Cricket genomes: the genomes of future food

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

1. Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, 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. Department of Molecular and Cellular Biology, Harvard University, Cambridge, USA

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

Abstract

Crickets are currently in focus as a possible source of animal protein for human consumption as an alternative to protein from vertebrate livestock. This practice could ease some of the challenges both of a worldwide growing population and of environmental issues. The two-spotted Mediterranean field cricket *Gryllus bimaculatus* has traditionally been consumed by humans in different parts of the world. Not only is this considered generally safe for human consumption, several studies also suggest that introducing crickets into one's diet may confer multiple health benefits. Moreover, *G. bimaculatus* has been widely used as a laboratory research model for decades in multiple scientific fields including evolution, developmental biology, neurobiology, and regeneration. Here we report the sequencing, assembly and annotation of the *G. bimaculatus* genome, and the annotation of the genome of the Hawaiian cricket *Laupala kohalensis*. The comparison of these two cricket genomes with those of 14 additional insects supports the hypothesis that a relatively small ancestral insect genome expanded to large sizes in many hemimetabolous lineages due to transposable element activity. Based on the ratio of observed versus expected CpG sites ($CpG_{o/e}$), we find higher conservation and stronger purifying selection 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.

Introduction

Multiple orthopteran species, and crickets in particular, are currently in focus as a source of animal protein for human consumption and for vertebrate livestock. Insect consumption, or entomophagy, is currently practiced in some populations, including some countries within Africa, Asia, and South America (Kouřimská & Adámková, 2016), but is relatively rare in most European and North American countries. The use of insects for human consumption and animal feeding could help both to decrease the emission of greenhouse gases, and to reduce the land extension considered necessary to feed the growing worldwide population. Crickets are especially attractive insects as a food source, since they are already found in the entomophagous diets of many countries (Van Huis et al., 2013), and possess high nutritional value. Crickets have a high proportion of protein for their body weight (>55%), and contain the essential linoleic acid as their most predominant fatty acid (Ghosh, Lee, Jung, & Meyer-Rochow, 2017; Kouřimská & Adámková, 2016; Van Huis et al., 2013).

The two-spotted Mediterranean field cricket *Gryllus bimaculatus* has traditionally been consumed in different parts of the world. In northeast Thailand, which recorded 20,000 insect farmers in 2011 (Hanboonsong, Jamjanya, & Durst, 2013), it is one of the most marketed and consumed insect species. Studies have reported no evidence for toxicological effects related to oral consumption of *G. bimaculatus* by humans (Ahn, Han, Kim, Hwang, & Yun, 2011; Ryu et al., 2016), neither were genotoxic effects detected using three different mutagenicity tests (Mi et al., 2005). A rare but known health risk associated with cricket consumption, however, is sensitivity and allergy to crickets (Pener, 2016; Ribeiro, Cunha, Sousa-Pinto, & Fonseca, 2018), especially in people allergic to seafood, shown by cross-allergies between *G. bimaculatus* and *Macrobrachium* prawns (Srinroch, Srisomsap, Chokchaichamnankit, Punyarit, & Phiriyangkul, 2015).

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. Water soluble compounds derived from ethanol extracts of whole adult *G. bimaculatus* applied to cultured mouse spleen cells were reported to stimulate the expression of multiple cytokines associated with immune cell proliferation and activation (Dong-Hwan et al., 2004). Rats treated with ethanol extracts of whole adult *G. bimaculatus* showed signs of reduced aging, including characteristic aging-associated gene expression profiles, and reduced levels of markers of DNA oxidative damage, (Ahn, Hwang, Yun, Kim, & Park, 2015). Glycosaminoglycans derived from *G. bimaculatus* were reported to elicit some anti-inflammatory effects in a rat model of chronic arthritis (Ahn, Han, Hwang, Yun, & Lee, 2014). Rats fed a diet including an ethanol extract of *G. bimaculatus* accumulated less abdominal fat and had lower serum glucose levels than control animals (Ahn, Kim, Kwon, Hwang, & Park, 2015). More recent studies suggest that *G. bimaculatus* powder has antidiabetic effects in rat models of Type I diabetes (Park, Lee, Lee, Hoang, & Chae, 2019) and protects against acute alcoholic liver damage in mice (Hwang et al., 2019). Beyond

rodent models, a study of healthy adult human subjects showed that the intake of 25g/day of the powdered cricket species *Grylloides sigillatus* supported growth of some probiotic microbiota, and correlated with reduced expression of the pro-imflammatory cytokine TNF-α (Stull et al., 2018).

Although crickets are becoming economically important players in the food industry, there are currently no publicly available annotated cricket genomes from any of these typically consumed species. Here, we present the 1.66-Gb genome assembly and annotation of *G. bimaculatus*, 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**).

G. bimaculatus has been widely used as a laboratory research model for decades, in scientific fields including neurobiology and neuroethology (Fisher et al., 2018; Huber, Moore, & Loher, 1989), evo-devo (Kainz, Ewen-Campen, Akam, & Extavour, 2011), developmental biology (Donoughe & Extavour, 2015), and regeneration (Mito & Noji, 2008). 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 (Kulkarni & Extavour, 2019), stable germline transgenic lines can be established (Shinmyo et al., 2004), and it has an extensive list of available experimental protocols ranging from behavioral to functional genetic analyses (Wilson Horch, Mito, Popadić, Ohuchi, & Noji, 2017).

We also report the first genome annotation for a second cricket species, the Hawaiian cricket *Laupala kohalensis*, whose genome assembly was recently made public (Blankers, Oh, Bombarely, & Shaw, 2018). Comparing these two cricket genomes 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 (Bewick, Vogel, Moore, & Schmitz, 2016; Bird, 1980), 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 (Mito & Noji, 2008) 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 hemimetabolous genomes, and in particular, to polyneopteran genomes, our assembly displays high-quality scores by a number of metrics (**Supplementary Table 1)**. Notably, the BUSCO scores (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) of this genome assembly at the arthropod and insect levels are 98.50% and 97.00% 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
BUSCO Score – Arthropoda	98.50%
BUSCO Score – Insecta	97.00%

Annotation of two cricket genomes

The publicly available 1.6-Gb genome assembly of the Hawaiian cricket *L. kohalensis* (Blankers, Oh, Bombarely, et al., 2018), although having lower assembly statistics than that of *G. bimaculatus* (N50=0.58 Mb, L90 = 3,483), 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 (Holt & Yandell, 2011), in which we combined *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 *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 (Jones et al., 2014) 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 graphic interface through which interested readers can access, search, BLAST and download the genome data and annotations (http://34.71.36.157:3838/).

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%
BUSCO-transcriptome Score – Insecta	92.30%	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 (Smit, Hubley, & Grenn, 2015) 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 two scaffolds that contained 1.75x-1.82x the density of repetitive content than the mean of the other N90 scaffolds **(Figure 1B)**. Transposable elements (TEs) accounted for 28.94% of this repetitive content in the *G. bimaculatus* genome, and for 34.50% of the repetitive content in the *L. kohalensis* genome. Although the overall proportion of repetitive content 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, while in *G. bimaculatus* LINEs made up only 8.88% of the genome. 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 (Bird, 1980). 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 2A**). 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 CpGo/e values in the honeybee *Apis mellifera*. This bee was the first insect for which evidence for DNA methylation was robustly described and studied (Elango, Hunt, Goodisman, & Yi, 2009; Y. Wang et al., 2006). We found that in *A. mellifera,* CpG-depleted genes were enriched for similar functions as those observed in cricket CpG-depleted genes (26 GO-terms were significantly enriched in both honeybee and crickets; **Supplementary Figure 3**). In the same way, high $CpG_{o/e}$ genes in both crickets and honeybee were enriched for similar functions (12 GO-terms commonly enriched; **Supplementary Figure 3)**.

Additionally, we observed that genes belonging to the low $CpG_{o/e}$ peak were more likely to have an orthologous gene in another insect species, and that ortholog was also more likely

to belong to the low $CpG_{o/e}$ peak (**Figure 2B and Supplementary Figure 4**). By contrast, genes with high $CpG_{o/e}$, were more likely to be species-specific, but if they had an ortholog in another species, this ortholog was also likely to have high $CpG_{o/e}$. This suggests 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 2C**), 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 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 (Misof et al., 2014). 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.

The *pickpocket* gene family appeared to be a significantly expanded gene family in crickets. Following the classification of *pickpocket* genes used in *Drosophila spp.* (Zelle, Lu, Pyfrom, & Ben-Shahar, 2013) 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* (Zelle et al., 2013). 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 (Adams et al., 1998). For example, in larvae, this gene is required for mechanical nociception (Zhong, Hwang, & Tracey, 2010), and for coordinating rhythmic locomotion (Ainsley et al., 2003). *ppk* is expressed in sensory neurons that also express the male sexual behavior determiner *fruitless (fru)* (Häsemeyer, Yapici, Heberlein, & Dickson, 2009; Pavlou & Goodwin, 2013; Rezával et al., 2012).

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 the RNA-seq libraries for the *G. bimaculatus* prothoracic ganglion (Fisher et al., 2018). This analysis detected expression (>5 FPKMs) of six *pickpocket* genes, four of them belonging to class V, in the *G. bimaculatus* nervous system. In the same RNA-seq libraries, we also detected the expression of *fru* **(Supplementary Table 3**).

Discussion

The importance of cricket genomes

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 (Y. H. Chen, Gols, & Benrey, 2015; Gepts, 2004; Thrall, Bever, & Burdon, 2010; Yamasaki et al., 2005). 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 (K. Chen & Gao, 2014). 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 (Qaim, 2009; Thrall et al., 2010). Crickets are naturally nutritionally rich (Ghosh et al., 2017), 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 REF. These challenges include sensitivity to common insect viruses, aggressive behavior resulting in cannibalism, complex mating rituals, and relatively slow growth rate.

An essential tool for any kind of genetic engineering is a high quality annotated reference genome, together with a deep understanding of the biology of the given species. Because *G. bimaculatus* has been used as a research model in multiple different scientific disciplines, including rearing for consumption, issues relevant to its biochemical composition (Ghosh et al., 2017), human health and safety (Ahn et al., 2011; Ryu et al., 2016), putative health benefits (Ahn et al., 2014; Ahn, Hwang, et al., 2015; Ahn, Kim, et al., 2015; Dong-Hwan et al., 2004; Hwang et al., 2019; Park et al., 2019), and processing techniques (Dobermann, Field, & Michaelson, 2019) have been extensively described. Thus, the genome of *G. bimaculatus* herein described, adds to this body of biological knowledge by providing invaluable information that will be required to maximize the potential of this cricket to become an increasingly significant part of the worldwide diet in the future.

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) (X. Wang et al., 2014) and the termites *Cryptotermes secundus* (18,162) (Harrison et al., 2018), *Macrotermes natalensis* (16,140) (Poulsen et al., 2014) and *Zootermopsis nevadensis*, (15,459) (Terrapon et al., 2014). 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, 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* (X. Wang et al., 2014). 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 (Chénais, Caruso, Hiard, & Casse, 2012; Kidwell, 2002).

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 (Hanrahan & Johnston, 2011). 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* (Camacho et al., 2015), is likely due to TE activity since the time of the last orthopteran ancestor.

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 (Gregory, 2002). Our observations are consistent with the first of these hypotheses.

Larger genome size correlates with slower developmental rates in some plants and animals, which is hypothesized to be due to a slower cell division rate (Gregory, 2002). Thus, one may speculate that the large proportion of TE-derived DNA in cricket genomes might negatively impact the developmental rate. If this repetitive DNA proves largely non-functional, then in principle, the developmental rate, an important factor for insect farming, might eventually be modifiable with the use of genetic engineering techniques.

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 (Lyko, Ramsahoye, & Jaenisch, 2000). The honeybee *A. mellifera* was one of the first insects for which functional DNA methylation was described (Y. Wang et al., 2006). Although this DNA modification was initially proposed to be associated with the eusociality of these bees

(Elango et al., 2009), subsequent studies showed that DNA methylation is widespread and present in different insect lineages independently of social behavior (Bewick et al., 2016). DNA methylation also occurs in other non-insect arthropods (Thomas et al., 2020).

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 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.

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 (Ylla, Piulachs, & Belles, 2018). These results in cockroaches, together with our observations, leads 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 (Adams et al., 1998). 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 (Liu, Johnson, & Welsh, 2003). 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 (Lee et al., 2017), pheromone detection (Averhoff, Richardson, Starostina, Kinser, & Pikielny, 1976), courtship behavior (Lu, LaMora, Sun, Welsh, & Ben-Shahar, 2012), and liquid clearance in the larval trachea (Liu et al., 2003).

In *D. melanogaster* adults, the abdominal ganglia mediate courtship and postmating behaviors through neurons expressing *ppk* and *fru* (Häsemeyer et al., 2009; Pavlou & Goodwin, 2013; Rezával et al., 2012). In *D. melanogaster* larvae, *ppk* expression in dendritic neurons is required to control the coordination of rhythmic locomotion (Ainsley et al., 2003). In crickets, the abdominal ganglia are responsible for determining song rhythm (Jacob & Hedwig, 2016). 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

391 that class V *pickpocket* genes could be involved in song rhythm determination in crickets
392 through their expression in abdominal ganglia.

393 This possibility is consistent with the results of multiple quantitative trait locus (QTL)
394 studies done in cricket species from the genus *Laupala*, which identified genomic regions
395 associated with mating song rhythm variations and female acoustic preference (Blankers,
396 Oh, & Shaw, 2018). The 179 scaffolds that the authors reported being within one LOD of the
397 seven QTL peaks, contained five *pickpocket* genes, three of them from class V and two from
398 class IV. One of the two class IV genes also appears within a QTL peak of a second experiment
399 (Blankers, Oh, Bombarely, et al., 2018; Shaw & Lesnick, 2009). Xu and Shaw (2019) found
400 that a scaffold in a region of LOD score 1.5 of one of their minor linkage groups (LG3) contains
401 *slowpoke*, a gene that affects song interpulse interval in *D. melanogaster,* and this scaffold
402 also contains two class III pickpocket genes (**Supplementary Table 4**).

403 In summary, the roles of *pickpocket* genes in controlling rhythmic locomotion, courtship
404 behavior, and pheromone detection in *D. melanogaster*, their appearance in genomic regions
405 associated with song rhythm variation in *Laupala*, and their expression in *G. bimaculatus*
406 abdominal ganglia, lead us to speculate that the expanded *pickpocket* gene family in cricket
407 genomes could be playing a role in regulating rhythmic wing movements and sound
408 perception, both of which are necessary for mating (Wilson Horch et al., 2017). We note that
409 Xu and Shaw (2019) hypothesized that song production in crickets is likely to be regulated
410 by ion channels, and that locomotion, neural modulation, and muscle development are all
411 involved in singing (Xu & Shaw, 2019). However, further experiments, which could take
412 advantage of the existing RNAi and genome modification protocols for *G. bimaculatus*
413 (Kulkarni & Extavour, 2019), will be required to test this hypothesis.

414

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

423

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.

Genome Assembly

Paired-end libraries were generated with insert sizes of 375 and 500 bp, and mate-pair library 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) (Kajitani et al., 2014). 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) (English et al., 2012).

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 (Ellinghaus, Kurtz, & Willhoeft, 2008), RepeatModeler/RepeatClassifier (www.repeatmasker.org/RepeatModeler), MITE tracker (Crescente, Zavallo, Helguera, & Vanzetti, 2018), TransposonPSI (http://transposonpsi.sourceforge.net), and the databases SINEBase (Vassetzky & Kramerov, 2013) and RepBase (Bao, Kojima, & Kohany, 2015). We removed redundancies from the library by merging sequences that were greater than 80% similar with usearch (Robert C. Edgar, 2010), and classified them with RepeatClassifier. Sequences classified as “unknown” were BLASTed (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 (Holt & Yandell, 2011) 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 (Zeng et al., 2013), an assembled prothoracic ganglion transcriptome (Fisher et al., 2018), and a genome-guided transcriptome generated with StringTie (Pertea et al., 2015) using 84 RNA-seq libraries (accession numbers: XXXX) mapped to the genome with HISAT2 (Kim, Langmead, & Salzberg, 2015). 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 (UniProt, 2019).

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 (Ter-Hovhannisyan, Lomsadze, Chernoff, & Borodovsky, 2008) in self-training mode, and with Augustus (Stanke & Waack, 2003) trained with BUSCO v3 (Simão et al., 2015). After each of the first two MAKER2 iterations, additional gene models were obtained with SNAP (Korf, 2004) trained with the annotated genes.

Functional annotations were obtained using InterProScan (Jones et al., 2014), which retrieved the InterProDomains, PFAM domains, and GO-terms. Additionally, we ran a series of BLAST rounds 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) were later BLASTed against all proteins from UniProtKB/TrEMBL, and those without a hit with E-value<1e-6 were BLASTed 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) (Simão et al., 2015) was used to assess the level

of completeness of the genome assemblies ('-m geno') as well as that of the gene annotations ('-m tran') 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 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 $CpG_{o/e}$ 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 (**Figures 2 & 4**).

The distributions of gene $CpG_{o/e}$ values per gene of the two crickets and the honeybee *A. mellifera* were fitted with a mixture of normal distributions using the mixtools R package (Benaglia, Chauveau, Hunter, & Young, 2009). 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 (Alexa & Rahnenfuhrer, 2019) with the weight01 algorithm and the Fisher statistic. GO-terms with a p-value<0.05 were plotted as word-clouds using the R package ggwordcloud (Pennec & Slowikowski, 2018) with the size of the word correlated with the proportion of the term within the set.

For each of the genes belonging to low and high $CpG_{o/e}$ categories in each of the three 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 (Lex, Gehlenborg, Strobelt, Vuillemot, & Pfister, 2014) 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. Then, the amino-acid alignments were transformed into codon-based nucleotide alignments using the Pal2Nal software (Suyama, Torrents, & Bork, 2006). 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 (Yang, 2007). 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 orthogroups (which we also refer to herein as "gene families") using OrthoFinder v2.3.3 (Emms & Kelly, 2019). 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 (Robert C Edgar, 2004) and the alignments trimmed with GBlocks (Castresana, 2000). The trimmed alignments were concatenated into a single meta-alignment that was used to infer the species tree with FastTree2 (Price, Dehal, & Arkin, 2010).

To calibrate the species tree, we used the "chronos" function from the R package ape v5.3 (Paradis & Schliep, 2019), 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 (Misof et al., 2014).

The gene family expansion/contraction analysis was done with the CAFE software (De Bie, Cristianini, Demuth, & Hahn, 2006). 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 (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 *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 (Thurmond et al., 2019). 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

The functional annotation of significantly expanded gene families in crickets allowed us to identify an orthogroup containing orthologs of *D. melanogaster pickpocket* class V genes.

Subsequently, we retrieved the 6 additional orthogroups containing the complete set of *pickpocket* genes in *D. melanogaster* according to FlyBase. The protein sequences of the members of the 7 Pickpocket orthogroups were aligned with MUSCLE, and the *pickpocket* gene tree obtained with FastTree. Following the *pickpocket* categorization described for *Drosophila spp.* (Zelle et al., 2013) and the obtained *pickpocket* gene tree, we classified the crickets *pickpocket* genes into classes from I to VI.

To check for evidence of expression *pickpocket* genes in the cricket nervous system, we used the 22 RNA-seq libraries from prothoracic ganglion (Fisher et al., 2018) of *G. bimaculatus* available at NCBI GEO (PRJNA376023). Reads were mapped against the *G. bimaculatus* genome with RSEM (Li & Dewey, 2011) using STAR (Dobin et al., 2013) as the mapping algorithm, and the number of expected counts and FPKMs was retrieved for each gene in each library. The FPKMs of the *pickpocket* genes and *fruitless* is shown in **Supplementary Table 3**. Genes with a sum of more than five FPKMs across all samples were considered to be expressed in *G. bimaculatus* prothoracic ganglion.

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 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 and TB 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 assembly and gene annotations for *Gryllus bimaculatus* were submitted to DDBJ and to NCBI under the accession number (**XXXXXX**). The scripts used for genome annotation and analysis are available at GitHub (https://github.com/guillemylla/Crickets_Genome_Annotation).).

598 # References
599

600 Adams, C. M., Anderson, M. G., Motto, D. G., Price, M. P., Johnson, W. A., & Welsh, M. J. (1998).
601 Ripped pocket and pickpocket, novel *Drosophila* DEG/ENaC subunits expressed in
602 early development and in mechanosensory neurons. In *The Journal of Cell Biology*
603 (Vol. 140, pp. 143-152).
604 Ahn, M. Y., Han, J. W., Hwang, J. S., Yun, E. Y., & Lee, B. M. (2014). Anti-inflammatory effect of
605 glycosaminoglycan derived from *Gryllus bimaculatus* (A type of cricket, insect) on
606 adjuvant-treated chronic arthritis rat model. In *Journal of Toxicology and*
607 *Environmental Health - Part A: Current Issues* (Vol. 77, pp. 1332-1345).
608 Ahn, M. Y., Han, J. W., Kim, S. J., Hwang, J. S., & Yun, E. Y. (2011). Thirteen-week oral dose
609 toxicity study of *G. bimaculatus* in sprague-dawley rats. In *Toxicological Research*
610 (Vol. 27, pp. 231-240).
611 Ahn, M. Y., Hwang, J. S., Yun, E. Y., Kim, M. J., & Park, K. K. (2015). Anti-aging effect and gene
612 expression profiling of aged rats treated with *G. bimaculatus* extract. In *Toxicological*
613 *Research* (Vol. 31, pp. 173-180).
614 Ahn, M. Y., Kim, M. J., Kwon, R. H., Hwang, J. S., & Park, K. K. (2015). Gene expression
615 profiling and inhibition of adipose tissue accumulation of *G. bimaculatus* extract in
616 rats on high fat diet. In *Lipids in Health and Disease* (Vol. 14, pp. 116).
617 Ainsley, J. A., Pettus, J. M., Bosenko, D., Gerstein, C. E., Zinkevich, N., Anderson, M. G., . . .
618 Johnson, W. A. (2003). Enhanced Locomotion Caused by Loss of the *Drosophila*
619 DEG/ENaC Protein Pickpocket1. In *Current Biology* (Vol. 13, pp. 1557-1563).
620 Alexa, A., & Rahnenfuhrer, J. (2019). topGO: Enrichment Analysis for Gene Ontology. *R*
621 *package version 2.36.0*.
622 Averhoff, W. W., Richardson, R. H., Starostina, E., Kinser, R. D., & Pikielny, C. W. (1976).
623 Multiple pheromone system controlling mating in *Drosophila melanogaster*. In
624 *Proceedings of the National Academy of Sciences of the United States of America* (Vol.
625 73, pp. 591-593).
626 Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive
627 elements in eukaryotic genomes. In *Mobile DNA* (Vol. 6, pp. 11).
628 Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. S. (2009). Mixtools: An R package for
629 analyzing finite mixture models. In *Journal of Statistical Software* (Vol. 32, pp. 1-29).
630 Bewick, A. J., Vogel, K. J., Moore, A. J., & Schmitz, R. J. (2016). Evolution of DNA Methylation
631 across Insects. In *Molecular Biology and Evolution* (Vol. 34, pp. 654-665).
632 Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. In *Nucleic*
633 *Acids Research* (Vol. 8, pp. 1499-1504).
634 Blankers, T., Oh, K. P., Bombarely, A., & Shaw, K. L. (2018). The genomic architecture of a
635 rapid Island radiation: Recombination rate variation, chromosome structure, and
636 genome assembly of the hawaiian cricket *Laupala*. In *Genetics* (Vol. 209, pp. 1329-
637 1344).
638 Blankers, T., Oh, K. P., & Shaw, K. L. (2018). The genetics of a behavioral speciation
639 phenotype in an Island system. In *Genes* (Vol. 9, pp. 346).
640 Camacho, J. P. M. M., Ruiz-Ruano, F. J., Martín-Blázquez, R., López-León, M. D., Cabrero, J.,
641 Lorite, P., . . . Bakkali, M. (2015). A step to the gigantic genome of the desert locust:
642 chromosome sizes and repeated DNAs. In *Chromosoma* (Vol. 124, pp. 263-275).
643 Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use
644 in Phylogenetic Analysis. In *Molecular Biology and Evolution* (Vol. 17, pp. 540-552).
645 Chen, K., & Gao, C. (2014). Targeted genome modification technologies and their
646 applications in crop improvements. *Plant Cell Reports, 33*(4), 575-583.
647 Chen, Y. H., Gols, R., & Benrey, B. (2015). Crop domestication and its impact on naturally
648 selected trophic interactions. *Annu Rev Entomol, 60*, 35-58.
649 Chénais, B., Caruso, A., Hiard, S., & Casse, N. (2012). The impact of transposable elements on
650 eukaryotic genomes: From genome size increase to genetic adaptation to stressful
651 environments. In *Gene* (Vol. 509, pp. 7-15).
652 Crescente, J. M., Zavallo, D., Helguera, M., & Vanzetti, L. S. (2018). MITE Tracker: an accurate
653 approach to identify miniature inverted-repeat transposable elements in large
654 genomes. In *BMC Bioinformatics* (Vol. 19, pp. 348).
655 De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a computational tool
656 for the study of gene family evolution. In *Bioinformatics* (Vol. 22, pp. 1269-1271).

Dobermann, D., Field, L. M., & Michaelson, L. V. (2019). Impact of heat processing on the nutritional content of *Gryllus bimaculatus* (black cricket). In *Nutrition Bulletin* (Vol. 44, pp. 116-122).

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. In *Bioinformatics* (Vol. 29, pp. 15-21).

Dong-Hwan, S., HWANG, S.-Y., HAN, J., KOH, S.-K., KIM, I., RYU, K. S., & YUN, C.-Y. (2004). Immune-Enhancing Activity Screening on Extracts from Two Crickets, *Gryllus bimaculatus* and *Teleogryllus emma*. In *Entomological Research* (Vol. 34, pp. 207-211).

Donoughe, S., & Extavour, C. G. (2015). Embryonic development of the cricket *Gryllus bimaculatus*. In *Developmental Biology* (Vol. 411, pp. 140-156).

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. In *Nucleic Acids Research* (Vol. 32, pp. 1792-1797).

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. In *Bioinformatics* (Vol. 26, pp. 2460-2461).

Elango, N., Hunt, B. G., Goodisman, M. A. D., & Yi, S. V. (2009). DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 106, pp. 11206-11211).

Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. In *BMC Bioinformatics* (Vol. 9, pp. 18).

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol, 20*(1), 238.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., . . . Gibbs, R. A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One, 7*(11), e47768.

Fisher, H. P., Pascual, M. G., Jimenez, S. I., Michaelson, D. A., Joncas, C. T., Quenzer, E. D., Christie, A.E. & 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. *PLoS One* (Vol. 13, pp. e0199070).

Gepts, P. (2004). Crop domestication as a long-term selection experiment. *Plant breeding reviews, 24*(2), 1-44.

Ghosh, S., Lee, S.-M., Jung, C., & Meyer-Rochow, V. (2017). Nutritional composition of five commercial edible insects in South Korea. *Journal of Asia-Pacific Entomology, 20*(2), 686-694.

Gregory, T. R. (2002). Genome size and developmental complexity. *Genetica, 115*(1), 131-146.

Hanboonsong, Y., Jamjanya, T., & Durst, P. B. (2013). Six-legged livestock : edible insect farming , collecting and marketing in Thailand. In *Office* (pp. 69).

Hanrahan, S. J., & Johnston, J. S. (2011). New genome size estimates of 134 species of arthropods. In *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* (Vol. 19, pp. 809-823).

Harrison, M. C., Jongepier, E., Robertson, H. M., Arning, N., Bitard-Feildel, T., Chao, H., . . . Bornberg-Bauer, E. (2018). Hemimetabolous genomes reveal molecular basis of termite eusociality. *Nature ecology & evolution*.

Häsemeyer, M., Yapici, N., Heberlein, U., & Dickson, B. J. (2009). Sensory Neurons in the *Drosophila* Genital Tract Regulate Female Reproductive Behavior. In *Neuron* (Vol. 61, pp. 511-518).

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

Huber, F., Moore, T. E. T. E., & Loher, W. (1989). Cricket behavior and neurobiology. 565.

Hwang, B. B., Chang, M. H., Lee, J. H., Heo, W., Kim, J. K., Pan, J. H., . . . Kim, J. H. (2019). The edible insect *Gryllus bimaculatus* protects against gut-derived inflammatory responses and liver damage in mice after acute alcohol exposure. In *Nutrients* (Vol. 11, pp. 857).

Jacob, P. F., & Hedwig, B. (2016). Acoustic signalling for mate attraction in crickets: Abdominal ganglia control the timing of the calling song pattern. In *Behavioural Brain Research* (Vol. 309, pp. 51-66).

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., . . . Hunter, S. (2014). InterProScan 5: Genome-scale Protein Function Classification. In *Bioinformatics* (pp. 1-5).

Kainz, F., Ewen-Campen, B., Akam, M., & Extavour, C. G. (2011). Notch/Delta signalling is not required for segment generation in the basally branching insect *Gryllus bimaculatus*. *Development, 138*(22), 5015-5026.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., . . . Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res, 24*(8), 1384-1395.

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. In *Genetica* (Vol. 115, pp. 49-63).

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*.

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

Kouřimská, L., & Adámková, A. (2016). Nutritional and sensory quality of edible insects. In *NFS Journal* (Vol. 4, pp. 22-26).

Kulkarni, A., & Extavour, C. G. (2019). The Cricket *Gryllus bimaculatus*: Techniques for Quantitative and Functional Genetic Analyses of Cricket Biology. In *Evo-Devo: Non-model Species in Cell and Developmental Biology* (Vol. 68, pp. 183-216): Springer.

Lee, M. J., Sung, H. Y., Jo, H., Kim, H.-W., Choi, M. S., Kwon, J. Y., & Kang, K. (2017). Ionotropic Receptor 76b Is Required for Gustatory Aversion to Excessive Na+ in *Drosophila*. In *Molecules and cells* (Vol. 40, pp. 787-795).

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014). UpSet: Visualization of intersecting sets. In *IEEE Transactions on Visualization and Computer Graphics* (Vol. 20, pp. 1983-1992).

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. In *BMC Bioinformatics* (Vol. 12, pp. 323).

Liu, L., Johnson, W. A., & Welsh, M. J. (2003). *Drosophila* DEG/ENaC pickpocket genes are expressed in the tracheal system, where they may be involved in liquid clearance. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 100, pp. 2128-2133).

Lu, B., LaMora, A., Sun, Y., Welsh, M. J., & Ben-Shahar, Y. (2012). ppk23-Dependent Chemosensory Functions Contribute to Courtship Behavior in *Drosophila melanogaster*. In M. B. Goodman (Ed.), *PLoS Genetics* (Vol. 8, pp. e1002587).

Lyko, F., Ramsahoye, B. H., & Jaenisch, R. (2000). DNA methylation in Drosophila melanogaster. *Nature, 408*, 538-540.

Mi, Y. A., Hye, J. B., In, S. K., Eun, J. Y., Seung, J. K., Hyung, S. K., . . . Byung, M. L. (2005). Genotoxic evaluation of the biocomponents of the cricket, *Gryllus bimaculatus*, using three mutagenicity tests. In *Journal of Toxicology and Environmental Health - Part A* (Vol. 68, pp. 2111-2118).

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., . . . Zhou, X. (2014). Phylogenomics resolves the timing and pattern of insect evolution. In *Science* (Vol. 346, pp. 763-767).

Mito, T., & Noji, S. (2008). The Two-Spotted Cricket *Gryllus bimaculatus*: An Emerging Model for Developmental and Regeneration Studies. In *CSH protocols* (Vol. 2008, pp. pdb.emo110).

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. In R. Schwartz (Ed.), *Bioinformatics* (Vol. 35, pp. 526-528).

Park, S. A., Lee, G. H., Lee, H. Y., Hoang, T. H., & Chae, H. J. (2019). Glucose-lowering effect of *Gryllus bimaculatus* powder on streptozotocin-induced diabetes through the AKT/mTOR pathway. In *Food Science and Nutrition* (Vol. 8, pp. 402-409).

Pavlou, H. J., & Goodwin, S. F. (2013). Courtship behavior in *Drosophila melanogaster*: towards a 'courtship connectome'. In *Current opinion in neurobiology* (Vol. 23, pp. 76-83).

Allergy to Crickets: A Review, 25 91-95 (2016).

Pennec, E., & Slowikowski, K. (2018). ggwordcloud: A Word Cloud Geom for 'ggplot2'. *R package version 0.5. 0. URL https://cran.r-project.org/package=ggwordcloud*.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. In *Nature Biotechnology* (Vol. 33, pp. 290-295).

Poulsen, M., Hu, H., Li, C., Chen, Z., Xu, L., Otani, S., . . . Zhang, G. (2014). Complementary symbiont contributions to plant decomposition in a fungus-farming termite. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 111, pp. 14500-14505).

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. In A. F. Y. Poon (Ed.), *PLoS One* (Vol. 5, pp. e9490).

Qaim, M. (2009). The Economics of Genetically Modified Crops. *Annual Review of Resource Economics, 1*(1), 665-694.

Rezával, C., Pavlou, H. J., Dornan, A. J., Chan, Y.-B., Kravitz, E. A., & Goodwin, S. F. (2012). Neural circuitry underlying *Drosophila* female postmating behavioral responses. In *Current Biology* (Vol. 22, pp. 1155-1165).

Allergic risks of consuming edible insects: A systematic review, 62 1700030 (2018).

Ryu, H. Y., Lee, S., Ahn, K. S., Kim, H. J., Lee, S. S., Ko, H. J., . . . Song, K. S. (2016). Oral toxicity study and skin sensitization test of a cricket. In *Toxicological Research* (Vol. 32, pp. 159-173).

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).

Shinmyo, Y., Mito, T., Matsushita, T., Sarashina, I., Miyawaki, K., Ohuchi, H., & Noji, S. (2004). piggyBac-mediated somatic transformation of the two-spotted cricket, *Gryllus bimaculatus*. In *Development, Growth and Differentiation* (Vol. 46, pp. 343-349).

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

Smit, A., Hubley, R., & Grenn, P. (2015). RepeatMasker Open-4.0. *RepeatMasker Open-4.0.7.*

Srinroch, C., Srisomsap, C., Chokchaichamnankit, D., Punyarit, P., & Phiriyangkul, P. (2015). Identification of novel allergen in edible insect, *Gryllus bimaculatus* and its cross-reactivity with Macrobrachium spp. allergens. In *Food Chemistry* (Vol. 184, pp. 160-166).

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. In *Bioinformatics* (Vol. 19, pp. ii215-ii225).

Stull, V. J., Finer, E., Bergmans, R. S., Febvre, H. P., Longhurst, C., Manter, D. K., . . . Weir, T. L. (2018). Impact of Edible Cricket Consumption on Gut Microbiota in Healthy Adults, a Double-blind, Randomized Crossover Trial. In *Scientific Reports* (Vol. 8, pp. 10762).

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. In *Nucleic Acids Research* (Vol. 34, pp. W609-W612).

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., & Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. In *Genome Research* (Vol. 18, pp. 1979-1990).

Terrapon, N., Li, C., Robertson, H. M., Ji, L., Meng, X., Booth, W., . . . Liebig, J. (2014). Molecular traces of alternative social organization in a termite genome. In *Nat Commun* (Vol. 5, pp. 3636).

Thomas, G. W. C., Dohmen, E., Hughes, D. S. T., Murali, S. C., Poelchau, M., Glastad, K., . . . Richards, S. (2020). Gene content evolution in the arthropods. *Genome Biol, 21*(1), 15.

Thrall, P. H., Bever, J. D., & Burdon, J. J. (2010). Evolutionary change in agriculture: the past, present and future. *Evol Appl, 3*(5-6), 405-408.

Thurmond, J., Goodman, J. L., Strelets, V. B., Attrill, H., Gramates, L. S., Marygold, S. J., . . . FlyBase, C. (2019). FlyBase 2.0: the next generation. *Nucleic Acids Res, 47*(D1), D759-D765.

UniProt, C. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res, 47*(D1), D506-D515.

Van Huis, A., Van Itterbeeck, J., Klunder, H., Mertens, E., Halloran, A., Muir, G., & Vantomme, P. (2013). *Edible insects: future prospects for food and feed security*: Food and agriculture organization of the United Nations (FAO).

838 Vassetzky, N. S., & Kramerov, D. A. (2013). SINEBase: a database and tool for SINE analysis.
839 In *Nucleic acids research* (Vol. 41, pp. D83-89).
840 Wang, X., Fang, X., Yang, P., Jiang, X., Jiang, F., Zhao, D., . . . Kang, L. (2014). The locust
841 genome provides insight into swarm formation and long-distance flight. In *Nature*
842 *communications* (Vol. 5, pp. 2957).
843 Wang, Y., Jorda, M., Jones, P. L., Maleszka, R., Ling, X., Robertson, H. M., . . . Robinson, G. E.
844 (2006). Functional CpG Methylation System in a Social Insect. In *Science* (Vol. 314,
845 pp. 645-647).
846 Wilson Horch, H., Mito, T., Popadić, A., Ohuchi, H., & Noji, S. (2017). The cricket as a model
847 organism: Development, regeneration, and behavior. In H. W. Horch, T. Mito, A.
848 Popadić, H. Ohuchi, & S. Noji (Eds.), *The Cricket as a Model Organism: Development,*
849 *Regeneration, and Behavior* (pp. 1-376). Tokyo.
850 Xu, M., & Shaw, K. L. (2019). The genetics of mating song evolution underlying rapid
851 speciation: Linking quantitative variation to candidate genes for behavioral
852 isolation. In *Genetics* (Vol. 211, pp. 1089-1104).
853 Yamasaki, M., Tenaillon, M. I., Bi, I. V., Schroeder, S. G., Sanchez-Villeda, H., Doebley, J. F., . . .
854 McMullen, M. D. (2005). A large-scale screen for artificial selection in maize
855 identifies candidate agronomic loci for domestication and crop improvement. *Plant*
856 *Cell, 17*(11), 2859-2872.
857 Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. In *Molecular*
858 *Biology and Evolution* (Vol. 24, pp. 1586-1591).
859 Ylla, G., Piulachs, M.-D., & Belles, X. (2018). Comparative transcriptomics in two extreme
860 neopterans reveal general trends in the evolution of modern insects. In *iScience* (Vol.
861 4, pp. 164-179).
862 Zelle, K. M., Lu, B., Pyfrom, S. C., & Ben-Shahar, Y. (2013). The genetic architecture of
863 degenerin/epithelial sodium channels in *Drosophila*. In *G3: Genes, Genomes, Genetics*
864 (Vol. 3, pp. 441-450).
865 Zeng, V., Ewen-Campen, B., Horch, H. W., Roth, S., Mito, T., & Extavour, C. G. (2013).
866 Developmental Gene Discovery in a Hemimetabolous Insect: De Novo Assembly and
867 Annotation of a Transcriptome for the Cricket *Gryllus bimaculatus*. In P. K. Dearden
868 (Ed.), *PLoS One* (Vol. 8, pp. e61479).
869 Zhong, L., Hwang, R. Y., & Tracey, W. D. (2010). Pickpocket Is a DEG/ENaC Protein Required
870 for Mechanical Nociception in *Drosophila* Larvae. In *Current Biology* (Vol. 20, pp.
871 429-434).

872

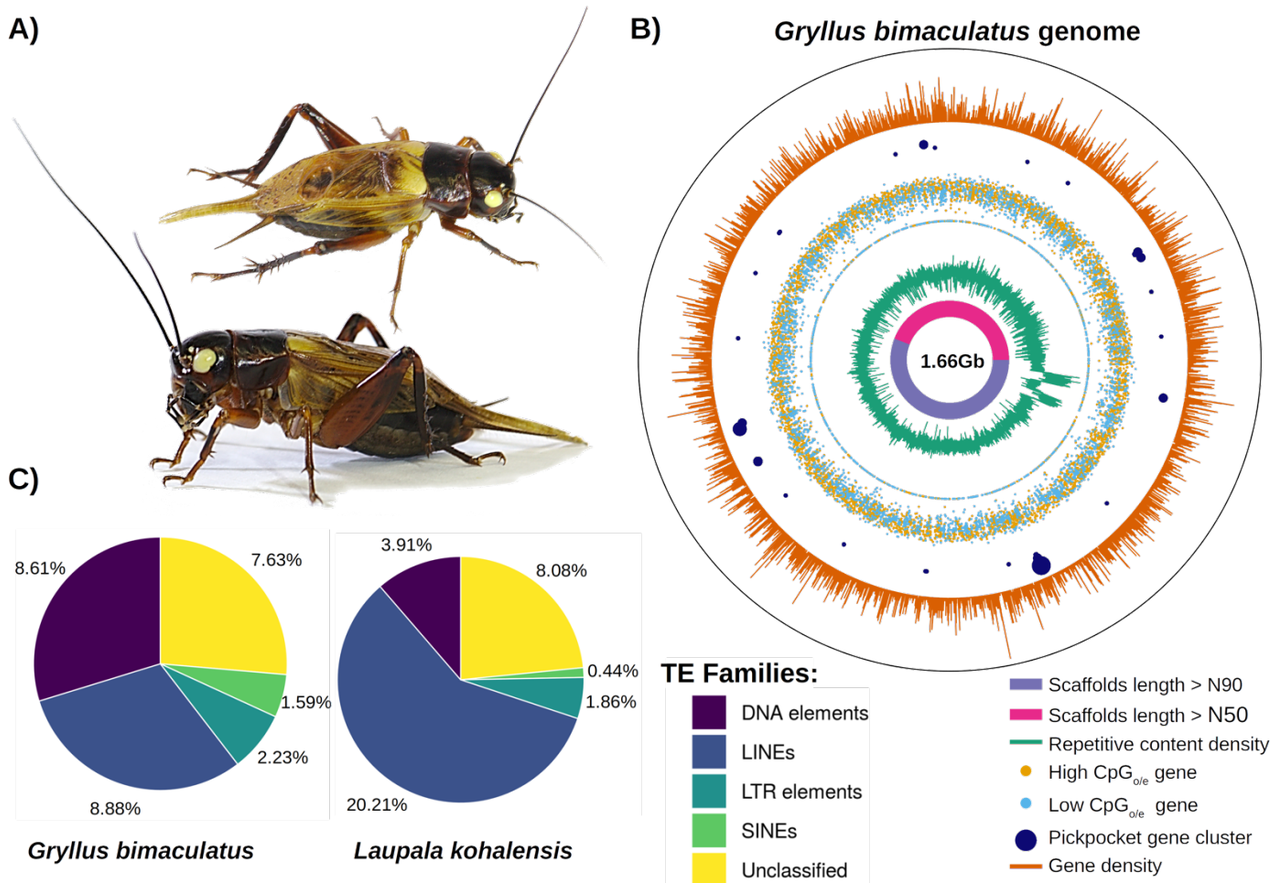


Figure 1: The *G. bimaculatus* genome. A) The cricket *G. bimaculatus* (top and side views), commonly called the two-spotted cricket, owes its name to the two yellow spots on the base of the forewings. **B)** Circular representation of the N50 (pink) and N90 (purple) scaffolds, repetitive content density (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 different families of transposable elements is different between the cricket species *G. bimaculatus* and *L. kohalensis*.

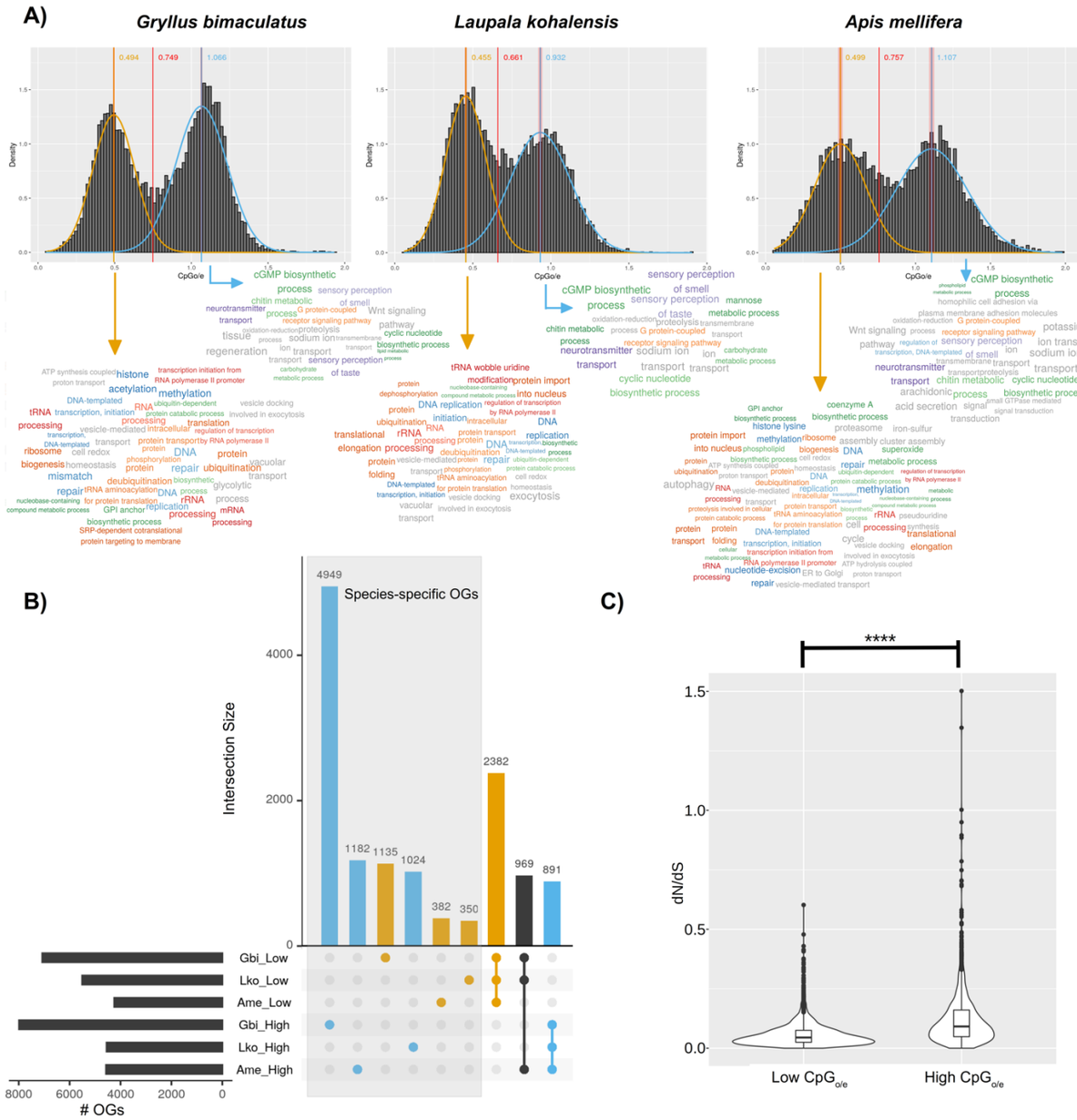
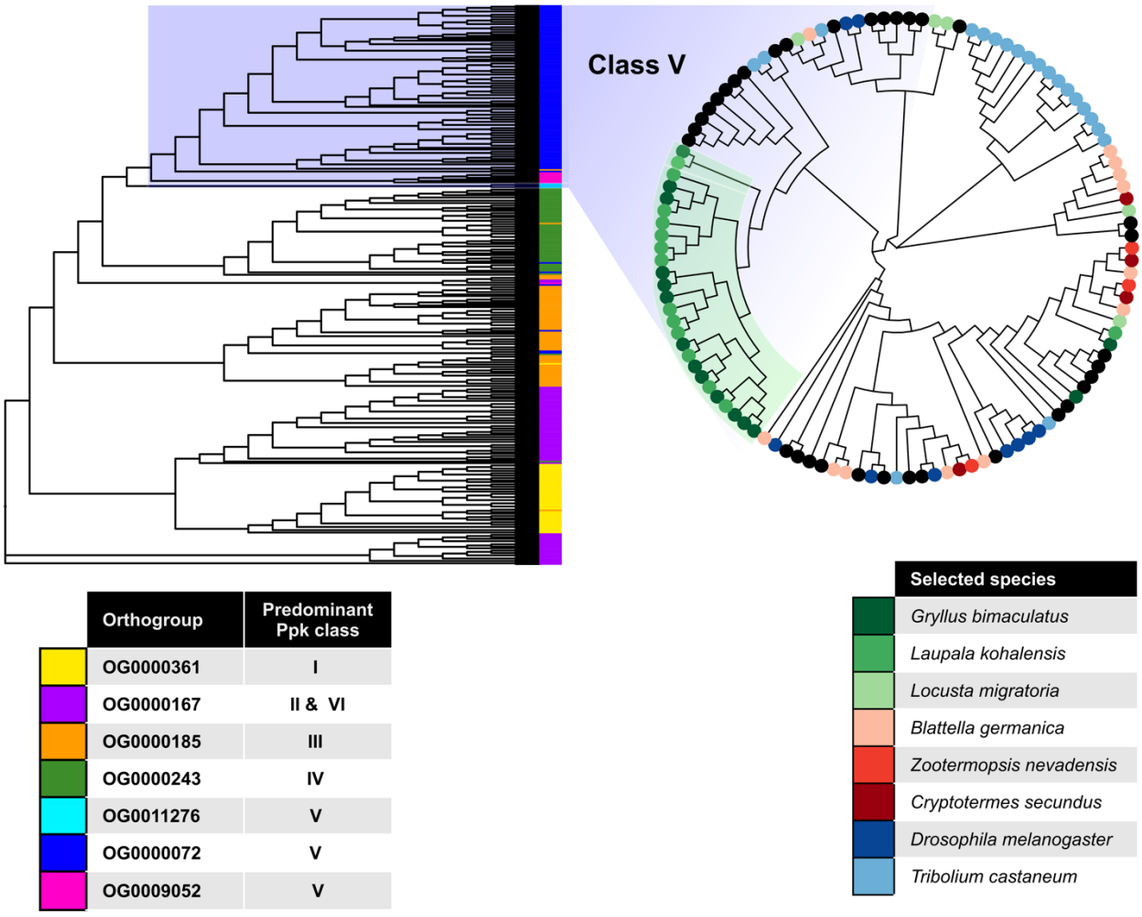


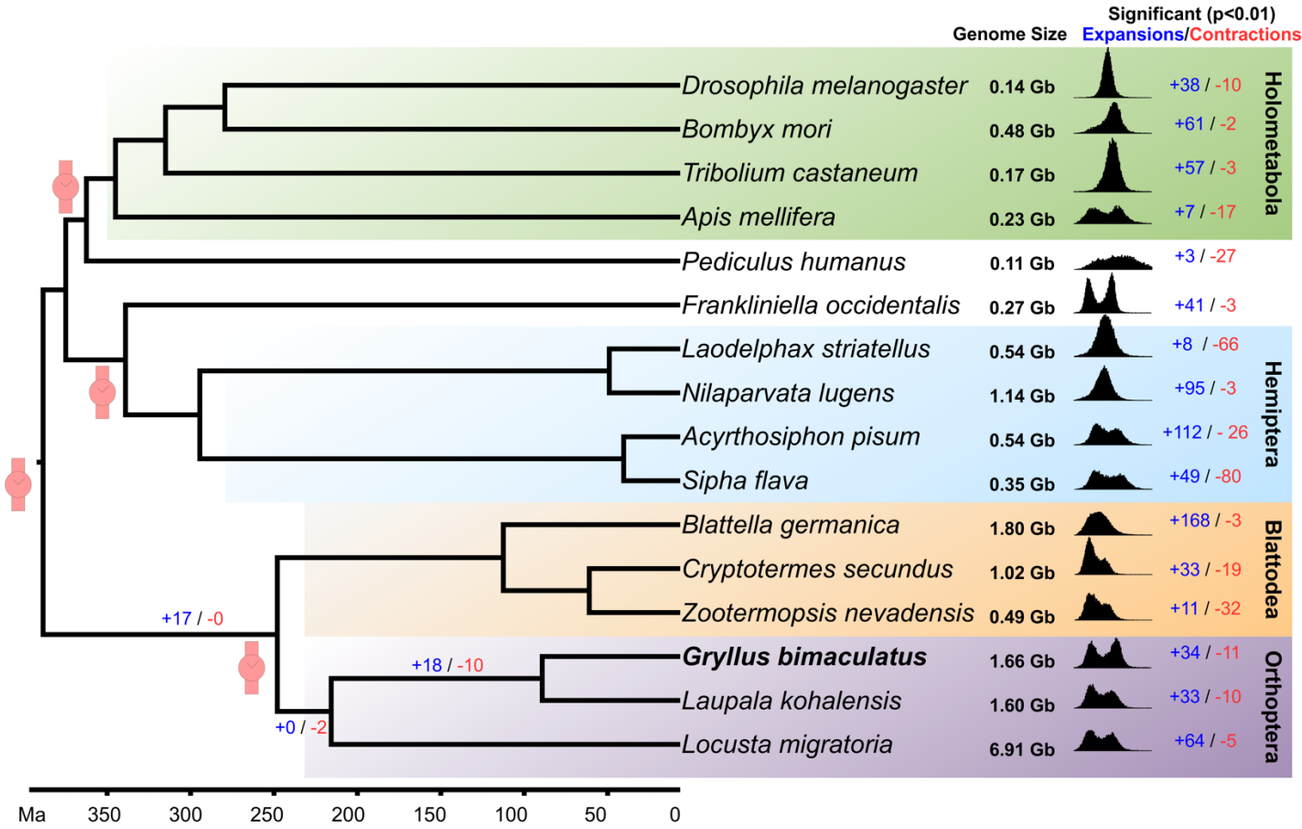
Figure 2: $CpG_{o/e}$ distribution across insects and functional analysis. A) The distribution of $CpG_{o/e}$ values within the CDS regions displays a bimodal distribution in both crickets and in the honeybee *A. mellifera*. We modeled each peak with a normal distribution and defined their intersection (red line) as a threshold to separate genes into high- and low- $CpG_{o/e}$ value categories. The significantly enriched (p-value<0.05) GO terms from the genes belonging to each group is represented as a word cloud with the font size of each term proportional to their fold change, and color-coded to indicate functional categories: red = RNA/transcription, blue = DNA/repair/histone/methylation, orange = protein/translation/ribosome, green = metabolism/catabolism/biosynthesis, purple = sensory/perception/neuro, and gray for others. **B)** UpSet plot showing the total number of orthogroups (OGs) belonging to each category (horizontal bar chart) and the number of them that are common across different categories (linked dots) or exclusive to a single category (unlinked dot). For all three insects, the species-specific OGs largely correspond to high $CpG_{o/e}$ value genes, while low $CpG_{o/e}$ value genes are more likely to have orthologs in the other species. This plot shows only the non-overlapping categories and the three most common combinations; the remaining possible combinations are shown in **Supplementary Figure 4**. **C)** 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.



898

899 **Figure 3: The *pickpocket* gene family class V is expanded in crickets**. *pickpocket* gene
900 tree with all the genes belonging to the seven OGs that contain the *D. melanogaster*
901 *pickpocket* genes. All OGs predominantly contain members of a single *ppk* family, except
902 OG0000167, which contains members of two *pickpocket* classes, II and VI. The *pickpocket*
903 class V (circular cladogram) was significantly expanded in crickets relative to other insects.

904



905

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