Insights into the genomic evolution of insects from cricket genomes

Guillem Ylla1*, Taro Nakamura1,2, Takehiko Itoh3, Rei Kajitani3, Atsushi Toyoda4,5, Sayuri Tomonari6, Tetsuya Bando7, Yoshiyasu Ishimaru6, Takahito Watanabe6, Masao Fuketa8, Yuji Matsuoka6,9, Austen A. Barnett1,10, Sumihare Noji6, Taro Mito6*, Cassandra G. Extavour1,11*

1. Department of Organismic and Evolutionary Biology, Harvard University, Cambridge MA, USA
2. Current address: National Institute for Basic Biology, Okazaki, Japan
3. School of Life Science and Technology, Tokyo Institute of Technology, Tokyo, Japan
4. Comparative Genomics Laboratory, National Institute of Genetics, Shizuoka, Japan
5. Advanced Genomics Center, National Institute of Genetics, Shizuoka, Japan
6. Department of Bioscience and Bioindustry, Tokushima University, Tokushima, Japan
7. Graduate School of Medicine, Pharmacology and Dentistry, Okayama University, Okayama, Japan
8. Graduate School of Advanced Technology and Science, Tokushima University, Tokushima, Japan
9. Current address: Department of Biological Sciences, National University of Singapore, Singapore
10. Current address: Department of Natural Sciences, DeSales University, Center Valley PA, USA
11. Department of Molecular and Cellular Biology, Harvard University, Cambridge MA, USA

* correspondence to guillemyllabou@gmail.com, mito.taro@tokushima-u.ac.jp and extavour@oeb.harvard.edu

CC-BY-NC-ND 4.0 International license available under a
Abstract

Most of our knowledge of insect genomes comes from Holometabolous species, which undergo the complete metamorphosis and have genomes under 2Gb with little signs of DNA methylation. In contrast, Hemimetabolous insects undergo the ancestral 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 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.
Introduction

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 which 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. This clade of insects includes nearly 90% of extant described insect species. 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 last transition to the sexually mature, winged adult, without major changes in body plan from hatching to adult. Many extant species maintained 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.

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 10 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). This has led to the hypothesis that there is a genome size threshold at 2 pg (~2 Gb) for holometabolan insect genomes. Studying genome size evolution the polyneopterans 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 Gb to 16.56 Gb. 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 studied polyneopteran species do. 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 (Figure 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. G. bimaculatus has been widely used as a laboratory research model for decades, in scientific fields including neurobiology and neuroethology, evo-devo, developmental biology, and regeneration. 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, stable germline transgenic lines can be established, and it has an extensive list of available experimental protocols ranging from behavioral to functional genetic analyses.

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, 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 of the cricket G. bimaculatus (Figure 1A). 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 polynopteran genomes, our assembly displays high quality in terms of contiguity (N50 and L50), and completeness (BUSCO scores) (Supplementary Table 1). Notably, the complete BUSCO scores 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 |

Ylla et al. Page 4 of 39
Annotation of two cricket genomes

The publicly available 1.6-Gb genome assembly of the Hawaiian cricket \textit{L. kohalensis}, although having lower assembly quality scores (N50=0.58 Mb, L90 = 3,483) than that of \textit{G. bimaculatus}, scores high in terms of completeness, with BUSCO scores of 99.3\% at the arthropod level and 97.80\% at the insect level (Supplementary Table 1).

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

To obtain functional insights into the annotated genes, we ran InterProScan\textsuperscript{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 \textit{G. bimaculatus} and \textit{L. kohalensis} respectively. We created a novel graphic interface through which interested readers can access, search, BLAST and download the genome data and annotations (http://gbimaculatusgenome.rc.fas.harvard.edu).

| 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 \textit{G. bimaculatus} and \textit{L. kohalensis}**

|                      | \textit{G. bimaculatus} | \textit{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% |

**Abundant Repetitive DNA**

We used RepeatMasker\(^2\) 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 File 1). In *G. bimaculatus* the repetitive content density was similar throughout the genome, with the exception of scaffolds shorter than 1Mb (L90), which make up 10% of the genome and have a high density of repetitive content and low gene density (Figure 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 transposable elements (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 (Figure 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 (CpGo/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. 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 (Figure 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 (Figure 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 (Figure 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, and the thrips *Frankliniella occidentalis*. 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 (Figure 3 and Supplementary Figure 3). Specifically, 23GO 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 CpG_o/e genes in all four insects were enriched for similar functions (8 GO-terms commonly enriched in all insects; Supplementary Figure 3).

Additionally, we observed that the proportion of species-specific genes was higher within the high CpG_o/e peak for all four insects (Figure 2C). In contrast, 86-96% of the genes belonging to the low CpG_o/e peak had an orthologous gene in at least one of the other studied insect species. Furthermore, we observed 2,182 orthogroups whose members always belonged to the low CpG_o/e peak in all four species, and 728 orthogroups whose members always belonged to the high CpG_o/e peak, indicating that orthologous genes are likely to share methylation state across these four insect species (Figure 2B and Supplementary Figure 4). Interestingly, 666 genes belonged to the low CpG_o/e peak in the three hemimetabolous
species (G. bimaculatus, L. kohalensis, and F. occidentallis), but to the high 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 CpG_{o/e} and low evolutionary rates, we compared the dN/dS values of 1-to-1 orthologous genes belonging to the same 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; Figure 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 orthogroups (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. The obtained species tree topology was in accordance with the currently understood insect phylogeny\(^\text{23}\). Then, we used the Misof et al. (2014) 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 File 2) revealed that the cricket-specific gene family expansions included pickpocket genes, which are involved in mechanosensation in Drosophila 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 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 (Figure 3; Supplementary Table 2). This gene tree allowed us to classify the different pickpocket genes in each of the 16 species. One orthogroup, 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.* we determined that the specific gene family expanded in crickets was pickpocket class V (Figure 3). In *D. melanogaster* this class contains eight genes: ppk (ppk1), rpk (ppk2), ppk5, ppk8, ppk12, ppk17, ppk26, and ppk28. 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 (Figure 1B). For instance, in *G. bimaculatus* nine of the 15 class V pickpocket genes were clustered within a region of 900Kb, 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 Table 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. For example, in larvae, this gene is required for mechanical nociception, and for coordinating rhythmic locomotion. *ppk* is expressed in sensory neurons that also express the male sexual behavior determiner *fruitless* (*fru*)..

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. 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 Table 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 Table 4).

**Discussion**

**The importance of cricket genomes**

Sequencing and analyzing genomes from underrepresented clades allow 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, 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 ten years later, with the publication of
the genome sequence and annotation of the Pea aphid (*Acyrthosiphon pisum*). When even
more recently, polypepiteran genome sequences became available, 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 an animal
species as tractable experimental models. *G. bimaculatus* is a common laboratory research
animal used in neuroethology, developmental and regeneration biology studies. 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 containing the
essential linoleic acid as their most predominant fatty acid. 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. Studies have reported
no evidence for toxicological effects related to oral consumption of *G. bimaculatus* by
humans, neither were genotoxic effects detected using three different mutagenicity
tests. A rare but known health risk associated with cricket consumption, however, is
sensitivity and allergy to crickets. 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. 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. 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. 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. Crickets are naturally nutritionally
rich, 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.
Comparing cricket genomes to other insect genomes

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)\(^2\) and the termites Cryptotermes secundus (18,162)\(^5\), Macrotermes natalensis (16,140)\(^6\) and Zootermopsis nevadensis, (15,459)\(^7\). 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, leading to higher assembly fragmentation, 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.

TEs and genome size evolution

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\(^2\). 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\(^6\)\(^,\)^\(^7\).

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\(^2\). 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\(^4\)\(^,\)^\(^5\), 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.\(^6\)\(^,\)^\(^6\) of massive and independent recent TE accumulation in four chromosome races of the grasshopper Vandienenella viatica.

There is a clear tendency of polyneopteran genomes to be much longer than those of the holometabolous genomes (Figure 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. Our observations are consistent with the first of these hypotheses.

DNA Methylation

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. The honeybee A. mellifera was one of the first insects for which functional DNA methylation was described. Although this DNA modification was initially proposed to be associated with the eusociality of these bees, subsequent studies showed that DNA methylation is widespread and present in different insect lineages independently of social behavior. DNA methylation also occurs in other non-insect arthropods.

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. 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, and functional analysis has shown that the DNA methyltransferase 1 is essential for early embryo development in this cockroach. 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.

pickpocket gene expansion

The pickpocket genes belong to the Degenerin/epithelial Na+ channel (DEG/ENaC) family, which were first identified in Caenorhabditis elegans as involved in mechanotransduction. 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.
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\(^{67}\), pheromone detection \(^{68}\), courtship behavior \(^{69}\), and liquid clearance in the larval trachea\(^{66}\).

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\(^{70}\). 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\(^{71}\). 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\(^{72}\). Xu and Shaw \(^{73}\) 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 \(^{73}\) 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\(^{73}\). 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
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, genomic DNA extracted from males contains the full set of chromosomes; males were therefore chosen for genomic DNA isolation.

Genome Assembly

Paired-end libraries were generated with insert sizes of 375 and 500 bp, and mate-pair libraries were generated with insert sizes of 3, 5, 10, and 20kb. 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)\(^75\). Scaffolding and gap closing were performed 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)\(^76\).

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\(^77\), RepeatModeler/RepeatClassifier (www.repeatmasker.org/RepeatModeler), MITE tracker\(^78\), TransposonPSI (http://transposonpsi.sourceforge.net), and the databases SINEBase\(^79\) and RepBase\(^80\). We removed redundancies from the library by merging sequences that were greater than 80% similar with usearch\(^81\), 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, transposable element, 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 Genes Annotation**

We performed genome annotations through three iterations of the MAKER2 (v2.31.8)
pipeline\(^{18}\) combining \textit{ab-initio} gene models and evidence-based models. For the \textit{G. bimaculatus} genome annotation, we provided the MAKER2 pipeline with the 43,595 \textit{G. bimaculatus} nucleotide sequences from NCBI, an assembled developmental transcriptome\(^{82}\), an assembled prothoracic ganglion transcriptome\(^9\), and a genome-guided transcriptome generated with StringTie\(^{83}\) using 30 RNA-seq libraries (accession numbers: DRA011174 and DDBJ DRA11117) mapped to the genome with HISAT2\(^{84}\). As alternative ESTs and protein sequences, we provided MAKER2 with 14,391 nucleotide sequences from \textit{L. kohalensis} available at NCBI, and an insect protein database obtained from UniProtKB/Swiss-Prot\(^\text{\textsuperscript{85}}\).

For the annotation of the \textit{L. kohalensis} genome, we ran the MAKER2 pipeline with the 14,391 \textit{L. kohalensis} nucleotide sequences from NCBI, the assembled \textit{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 \textit{G. bimaculatus} genome.

For both crickets, we generated \textit{ab-initio} gene models with GeneMark-ES\(^\text{\textsuperscript{86}}\) in self-training mode, and with Augustus\(^\text{\textsuperscript{87}}\) trained with BUSCO v3\(^\text{\textsuperscript{17}}\). After each of the first two MAKER2 iterations, additional gene models were obtained with SNAP\(^\text{\textsuperscript{88}}\) trained with the annotated genes.

Functional annotations were obtained using InterProScan\(^\text{\textsuperscript{19}}\), 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 all 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 Figures 1 & 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) 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.

CpG₀/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 Table 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 CpG₀/e value per gene was computed as the observed frequency of CpGs (fCpG) divided by the product of C and G frequencies (fC and fG) fCpG/fC*fG in the longest CDS per gene for each of the 16 studied insects. CpG₀/e values larger than zero and smaller than two were retained and represented as density plots (Figures 2 & 4).

The distributions of gene CpG₀/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. 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 CpG₀/e and high CpG₀/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 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 Figure 3, and a subset of them with p-value<0.0001 are shown in Figure 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 CpG₀/e categories in each of the four insect species, we retrieved their orthogroup identifier from our gene family analysis, allowing us to assign putative methylation status to orthogroups in each insect. Then we used the UpSet R package to compute and display the number of orthogroups 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=5,728) with MUSCLE (v3.8.31). Then, the amino-acid alignments were transformed into codon-based nucleotide alignments using the
Pal2Nal software\textsuperscript{92}. 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\textsuperscript{93}. 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\textsubscript{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 orthogroups (which we also refer to herein as "gene families") using OrthoFinder v2.3.3\textsuperscript{94}. The orthogroups (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 orthogroup were aligned with MUSCLE \textsuperscript{95} and the alignments trimmed with GBlocks (\texttt{-t=p \ -b4=5 \ -b5=a})\textsuperscript{96}. The trimmed alignments were concatenated into a single meta-alignment that was used to infer the species tree with FastTree2 (FastTreeMP \--gamma)\textsuperscript{97}.

To calibrate the species tree, we used the "chronos" function from the R package ape \textsuperscript{v5.3}\textsuperscript{98}, 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\textsuperscript{23}.

The gene family expansion/contraction analysis was done with the CAFE software\textsuperscript{99}. We ran CAFE using the calibrated species tree and the table generated by OrthoFinder with the number of genes belonging to each orthogroup in each insect. Following the CAFE manual, we first calculated the birth-death parameters with the orthogroups 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 (\textit{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 orthogroups of interest, we first obtained the \textit{D. melanogaster} identifiers of the proteins within each orthogroup, and retrieved the FlyBase Symbol and the FlyBase gene summary per gene using the FlyBase API\textsuperscript{100}. Additionally, we ran InterProScan on all the proteins of each orthogroup and retrieved all PFAM motifs and the GO terms together with their descriptors. All of this information was summarized in tabulated files (Supplementary File 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 orthogroup 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 additional orthogroups containing the complete set of pickpocket genes in *D. melanogaster*, and we assigned to each orthogroup the pickpocket class to which most of its *D. melanogaster* genes belonged according to Zelle and colleagues (Supplementary Table 2). The protein sequences of all the members of the seven Pickpocket orthogroups were aligned with MUSCLE, and the pickpocket gene tree obtained with FastTree2 (FastTreeMP --gamma). The tips of the tree were colored based on the orthogroup to which they belong. A subset of the tree containing all the orthogroups that compose the entire pickpocket class V family was displayed as a circular cladogram (Figure 3), 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 of *G. bimaculatus* available at NCBI GEO (PRJNA376023). Reads were mapped against the *G. bimaculatus* genome with RSEM using STAR 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 Table 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 (DRR001985, DRR001986), using the same methodology.

**Acknowledgments**

This work was supported by Harvard University and MEXT KAKENHI (No. 221S0002; 26292176; 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 statement**

GY, SN, TM and CE designed experiments; TI and AT conducted sequencing by HiSeq and assembling short reads using the Platanus assembler; ST, YI, TW, MF and YM performed DNA isolation, gap closing of contigs and manual annotation; GY, TN, ST, TB and AAB conducted all other experiments and analyses; TM and CE funded the project; GY and CE wrote the paper with input from all authors.
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). The genome assembly and annotations can also be accessed and browsed at [http://gbimaculatusgenome.rc.fas.harvard.edu](http://gbimaculatusgenome.rc.fas.harvard.edu).

Code availability

The scripts used for genome annotation and analysis are available at GitHub ([https://github.com/guillemylla/Crickets_Genome_Annotation](https://github.com/guillemylla/Crickets_Genome_Annotation)).

Competing interests

The authors declare no competing interests.

References

1. Belles, X. in *Encyclopedia of Life Sciences* 1-11 (John Wiley & Sons, Ltd, Chichester, 2011).
2. Engel, M. S. & Grimaldi, D. A. in *Nature* Vol. 427 627-630 (2004).
3. Gregory, T. R. Genome size and developmental complexity. *Genetica* **115**, 131-146, doi:10.1023/a:1016032400147 (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, doi:10.1007/s00412-014-0499-0 (2015).
5. Bewick, A. J., Vogel, K. J., Moore, A. J. & Schmitz, R. J. in *Molecular Biology and Evolution* Vol. 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, doi:10.1093/gbe/evy066 (2018).
7. Blankers, T., Oh, K. P., Bombarely, A. & Shaw, K. L. in *Genetics* Vol. 209 1329-1344 (Genetics, 2018).
8. Huber, F., Moore, T. E. T. E. & Loher, W. Cricket behavior and neurobiology. 565 (1989).
9. Fisher, H. P. *et al.* in *PLoS ONE* Vol. 13 (ed Silvana Allodi) e0199070 (Public Library of Science, 2018).
10. Kainz, F., Ewen-Campen, B., Akam, M. & Extaur, C. G. Notch/Delta signalling is not required for segment generation in the basally branching insect *Gryllus bimaculatus*. *Development* **138**, 5015-5026, doi:10.1242/dev.073395 (2011).
11. Donoughe, S. & Extaur, C. G. in *Developmental Biology* Vol. 411 140-156 (Academic Press, 2015).
12. Mito, T. & Noji, S. in *CSH protocols* Vol. 2008 pdb.emo110 (Cold Spring Harbor Laboratory Press, 2008).
13. Kulkarni, A. & Extaur, C. G. in *Evo-Devo: Non-model Species in Cell and Developmental Biology* Vol. 68 183-216 (Springer, 2019).
14. Shinmyo, Y. *et al.* in *Development, Growth and Differentiation* Vol. 46 343-349 (John Wiley & Sons, 2004).
15. Wilson Horch, H., Mito, T., Popadić, A., Ohuchi, H. & Noji, S. in *The Cricket as a Model Organism: Development, Regeneration, and Behavior* (eds Hadley Wilson Horch et al.) 1-376 (Springer Japan, Tokyo, 2017).
16. Bird, A. P. in *Nucleic Acids Research* Vol. 8 1499-1504 (1980).
A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell* **17**, 2859-2872, doi:10.1105/tpc.105.037242 (2005).
712 51 Chen, Y. H., Gols, R. & Benrey, B. Crop domestication and its impact on naturally
713 selected trophic interactions. *Annu Rev Entomol* **60**, 35-58, doi:10.1146/annurev-
714 ento-010814-020601 (2015).
715 52 Gepts, P. Crop domestication as a long-term selection experiment. *Plant breeding
716 reviews* **24**, 1-44 (2004).
717 53 Chen, K. & Gao, C. Targeted genome modification technologies and their applications
718 in crop improvements. *Plant Cell Reports* **33**, 575-583, doi:10.1007/s00299-013-
719 1539-6 (2014).
720 54 Qaim, M. The Economics of Genetically Modified Crops. *Annual Review of Resource
721 Economics* **1**, 665-694, doi:10.1146/annurev.resource.050708.144203 (2009).
722 55 Harrison, M. C. *et al.* Hemimethabolous genomes reveal molecular basis of termite
eusociality. *Nature ecology & evolution*, doi:10.1038/s41559-017-0459-1 (2018).
723 56 Kidwell, M. G. in *Genetica* Vol. **115**, 49-63 (Kluwer Academic Publishers, 2002).
724 57 Chénais, B., Caruso, A., Hiard, S. & Casse, N. in *Gene* Vol. **509**, 7-15 (Elsevier, 2012).
725 58 Hanrahan, S. J. & Johnston, J. S. in *Chromosome research : an international journal on
the molecular, supramolecular and evolutionary aspects of chromosome biology* Vol.
726 19, 809-823 (2011).
727 59 Westerman, M., Barton, N. & Hewitt, G. M. Differences in DNA content between two
728 chromosomal races of the grasshopper Podisma pedestris. *Heredity* **58**, 221-228
729 (1987).
730 60 Palacios-Gimenez, O. M. *et al.* Comparative analysis of morabine grasshopper
731 genomes reveals highly abundant transposable elements and rapidly proliferating
732 satellite DNA repeats. *BMC Biology* **18**, doi:10.1186/s12915-020-00925-x (2020).
733 61 Lyko, F., Ramsahoye, B. H. & Jaenisch, R. DNA methylation in Drosophila
734 melanogaster. *Nature* **408**, 538-540, doi:10.1038/35046205 (2000).
735 62 Thomas, G. W. C. *et al.* Gene content evolution in the arthropods. *Genome Biol
21, 15, doi:10.1186/s13059-019-1925-7 (2020).
736 63 Lewis, S. H. *et al.* Widespread conservation and lineage-specific diversification of
737 genome-wide DNA methylation patterns across arthropods. *PLoS Genet* **16**, e1008864, doi:10.1371/journal.pgen.1008864 (2020).
738 64 Ylla, G., Piulachs, M.-D. & Belles, X. in *iScience* Vol. **4** 164-179 (Elsevier, 2018).
739 65 Ventós-Alfonso, A., Ylla, G., Montaña, J.-C. & Belles, X. DNMT1 promotes genome
740 methylation and early embryo development in cockroaches. *iScience*, 101778,
doi:10.1016/j.isci.2020.101778 (2020).
741 66 Liu, L., Johnson, W. A. & Welsh, M. J. in *Proceedings of the National Academy of
Sciences of the United States of America* Vol. **100**, 2128-2133 (National Academy of
Sciences, 2003).
742 67 Lee, M. J. *et al.* in *Molecules and cells* Vol. **40** 787-795 (Korean Society for Molecular
743 and Cellular Biology, 2017).
744 68 Averhoff, W. W., Richardson, R. H., Starostina, E., Kinser, R. D. & Pikielny, C. W. in
745 Proceedings of the National Academy of Sciences of the United States of America* Vol.
746 **73** 591-593 (National Academy of Sciences, 1976).
747 69 Lu, B., LaMora, A., Sun, Y., Welsh, M. J. & Ben-Shahar, Y. in *PLoS Genetics* Vol. **8** (ed
748 Miriam B. Goodman) e100287 (Public Library of Science, 2012).
749 70 Jacob, P. F. & Hedwig, B. in *Behavioural Brain Research* Vol. **309** 51-66 (Elsevier,
750 2016).
751 71 Blankers, T., Oh, K. P. & Shaw, K. L in *Genes* Vol. **9** 346 (Multidisciplinary Digital
752 Publishing Institute, 2018).
753 72 Shaw, K. L. & Lesnick, S. C. in *Proceedings of the National Academy of Sciences* Vol.
754 **106** 9737-9742 (National Academy of Sciences, 2009).
755 73 Xu, M. & Shaw, K. L in *Genetics* Vol. **211** 1089-1104 (2019).
756 74 Yoshimura, A., Nakata, A., Mito, T. & Noji, S. in *Cytogenetic and Genome Research* Vol.
757 **112** 329-336 (2006).
758 75 Kajitani, R. *et al.* Efficient de novo assembly of highly heterozygous genomes from
759 whole-genome shotgun short reads. *Genome Res* **24**, 1384-1395,
doi:10.1101/gr.170720.113 (2014).
759 76 English, A. C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS
760 long-read sequencing technology. *PLoS One* **7**, e47768,
doi:10.1371/journal.pone.0047768 (2012).
761 77 Ellinghaus, D., Kurtz, S. & Willhoef't, U. in *BMC Bioinformatics* Vol. **9** 18 (BioMed
762 Central, 2008).
Crescente, J. M., Zavallo, D., Helguera, M. & Vanzetti, L. S. in *BMC Bioinformatics* Vol. 19, 348 (BioMed Central, 2018).

Vassetzky, N. S. & Kramerov, D. A. in *Nucleic Acids Research* Vol. 41, 2460-2461 (2013).

Edgar, R. C. in *Bioinformatics* Vol. 26, 2460-2461 (Oxford University Press, 2013).

Bao, W., Kojima, K. K. & Kohany, O. in *Mobile DNA* Vol. 6, 2460-2461 (BioMed Central, 2015).

Edgar, R. C. in *Bioinformatics* Vol. 26, 2460-2461 (2010).

Zeng, V. *et al.* in *PLoS ONE* Vol. 8 (ed Peter K. Dearden) e61479 (Public Library of Science, 2013).

Pertea, M. *et al.* in *Nature Biotechnology* Vol. 33, 290-295 (Nature Publishing Group, 2015).

Edgar, R. C. in *Nucleic Acids Research* Vol. 32, 1792-1797 (Oxford University Press, 2004).

Price, M. N., Dehal, P. S. & Arkin, A. P. in *PLoS ONE* Vol. 5, e9490 (Public Library of Science, 2010).

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, doi:10.1242/dev.084590 (2013).
Figure 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<sub>o/e</sub> 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.
Figure 2: CpG<sub>o/e</sub> distribution across insects. A) The distribution of CpG<sub>o/e</sub> values within the CDS regions displays a bimodal distribution in the two crickets, 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<sub>o/e</sub> value categories represented in yellow and blue color 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<sub>o/e</sub>) across the four insect species. The largest intersection corresponds to 2,182 OGs whose genes have low CpG<sub>o/e</sub> in the four insect species, followed by the 728 OGs whose genes have high CpG<sub>o/e</sub> levels in all four species, and 666 OGs whose genes have low CpG<sub>o/e</sub> in the three hemimetabolous species and high CpG<sub>o/e</sub> in the holometabolous honeybee. Extended plot with 50 intersections is shown in Supplementary Figure 4. C) Percentage of species-specific genes within low CpG<sub>o/e</sub> (yellow) and high CpG<sub>o/e</sub> (blue) in each insect, indicating that more such genes tend to have high CpG<sub>o/e</sub> values. D) One-to-one orthologous genes with low CpG<sub>o/e</sub> values in both crickets have significantly lower dN/dS values than genes with high CpG<sub>o/e</sub> values.
Figure 3: Functional analysis of high- and low-CpG<sub>o/e</sub> genes: Enriched GO terms with a p-value<0.00001 in at least one of the eight categories, which are high CpG<sub>o/e</sub> and low CpG<sub>o/e</sub> genes of G. bimaculatus, L. kohalensis, F. occidentalis, and A. mellifera. The dot diameter is proportional to the percentage of significant genes with the 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 Figure 3.
Figure 4: 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 orthogroup OG0000167 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.
Figure 5: 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. \(23\), 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 Table 1).
Supplementary Materials

Insights into the genomic evolution of insect from Cricket genomes

Guillem Ylla, Taro Nakamura, Takehiko Itoh, Rei Kajitani, Atsushi Toyoda, Sayuri Tomonari, Tetsuya Bando, Yoshiyasu Ishimaru, Takahito Watanabe, Masao Fuketa, Yuji Matsuoka, Austen A. Barnett, Sumihare Noji, Taro Mito, Cassandra G. Extavour

These Supplementary Materials consist of the following:

- Supplementary Figures 1 – 4 (this document)
- Supplementary File 1 (this document)
- Supplementary File 2 (“Supplementary_File_2.xls”)
- Supplementary Table 1 (“Supplementary_Table_1.xls”)
- Supplementary Table 2 (this document)
- Supplementary Table 3 and 4 (“Supplementary_Table_3-4.xls”)
- Supplementary Table 5 (this document)
- Supplementary References (this document)
Supplementary Figure 1: Schematic of *G. bimaculatus* genome annotation pipeline.

Rectangles represent data inputs: yellow rectangles represent *G. bimaculatus* data; purple rectangles represent data from other species or databases. Diamonds represent computational processes: gray diamonds indicate processes executed a single time; non-gray diamonds of the same color indicate the same process. Circles indicate outputs: blue circles indicate quality controls; green circles indicate annotations. Scripts available at GitHub https://github.com/guillemylla/Crickets_Genome_Annotation.
Supplementary Figure 2: Scheme of *L. kohalensis* genome annotation pipeline. All symbols as per Supplementary Figure 1.
Supplementary Figure 3: Enriched GO-terms among genes with high or low CpGo/e levels. This plot shows the enriched GO terms with p-value < 0.05 in at least one of the eight categories which are the high CpGo/e and low CpGo/e genes of *G. bimaculatus* (Gbi), *L. kohalensis* (Lko), *F. occidentalis* (Foc), and *A. mellifera* (Ame). The dot diameter is proportional the percentage of significant genes with the 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.
Supplementary Figure 4: UpSet plot of orthologous genes within the high and low CpG\textsubscript{o/e} value categories. Top 50 intersections of orthogroups (OGs) that are common across the eight different categories, which are the high CpG\textsubscript{o/e} and low CpG\textsubscript{o/e} genes for *G. bimaculatus* (Gbi), *L. kohalensis* (Lko), *F. Occidentalis* (Foc), and *A. mellifera* (Ame). Blue color indicates OGs that contain genes that only belong to the high CpG\textsubscript{o/e} peak, and yellow OGs contain genes that only belong to the low CpG\textsubscript{o/e} peak.
Supplementary File 1: RepeatMasker summaries. Report of the repeat content in the genomes of *G. bimaculatus* and *L. kohalensis* generated by RepeatMasker using custom libraries.

**Gryllus bimaculatus**

-------------------------------------------------------

file name: Gbimaculatus_Gap_filled.fasta
sequences: 47877

| total length: 1658007496 bp (1601517380 bp excl N/X-runs) |
| GC level: 39.93 % |
| bases masked: 558652201 bp (33.69 %) |

-------------------------------------------------------

| number of elements | length (bp) | percentage occupied of sequence |
|--------------------|-------------|---------------------------------|
| SINEs 138895       | 26406967    | 1.59 %                          |
| ALUs 6             | 9564        | 0.00 %                          |
| MIRs 0             | 0           | 0.00 %                          |
| LINEs 454301       | 147302087   | 8.88 %                          |
| LINE1 1803         | 826764      | 0.05 %                          |
| LINE2 115576       | 32029561    | 1.93 %                          |
| L3/CR1 18286       | 6358119     | 0.38 %                          |
| LTR elements 131656| 36970251    | 2.23 %                          |
| ERVL 92            | 44183       | 0.00 %                          |
| ERVL-MaLRs 0       | 0           | 0.00 %                          |
| ERV_classI 11451   | 2441461     | 0.15 %                          |
| ERV_classII 980    | 401749      | 0.02 %                          |
| DNA elements 500741| 142828465   | 8.61 %                          |
| hAT-Charlie 11512  | 4094376     | 0.25 %                          |
| TcMar-Tigger 2039  | 537995      | 0.03 %                          |
| Unclassified 367653| 126552078   | 7.63 %                          |
| Total interspersed repeats: 480059848 bp | 28.95 % |

Small RNA: 2562 1002728 bp 0.06 %

Satellites: 31087 7528498 bp 0.45 %

Simple repeats: 769175 77632578 bp 4.68 %

Low complexity: 85129 6215377 bp 0.37 %

-------------------------------------------------------

**Laupala kohalensis**

-------------------------------------------------------

file name: GCA_002313205.1_ASM231320v1_genomic.fna
sequences: 148784

| total length: 1595214429 bp (1563778341 bp excl N/X-runs) |
| GC level: 35.58 % |

937
bases masked: 566518287 bp (35.51 %)

| number of elements | length occupied of sequence |
|--------------------|-----------------------------|
| SINEs: 29510       | 7083717 bp                  |
| ALUs: 304          | 101257 bp                   |
| MIRs: 1248         | 430584 bp                   |
| LINEs: 1035151     | 322470849 bp                |
| LINE1: 941         | 367057 bp                   |
| LINE2: 584526      | 167380843 bp                |
| L3/CRI: 10257      | 4624100 bp                  |
| LTR elements: 57347 | 29690552 bp                |
| ERVL: 231          | 43500 bp                    |
| ERVL-MaLRs: 0      | 0 bp                        |
| ERV_classI: 1821   | 585650 bp                   |
| ERV_classII: 389   | 125302 bp                   |
| DNA elements: 189815 | 62384975 bp              |
| hAT-Charlie: 15008 | 5154516 bp                  |
| TcMar-Tigger: 8896 | 2459752 bp                  |
| Unclassified: 409303 | 128822550 bp              |
| Total interspersed repeats: 550452643 bp | 34.51 % |
| Small RNA: 13816 | 3005585 bp                  |
| Satellites: 2088   | 882748 bp                   |
| Simple repeats: 307925 | 19782955 bp              |
| Low complexity: 48386 | 2381730 bp               |

938
939
**Supplementary File 2: Gene family expansions in crickets.** Gene families (Orthogroups) significantly expanded in the lineage leading to crickets (tab 1), expanded in *G. bimaculatus* (tab 2), and expanded in *L. kohalensis* (tab 3). For each expanded orthogroup (OG), we report the expansion size as the number of genes gained, and the functional information about the OG. The functional information consists of the list of PFAMs and GO terms associated with the genes within the OG, and the list of *D. melanogaster* genes within the OG with their FlyBase summaries.

See file "Supplementary_File_2_GeneExpansions.xls"

**Supplementary Table 1: Genome assembly information for the 16 insect genomes analyzed.** For each genome, we show the database that the assembly was retrieved from, the assembly file name, the accession code, the assembly statistics obtained with assembly-stats software (https://github.com/sanger-pathogens/assembly-stats) and the BUSCO v3.1.0 reports at Arthropoda and Insecta levels.

See file "Supplementary_Table_1_GenomeStats.xls"
**Supplementary Table 2:** The orthogroups (OG) containing the 31 *D. melanogaster* pickpocket genes, with their FlyBase ID, symbol, and class according to Zelle et al. (2013).

| OG              | FlyBase ID     | gene symbol | Zelle et al. (2013) class |
|-----------------|----------------|-------------|---------------------------|
| OG0000361.fa    | FBgn0034965    | ppk29       | I                         |
| OG0000361.fa    | FBgn0039424    | ppk15       | I                         |
| OG0000361.fa    | FBgn0051065    | ppk31       | I                         |
| OG0000361.fa    | FBgn0053508    | ppk13       | I                         |
| OG00009052.fa   | FBgn0032602    | ppk17       | V                         |
| OG0000185.fa    | FBgn0039675    | ppk21       | III                       |
| OG0000185.fa    | FBgn0039677    | ppk30       | III                       |
| OG0000185.fa    | FBgn0039679    | ppk19       | III                       |
| OG0000185.fa    | FBgn0065109    | ppk11       | IV                        |
| OG0000185.fa    | FBgn0039676    | ppk20       | III                       |
| OG0000185.fa    | FBgn0031802    | ppk7        | III                       |
| OG0000185.fa    | FBgn0031803    | ppk14       | III                       |
| OG0000072.fa    | FBgn0022981    | rpk / ppk2  | V                         |
| OG0000072.fa    | FBgn0034730    | ppk12       | V                         |
| OG0000072.fa    | FBgn0052792    | ppk8        | V                         |
| OG0000072.fa    | FBgn0053289    | ppk5        | V                         |
| OG0000072.fa    | FBgn0020258    | ppk / ppk1  | V                         |
| OG0000072.fa    | FBgn0265001    | ppk18       | IV                        |
| OG0000072.fa    | FBgn0030795    | ppk28       | V                         |
| OG0000072.fa    | FBgn0035785    | ppk26       | V                         |
| OG0011276.fa    | FBgn0035458    | ppk27       | IV                        |
| OG0000243.fa    | FBgn0034489    | ppk6        | IV                        |
| OG0000243.fa    | FBgn0039839    | ppk24       | IV                        |
| OG0000243.fa    | FBgn0051105    | ppk22       | IV                        |
| OG0000243.fa    | FBgn0065108    | ppk16       | IV                        |
| OG0000243.fa    | FBgn0024319    | Nach / ppk4 | IV                        |
| OG0000167.fa    | FBgn0050181    | ppk3        | II                        |
| OG0000167.fa    | FBgn0053349    | ppk25       | II                        |
| OG0000167.fa    | FBgn0065110    | ppk10       | II                        |
| OG0000167.fa    | FBgn0085398    | ppk9        | II                        |
| OG0000167.fa    | FBgn0030844    | ppk23       | VI                        |
Supplementary Table 3: *pickpocket* gene expression levels in the *G. bimaculatus* prothoracic ganglion. Expression in TPMs of *fruitless* and *pickpocket* genes in each RNA-seq library generated from adult male prothoracic ganglia previously generated by Fisher and colleagues (2018). Genes with read sum across samples > 20 TPMs across samples are highlighted.

Supplementary Table 4: *pickpocket* gene expression levels in the *G. bimaculatus* embryo and regenerating legs. Expression in TPMs of *fruitless* and *pickpocket* genes in the aggregated embryo RNA-seq dataset, control legs and regenerating legs. Genes with read sum across samples > 20 TPMs across samples in the prothoracic ganglion (Supplementary Table 3) are highlighted.

See file “Supplementary_Table_3-4.xls”
**Supplementary Table 5: pickpocket genes present in previous QTL analyses examining the genetic basis for sound-based cricket courtship behavior variation.** Genomic position information for the *L. kohalensis* pickpocket genes found in linkage groups (LG) in previously published QTL analyses (Blankers, Oh & Shaw 2018; Shaw & Lesnick, 2009) examining mating song rhythm variations and female acoustic preference in the genus *Laupala*.

| Scaff names Shaw | Scaff Names NCBI | start  | end    | width  | strand | Name     | Ppk class | LG | proximity | LG | LG |
|------------------|------------------|--------|--------|--------|--------|----------|-----------|----|-----------|----|----|
| Lko057S000409    | NNCF01126148.1   | 1083057| 1116038| 32982  | +      | Lko_01144| Class IV  | 1  | LOD1      | 1  |    |
| Lko057S000550    | NNCF01126289.1   | 666338 | 667949 | 1612   | -      | Lko_06470| Class IV  | 3  | LOD2      |    |    |
| Lko057S005538    | NNCF01131273.1   | 20948  | 31450  | 10503  | -      | Lko_31867| Class V   | 4  | LOD1      |    |    |
| Lko057S005538    | NNCF01131273.1   | 6676   | 8154   | 1479   | -      | Lko_31866| Class V   | 4  | LOD1      |    |    |
| Lko057S005538    | NNCF01131273.1   | 43198  | 60736  | 17539  | -      | Lko_31869| Class V   | 4  | LOD1      |    |    |
| Lko057S00206     | NNCF01125945.1   | 353321 | 357106 | 3786   | -      | Lko_06341| Class III | 3  |          |    |    |
| Lko057S00206     | NNCF01125945.1   | 404113 | 432386 | 28274  | -      | Lko_06342| Class III | 3  |          |    |    |
Supplementary References

Blankers, T., Oh, K. P., Bombarely, A., & Shaw, K. L. (2018). The genomic architecture of a rapid Island radiation: Recombination rate variation, chromosome structure, and genome assembly of the Hawaiian cricket *Laupala*. In *Genetics* (Vol. 209, pp. 1329-1344).

Blankers, T., Oh, K. P., & Shaw, K. L. (2018). The genetics of a behavioral speciation phenotype in an Island system. In *Genes* (Vol. 9, pp. 346).

Fisher, H. P., Pascual, M. G., Jimenez, S. I., Michaelson, D. A., Joncas, C. T., Quenzer, E. D., . . . Horch, H. W. (2018). De novo assembly of a transcriptome for the cricket *Gryllus bimaculatus* prothoracic ganglion: An invertebrate model for investigating adult central nervous system compensatory plasticity. In *PLoS One* (Vol. 13, pp. e0199070).

Shaw, K. L., & Lesnick, S. C. (2009). Genomic linkage of male song and female acoustic preference QTL underlying a rapid species radiation. In *Proceedings of the National Academy of Sciences* (Vol. 106, pp. 9737-9742).

Xu, M., & Shaw, K. L. (2019). The genetics of mating song evolution underlying rapid isolation. In *Genetics* (Vol. 211, pp. 1089-1104).

Zelle, K. M., Lu, B., Pyfrom, S. C., & Ben-Shahar, Y. (2013). The genetic architecture of degenerin/epithelial sodium channels in *Drosophila*. In *G3: Genes, Genomes, Genetics* (Vol. 3, pp. 441-450).