communications biology

ARTICLE

https://doi.org/10.1038/s42003-021-02197-9 OPEN

Insights into the genomic evolution of insects from cricket genomes

Guillem Ylla[1✉], Taro Nakamura[1,9], Takehiko Itoh[2], Rei Kajitani[2], Atsushi Toyoda[3,4], Sayuri Tomonari[5], Tetsuya Bando[6], Yoshiyasu Ishimaru[5], Takahito Watanabe[5], Masao Fuketa[7], Yuji Matsuoka[5,10], Austen A. Barnett[1,11], Sumihare Noji[5], Taro Mito[5✉] & Cassandra G. Extavour[1,8✉]

Most of our knowledge of insect genomes comes from Holometabolous species, which undergo complete metamorphosis and have genomes typically under 2 Gb with little signs of DNA methylation. In contrast, Hemimetabolous insects undergo the presumed ancestral process of incomplete metamorphosis, and have larger genomes with high levels of DNA methylation. Hemimetabolous species from the Orthopteran order (grasshoppers and crickets) have some of the largest known insect genomes. What drives the evolution of these unusual insect genome sizes, remains unknown. Here we report the sequencing, assembly and annotation of the 1.66-Gb genome of the Mediterranean field cricket *Gryllus bimaculatus*, and the annotation of the 1.60-Gb genome of the Hawaiian cricket *Laupala kohalensis*. We compare these two cricket genomes with those of 14 additional insects and find evidence that hemimetabolous genomes expanded due to transposable element activity. Based on the ratio of observed to expected CpG sites, we find higher conservation and stronger purifying selection of methylated genes than non-methylated genes. Finally, our analysis suggests an expansion of the *pickpocket* class V gene family in crickets, which we speculate might play a role in the evolution of cricket courtship, including their characteristic chirping.

[1] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. [2] School of Life Science and Technology, Tokyo Institute of Technology, Tokyo, Japan. [3] Comparative Genomics Laboratory, National Institute of Genetics, Shizuoka, Japan. [4] Advanced Genomics Center, National Institute of Genetics, Shizuoka, Japan. [5] Department of Bioscience and Bioindustry, Tokushima University, Tokushima, Japan. [6] Graduate School of Medicine, Pharmacology and Dentistry, Okayama University, Okayama, Japan. [7] Graduate School of Advanced Technology and Science, Tokushima University, Tokushima, Japan. [8] Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA. [9] Present address: National Institute for Basic Biology, Okazaki, Japan. [10] Present address: Department of Biological Sciences, National University of, Singapore, Singapore. [11] Present address: Department of Natural Sciences, DeSales University, Center Valley, PA, USA. ✉email: guillemyllabou@gmail.com; mito.taro@tokushima-u.ac.jp; extavour@oeb.harvard.edu

Much of what we know about insect genome structure and evolution comes from examination of the genomes of insects belonging to a single clade, the Holometabola. This group includes species such as flies and beetles, and is characterized by undergoing complete, or holometabolous, metamorphosis. In these insects, the product of embryogenesis is a larva, which then undergoes an immobile stage called a pupa or chrysalis, during which the larval body plan is abandoned and the new, adult body plan is established. Following the pupal stage, the adult winged insect emerges[1]. This clade of insects includes nearly 90% of extant described insect species[2]. Members of this clade have become prominent model organisms for laboratory research, including the genetic model *Drosophila melanogaster*. Thus, a large proportion of our knowledge of insect biology, genetics, development, and evolution is based on studies of this clade.

Before the evolution of holometabolous metamorphosis, insects developed through incomplete or hemimetabolous metamorphosis. This mode of development is characterized by a generation of the final adult body plan during embryogenesis, followed by gradual physical growth of the hatchling through nymphal stages until the final transition to the sexually mature, winged adult, without major changes in body plan from hatchling to adult[1]. Many extant species maintain this presumed ancestral type of metamorphosis, including crickets, cockroaches, and aphids. Among hemimetabolous insects, most of our current genomic data is from the order Hemiptera (true bugs), which is the sister group to the Holometabola. For the remaining 15 hemimetabolous orders, genomic data remain scarce. For example, at the time of writing, out of these 15 orders, only nine of them contain a species with an available genome assembly in NCBI. Within these nine orders, there are only 39 species with available genomes, including the herein reported *Gryllus bimaculatus* genome. By contrast, there are 49 genomes of species with available genomes from the order Hemiptera alone, and 601 genomes of holometabolous species (Supplementary Table 1).

Based on data available to date, genome size and genome methylation show unexplained variation across insects. While most holometabolan species have relatively small genomes (0.2–1.5 pg), hemimetabolous species, and specifically polyneopterans (a taxon comprising ten major hemimetabolous orders of winged insects with fan-like extensions of the hind wings), display a much larger range of genome sizes (up to 8 pg)[3]. This has led to the hypothesis that there is a genome size threshold at 2 pg (~2 Gb) for holometabolan insect genomes[3]. Studying genome size evolution in the polyneopteran order Orthoptera (crickets, grasshoppers, locusts, and katydids) offers a valuable opportunity to investigate potential mechanisms of genome size evolution, as it includes species that have similar predicted gene counts, but have genomes ranging from 1.25 to 16.56 Gb[4]. With respect to the level of CpG DNA methylation, only a few holometabolous species display evidence of genome-wide DNA methylation at CpG sites, whereas 30 out of 34 analyzed polyneopteran species do[5,6]. However, the role of DNA methylation in polyneopteran species, and why it appears to have been lost in many holometabolans, is not clear.

Here, we present the 1.66-Gb genome assembly and annotation of *G. bimaculatus* (Orthoptera), commonly known as the two-spotted cricket, a name derived from the two yellow spots found on the base of the forewings of this species (Fig. 1a). We also report the first genome annotation for a second cricket species, the Hawaiian cricket *Laupala kohalensis*, whose genome assembly was recently made public[7]. *G. bimaculatus* has been widely used

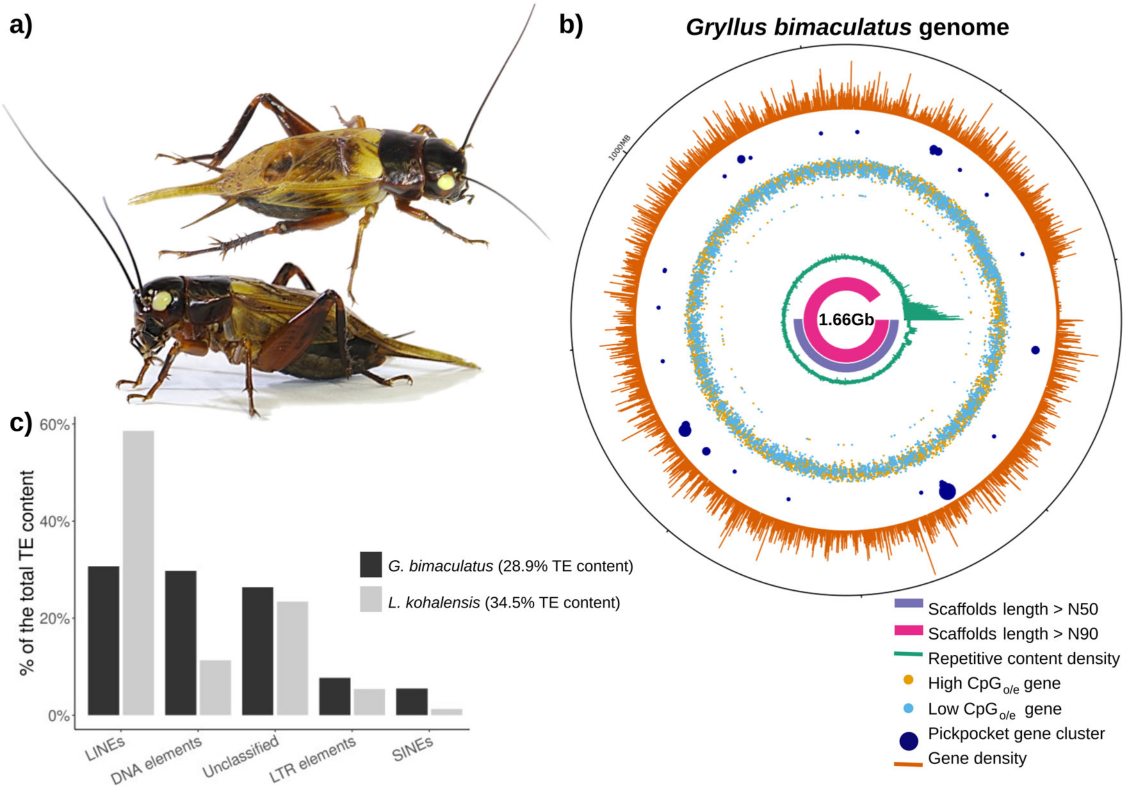

Fig. 1 The *G. bimaculatus* genome. a The cricket *G. bimaculatus* (top and side views of an adult male), commonly called the two-spotted cricket, owes its name to the two yellow spots on the base of the forewings. **b** Circular representation of the *G. bimaculatus* genome, displaying the N50 (pink) and N90 (purple) scaffolds, repetitive content density (green), the high- (yellow) and low- (light blue) CpG$_{o/e}$ value genes, *pickpocket* gene clusters (dark blue), and gene density (orange). **c** The proportion of the genome made up of transposable elements (TEs) is similar between *G. bimaculatus* and *L. kohalensis* (28.9 and 34.5%, respectively), but the specific TE family composition varies widely between the two species.

as a laboratory research model for decades, in scientific fields including neurobiology and neuroethology[8,9], evo-devo[10], developmental biology[11], and regeneration[12]. Technical advantages of this cricket species as a research model include the fact that *G. bimaculatus* does not require cold temperatures or diapause to complete its life cycle, it is easy to rear in laboratories since it can be fed with generic insect or other pet foods, it is amenable to RNA interference (RNAi) and targeted genome editing[13], stable germline transgenic lines can be established[14], and it has an extensive list of available experimental protocols ranging from behavioral to functional genetic analyses[15].

Comparing the two cricket genomes annotated here with those of 14 other insect species allowed us to identify three interesting features of these cricket genomes, some of which may relate to their unique biology. First, the differential transposable element (TE) composition between the two cricket species suggests abundant TE activity since they diverged from a last common ancestor, which our results suggest occurred circa 89.2 million years ago (Mya). Second, based on gene CpG depletion, an indirect but robust method to identify typically methylated genes[5,16], we find higher conservation of typically methylated genes than of non-methylated genes. Finally, our gene family expansion analysis reveals an expansion of the *pickpocket* class V gene family in the lineage leading to crickets, which we speculate might play a relevant role in cricket courtship behavior, including their characteristic chirping.

Results

***Gryllus bimaculatus* genome assembly**. We sequenced, assembled, and annotated the 1.66 Gb haploid genome of the white-eyed mutant strain[12] of the cricket *G. bimaculatus* (Fig. 1a). The 1.66 Gb size predicted by the assembly is similar in size to a previous estimation of 1.68 Gb, obtained from a k-mer analysis performed on an independent dataset (Supplementary Note 1). 50% of the genome is contained within the 71 longest scaffolds (L50), the shortest of them having a length of 6.3 Mb (N50), and 90% of the genome is contained within 307 scaffolds (L90). In comparison to other polyneopteran genomes, our assembly displays high quality in terms of contiguity (N50 and L50), and completeness (BUSCO scores) (Supplementary Data 1). Notably, the percentage of complete BUSCO genes[17] of this genome assembly at the arthropod and insect levels are 98.50% (C:98.5% [S:97.2%, D:1.3%], F:0.4%, M:1.1%, $n = 1066$) and 97.00% (C:97.0% [S:95.2%, D:1.8%], F:0.8%, M:2.2%, $n = 1658$), respectively, indicating high completeness of this genome assembly (Table 1). The low percentage of duplicated BUSCO genes (1.31–1.81%) suggests that putative artifactual genomic duplication due to mis-assembly of heterozygotic regions is unlikely.

Table 1 *Gryllus bimaculatus* genome assembly statistics.

Number of scaffolds	47,877
Genome length (nt)	1,658,007,496
Genome length (Gb)	1.66
Avg. scaffold size (Kb)	34.63
N50 (Mb)	6.29
N90 (Mb)	1.04
L50	71
L90	307
Complete BUSCO Score - Arthropoda	98.50%
Complete BUSCO Score - Insecta	97.00%

Table 2 Genome annotation summary for the crickets *G. bimaculatus* and *L. kohalensis*.

	G. bimaculatus	***L. kohalensis***
Annotated protein-coding genes	17,871	12,767
Annotated transcripts	28,529	13,078
% With InterPro ID	59.56%	72.52%
% With GO-terms	38.66%	47.03%
% With PFAM motif	62.44%	76.59%
% With significant BLASTP hit	73.64%	93.23%
Complete BUSCO-proteome Score - Insecta	90.50%	87.20%
Repetitive content	33.69%	35.51%
TE content	28.94%	34.50%
GC level	39.93%	35.58%

Annotation of two cricket genomes. The publicly available 1.6-Gb genome assembly of the Hawaiian cricket *L. kohalensis*[7], although having lower assembly quality scores (N50 = 0.58 Mb, L90 = 3483) than that of *G. bimaculatus*, scores high in terms of completeness, with complete BUSCO scores of 99.3% at the arthropod level and 97.80% at the insect level (Supplementary Data 1).

Using three iterations of the MAKER2 pipeline[18], in which we combined ab initio and evidence-based gene models, we annotated the protein-coding genes in both cricket genomes (Supplementary Figs. 1 and 2). We identified 17,871 coding genes and 28,529 predicted transcripts for *G. bimaculatus*, and 12,767 coding genes and 13,078 transcripts for *L. kohalensis* (Table 2).

To obtain functional insights into the annotated genes, we ran InterProScan[19] for all predicted protein sequences and retrieved their InterPro ID, PFAM domains, and Gene Ontology (GO) terms (Table 2). In addition, we retrieved the best significant BLASTP hit (E-value <1e-6) for 70–90% of the proteins. Taken together, these methods predicted functions for 75 and 94% of the proteins annotated for *G. bimaculatus* and *L. kohalensis*, respectively. We created a novel graphical interface through which interested readers can access, search, BLAST, and download the genome data and annotations (http://gbimaculatusgenome.rc.fas.harvard.edu).

Abundant repetitive DNA. We used RepeatMasker[20] to determine the degree of repetitive content in the cricket genomes, using specific custom repeat libraries for each species. This approach identified 33.69% of the *G. bimaculatus* genome, and 35.51% of the *L. kohalensis* genome, as repetitive content (Supplementary Tables 2 and 3). In *G. bimaculatus* the repetitive content density was similar throughout the genome, with the exception of scaffolds shorter than 1 Mb (L90), which make up 10% of the genome and have a high density of repetitive content and low gene density (Fig. 1b). Because the repetitive content makes genome assemblies more challenging, as observed for the shortest scaffolds of *G. bimaculatus*, we cannot rule out the possibility that the lower contiguity of the *L. kohalensis* genome could lead us to underestimate its repetitive content. This caveat notwithstanding, we observed that TEs accounted for 28.94% of the *G. bimaculatus* genome, and for 34.50% of the *L. kohalensis* genome. Although the overall proportion of genome made up of TEs was similar between the two cricket species, the proportion of each specific TE class varied greatly (Fig. 1c). In *L. kohalensis* the most abundant TE type was long interspersed elements (LINEs), accounting for 20.21% of the genome and 58.58% of the total TE content, while in *G. bimaculatus* LINEs made up only 8.88% of the genome and 30.68% of the total TE content. The specific LINE subtypes LINE1 and LINE3 appeared at a similar frequency in both cricket genomes (<0.5%), while the LINE2 subtype was over five times more represented in *L. kohalensis*, covering 10% of the genome (167 Mb). On the other hand, DNA transposons accounted for 8.61% of the *G. bimaculatus* genome, but only for 3.91% of the *L. kohalensis* genome.

DNA methylation. CpG depletion, calculated as the ratio between observed versus the expected incidence of a cytosine followed by a guanine ($CpG_{o/e}$), is considered a reliable indicator of DNA methylation. This is because spontaneous C to T mutations occur more frequently on methylated CpGs than unmethylated CpGs[16]. Thus, genomic regions that undergo methylation are eventually CpG-depleted. We calculated the $CpG_{o/e}$ value for each predicted protein-coding gene for the two cricket species. In both species, we observed a clear bimodal distribution of $CpG_{o/e}$ values (Fig. 2a). One interpretation of this distribution is that the peak corresponding to lower $CpG_{o/e}$ values contains genes that are typically methylated, and the peak of higher $CpG_{o/e}$ contains genes that do not undergo DNA methylation. Under this interpretation, some genes have non-random differential DNA methylation in crickets. To quantify the genes in the two putative methylation categories, we set a $CpG_{o/e}$ threshold as the value of the point of intersection between the two normal distributions (Fig. 2a). After applying this cutoff, 44% of *G. bimaculatus* genes and 45% of *L. kohalensis* genes were identified as CpG-depleted.

A GO enrichment analysis of the genes above and below the $CpG_{o/e}$ threshold defined above revealed clear differences in the predicted functions of genes belonging to each of the two categories. Strikingly, however, genes in each threshold category had functional similarities across the two cricket species (Fig. 3). Genes with low $CpG_{o/e}$ values, which are likely those undergoing methylation, were enriched for functions related to DNA replication and regulation of gene expression (including transcriptional, translational, and epigenetic regulation), while genes with high $CpG_{o/e}$ values, suggesting little or no methylation, tended to have functions related to metabolism, catabolism, and sensory systems.

To assess whether the predicted distinct functions of high- and low- $CpG_{o/e}$ value genes were specific to crickets, or were a potentially more general trend of insects with DNA methylation systems, we analyzed the predicted functions of genes with different $CpG_{o/e}$ values in the honeybee *Apis mellifera*, the first insect for which evidence for DNA methylation was robustly described and studied[21,22], and the thrips *Frankliniella occidentalis*. We chose these insects because they are relatively distant

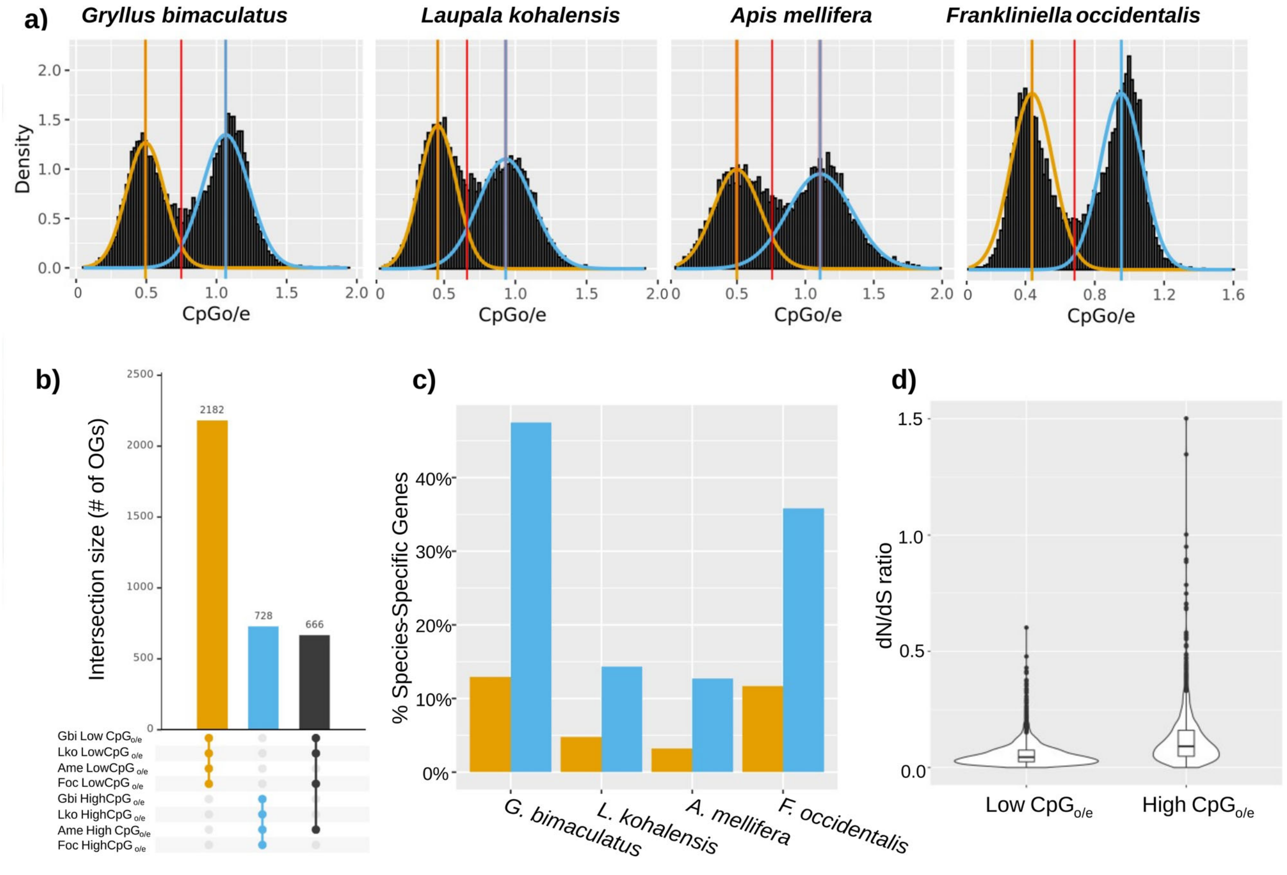

Fig. 2 $CpG_{o/e}$ bimodal distribution across distant insects. a The distribution of $CpG_{o/e}$ values within the CDS regions displays a bimodal distribution in the two cricket species studied here, as well as in the honeybee *A. mellifera* and the thrips *F. occidentalis*. We modeled each peak with a normal distribution and defined their intersection (red line) as a threshold to separate genes into low- and high- $CpG_{o/e}$ value categories, represented in yellow and blue respectively. **b** UpSet plot showing the top three intersections (linked dots) in terms of the number of orthogroups (OGs) commonly present in the same category (low- and high- CpG $_{o/e}$) across the four insect species. The largest intersection corresponds to 2182 OGs whose genes have low $CpG_{o/e}$ in the four insect species, followed by the 728 OGs whose genes have high $CpG_{o/e}$ levels in all four species, and 666 OGs whose genes have low $CpG_{o/e}$ in the three hemimetabolous species and high $CpG_{o/e}$ in the holometabolous honeybee. Extended plot with 50 intersections is shown in Supplementary Fig. 4. **c** Percentage of species-specific genes within low $CpG_{o/e}$ (yellow) and high $CpG_{o/e}$.(blue) categories in each insect, indicating that more such genes tend to have high $CpG_{o/e}$ values. **d** One-to-one orthologous genes with low $CpG_{o/e}$ values in both crickets have significantly lower dN/dS values than genes with high $CpG_{o/e}$ values.

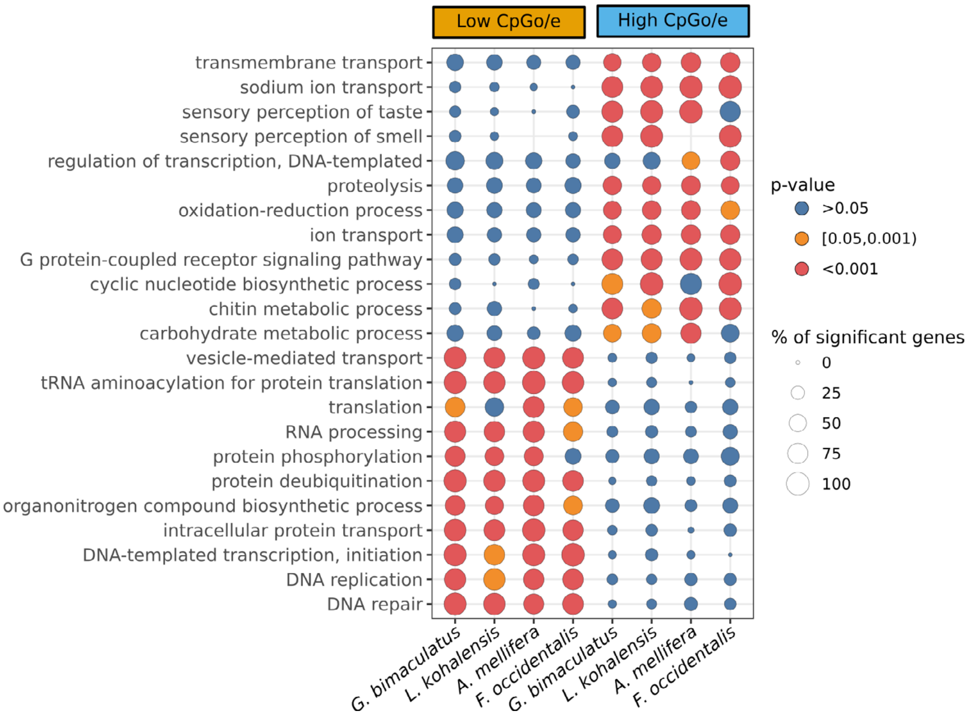

Fig. 3 Functional differences between high- and low- $\mathrm{CpG_{o/e}}$ genes. Enriched GO terms with a *p* value <0.00001 in at least one of the eight categories, which are high $\mathrm{CpG_{o/e}}$ and low $\mathrm{CpG_{o/e}}$ genes of *G. bimaculatus*, *L. kohalensis*, *F. occidentalis*, and *A. mellifera*, respectively. The dot diameter is proportional to the percentage of significant genes with that GO term within the gene set. The dot color represents the *p* value level: blue >0.05, orange [0.05, 0.001), red <0.001. Extended figure with all significant GO terms (*p* value <0.05) available as Supplementary Fig. 3.

relatives of crickets, sharing a last common ancestor with crickets circa 390 Mya[23]. Thrips are hemimetabolous, while bees are holometabolous, and both show a clear $\mathrm{CpG_{o/e}}$ bimodal distribution. We found that in both *F. occidentalis* and *A. mellifera*, CpG-depleted genes were enriched for similar functions as those observed in cricket CpG-depleted genes (Fig. 3 and Supplementary Fig. 3). Specifically, 23 GO terms were significantly enriched in all four studied insects, and 15 additional GO terms were significantly enriched in the three hemimetabolous insects. In the same way, high $\mathrm{CpG_{o/e}}$ genes in all four insects were enriched for similar functions (8 GO-terms commonly enriched in all insects; Supplementary Fig. 3).

Additionally, we observed that the proportion of species-specific genes was higher within the high $\mathrm{CpG_{o/e}}$ peak for all four insects (Fig. 2c). In contrast, 86–96% of the genes belonging to the low $\mathrm{CpG_{o/e}}$ peak had an orthologous gene in at least one of the other studied insect species. Furthermore, we observed 2182 orthogroups (OGs) whose members always belonged to the low $\mathrm{CpG_{o/e}}$ peak in all four species, and 728 OGs whose members always belonged to the high $\mathrm{CpG_{o/e}}$ peak, indicating that orthologous genes are likely to share methylation state across these four insect species (Fig. 2b and Supplementary Fig. 4). Interestingly, 666 genes belonged to the low $\mathrm{CpG_{o/e}}$ peak in the three hemimetabolous species (*G. bimaculatus, L. kohalensis*, and *F. occidentallis)*, but to the high $\mathrm{CpG_{o/e}}$ peak in the holometabolous *A. mellifera*.

Taken together, these results suggest that genes that are typically methylated tend to be more conserved across species, which could imply low evolutionary rates and strong selective pressure. To test this hypothesized relationship between low $\mathrm{CpG_{o/e}}$ values and low evolutionary rates, we compared the dN/dS values of 1-to-1 orthologous genes belonging to the same $\mathrm{CpG_{o/e}}$ peak between the two cricket species. We found that CpG-depleted genes in both crickets had significantly lower dN/dS values than non-CpG-depleted genes (*p* value <0.05; Fig. 2d), consistent with stronger purifying selection on CpG-depleted genes.

Phylogenetics and gene family expansions. To study the genome evolution of these cricket lineages, we compared the two cricket genomes with those of 14 additional insects, including members of all major insect lineages with special emphasis on hemimetabolous species. For each of these 16 insect genomes, we retrieved the longest protein per gene and grouped them into OGs, which we called "gene families" for the purpose of this analysis. The 732 OGs containing a single protein per insect, namely single copy orthologs, were used to infer a phylogenetic tree for these 16 species (Fig. 4). The obtained species tree topology was in accordance with the currently understood insect phylogeny[23]. Then, we used the Misof et al. (2014)[23] dated phylogeny to calibrate our tree on four different nodes, which allowed us to estimate that the two cricket species diverged circa 89.2 million years ago.

Our gene family expansion/contraction analysis using 59,516 OGs identified 18 gene families that were significantly expanded (*p* value <0.01) in the lineage leading to crickets. In addition, we identified a further 34 and 33 gene family expansions specific to *G. bimaculatus* and *L. kohalensis*, respectively. Functional analysis of these expanded gene families (Supplementary Data 2) revealed that the cricket-specific gene family expansions included *pickpocket* genes, which are involved in mechanosensation in *D. melanogaster* as described in the following section.

Expansion of *pickpocket* genes. In *D. melanogaster*, the complete *pickpocket* gene repertoire is composed of 6 classes containing 31 genes. We found cricket orthologs of all 31 *pickpocket* genes across seven of our OGs, and each OG predominantly contained

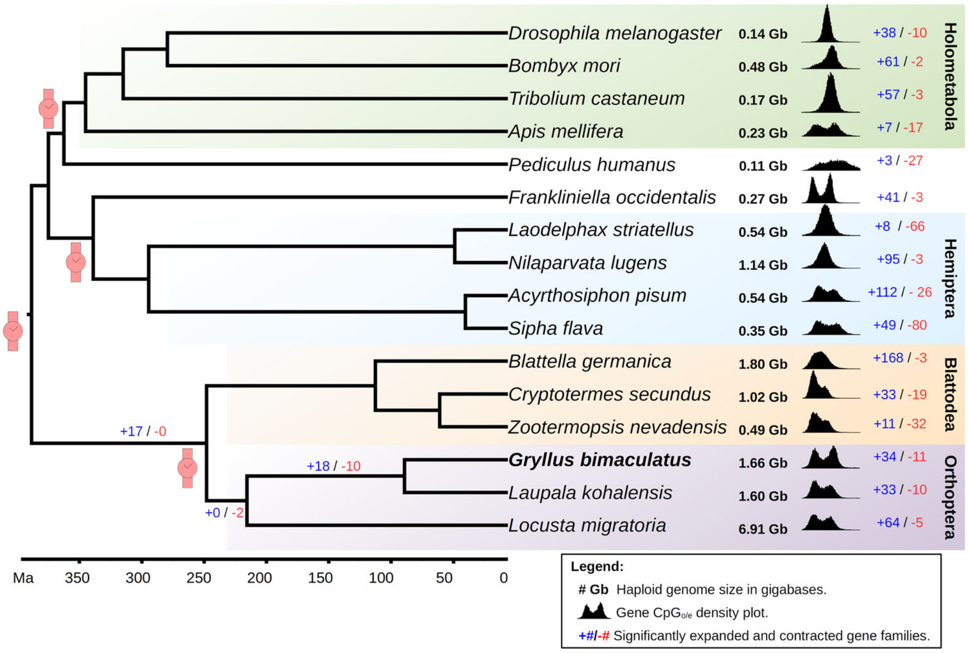
Drosophila melanogaster 0.14 Gb +38 / -10
Bombyx mori 0.48 Gb +61 / -2
Tribolium castaneum 0.17 Gb +57 / -3
Apis mellifera 0.23 Gb +7 / -17
Holometabola
Pediculus humanus 0.11 Gb +3 / -27
Frankliniella occidentalis 0.27 Gb +41 / -3
Laodelphax striatellus 0.54 Gb +8 / -66
Nilaparvata lugens 1.14 Gb +95 / -3
Acyrthosiphon pisum 0.54 Gb +112 / - 26
Sipha flava 0.35 Gb +49 / -80
Hemiptera
Blattella germanica 1.80 Gb +168 / -3
Cryptotermes secundus 1.02 Gb +33 / -19
Zootermopsis nevadensis 0.49 Gb +11 / -32
Blattodea
Gryllus bimaculatus 1.66 Gb +34 / -11
Laupala kohalensis 1.60 Gb +33 / -10
Locusta migratoria 6.91 Gb +64 / -5
Orthoptera
+17 / -0
+18 / -10
+0 / -2
Ma 350 300 250 200 150 100 50 0
Legend:
# Gb Haploid genome size in gigabases.
Gene CpG$_{o/e}$ density plot.
+#/-# Significantly expanded and contracted gene families.

Fig. 4 Cricket genomes in the context of insect evolution. A phylogenetic tree including 16 insect species calibrated at four different time points (red watch symbols) based on Misof et al. (2014)[23], suggests that *G. bimaculatus* and *L. kohalensis* diverged ca. 89.2 Mya. The number of expanded (blue text) and contracted (red text) gene families is shown for each insect, and for the branches leading to crickets. The density plots show the CpG$_{o/e}$ distribution for all genes for each species. The genome size in Gb was obtained from the genome fasta files (Supplementary Data 1).

members of a single *pickpocket* class. We used all the genes belonging to these 7 OGs to build a *pickpocket* gene tree, using the predicted *pickpocket* orthologs from 16 insect species (Fig. 5; Supplementary Table 4). This gene tree allowed us to classify the different *pickpocket* genes in each of the 16 species.

One OG, which contained eight members of the *pickpocket* gene family of *D. melanogaster*, appeared to be significantly expanded to 14 or 15 members in crickets. Following the classification of *pickpocket* genes used in *Drosophila spp.*[24] we determined that the specific gene family expanded in crickets was *pickpocket* class V (Fig. 5). In *D. melanogaster* this class contains eight genes: *ppk* (*ppk1*), *rpk* (*ppk2*), *ppk5*, *ppk8*, *ppk12*, *ppk17*, *ppk26*, and *ppk28*[24]. Our analysis suggests that the class V gene family contains 15 and 14 genes in *G bimaculatus* and *L. kohalensis*, respectively. In contrast, their closest analyzed relative, the locust *Locusta migratoria*, has only five such genes.

The *pickpocket* genes in crickets tended to be grouped in genomic clusters (Fig. 1b). For instance, in *G. bimaculatus* nine of the 15 class V *pickpocket* genes were clustered within a region of 900 Kb, and four other genes appeared in two groups of two. In the *L. kohalensis* genome, although this genome is more fragmented than that of *G. bimaculatus* (Supplementary Data 1), we observed five clusters containing between two and five genes each.

In *D. melanogaster*, the *pickpocket* gene *ppk1* belongs to class V and is involved in functions related to stimulus perception and mechanotransduction[25]. For example, in larvae, this gene is required for mechanical nociception[26], and for coordinating rhythmic locomotion[27]. *ppk* is expressed in sensory neurons that also express the male sexual behavior determiner *fruitless (fru)*[28–30].

To determine whether *pickpocket* genes in crickets are also expressed in the nervous system, we checked for evidence of expression of *pickpocket* genes in the publicly available RNA-seq libraries for the *G. bimaculatus* prothoracic ganglion[9]. This analysis detected expression (>20 transcripts per kilobase million, TPMs) of five *pickpocket* genes, four of them belonging to class V, in the *G. bimaculatus* nervous system. In the same ganglionic RNA-seq libraries, we also detected the expression of *fru* (Supplementary Data 3). Out of the four *pickpocket* genes, only one was detected in embryonic RNA-seq libraries. All four genes together with *fru* were detected in wild type leg transcriptomes, and their expression was found to be higher than wild type in a transcriptome from regenerating legs (Supplementary Data 4).

Discussion

Sequencing and analyzing genomes from underrepresented clades allows us to get a more complete picture of genome diversity across the tree of life, and can provide insights regarding their evolution. Since the first sequenced insect genome, that of *D. melanogaster*, was made publicly available in 2000[31], the field of holometabolous genomics has flourished, and this clade became the main source of subsequent genomic information for insects. The first hemimetabolous genome was not available until 10 years later, with the publication of the genome sequence and annotation of the Pea aphid (*Acyrthosiphon pisum*)[32]. When even more recently, polyneopteran genome sequences became available[33–36], some of their distinct characteristics, such as their length and DNA methylation profiles, began to be appreciated. Genome data are also very important as they can help establish species as tractable experimental models. *G. bimaculatus* is a common laboratory research animal used in neuroethology, developmental and regeneration biology studies[12,15]. It is our hope that the availability of the annotated genome presented here will encourage other researchers to adopt this cricket as a model organism, and facilitate development of new molecular genetic manipulation tools.

Moreover, we note that crickets are currently in focus as a source of animal protein for human consumption and for vertebrate livestock. Crickets possess high nutritional value, having a high proportion of protein for their body weight (>55%), and

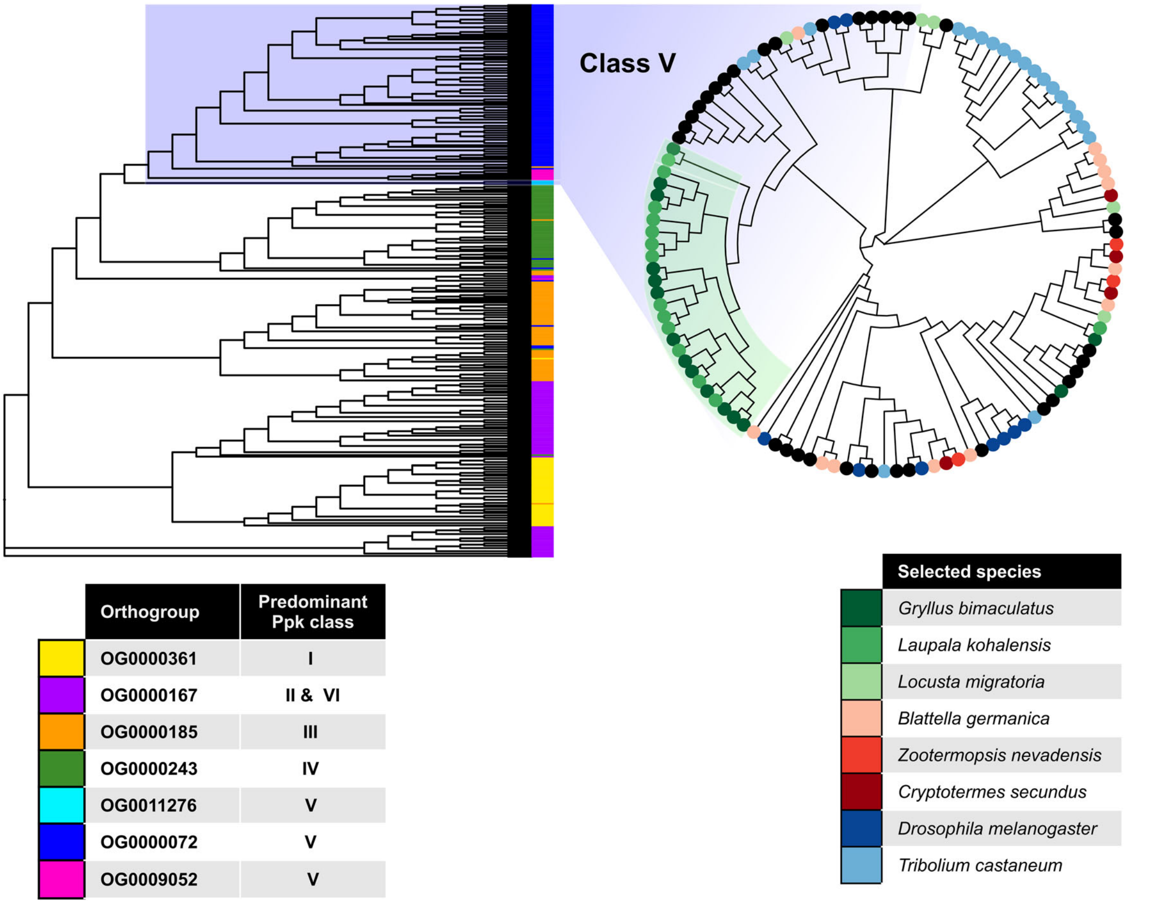

Fig. 5 The *pickpocket* gene family class V is expanded in crickets. *pickpocket* gene tree with all the genes belonging to the seven OGs that contain the *D. melanogaster pickpocket* genes. All OGs predominantly contain members of a single *ppk* family. The OG0000167 orthogroup contains members of two *pickpocket* classes, II and VI. The orthogroup OG0000072 containing most *pickpocket* class V genes (circular cladogram) was significantly expanded in crickets relative to other insects.

containing the essential linoleic acid as their most predominant fatty acid[37–39]. Specifically, the cricket *G. bimaculatus* has traditionally been consumed in different parts of the world including northeast Thailand, which recorded 20,000 insect farmers in 2011[40]. Studies have reported no evidence for toxicological effects related to oral consumption of *G. bimaculatus* by humans[41,42], neither were genotoxic effects detected using three different mutagenicity tests[43]. A rare but known health risk associated with cricket consumption, however, is sensitivity and allergy to crickets[44,45]. Nevertheless, not only is the cricket *G. bimaculatus* considered generally safe for human consumption, several studies also suggest that introducing crickets into one's diet may confer multiple health benefits[46–48]. Crickets might therefore be part of the solution to the problem of feeding a worldwide growing population in a sustainable way. However, most of the crops and livestock that humans eat have been domesticated and subjected to strong artificial selection for hundreds or even thousands of years to improve their characteristics most desirable for humans, including size, growth rate, stress resistance, and organoleptic properties[49–52]. In contrast, to our knowledge, crickets have never been selected based on any food-related characteristic. The advent of genetic engineering techniques has accelerated domestication of some organisms[53]. These techniques have been used, for instance, to improve the nutritional value of different crops, or to make them tolerant to pests and climate stress[49,54]. Crickets are naturally nutritionally rich[39], but in principle, their nutritional value could be further improved, for example by increasing vitamin content or Omega-3 fatty acids proportion. In addition, other issues that present challenges to cricket farming could potentially be addressed by targeted genome modification, which can be achieved in *G. bimaculatus* using Zinc finger nucleases, TALENs, or CRISPR/Cas9. These challenges include sensitivity to common insect viruses, aggressive behavior resulting in cannibalism, complex mating rituals, and relatively slow growth rate.

The annotation of these two cricket genomes was done by combining de novo gene models, homology-based methods, and the available RNA-seq and ESTs. This pipeline allowed us to predict 17,871 genes in the *G. bimaculatus* genome, similar to the number of genes reported for other hemimetabolous insect genomes including the locust *L. migratoria* (17,307)[33] and the termites *Cryptotermes secundus* (18,162)[34], *Macrotermes natalensis* (16,140)[36], and *Zootermopsis nevadensis*, (15,459)[35]. We speculate that the slightly lower number of protein-coding genes annotated in *L. kohalensis* (12,767) may be due to the lesser amount of RNA-seq data available for this species and the higher fragmentation of its assembly, which challenges gene annotation. Nevertheless, the BUSCO scores are similar between the two crickets, and the proportion of annotated proteins with putative orthologous genes in other species (proteins with significant BLAST hits; see Methods) for *L. kohalensis* is higher than for *G. bimaculatus*. This suggests the possibility that we may have successfully annotated most conserved genes, but that highly derived or species-specific genes might be missing from our annotations.

Approximately 35% of the genome of both crickets corresponds to repetitive content. This is substantially less than the 60% reported for the genome of *L. migratoria*[33]. This locust genome is one of the largest sequenced insect genomes to date (6.5 Gb) but has a very similar number of annotated genes (17,307) to those we report for crickets. We hypothesize that the large genome size difference between these orthopteran species is due to the TE content, which has also been correlated with genome size in multiple eukaryote species[55,56].

Furthermore, we hypothesize that the differences in the TE composition between the two crickets are the result of abundant and independent TE activity since their divergence around 89.2 Mya. This, together with the absence of evidence for large genome duplication events in this lineage, leads us to hypothesize that the ancestral orthopteran genome was shorter than those of the crickets studied here (1.6 Gb for *G. bimaculatus* and 1.59 Gb for *L. kohalensis*), which are in the lowest range of orthopteran genome sizes[57]. In summary, we propose that the wide range of genome sizes within Orthoptera, reaching as high as 8.55 Gb in the locust *Schistocerca gregaria*, and 16.56 Gb in the grasshopper *Podisma pedestris*[4,58], is likely due to TE activity since the time of the last orthopteran ancestor. These observations are consistent with the results reported by Palacios-Gimenez et al. (2020)[59] of massive and independent recent TE accumulation in four chromosome races of the grasshopper *Vandiemenella viatica*.

There is a clear tendency of polyneopteran genomes to be much longer than those of the holometabolous genomes (Fig. 4). Two currently competing hypotheses are that (1) the ancestral insect genome was small, and was expanded outside of Holometabola, and (2) the ancestral insect genome was large, and it was compressed in the Holometabola[3]. Our observations are consistent with the first of these hypotheses.

Most holometabolan species, including well-studied insects like *D. melanogaster* and *Tribolium castaneum*, do not perform DNA methylation, or they do it at very low levels[6,60]. The honeybee *A. mellifera* was one of the first insects for which functional DNA methylation was described[21]. Although this DNA modification was initially proposed to be associated with the eusociality of these bees[22], subsequent studies showed that DNA methylation is widespread and present in different insect lineages independently of social behavior[5]. DNA methylation also occurs in other non-insect arthropods[61].

While the precise role of DNA methylation in gene expression regulation remains unclear, our analysis suggests that cricket CpG-depleted genes (putatively hypermethylated genes) show signs of purifying selection, tend to have orthologs in other insects, and are involved in basic biological functions related to DNA replication and the regulation of gene expression. These enriched functions are in agreement with previous observations that DNA methylated genes in arthropods tend to perform housekeeping functions[6,62]. These predicted functions differ from those of the non-CpG-depleted genes (putatively hypomethylated genes), which appear to be involved in signaling pathways, metabolism, and catabolism. These predicted functional categories may be conserved from crickets over circa 345 million years of evolution, as we also detect the same pattern in the honeybee and a thrips species.

Taken together, these observations suggest a potential relationship between DNA methylation, sequence conservation, and function for many cricket genes. Nevertheless, based on our data, we cannot determine whether the methylated genes are highly conserved because they are methylated, or because they perform basic functions that may be regulated by DNA methylation events. In the cockroach *Blattella germanica*, DNA methyltransferase enzymes and genes with low $CpG_{o/e}$ values show an expression peak during the maternal to zygotic transition[63], and functional analysis has shown that the DNA methyltransferase 1 is essential for early embryo development in this cockroach[64]. These results in cockroaches, together with our observations, lead us to speculate that at least in Polyneopteran species, DNA methylation might contribute to the maternal zygotic transition by regulating essential genes involved in DNA replication, transcription, and translation.

The *pickpocket* genes belong to the Degenerin/epithelial Na^+ channel (DEG/ENaC) family, which were first identified in *Caenorhabditis elegans* as involved in mechanotransduction[25]. The same family of ion channels was later found in many multicellular animals, with a diverse range of functions related to mechanoreception and fluid–electrolyte homeostasis[65]. Most of the information on their roles in insects comes from studies in *D. melanogaster*. In this fruit fly, *pickpocket* genes are involved in neural functions including NaCl taste[66], pheromone detection[67], courtship behavior[68], and liquid clearance in the larval trachea[65].

In *D. melanogaster* adults, the abdominal ganglia mediate courtship and postmating behaviors through neurons expressing *ppk* and *fru*[28–30]. In *D. melanogaster* larvae, *ppk* expression in dendritic neurons is required to control the coordination of rhythmic locomotion[27]. In crickets, the abdominal ganglia are responsible for determining song rhythm[69]. Moreover, we find that in *G. bimaculatus*, both *ppk* and *fru* gene expression are detectable in the adult prothoracic ganglion. These observations suggest the possibility that class V *pickpocket* genes could be involved in song rhythm determination in crickets through their expression in abdominal ganglia.

This possibility is consistent with the results of multiple quantitative trait locus (QTL) studies done in cricket species from the genus *Laupala*, which identified genomic regions associated with mating song rhythm variations and female acoustic preference[70]. The 179 scaffolds that the authors reported being within one logarithm of the odds (LOD) of the seven QTL peaks, contained five *pickpocket* genes, three of them from class V and two from class IV. One of the two class IV genes also appears within a QTL peak of a second experiment[7,71]. Xu and Shaw (2019)[72] found that a scaffold in a region of LOD score 1.5 of one of their minor linkage groups (LG3) contains *slowpoke*, a gene that affects song interpulse interval in *D. melanogaster*, and this scaffold also contains two class III pickpocket genes (Supplementary Table 5).

In summary, the roles of *pickpocket* genes in controlling rhythmic locomotion, courtship behavior, and pheromone detection in *D. melanogaster*, their appearance in genomic regions associated with song rhythm variation in *Laupala*, and their expression in *G. bimaculatus* abdominal ganglia, lead us to speculate that the expanded *pickpocket* gene family in cricket genomes could be playing a role in regulating rhythmic wing movements and sound perception, both of which are necessary for mating[15]. We note that Xu and Shaw (2019)[72] hypothesized that song production in crickets is likely to be regulated by ion channels, and that locomotion, neural modulation, and muscle development are all involved in singing[72]. However, further experiments, which could take advantage of the existing RNAi and genome modification protocols for *G. bimaculatus*[13], will be required to test this hypothesis.

In conclusion, the *G. bimaculatus* genome assembly and annotation presented here is a source of information and an essential tool that we anticipate will enhance the status of this cricket as a modern functional genetics research model. This genome may also prove useful to the agricultural sector, and could allow improvement of cricket nutritional value, productivity, and reduction of allergen content. Annotating a second cricket genome, that of *L. kohalensis*, and comparing the two genomes, allowed us to unveil possible synapomorphies of cricket genomes, and to suggest potentially general evolutionary trends of insect genomes.

Materials and Methods

DNA isolation. The *G. bimaculatus* white-eyed mutant strain was reared at Tokushima University, at 29 ± 1 °C and 30–50% humidity under a 10-h light, 14-h dark photoperiod. Testes of a single male adult of the *G. bimaculatus* white-eyed mutant strain were used for DNA isolation and short-read sequencing. We used DNA from testes of an additional single individual to make a long-read PacBio sequencing library to close gaps in the genome assembly. Because sex differentiation in the cricket *G. bimaculatus* is determined by the XX/XO system[73], genomic DNA extracted from males contains the full set of chromosomes. Besides, testes contain a large number of nuclei and are easily isolated for DNA extraction. Male testes were therefore chosen for genomic DNA isolation.

Genome assembly. Paired-end libraries with an average read length of 100 bp were generated with insert sizes of 375 and 500 bp, and mate-pair libraries were generated with insert sizes of 3, 5, 10, and 20 kb. Libraries were sequenced using the Illumina HiSeq 2000 and HiSeq 2500 sequencing platforms. This yielded a total of 127.4 Gb of short read paired-end data, that was subsequently assembled using the de novo assembler Platanus (v. 1.2.1)[74]. Scaffolding and gap closing were also performed with Platanus using total 138.2 Gb of mate-pair data. A further gap closing step was performed using long reads generated by the PacBio RS system. The 4.3 Gb of PacBio subread data were used to fill gaps in the assembly using PBjelly (v. 15.8.24)[75].

Repetitive content masking. We generated a custom repeat library for each of the two cricket genomes by combining the outputs from homology-based and de novo repeat identifiers, including the LTRdigest together with LTRharvest[76], RepeatModeler/RepeatClassifier (www.repeatmasker.org/RepeatModeler), MITE tracker[77], TransposonPSI (http://transposonpsi.sourceforge.net), and the databases SINEBase[78] and RepBase[79]. We removed redundancies from the library by merging sequences that were greater than 80% similar with usearch[80], and classified them with RepeatClassifier. Sequences classified as "unknown" were searched with BLASTX against the 9,229 reviewed proteins of insects from UniProtKB/Swiss-Prot. Those sequences with a BLAST hit (E-value <1e-10) against a protein not annotated as a transposase, TE, copia protein, or transposon were removed from the custom repeat library. The custom repeat library was provided to RepeatMasker version open-4.0.5 to generate the repetitive content reports, and to the MAKER2 pipeline to mask the genome.

Protein-coding gene annotation. We performed genome annotations through three iterations of the MAKER2 (v2.31.8) pipeline[18] combining ab initio gene models and evidence-based models. For the *G. bimaculatus* genome annotation, we provided the MAKER2 pipeline with the 43,595 *G. bimaculatus* nucleotide sequences from NCBI, an assembled developmental transcriptome[81], an assembled prothoracic ganglion transcriptome[9], and a genome-guided transcriptome generated with StringTie[82] using 30 RNA-seq libraries (accession numbers: DRA011174 and DDBJ DRA11117) mapped to the genome with HISAT2[83]. As alternative ESTs and protein sequences, we provided MAKER2 with 14,391 nucleotide sequences from *L. kohalensis* available at NCBI, and an insect protein database obtained from UniProtKB/Swiss-Prot[84].

For the annotation of the *L. kohalensis* genome, we ran the MAKER2 pipeline with the 14,391 *L. kohalensis* nucleotide sequences from NCBI, the assembled *G. bimaculatus* developmental and prothoracic ganglion transcriptomes described above, and the 43,595 NCBI nucleotide sequences. As protein databases, we provided the insect proteins from UniProtKB/Swiss-Prot plus the proteins that we annotated in the *G. bimaculatus* genome.

For both crickets, we generated ab initio gene models with GeneMark-ES[85] in self-training mode, and with Augustus[86] trained with BUSCO v3[17]. After each of the first two MAKER2 iterations, additional gene models were obtained with SNAP[87] trained with the annotated genes.

Functional annotations were obtained using InterProScan[19], which retrieved the InterProDomains, PFAM domains, and GO-terms. Additionally, we ran a series of BLAST rounds from more specific to more generic databases, to assign a descriptor to each transcript based on the best BLAST hit. The first round of BLAST was against the reviewed insect proteins from UniProtKB/Swiss-Prot. Proteins with no significant BLAST hits (E-value <1e-6) went to a second round against insect proteins from UniProtKB/TrEMBL, and those without a hit with E-value <1e-6 were used in the final round of BLAST against all proteins from UniProtKB/Swiss-Prot.

A detailed pipeline scheme is available in Supplementary Figs. 1 and 2, and the annotation scripts are available on GitHub (https://github.com/guillemylla/Crickets_Genome_Annotation).

Quality assessment. Genome assembly statistics were obtained with assembly-stats (https://github.com/sanger-pathogens/assembly-stats). BUSCO (v3.1.0)[17] was used to assess the level of completeness of the genome assemblies ("-m geno") as well as that of the gene annotations ("-m prot") at both arthropod ("arthropoda_odb9") and insect ("insecta_odb9") levels.

$CpG_{o/e}$ analysis. We used the genome assemblies and their gene annotations from this study for the two cricket species, and retrieved publicly available annotated genomes from the other 14 insect species (Supplementary Data 1). The gene annotation files (in gff format) were used to obtain the amino acid and CDS sequences for each annotated protein-coding gene per genome using gffread, with options "-y" and "-x", respectively. The $CpG_{o/e}$ value per gene was computed as the observed frequency of CpGs (f_{CpG}) divided by the product of C and G frequencies (f_C and f_G) $f_{CpG}/f_C{*}f_G$ in the longest CDS per gene for each of the 16 studied insects. $CpG_{o/e}$ values larger than zero and smaller than two were retained and represented as density plots (Figs. 2, 4).

The distributions of gene $CpG_{o/e}$ values per gene of the two crickets, the honeybee *A. mellifera*, and the thrips *F. occidentalis*, were fitted with a mixture of normal distributions using the mixtools R package[88]. This allowed us to obtain the mean of each distribution, the standard errors, and the interception point between the two distributions, which was used to categorize the genes into low $CpG_{o/e}$ and high $CpG_{o/e}$ bins. For these two bins of genes, we performed a GO enrichment analysis (based on GO-terms previously obtained using InterProScan) of Biological Process terms using the TopGO package[89] with all genes as universe, minimum node size of 10, the weight01 algorithm, and the Fisher statistic. The GO terms with a p value <0.05 were considered significantly enriched. Those GO terms significantly enriched in at least one gene set are shown in Supplementary Fig. 3, and a subset of them with p value <0.0001 are shown in Fig. 3. In both figures, the size of the circle represents the percentage of enriched genes inside the set compared to all genes with the given GO term.

For each of the genes belonging to low and high $CpG_{o/e}$ categories in each of the four insect species, we retrieved their OG identifier from our gene family analysis, allowing us to assign putative methylation status to OGs in each insect. Then we used the UpSet R package[90] to compute and display the number of OGs exclusive to each combination as an UpSet plot.

dN/dS analysis. We first aligned the longest predicted protein product of the single copy orthologs of all protein-coding genes between the two crickets ($N = 5728$) with MUSCLE (v3.8.31). Then, the amino acid alignments were transformed into codon-based nucleotide alignments using the Pal2Nal software[91]. The resulting codon-based nucleotide alignments were used to calculate the pairwise dN/dS for each gene pair with the yn00 algorithm implemented in the PAML package[92]. Genes with dN or dS >2 were discarded from further analysis. The Wilcoxon–Mann–Whitney statistical test was used to compare the dN/dS values between genes with high and low $CpG_{o/e}$ values in both insects.

Gene family expansions and contractions. Using custom Python scripts (see https://github.com/guillemylla/Crickets_Genome_Annotation) we obtained the longest predicted protein product per gene in each of the 16 studied insect species and grouped them into OGs (which we also refer to herein as "gene families") using OrthoFinder v2.3.3[93]. The OGs determined by OrthoFinder that contained a single gene per insect, namely putative one-to-one orthologs, were used for phylogenetic reconstruction. The proteins within each OG were aligned with MUSCLE[94] and the alignments trimmed with GBlocks ($-t = p$ $-b4 = 5$ $-b5 = a$)[95]. The trimmed alignments were concatenated into a single meta-alignment that was used to infer the species tree with FastTree2 (FastTreeMP –gamma)[96].

To calibrate the species tree, we used the "chronos" function from the R package ape v5.3[97], setting the common node between Blattodea and Orthoptera at 248 million years (my), the origin of Holometabola at 345 my, the common node between Hemiptera and Thysanoptera at 339 my, and the ancestor of hemimetabolous and holometabolous insects (root of the tree) at between 385 and 395 my. These time points were obtained from a phylogeny published that was calibrated with several fossils[23].

The gene family expansion/contraction analysis was done with the CAFE software[98]. We ran CAFE using the calibrated species tree and the table generated by OrthoFinder with the number of genes belonging to each OG in each insect. Following the CAFE manual, we first calculated the birth-death parameters with the OGs having less than 100 genes. We then corrected them by assembly quality and calculated the gene expansions and contractions for both large (>100 genes) and small (≤100) gene families. This allowed us to identify gene families that underwent a significant (p value <0.01) gene family expansion or contraction on each branch of the tree. We proceeded to obtain functional information from those families expanded on our branches of interest (i.e., the origin of Orthoptera, the branch leading to crickets, and the branches specific to each cricket species.). To functionally annotate the OGs of interest, we first obtained the *D. melanogaster* identifiers of the proteins within each OG, and retrieved the FlyBase Symbol and the FlyBase gene summary per gene using the FlyBase API[99]. Additionally, we ran InterProScan on all the proteins of each OG and retrieved all PFAM motifs and the GO terms together with their descriptors. All of this information was summarized in tabulated files (Supplementary Data 2), which we used to identify gene expansions with potentially relevant functions for insect evolution.

***pickpocket* gene family expansion**. Among the expanded gene families in crickets, we identified an OG containing seven out of the eight *D. melanogaster pickpocket* class V genes, leading us to interpret that the *pickpocket* class V was significantly expanded in crickets. Subsequently, we retrieved the six additional OGs containing the complete set of *pickpocket* genes in *D. melanogaster*, and we assigned to each OG the *pickpocket* class to which most of its *D. melanogaster* genes belonged according to Zelle et al. [24] (Supplementary Table 4). The protein sequences of all the members of the seven Pickpocket OGs were aligned with MUSCLE, and the *pickpocket* gene tree obtained with FastTree2 (FastTreeMP --gamma). The tips of the tree were colored based on the OG to which they belong. A subset of the tree containing all the OGs that compose the entire *pickpocket* class V family was displayed as a circular cladogram (Fig. 5), revealing an independent expansion of this family in *T. castaneum*.

To check for evidence of expression *pickpocket* genes in the cricket nervous system, we used the 21 RNA-seq libraries from prothoracic ganglion[9] of *G. bimaculatus* available at NCBI GEO (PRJNA376023). Reads were mapped against the *G. bimaculatus* genome with RSEM[100] using STAR[101] as the mapping algorithm, and the number of expected counts and TPMs were retrieved for each gene in each library. The TPMs of the *pickpocket* genes and *fruitless* are shown in Supplementary Data 3. Genes with a sum of more than 20 TPMs across all samples were considered to be expressed in *G. bimaculatus* prothoracic ganglion. We further analyzed the *pickpocket* expression in the aggregated embryo RNA-seq dataset (DRA011174) and normal and regenerating legs RNA-seq dataset[102] (DRR001985 and DRR001986), using the same methodology (Supplementary Data 3).

Statistics and reproducibility. Statistical test used are described in the corresponding materials and methods section together with the *p* value cutoffs, which are also described in the results sections and corresponding figure captions. Furthermore, to allow the reproducibility of all our analysis and results, all the data has been made available in public databases and the scripts developed for our analysis are available in GitHub as described in the data availability section.

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

Data availability

The genome sequencing reads, RNA-seq reads, and the genome assembly for *Gryllus bimaculatus* were submitted to DDBJ and to NCBI under the accession number (PRJDB10609), and the genome assembly received the GenBank accession number GCA_017312745.1. The genome assembly and annotations can also be accessed and browsed at http://gbimaculatusgenome.rc.fas.harvard.edu. The genome annotation files for the two crickets, *G. bimaculatus* and *L. kohalensis* have also been made available through FigShare (https://figshare.com/projects/Gryllus_bimaculatus_and_Laupala_kohalensis_genome_annotations/101402). The source data underlying the main figures of this text is accessible via FigShare (https://figshare.com/projects/Source_data_for_the_figures_of_Ylla_et_al_2021/101423).

Code availability

The scripts used for genome annotation and analysis are available at GitHub (https://github.com/guillemylla/Crickets_Genome_Annotation) under GNU GPLv3 license. The code has been deposited to Zenodo (https://doi.org/10.5281/zenodo.4648022).

Received: 9 August 2020; Accepted: 16 April 2021;
Published online: 14 June 2021

References

1. Belles, X. *Origin and Evolution of Insect Metamorphosis* (John Wiley & Sons, Ltd, 2011).
2. Engel, M. S. & Grimaldi, D. A. New light shed on the oldest insect. *Nature* **427**, 627–630 (2004).
3. Gregory, T. R. Genome size and developmental complexity. *Genetica* **115**, 131–146 (2002).
4. Camacho, J. P. et al. A step to the gigantic genome of the desert locust: chromosome sizes and repeated DNAs. *Chromosoma* **124**, 263–275 (2015).
5. Bewick, A. J., Vogel, K. J., Moore, A. J. & Schmitz, R. J. Evolution of DNA methylation across insects. *Mol. Biol. Evol.* **34**, 654–665 (2016).
6. Provataris, P., Meusemann, K., Niehuis, O., Grath, S. & Misof, B. Signatures of DNA methylation across insects suggest reduced DNA methylation levels in Holometabola. *Genome Biol. Evol.* **10**, 1185–1197 (2018).
7. Blankers, T., Oh, K. P., Bombarely, A. & Shaw, K. L. The genomic architecture of a rapid Island radiation: recombination rate variation, chromosome structure, and genome assembly of the hawaiian cricket Laupala. *Genetics* **209**, 1329–1344 (2018).
8. Huber, F., Moore, T. E. & Loher, W. *Cricket Behavior and Neurobiology* (Comstock Pub. Associates, 1989).
9. Fisher, H. P. et al. De novo assembly of a transcriptome for the cricket Gryllus bimaculatus prothoracic ganglion: an invertebrate model for investigating adult central nervous system compensatory plasticity. *PLoS ONE* **13**, e0199070 (2018).
10. Kainz, F., Ewen-Campen, B., Akam, M. & Extavour, C. G. Notch/Delta signalling is not required for segment generation in the basally branching insect Gryllus bimaculatus. *Development* **138**, 5015–5026 (2011).
11. Donoughe, S. & Extavour, C. G. Embryonic development of the cricket Gryllus bimaculatus. *Developmental Biol.* **411**, 140–156 (2015).
12. Mito, T. & Noji, S. The two-spotted cricket Gryllus bimaculatus: an emerging model for developmental and regeneration studies. *CSH Protoc.* **2008**, pdb.emo110 (2008).
13. Kulkarni, A. & Extavour, C. G. The Cricket Gryllus bimaculatus: Techniques for Quantitative and Functional Genetic Analyses of Cricket Biology. In *Results and Problems in Cell Differentiation, Volume 68: Evo-Devo: Non-model Species in Cell and Developmental Biology* (eds. Tworzydlo, W. & Bilinski, S. M.) **68**, 183–216 (Springer, 2019).
14. Shinmyo, Y. et al. piggyBac-mediated somatic transformation of the two-spotted cricket, Gryllus bimaculatus. *Dev. Growth Differ.* **46**, 343–349 (2004).
15. Wilson Horch, H., Mito, T., Popadić, A., Ohuchi, H. & Noji, S. *The Cricket as a Model Organism: Development, Regeneration, and Behavior.* (Springer, 2017).
16. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504 (1980).
17. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
18. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinforma.* **12**, 491 (2011).
19. Jones P., et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
20. Smit, A., Hubley, R. & Grenn, P. *RepeatMasker Open-4.0* (2015).
21. Wang, Y. et al. Functional CpG methylation system in a social insect. *Science* **314**, 645–647 (2006).
22. Elango, N., Hunt, B. G., Goodisman, M. A. D. & Yi, S. V. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, Apis mellifera. *Proc. Natl Acad. Sci. USA* **106**, 11206–11211 (2009).
23. Misof, B. et al. Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–767 (2014).
24. Zelle, K. M., Lu, B., Pyfrom, S. C. & Ben-Shahar, Y. The genetic architecture of degenerin/epithelial sodium channels in Drosophila. *G3* **3**, 441–450 (2013).
25. Adams, C. M. et al. Ripped pocket and pickpocket, novel Drosophila DEG/ENaC subunits expressed in early development and in mechanosensory neurons. *J. Cell Biol.* **140**, 143–152 (1998).
26. Zhong, L., Hwang, R. Y. & Tracey, W. D. Pickpocket is a DEG/ENaC protein required for mechanical nociception in Drosophila larvae. *Curr. Biol.* **20**, 429–434 (2010).
27. Ainsley, J. A. et al. Enhanced locomotion caused by loss of the Drosophila DEG/ENaC protein pickpocket1. *Curr. Biol.* **13**, 1557–1563 (2003).
28. Häsemeyer, M., Yapici, N., Heberlein, U. & Dickson, B. J. Sensory neurons in the Drosophila genital tract regulate female reproductive behavior. *Neuron* **61**, 511–518 (2009).
29. Rezával, C. et al. Neural circuitry underlying Drosophila female postmating behavioral responses. *Curr. Biol.* **22**, 1155–1165 (2012).
30. Pavlou, H. J. & Goodwin, S. F. Courtship behavior in Drosophila melanogaster: towards a 'courtship connectome'. *Curr. Opin. Neurobiol.* **23**, 76–83 (2013).
31. Adams, M. D. et al. The genome sequence of Drosophila melanogaster. *Science* **287**, 2185–2195 (2000).
32. Elsik, C. G. The pea aphid genome sequence brings theories of insect defense into question. *Genome Biol.* **11**, 106 (2010).
33. Wang, X. et al. The locust genome provides insight into swarm formation and long-distance flight. *Nat. Commun.* **5**, 2957 (2014).
34. Harrison, M. C. et al. Hemimetabolous genomes reveal molecular basis of termite eusociality. *Nat. Ecol. Evol.* **2**, 557–566 (2018).
35. Terrapon, N. et al. Molecular traces of alternative social organization in a termite genome. *Nat. Commun.* **5**, 3636 (2014).
36. Poulsen, M. et al. Complementary symbiont contributions to plant decomposition in a fungus-farming termite. *Proc. Natl Acad. Sci. USA* **111**, 14500–14505 (2014).
37. Kouřimská, L. & Adámková, A. Nutritional and sensory quality of edible insects. *NFS J.* **4**, 22–26 (2016).
38. Van Huis, A. et al. *Edible Insects: Future Prospects for Food and Feed Security* (FAO, 2013).
39. Ghosh, S., Lee, S.-M., Jung, C. & Meyer-Rochow, V. Nutritional composition of five commercial edible insects in South Korea. *J. Asia-Pac. Entomol.* **20**, 686–694 (2017).

40. Hanboonsong, Y., Jamjanya, T. & Durst, P. B. *Six-legged Livestock: Edible Insect Farming, Collecting and Marketing in Thailand* (FAO, 2013).
41. Ryu, H. Y. et al. Oral toxicity study and skin sensitization test of a cricket. *Toxicological Res.* **32**, 159–173 (2016).
42. Ahn, M. Y., Han, J. W., Kim, S. J., Hwang, J. S. & Yun, E. Y. Thirteen-week oral dose toxicity study of G. bimaculatus in sprague-dawley rats. *Toxicological Res.* **27**, 231–240 (2011).
43. Mi, Y. A. et al. Genotoxic evaluation of the biocomponents of the cricket, Gryllus bimaculatus, using three mutagenicity tests. *J. Toxicol. Environ. Health Part A* **68**, 2111–2118 (2005).
44. Pener, M. P. Allergy to crickets: a review. *J. Orthoptera Res.* **25**, 91–95 (2016).
45. Ribeiro, J. C., Cunha, L. M., Sousa-Pinto, B. & Fonseca, J. Allergic risks of consuming edible insects: a systematic review. *Mol. Nutr. Food Res.* **62**, 1700030 (2018).
46. Ahn, M. Y., Hwang, J. S., Yun, E. Y., Kim, M. J. & Park, K. K. Anti-aging effect and gene expression profiling of aged rats treated with G. bimaculatus extract. *Toxicological Res.* **31**, 173–180 (2015).
47. Park, S. A., Lee, G. H., Lee, H. Y., Hoang, T. H. & Chae, H. J. Glucose-lowering effect of Gryllus bimaculatus powder on streptozotocin-induced diabetes through the AKT/mTOR pathway. *Food Sci. Nutr.* **8**, 402–409 (2019).
48. Hwang, B. B. et al. The edible insect Gryllus bimaculatus protects against gut-derived inflammatory responses and liver damage in mice after acute alcohol exposure. *Nutrients* **11**, 857 (2019).
49. Thrall, P. H., Bever, J. D. & Burdon, J. J. Evolutionary change in agriculture: the past, present and future. *Evol. Appl.* **3**, 405–408 (2010).
50. Yamasaki, M. et al. A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell* **17**, 2859–2872 (2005).
51. Chen, Y. H., Gols, R. & Benrey, B. Crop domestication and its impact on naturally selected trophic interactions. *Annu. Rev. Entomol.* **60**, 35–58 (2015).
52. Gepts, P. Crop domestication as a long-term selection experiment. *Plant Breed. Rev.* **24**, 1–44 (2004).
53. Chen, K. & Gao, C. Targeted genome modification technologies and their applications in crop improvements. *Plant Cell Rep.* **33**, 575–583 (2014).
54. Qaim, M. The economics of genetically modified crops. *Annu. Rev. Resour. Econ.* **1**, 665–694 (2009).
55. Kidwell, M. G. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**, 49–63 (2002).
56. Chénais, B., Caruso, A., Hiard, S. & Casse, N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene* **509**, 7–15 (2012).
57. Hanrahan, S. J. & Johnston, J. S. New genome size estimates of 134 species of arthropods. *Chromosome Res.* **19**, 809–823 (2011).
58. Westerman, M., Barton, N. & Hewitt, G. M. Differences in DNA content between two chromosomal races of the grasshopper Podisma pedestris. *Heredity* **58**, 221–228 (1987).
59. Palacios-Gimenez, O. M. et al. Comparative analysis of morabine grasshopper genomes reveals highly abundant transposable elements and rapidly proliferating satellite DNA repeats. *BMC Biol.* **18**, 199 (2020).
60. Lyko, F., Ramsahoye, B. H. & Jaenisch, R. DNA methylation in Drosophila melanogaster. *Nature* **408**, 538–540 (2000).
61. Thomas, G. W. C. et al. Gene content evolution in the arthropods. *Genome Biol.* **21**, 15 (2020).
62. Lewis, S. H. et al. Widespread conservation and lineage-specific diversification of genome-wide DNA methylation patterns across arthropods. *PLoS Genet.* **16**, e1008864 (2020).
63. Ylla, G., Piulachs, M.-D. & Belles, X. Comparative transcriptomics in two extreme neopterans reveal general trends in the evolution of modern insects. *iScience* **4**, 164–179 (2018).
64. Ventós-Alfonso, A., Ylla, G., Montañes, J.-C. & Belles, X. DNMT1 promotes genome methylation and early embryo development in cockroaches. *iScience* **23**, 101778 (2020).
65. Liu, L., Johnson, W. A. & Welsh, M. J. Drosophila DEG/ENaC pickpocket genes are expressed in the tracheal system, where they may be involved in liquid clearance. *Proc. Natl Acad. Sci. USA* **100**, 2128–2133 (2003).
66. Lee, M. J. et al. Ionotropic receptor 76b is required for gustatory aversion to excessive Na^+ in Drosophila. *Mol. Cells* **40**, 787–795 (2017).
67. Averhoff, W. W., Richardson, R. H., Starostina, E., Kinser, R. D. & Pikielny, C. W. Multiple pheromone system controlling mating in Drosophila melanogaster. *Proc. Natl Acad. Sci. USA* **73**, 591–593 (1976).
68. Lu, B., LaMora, A., Sun, Y., Welsh, M. J. & Ben-Shahar, Y. ppk23-dependent chemosensory functions contribute to courtship behavior in Drosophila melanogaster. *PLoS Genet.* **8**, e1002587 (2012).
69. Jacob, P. F. & Hedwig, B. Acoustic signalling for mate attraction in crickets: abdominal ganglia control the timing of the calling song pattern. *Behav. Brain Res.* **309**, 51–66 (2016).
70. Blankers, T., Oh, K. P. & Shaw, K. L. The genetics of a behavioral speciation phenotype in an Island system. *Genes* **9**, 346 (2018).
71. Shaw, K. L. & Lesnick, S. C. Genomic linkage of male song and female acoustic preference QTL underlying a rapid species radiation. *Proc. Natl Acad. Sci.* **106**, 9737–9742 (2009).
72. Xu, M. & Shaw, K. L. The genetics of mating song evolution underlying rapid speciation: linking quantitative variation to candidate genes for behavioral isolation. *Genetics* **211**, 1089–1104 (2019).
73. Yoshimura, A., Nakata, A., Mito, T. & Noji, S. The characteristics of karyotype and telomeric satellite DNA sequences in the cricket, Gryllus bimaculatus (Orthoptera, Gryllidae). *Cytogenet. Genome Res.* **112**, 329–336 (2006).
74. Kajitani, R. et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).
75. English, A. C. et al. Mind the gap: upgrading genomes with Pacific biosciences RS long-read sequencing technology. *PLoS ONE* **7**, e47768 (2012).
76. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinforma.* **9**, 18 (2008).
77. Crescente, J. M., Zavallo, D., Helguera, M. & Vanzetti, L. S. MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinforma.* **19**, 348 (2018).
78. Vassetzky, N. S. & Kramerov, D. A. SINEBase: a database and tool for SINE analysis. *Nucleic Acids Res.* **41**, D83–D89 (2013).
79. Bao, W., Kojima, K. K. & Kohany, O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
80. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
81. Zeng, V. et al. Developmental gene discovery in a hemimetabolous insect: de novo assembly and annotation of a transcriptome for the cricket Gryllus bimaculatus. *PLoS ONE* **8**, e61479 (2013).
82. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
83. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
84. UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
85. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).
86. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215–ii225 (2003).
87. Korf, I. Gene finding in novel genomes. *BMC Bioinforma.* **5**, 59 (2004).
88. Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. S. Mixtools: an R package for analyzing finite mixture models. *J. Stat. Softw.* **32**, 1–29 (2009).
89. Alexa, A. & Rahnenfuhrer, J. topGO: enrichment analysis for gene ontology. *R package version 2.42.0* (2019).
90. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: visualization of intersecting sets. *IEEE Trans. Vis. Computer Graph.* **20**, 1983–1992 (2014).
91. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
92. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
93. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
94. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
95. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
96. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
97. Paradis, E. & Schliep, K. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
98. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
99. Thurmond, J. et al. FlyBase 2.0: the next generation. *Nucleic Acids Res.* **47**, D759–D765 (2019).
100. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinforma.* **12**, 323 (2011).
101. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
102. Bando, T. et al. Analysis of RNA-Seq data reveals involvement of JAK/STAT signalling during leg regeneration in the cricket Gryllus bimaculatus. *Development* **140**, 959–964 (2013).

Acknowledgements

This work was supported by Harvard University and MEXT KAKENHI (No. 221S0002; 26292176; and 17H03945). The computational infrastructure in the cloud used for the genome analysis was funded by AWS Cloud Credits for Research. The authors are grateful to Hiroo Saihara for his support in the management of a genome data server at Tokushima University.

Author contributions

G.Y., S.N., T.M., and C.G.E. designed experiments; T.I. and A.T. conducted sequencing by HiSeq and assembling short reads using the Platanus assembler; S.T., Y.I., T.W., M.F., and Y.M. performed DNA isolation, gap closing of contigs, and manual annotation; G.Y., T.N., S.T., T.B., and A.A.B. conducted all other experiments and analyses; T.M. and C.G. E. funded the project; G.Y. and C.G.E. wrote the paper with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-021-02197-9.

Correspondence and requests for materials should be addressed to G.Y., T.M. or C.G.E.