enzymes from three glycoside hydrolase families in the cerambycid beetle *Apriona japonica*. *Insect Biochemistry and Molecular Biology*, 49, 1–13.

Petitpierre, E., Segarra, C. & Juan, C. (1993) Genome size and chromosomal evolution in leaf beetles (coleoptera, Chrysomelidae). *Hereditas*, 119, 1–6.

Quinlan, A.R. (2014) BEDTools: the swiss-Army tool for genome feature analysis. *Current Protocols in Bioinformatics*, 47 11 12 1-34.

Rang, F.J., Kloosterman, W.P. & de Ridder, J. (2018) From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, 19, 90.

Roach, M.J., Schmidt, S.A. & Borneman, A.R. (2018) Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, 19, 460.

Schell, T., Feldmeyer, B., Schmidt, H., Greshake, B., Tills, O., Truebano, M. et al. (2017) An annotated draft genome for *Radix auricularia* (Gastropoda, Mollusca). *Genome Biology and Evolution*, 9, 585–592.

Scully, E.D., Hoover, K., Carlson, J.E., Tien, M. & Geib, S.M. (2013) Midgut transcriptome profiling of *Anoplophora glabripennis*, a lignocellulose degrading cerambycid beetle. *BMC Genomics*, 14, 850.

Shah, A., Hoffman, J.I. & Schielzeth, H. (2020) Comparative analysis of genomic repeat content in Gomphocerine grasshoppers reveals expansion of satellite DNA and Helitrons in species with unusually large genomes. *Genome Biology and Evolution*, 12, 1180–1193.

Shen, W., Le, S., Li, Y. & Hu, F. (2016) SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One*, 11, e0163962.

Shin, N.R., Doucet, D. & Pauchet, Y. (2022) Duplication of horizontally acquired GH5_2 enzymes played a central role in the evolution of longhorned beetles. *Molecular Biology and Evolution*, 39, msac128.

Shin, N.R., Shin, S., Okamura, Y., Kirsch, R., Lombard, V., Svacha, P. et al. (2021) Larvae of longhorned beetles (coleoptera; Cerambycidae) have evolved a diverse and phylogenetically conserved array of plant cell wall degrading enzymes. *Systematic Entomology*, 46, 784–797.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31, 3210–3212.

So, W.L., Nong, W., Xie, Y., Baril, T., Ma, H.Y., Qu, Z. et al. (2022) Myriapod genomes reveal ancestral horizontal gene transfer and hormonal gene loss in millipedes. *Nature Communications*, 13, 3010.

Song, N., Yin, X., Zhao, X., Chen, J. & Yin, J. (2018) Reconstruction of mitogenomes by NGS and phylogenetic implications for leaf beetles. *Mitochondrial DNA Part A*, 29, 1041–1050.

Talla, V., Suh, A., Kalsoom, F., Dinca, V., Vila, R., Friberg, M. et al. (2017) Rapid increase in genome size as a consequence of transposable element hyperactivity in Wood-white (Leptidea) butterflies. *Genome Biology and Evolution*, 9, 2491–2505.

Tempel, S. (2012) Using and understanding RepeatMasker. *Methods in Molecular Biology*, 859, 29–51.

Van Dam, M.H., Cabras, A.A., Henderson, J.B., Rominger, A.J., Pérez Estrada, C., Omer, A.D. et al. (2021) The Easter egg weevil (*Pachyrhynchus*) genome reveals syntenic patterns in coleoptera across 200 million years of evolution. *PLoS Genetics*, 17, e1009745.

Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. (2017) Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Research*, 27, 737–746.

Vega, F.E., Brown, S.M., Chen, H., Shen, E., Nair, M.B., Ceja-Navarro, J.A. et al. (2015) Draft genome of the most devastating insect pest of coffee worldwide: the coffee berry borer, *Hypothenemus hampei*. *Scientific Reports*, 5, 12525.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G. et al. (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, 35, 543–548.

Wick, R.R., Judd, L.M. & Holt, K.E. (2019) Performance of neural network basecalling tools for Oxford nanopore sequencing. *Genome Biology*, 20, 129.

Wilson, J.J. (2012) DNA barcodes for insects. *Methods in Molecular Biology*, 858, 17–46.

Wood, D.E., Lu, J. & Langmead, B. (2019) Improved metagenomic analysis with kraken 2. *Genome Biology*, 20, 257.

Zhang, S.Q., Che, L.H., Li, Y., Liang, D., Pang, H., Ślipiński, A. et al. (2018) Evolutionary history of coleoptera revealed by extensive sampling of genes and species. *Nature Communications*, 9, 205.

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z. et al. (2018) dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, 46, W95–W101.

Zhao, T., Holmer, R., de Bruijn, S., Angenent, G.C., van den Burg, H.A. & Schranz, M.E. (2017) Phylogenomic synteny network analysis of MADS-box transcription factor genes reveals lineage-specific transpositions, ancient tandem duplications, and deep positional conservation. *Plant Cell*, 29, 1278–1292.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Data S1. Supporting Information.

How to cite this article: Shin, N.R., Okamura, Y., Kirsch, R. & Pauchet, Y. (2023) Genome sequencing provides insights into the evolution of gene families encoding plant cell wall-degrading enzymes in longhorned beetles. *Insect Molecular Biology*, 32(5), 469–483. Available from: https://doi.org/10.1111/imb.12844