File S1 Supplemental method details and results

Methods

Near isogenic line (NIL) construction

NILs were constructed by hybridizing isofemale *L. paranigra* and *L. kohalensis* lines. The resulting F_1 offspring were backcrossed to *L. kohalensis* for four generations. Only those backcross offspring carrying the *L. paranigra* allele at the marker previously found to be linked to the QTL for pulse rate variation on linkage group 5 (LG5) were used in the next generation of backcrossing (for details see Wiley et al. 2012). Fourth-generation backcross offspring were intercrossed to generate three independent NILs for QTL4 (NIL4B, 4C, and 4E).

DNA extraction, library preparation and sequencing

Genomic DNA from grandparents (for NIL4B only) and F_2 offspring were extracted using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA, USA) following the recommended protocol for animal tissues with a modification of doubling the amount of proteinase K and lysis buffers as well as repeating wash step 2. We examined the integrity and purity of the extracted DNA by agarose gel electrophoresis. DNA for library preparation was quantified with a QuantiFluor dsDNA system (Promega Corp., Madison, WI, USA). One hundred nano-grams of genome DNA from each sample was digested using the restriction enzyme PstI (New England BioLabs, Beverly, MA, USA) and 96-plex GBS libraries were prepared according to the protocol described by Elshire et al. (2011). Libraries were sequenced on the Illumina HiSeq2000 platform using single end sequencing. Library preparation and sequencing were performed at the Genomic Diversity Facility at Cornell University.

Linkage mapping

Maps of autosomal linkage groups were calculated using the regression algorithm with Kosambi mapping function. Specifically, we used linkages with maximum recombination frequency of 0.4 and a log-of-odd (LOD) score higher than 1, as well as a goodness-of-fit jump threshold for removal of markers at 4. Potentially erroneous markers that indicate highly improbable double recombination events and/or cause high stress on the map were excluded until all nearest neighbor fit was below 3.5 cM. Using more relaxed or more stringent parameter values for mapping did not change marker order or cause large difference in map distance estimation. The marker order of the final regression maps were also consistent with maps calculated using the maximum likelihood algorithm.

A group of 18 markers grouped to LG5 in 4C.9 exhibit high recombination fraction with the rest of the markers. They form a separate group when markers were grouped using recombination fractions with a threshold of 0.25. Five of the 18 markers are located on the same scaffolds that are on the final linkage map of LG5. These 18 markers likely reside in regions that are not homogeneous in the parents and double recombination events have occurred. We constructed a linkage map with these 18 markers using the same method stated above.

QTL mapping

Although we are focusing on QTL on LG5 in this study, we included all autosomal linkage groups in QTL mapping to account for potential QTL on other linkage groups that can partition error and increase power for detecting QTL on LG5 in the QTL models (Broman and Sen 2009). We simulated missing genotype data using 20000 multiple imputations. Genotype probability was calculated at a step size of 0.2 cM and genotyping error rate of 0.1% under Kosambi map function. We performed standard interval mapping, two QTL scan and multiple QTL mapping in each family. Multiple QTL mapping was conducted using forward selection with backward elimination. Specifically, we began with a single-QTL model at the location of the significant QTL with the highest LOD score from standard interval mapping. We then scanned for additional QTL genome wide. Significant QTL with the highest LOD score was added to the subsequent model in each round. The additional QTL and its interaction with previous QTL were accepted only when: (1) the LOD score of the main effect or the interaction term from ANOVA when each term is dropped from the model exceed the threshold for additive effect and interaction term respectively, and (2) the increase of LOD score of the overall new model compared to the previous model is greater than the penalty for the additional degrees of freedom. The penalty value controls the rate of including extraneous terms in the model at a target rate (here, 5%). For additive models, we used the main-effect penalty and for models with interactions, the heavy interaction penalty that controls false positive rate for models of any size was used. Because the final models in all families included more than one QTL, we did not use the light interaction penalty that only controls false positive rate for including a second, interaction term in single QTL models. QTL locations were refined after each step with an iterative maximum likelihood algorithm. We repeated the process until no significant additional QTL or QTL interaction can be found. LOD threshold for standard interval mapping was calculated from 20000 permutations using the maximum likelihood method and LOD thresholds for multiple QTL mapping were calculated from 1000 permutations using Haley-Knott regression. Penalties for main effect and interaction terms were calculated according to Broman and Sen (2009). We estimated 1.5-LOD support intervals for significant QTL in the final multiple QTL model using the default algorism in R/qtl that links all markers within 1.5-LOD drop from the peak in a continuous interval. We also estimated QTL effect, proportion of parental species difference explained and the proportion of F_2 variance explained from the final multiple QTL models. The proportion of the phenotypic differences between the two parental species explained by a QTL was calculated as the additive effect (i.e., effect of substituting one allele) divided by the total difference in mean phenotypic values of the two parental species and thus, this estimate has a maximum value of 50%. The proportion of F_2 variance explained by a QTL is calculated from ANOVA tests dropping one QTL at a time from the model. When two linked QTL are identified on the same linkage group, the proportion of F_2 variance explained by the major QTL is estimated from a model including the major QTL only, and the proportion of F_2 variance explained by the minor QTL is estimated by an ANOVA test dropping the minor QTL from the final multiple QTL model.

QTL region coverage estimation

We estimated the coverage of the genomic region in the confidence interval of the major QTL in 4C.9. Using markers on the same scaffold in 4C.9 map, we estimated the mean physical distance covered per centiMorgan (cM) by the map. By comparing the 1.5-LOD confidence interval width and the total physical distance covered by scaffolds within the confidence interval, we estimated the proportion of genomic region covered by scaffolds on our linkage map.

RNA sequencing and assembly of *L. cerasina* transcriptome

To provide RNA evidence for gene prediction in the QTL region, we conducted RNA-sequencing of four adults (2 sampled in the morning and 2 sampled in the afternoon) of each sex, one immature male and one immature female *L. cerasina* (for details of RNA-sequencing see Blankers et al. 2018b). All individuals were first generation offspring of field caught individuals. RNA was extracted using the Qiagen RNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA) from whole body. Pooled library for each sex was constructed. Paired-end sequencing of the libraries was done in a single lane on the Illumina HiSeq 2000 platform. We trimmed adaptor sequence in the raw reads using cutadapt 1.14 (Martin 2011). No read included bases for which the Phred score was below 30. Therefore, no read trimming for base call quality was performed. Processed reads were mapped to *L. kohalensis* genome reference using TopHat-2.0.13 (Trapnell et al. 2009) and assembled into transcriptome using Cufflinks-2.2.1 (Trapnell et al. 2010) using default settings.

Gene prediction and functional annotation

Gene prediction was done using the Maker pipeline (Cantarel et al. 2008). Because of the large genome assembly size (1595.21 Mb), we generated a *L. kohalensis* specific repeat library with RepeatModeler-1.0.10 (http://www.repeatmasker.org/RepeatModeler/) using 200 longest scaffolds in the genome reference as well as the five scaffolds to be annotated. Together these scaffolds cover 21.8% (347.75 Mb) of the total reference genome size. The repeat library was used to mask the sequences of the five focal scaffolds in RepeatMasker-open-4-0-7 (http://www.repeatmasker.org). Interspersed repeats were hard masked and low complexity repeats were soft masked. We then performed two rounds of training with SNAP (Korf 2004) and Augustus-3.2.3 (Stanke & Morgenstern 2005) for 10 longest scaffolds and 7 scaffolds residing in the 1.5 LOD confidence interval in 4C.5, 4C.9 and 4E.1 in a bootstrap manner with RNA and protein evidence in the first two rounds. For RNA, EST and protein evidence, we used published *L. kohalensis* ESTs (Danley et al. 2007) and transcriptomes of three co-familial cricket species *Gryllus rubens* (Berdan et al. 2016), *Gryllus bimaculatus* (Zeng et al. 2013) and *Teleogryllus oceanicus* (Bailey et al. 2013), the co-generic *L. cerasina* transcriptome assembled herein and the Swiss-Prot protein database (Apweiler et al. 2004). The second round gene model outputs from both SNAP and Augustus were used to predict gene structures of the 5 focal scaffolds in the third round.

Candidate gene identification and SNP effect annotation

For any predicted gene that fulfills our three criteria for being the potential causal gene but its identity is uncertain because the top 20 significant blast hits include different members of the same gene family, we infer the most likely gene identity using both sequence alignment and phylogenetic relationship inference. Because this situation only applies to one predicted gene, the putative cyclic nucleotide-gated ion channel-like gene on scaffold S001371 (gene #20 in3), we explain our procedure to infer the identity of this gene specifically. For both sequence alignment and phylogenetic inference, we first identified and downloaded sequences of representative genes in all subfamilies (CNG subunit A1-4, B1 and B3) and one representative gene from the sister gene family (hyperpolarization-activated cyclic nucleotide-gated ion channel gene family) of the indicted gene family from fruit fly *Drosophila melanogaster,* zebrafish *Danio rerio*, house mouse *Mus musculus* and human *Homo sapiens* from NCBI (Table S6). Because the above genes from model organisms are biased towards vertebrates, we also independently identified putative homologs of the predicted *Laupala* gene in all animals by collecting protein sequences from NCBI HomoloGene database using search term "cngl" and from the nr database of any animal protein sequences whose names include keywords "cyclic nucleotide-gated channel-like". Conserved domains in all collected sequences were identified in SMART v.8 (Letunic and Bork 2017) using both SMART and PFAM domain databases. Only proteins with the same conserved domain architecture as the predicted gene (from N-terminus to C-terminus: Ion_trans, Ion_trans 2, and cNMP binding, with Ion_trans2 nested within Ion_trans), were retained as putative homologs of the predicted *Laupala* gene (see Table S6).

We then conducted sequence alignment between the predicted *Laupala* gene and all collected protein sequences with Exonerate 2.2.0 (Slater and Birney 2005) using the protein2genome model in local, exhaustive alignment mode. In addition, we performed multiple alignment of the amino acid sequences of conserved domains in all collected protein sequences in MUSCLE v3.8 (Edgar 2004) and constructed a maximum likelihood tree using the LG model with 500 bootstraps in PhyML 3.0 (Guindon & Gascuel 2003). If the proteins have both top alignment scores in Exonerate alignment and group in the same clade as the predicted gene in *Laupala*, we infer our predicted gene as a member of that sub-gene family.

We manually annotated SNPs identified from WGS using the alignment output with the highest alignment score in Exonerate. ~~When a non-synonymous SNP alternative homozygous for the parental RIL and *L. kohalensis* lines is detected, we estimated the effect of the resulting amino acid substitution using PROVEAN Protein (Choi and Chan 2015), a program that is not restricted to model organisms. In one case where a non-synonymous SNP is located within a conserved domain, multiple alignment of the conserved domain from the identified putative homologs of the predicted gene in Animalia was performed in MUSCLE v3.8.~~

Results

Linkage mapping details

A total of 431, 346, 508, 770, 769 and 301 SNP markers that passed quality and segregation distortion filters were grouped into 7 linkage groups in 4B.1, 4B.2, 4B.3, 4C.5, 4C.9 and 4E.1 respectively. After marker selection based on mean depth, missing genotype, physical location on the scaffolds, and nearest neighbor fit on the map, the final linkage maps of these six families contained 180, 112, 139, 159, 204 and 97 markers respectively, of which between 17 and 73 mapped to LG5 (Table 2).

QTL region coverage

In 4C.9, the 1.5-LOD confidence interval spans 1.062 cM when expanded to the nearest markers and 0.8 cM when not expanded to markers. The mean physical distance covered by one centiMorgan on the map is 1784.43 kb. The estimated genomic region covered by the confidence interval is thus 1895.06 kb and 1427.54 kb when the confidence interval is and is not expanded to markers respectively. For the confidence interval expanded to markers, the actual physical distance covered by scaffolds within the confidence interval is 2331.06 and 2291.55 kb for the positive and negative orientation of S003497, respectively (the orientation of the other flanking scaffold S000933 is known from linkage map). For the confidence interval not expanded to markers, the actual physical distance covered by the linkage map is 1432.61 kb. In all three scenarios our linkage map has >100% coverage, and has saturated the genomic region within the confidence interval of the QTL.

Literature Cited

Broman, K. W., and Ś. Sen, 2009 A Guide to QTL Mapping with R/qtl. Springer, New York.

Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792-1797.

Letunic, I., and P. Bork, 2017 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. 46: D493-D496.

Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 17: 10-12.

Smit, A. F. A., R. Hubley and P. Green, 2013 RepeatMasker Open-4.0. 2013-2015. <http://www.repeatmasker.org>

Trapnell, C., L. Pachter and S. L. Salzberg, 2009 TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105-1111.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan et al., 2010 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28: 511.