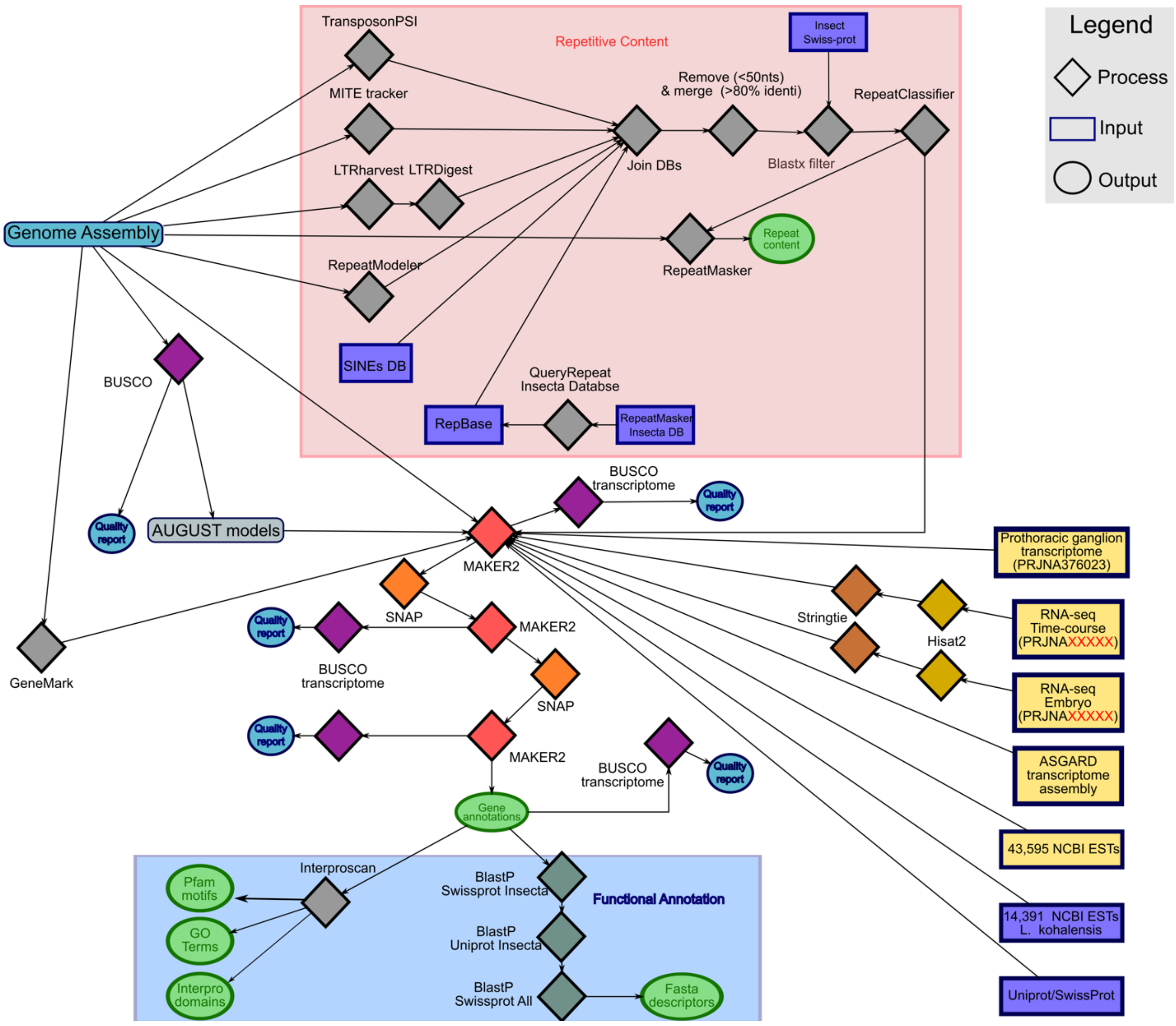
Supplementary Materials for

Cricket genomes: the genomes of future food

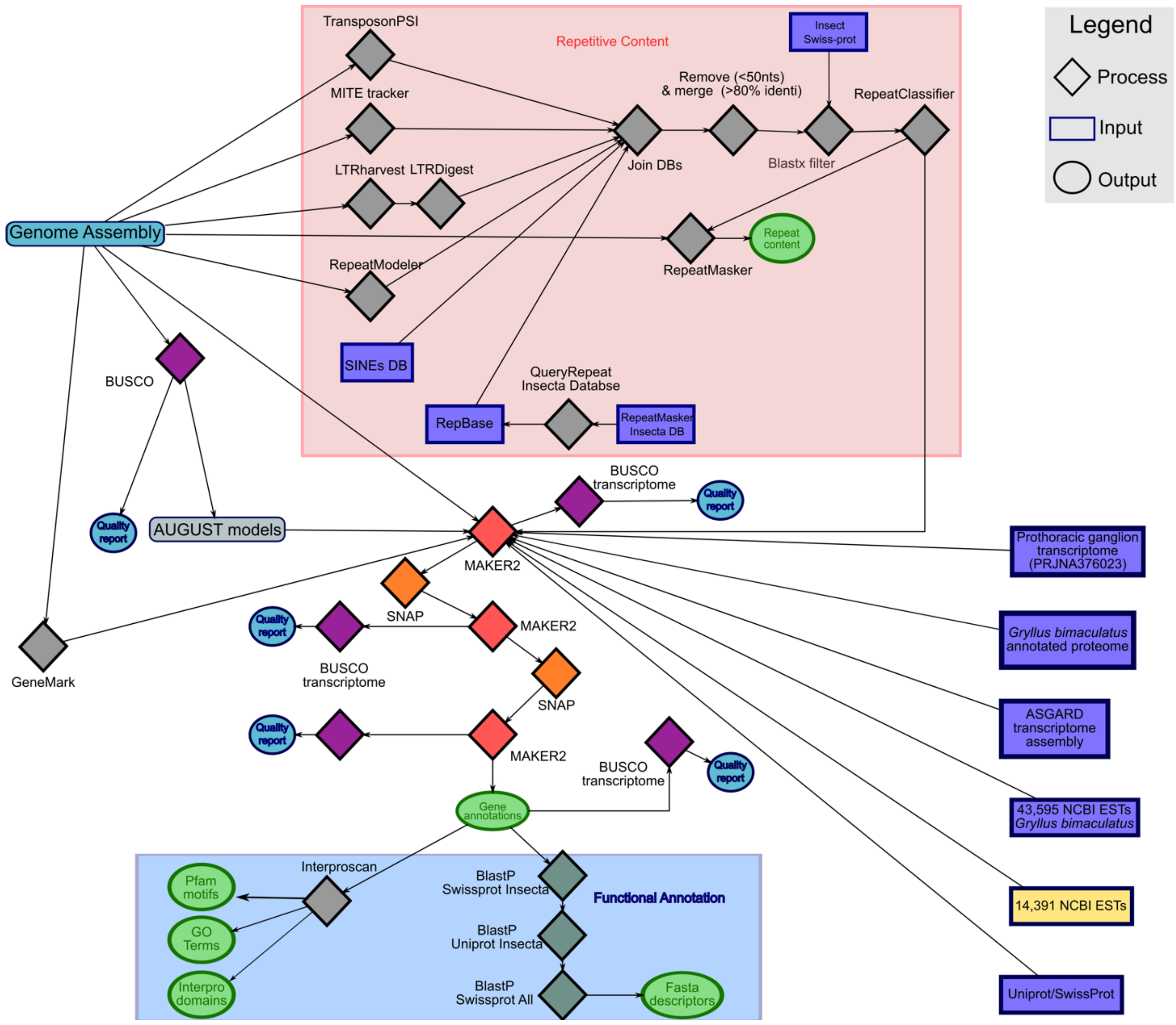
Guillem Ylla, Taro Nakamura, Takehiko Itoh, Rei Kajitani, Atsushi Toyoda, Sayuri Tomonari, Tetsuya Bando, Yoshiyasu Ishimaru, Takahito Watanabe, Masao Fuketa, Yuji Matsuoka, Sumihare Noji, Taro Mito, Cassandra G. Extavour

These Supplementary Materials consist of the following:

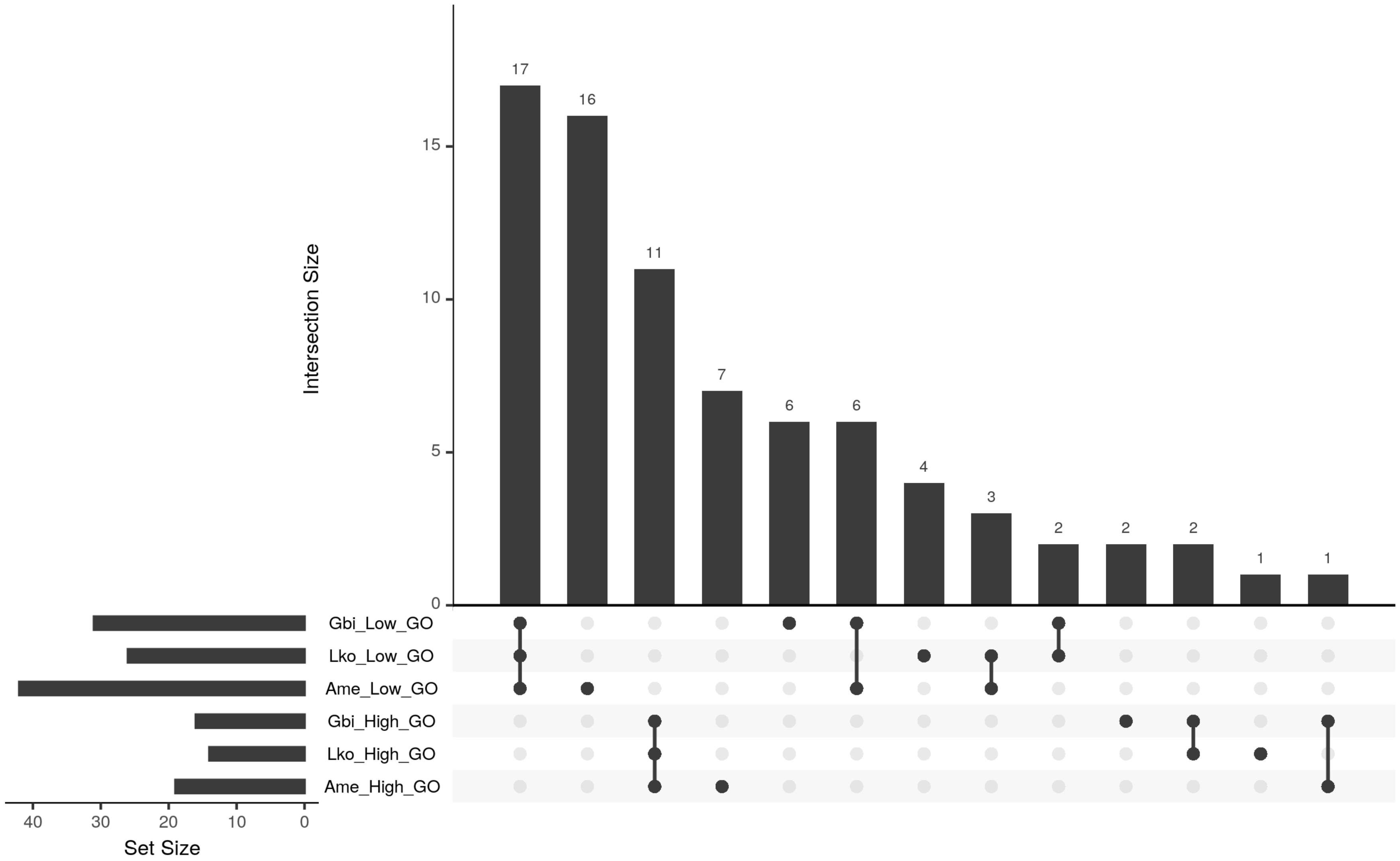
- Supplementary Figures 1 – 4 (this document)
- Supplementary File 1 (this document)
- Supplementary File 2 (“Supplementary_File_2_GeneExpansions.xls”)
- Supplementary Table 1 (“Supplementary_Table_1_GenomeStats.xls”)
- Supplementary Table 2 (this document)
- Supplementary Table 3 (“Supplementary_Table_3_TablePpkExpression.xls”)
- Supplementary Table 4 (this document)
- Supplementary References (this document)



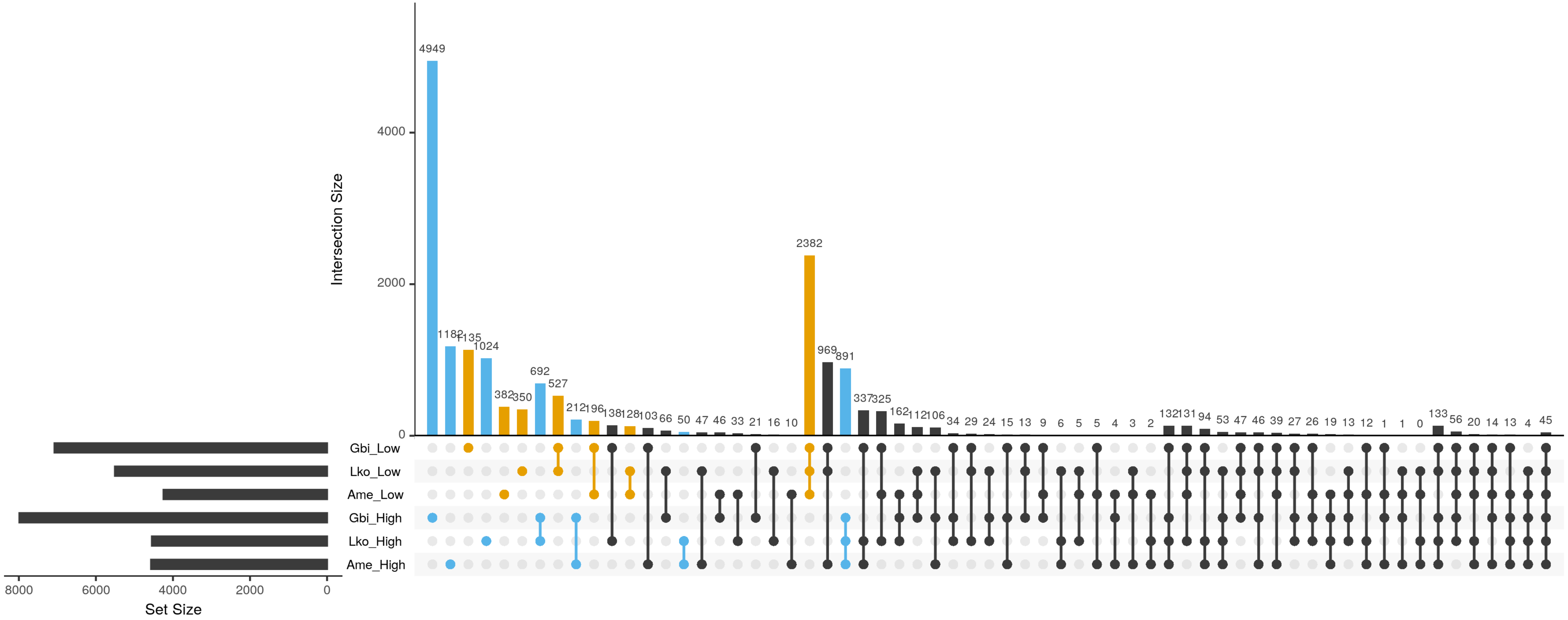
Supplementary Figure 1: Schematic of *G. bimaculatus* genome annotation pipeline. Rectangles represent data inputs: yellow rectangles represent *G. bimaculatus* data; purple rectangles represent data from other species or databases. Diamonds represent computational processes: gray diamonds indicate processes 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: UpSet plot of enriched GO-terms among genes with unusually high or low CpG levels. This plot shows the intersections between the enriched GO-terms of six different categories, which are the high CpG$_{o/e}$ and low CpG$_{o/e}$ genes for each of *G. bimaculatus* (Gbi), *L. kohalensis* (Lko) and *A. mellifera* (Ame).



Supplementary Figure 4: UpSet plot of orthologous genes in crickets and honeybees. Complete UpSet plot (all combinations are shown) of the number of orthogroups (OGs) that are common across the different the 6 different categories, which are the high $CpG_{o/e}$ and low $CpG_{o/e}$ genes for *G. bimaculatus* (Gbi), *L. kohalensis* (Lko) and *A. mellifera* (Ame).

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      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:           138895     26406967 bp    1.59 %
      ALUs            6         9564 bp    0.00 %
      MIRs            0            0 bp    0.00 %

LINEs:           454301    147302087 bp    8.88 %
      LINE1        1803       826764 bp    0.05 %
      LINE2      115576     32029561 bp    1.93 %
      L3/CR1      18286      6358119 bp    0.38 %

LTR elements:    131656     36970251 bp    2.23 %
      ERVL           92        44183 bp    0.00 %
      ERVL-MaLRs      0            0 bp    0.00 %
      ERV_classI  11451      2441461 bp    0.15 %
      ERV_classII   980       401749 bp    0.02 %

DNA elements:    500741    142828465 bp    8.61 %
     hAT-Charlie  11512      4094376 bp    0.25 %
     TcMar-Tigger  2039       537995 bp    0.03 %

Unclassified:    367653    126552078 bp    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 %
```

```
bases masked:  566518287 bp ( 35.51 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:            29510      7083717 bp    0.44 %
      ALUs          304       101257 bp    0.01 %
      MIRs         1248       430584 bp    0.03 %

LINEs:          1035151    322470849 bp   20.21 %
      LINE1         941       367057 bp    0.02 %
      LINE2      584526    167380843 bp   10.49 %
      L3/CR1      10257      4624100 bp    0.29 %

LTR elements:     57347     29690552 bp    1.86 %
      ERVL          231        43500 bp    0.00 %
      ERVL-MaLRs      0            0 bp    0.00 %
      ERV_classI   1821       585650 bp    0.04 %
      ERV_classII   389       125302 bp    0.01 %

DNA elements:    189815     62384975 bp    3.91 %
     hAT-Charlie  15008      5154516 bp    0.32 %
     TcMar-Tigger  8896      2459752 bp    0.15 %

Unclassified:    409303    128822550 bp    8.08 %

Total interspersed repeats:550452643 bp   34.51 %


Small RNA:        13816      3005585 bp    0.19 %

Satellites:        2088       882748 bp    0.06 %
Simple repeats:  307925     19782955 bp    1.24 %
Low complexity:   48386      2381730 bp    0.15 %
==================================================
```

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 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	Dmel symbol	Zelle 2013 class
OG0000361.fa	FBgn0034965	*ppk29*	I
OG0000361.fa	FBgn0039424	*ppk15*	I
OG0000361.fa	FBgn0051065	*ppk31*	I
OG0000361.fa	FBgn0053508	*ppk13*	I
OG0009052.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 FPKMs of *fruitless* and *ppk* 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 > 5 FPKMs across samples are highlighted.

See file "Supplementary_Table_3_TablePpkExpression.xls"

Supplementary Table 4: *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*.

								Table S3 and S6 (Blankers, Oh, & Shaw, 2018)		Table S4 (Blankers, Oh, Bombarely, & Shaw, 2018)	Table 2 (Xu and Shaw, 2019)
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		
Lko057S000206	NNCF01125945.1	353321	357106	3786	-	Lko_06341	Class III				3
Lko057S000206	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 S. Allodi (Ed.), *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 speciation: Linking quantitative variation to candidate genes for behavioral 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).