Israel Journal of Ecology & Evolution

Genome-wide analysis revisits incipient sympatric and allopatric speciation in a beetle

Wei Hong[a,*], Kexin Li[b,c,*], Kamal Sharaf[c], Xiaoying Song[b,c], Tomàš Pavlìcek[c], Huabin Zhao[a,d,**,***], and Eviatar Nevo[c,**,***]

[a]*Department of Ecology, Tibetan Centre for Ecology and Conservation at WHU-TU, Hubei Key Laboratory of Cell Homeostasis, College of Life Sciences, Wuhan University, Wuhan 430072, China*

[b]*State Key Laboratory of Grassland Agro-ecosystem, Institute of Innovation Ecology, Lanzhou University, Lanzhou 730000, China*

[c]*Institute of Evolution, University of Haifa, Haifa 3498838, Israel*

[d]*College of Science, Tibet University, Lhasa 850000, China*

[*]W.H. and K.L. contributed equally to this work.

[**]H.Z. and E.N. contributed equally to this work.

[***]Corresponding authors. E-mails: huabinzhao@whu.edu.cn / nevo@research.haifa.ac.il

Supplementary material

Text S1

Materials and methods

Ethics statement and sample collection

Ethics statement

All the experiments on the grain beetle (*Oryzaephilus surinamensis*) were conducted following the rules and guidelines on animal experiments in Israel and China. Experimental protocols were reviewed and approved by the Ethics Committee for Animal Experimentation of University of Haifa and Wuhan University.

Sample collection

Adults of the grain beetle were collected from each site of the south-facing slope (SFS) and north-facing slope (NFS) in Evolution Canyon (EC-I), and the grain silo (S), the latter of which is 26 km from EC-I, in lower Nahal Oren, Mount Carmel, Israel (32°43'N, 34°58'E) in May, 2014. After being captured, beetles were inserted in 95% ethanol (vol/vol) or RNAlater RNA Stabilization Reagent (QIAGEN) directly, and stored at -80°C until further molecular analysis.

DNA library construction and sequencing

Library construction and sequencing of reference genome

A total of approximately 1.3 μg genomic DNA was extracted from the whole body of a male beetle from the SFS population, which was stored in ethanol. Before DNA extraction, the beetle was dried at room temperature in order to remove the ethanol. We used TIANamp Genomic DNA Kit (TIANGEN) to isolate the genomic DNA according to the manufacturer's protocol. An Illumnia

paired-end library with an insert size of ~500bp was constructed following the manufacturer's instructions (Illumina), and 2×300 bp paired-end sequencing was carried out on the Illumina Miseq platform. After removing adaptors and PCR duplications, the resulting clean reads were processed for the genome assembly.

Library construction and population sequencing

Genomic DNA for genome resequencing was extracted from each of the 24 beetles from the 3 populations with the same kit and protocol as reference genomic DNA extraction. At least 0.5μg DNA from each individual was provided for constructing Illumina paired-end libraries with an insert size of 500 bp. Libraries were constructed following the manufacturer's instructions (Illumina). For each individual, 2×125 bp paired-end sequencing was performed on the Illumina Hiseq 2000 platform. Adaptors and PCR duplications were removed before subsequent analysis.

Transcriptome sequencing and assembly

Thirteen beetles from the NFS population, which had been stored in RNAlater, were used for RNA extraction.After washing with RNase free water; all beetles were grinded together in liquid nitrogen.The total RNA of grain beetles was isolated with RNAiso Plus Total RNA Extraction Reagent Kit (TAKARA), following the manufacturer's protocol. Approximately 3.15 μg RNA was provided for constructing an Illumina paired-end library. mRNA was enriched by oligo(dT)-attached magnetic beads. Following mRNA fragmentation, cDNA synthesis, end repair, 3' adenylation, adaptor ligation, and PCR enrichment, the 200bp insert size RNA sequencing (RNA-Seq) library was constructed using the Illumina TruSeq RNA sample preparation kit, as described elsewhere [1, 2].

Massive parallel sequencing was performed on the Illumina HiSeq 2000 platform, and 2×125bp paired-end reads were generated. Adaptors and PCR duplications were removed from raw reads. Subsequently, we removed low-quality reads, including those with average base quality <15, those with >50% having a base quality score <10, and those with >10% unidentified nucleotides (N). Additionally, we trimmed 13 bases from the 5' end of each read to minimize problems associated with low-quality ends. We used the Trinity [3] to assemble the transcriptome with default parameters. We obtained 35.2Mbp clean reads for the transcriptome sequencing, which were assembled to 35,241 contigs with a N50 value of 6,458bp (Supplementary Table S1).

Genome assembly

We used the k-mer method to estimate genome size. We obtained the 17-mer depth distribution with jellyfish 2.0.0 [4].Genome size was estimated from the total number of k-mers divided by the peak depth of k-mer distribution graph. To improve assembly quality, we merged the 2×300 paired-end reads to generate ~500bp long reads with FLASH-1.2.11 [5]. Low-quality reads, which are characterized with >50% having a base quality score <5 or with >10% unidentified nucleotides (N), were removed. In order to remove sequencing ends, we trimmed 5 bases from 5' end and 75 bases from 3' end of paired-end reads, and trimmed 10 bases from 5' end and 5 bases from 3' end of long reads. We used the Celera Assembler version 8.3rc2 [6] to assemble contigs and scaffolds with default parameters. Scaffolds containing microbe sequences or shorter than 1000bp were removed from the raw assembly. Core eukaryotic genes were predicted and G evaluated by CEGMA [7], in order to measure the assembly completeness.

Genome annotation

Repeat annotation

Known repetitive elements were identified using RepeatMasker [8]. We generated consensus sequences and classification information for each repeat family with default parameters. Tandem repeats were identified by TandemRepeatFinder [9].

Protein alignment

Protein sequences came from a protein database (SwissProt) and proteomes of five insects from Genbank (*Drosophila melanogaster* [release v5.48], *Bombyx mori* [ASM15162v1], *Apis mellifera* [Amel_4.5], *Tribolium castaneum* [Tcas5.2], *Dendroctonus ponderosae* [DendPond_male_1.0]), were first aligned to the grain beetle genome to identify conserved genes. We aligned these sequences with the grain beetle genome with the e-value cutoff of 1e-5 and gapped alignment allowed. Subsequently, we extracted matched genomic regions and used GeneWise v2.4.1 [10] to identify exon/intron boundaries, with modeled spliced sites (--nosplice_gtag).

Transcript alignments

Low-quality transcripts, vector sequences and poly-A tails were removed from the transcriptome. Filtered transcripts were reconstructed and aligned to the grain beetle genome, aiming to identify putative coding regions.

Ab-initio prediction

Transcripts containing complete open reading frames (ORFs) were extracted. All the ORFs were translated and compared with SwissProt database by BLASTP, with the e-value cutoff of 1e-5. ORFs covering more than 95% of their best hits against SwissProt sequences were regarded as high-quality ORFs. Transcripts containing high-quality ORFs were used as the training set of three ab-initio predictors: GlimmerHMM v3.0.4 [11], geneid v1.4 [12], and augustus v3.0.2 [13]. The gene models were created by running three predictors on the genome separately.

Integration of predictions

The EVidenceModeler (EVM) v1.1.1 [14] software combines gene predictions from different sources into weighted consensus gene structures. We integrated gene predictions generated by protein alignments, transcript alignments, and ab-initio predictions by EVM with default parameters.

Gene function annotation

In order to annotate the function of grain beetle genes, we searched the grain beetle genes against SwissProt database by BLASTX (e-value <1e-5). The basic information including name and function of each gene was added to annotation. Genes that cannot find best hit in SwissProt were deleted from the annotation. In order to remove microbe sequences, scaffolds that only contain gene models with best matches to microbes were discarded in all analyses.

Genome mapping and variation calling

Low-quality reads with an average base quality <30, or with >50% having a base quality score <10, or with >10% unidentified nucleotides (N) were removed from genome resequencing data.

Subsequently, 17 bases were removed from 5' end of each reads. Filtered high-quality reads mapping and variation calling followed an earlier study [15]. The high-quality SNPs (base quality ≥20, mapping quality ≥20, coverage depth ≥130 and ≤785, root mean square (RMS) of mapping quality ≥10, the distance of adjacent SNPs ≥5 bp) were retained for further analysis. SNPs deviating from Hardy-Weinberg equilibrium ($P < 0.05$) were excluded from subsequent analysis.

Population structure analysis

All SNPs of the whole genome and SNPs in regions with the highest 5% of F_{ST} values were used to investigate population structure separately. We performed population structure analysis as described previously [15]. The nonparametric principal component analysis (PCA) was performed with the parameter "-pca2." According to the ecological information, when estimating individual ancestry and admixture proportions, the probable number of ancestral populations was assumed to be 3. The neighbor-joining phylogenetic tree was reconstructed using the nucleotide p-distance matrix, and the reliability of the tree was evaluated with 1000 bootstrap replications.

Genetic diversity and recombination rate

Genetic diversity was estimated by measuring the Watterson's θ [16]. For each population, we undertook a sliding window analysis, with a window size of 2 kb and a step size of 1 kb. We calculated θ for each window, and the mean value of θ was considered as the whole genome genetic diversity. The significance of difference in θ between two populations was examined with the Mann-Whitney U test. Recombination rate of each individual was calculated by mlRho v.2.8 [17]

with the parameters -m 1000 -M 2000. The significance of difference in the average recombination rate between two populations was evaluated with the student's *t*-test.

Linkage disequilibrium (LD) analysis

To estimate the LD patterns between the two populations, we used the program Beagle v4.0 [18] to phase the genotypes into associated haplotypes with the command "gtgl". The correlation coefficient (r^2) between any two loci was calculated using VCFtools with the "hap-r2" option and default parameters. Average r^2 was calculated in a 2-kb window for pairwise SNPs with a custom written Perl script and was plotted against physical distance in base pairs with R.

Detection of population-specific putatively selected genes

Weir and Cockerham's fixation index (F_{ST}) [19] and Tajima's *D* [20] were chosen as the indicators of population specific selection. F_{ST} values between two populations were calculated by VCFtools with 2-kb of window size and 1-kb of step size. Nucleotide divergence ($\theta\pi$) of each population was calculated under the same sliding window parameters, and then Tajima's *D* value of each window was calculated based on $\theta\pi$ using Tajima's formula [20]. Windows that shared the highest 5% of F_{ST} values between two populations and lowest 5% Tajima's *D* values in one population were recognized as population-specific putatively selected regions (PSRs). Genes that were located in selected regions were regarded as population-specific putatively selected genes (PSGs). Spatial autocorrelation analysis of PSRs followed the method of a previous research [21]; the windows of each scaffold were shuffled 1000 times to estimate the expected spatial autocorrelation level. We used a simple approach to test the number of PSGs. In each population, we randomly sampled the same amount of

windows as population-specific PSRs and repeat 1000 times to evaluate the expected number of genes. Z-test was used to test the significance.

Detection of population-specific transposable elements (TEs)

We conducted a simple approach to detect population-specific TEs. First, we summarized the average depth of coverage of each TE region in each individual by counting the depth of each base in the TE region with SAMtools v1.3.1 [22]. For each individual, TEs that have an average depth equal or higher than 3× were regarded as present in this individual. Otherwise, TEs that have an average depth lower than 3× were regarded as absent in this individual. TEs may affect gene functions by inserting into any components of a gene. For instance, TEs inserted into exons or regulatory elements may cause a premature stop codon or gene silence. By contrast, conserved TEs may be important parts of some regulatory elements; the absence of these TEs also can impact gene functions. Hence, we defined the TEs present in two or more individuals from one population but absent in all individuals from another population, or absent in two or more individuals from one population but present in all individuals from another population, as population-specific TEs. Genes that have overlapped with population-specific TE regions were recognized as population-specific-TE related genes. We compared the SFS population with the NFS population and the wild population with S population, respectively.

Functional enrichment analysis

Functional enrichment analysis of gene ontology (GO) terms and Kyoto encyclopedia of genes and genomes **(**KEGG) pathways was conducted with the method previously described [15]. Putatively

selected genes and population-specific-TE related genes were chosen for enrichment analysis of the significant overrepresentation of GO terms and KEGG pathways, and the whole gene set of the grain beetle genome was selected as the background. We undertook Benjamini-corrected modified Fisher's exact tests to examine the significance of functional enrichment between various gene sets, and *P*-values less than 0.05 were considered significant.

Data availability

All the genome assembly, annotation and re-sequencing data were under NCBI BioProject PRJNA356192.

References

1. Wang S-W, Liu S-C, Sun H-L, Huang T-Y, Chan C-H, Yang C-Y, Yeh H-I, Huang Y-L, Chou W-Y, Lin Y-M: CCL5/CCR5 axis induces vascular endothelial growth factor-mediated tumor angiogenesis in human osteosarcoma microenvironment. *Carcinogenesis* 2014:bgu218.
2. Wang K, Hong W, Jiao H, Zhao H: Transcriptome sequencing and phylogenetic analysis of four species of luminescent beetles. *Sci Rep* 2017, In press.
3. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q: Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011, 29(7):644-652.
4. Marçais G, Kingsford C: A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011, 27(6):764-770.
5. Magoč T, Salzberg SL: FLASH: fast length adjustment of short reads to improve genome

assemblies. *Bioinformatics* 2011, 27(21):2957-2963.

6. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA: A whole-genome assembly of *Drosophila*. *Science* 2000, 287(5461):2196-2204.

7. Parra G, Bradnam K, Korf I: CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 2007, 23(9):1061-1067.

8. RepeatMasker 4.0 [http://repeatmasker.org/]

9. Benson G: Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999, 27(2):573-580.

10. Birney E, Durbin R: Using GeneWise in the *Drosophila* annotation experiment. *Genome Res* 2000, 10(4):547-548.

11. Majoros WH, Pertea M, Salzberg SL: TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 2004, 20(16):2878-2879.

12. Blanco E, Parra G, Guigó R: Using geneid to identify genes. *Current Protocols in Bioinfomatics* 2007, 4.3:1-28.

13. Keller O, Kollmar M, Stanke M, Waack S: A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* 2011, 27(6):757-763.

14. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* 2008, 9(1):R7.

15. Li K, Hong W, Jiao H, Wang G-D, Rodriguez KA, Buffenstein R, Zhao Y, Nevo E, Zhao H: Sympatric speciation revealed by genome-wide divergence in the blind mole rat *Spalax*. *Proc*

Natl Acad Sci U S A 2015, 112(38):11905-11910.

16. Watterson G: On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 1975, 7(2):256-276.

17. Haubold B, Pfaffelhuber P, Lynch M: mlRho-a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Mol Ecol* 2010, 19(s1):277-284.

18. Browning SR, Browning BL: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 2007, 81(5):1084-1097.

19. Weir BS, Cockerham CC: Estimating F-statistics for the analysis of population structure. *Evolution* 1984, 38(6):1358-1370.

20. Tajima F: Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989, 123(3):585-595.

21. Wang K, Hu Q, Ma H, Wang L, Yang Y, Luo W, Qiu Q: Genome-wide variation within and between wild and domestic yak. *Mol Ecol Resour* 2014, 14(4):794-801.

22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: The sequence alignment/map format and SAMtools. *Bioinformatics* 2009, 25(16):2078-2079.

Figure S1. Distribution of the beetle *Oryzaephilus surinamensis* across the world.

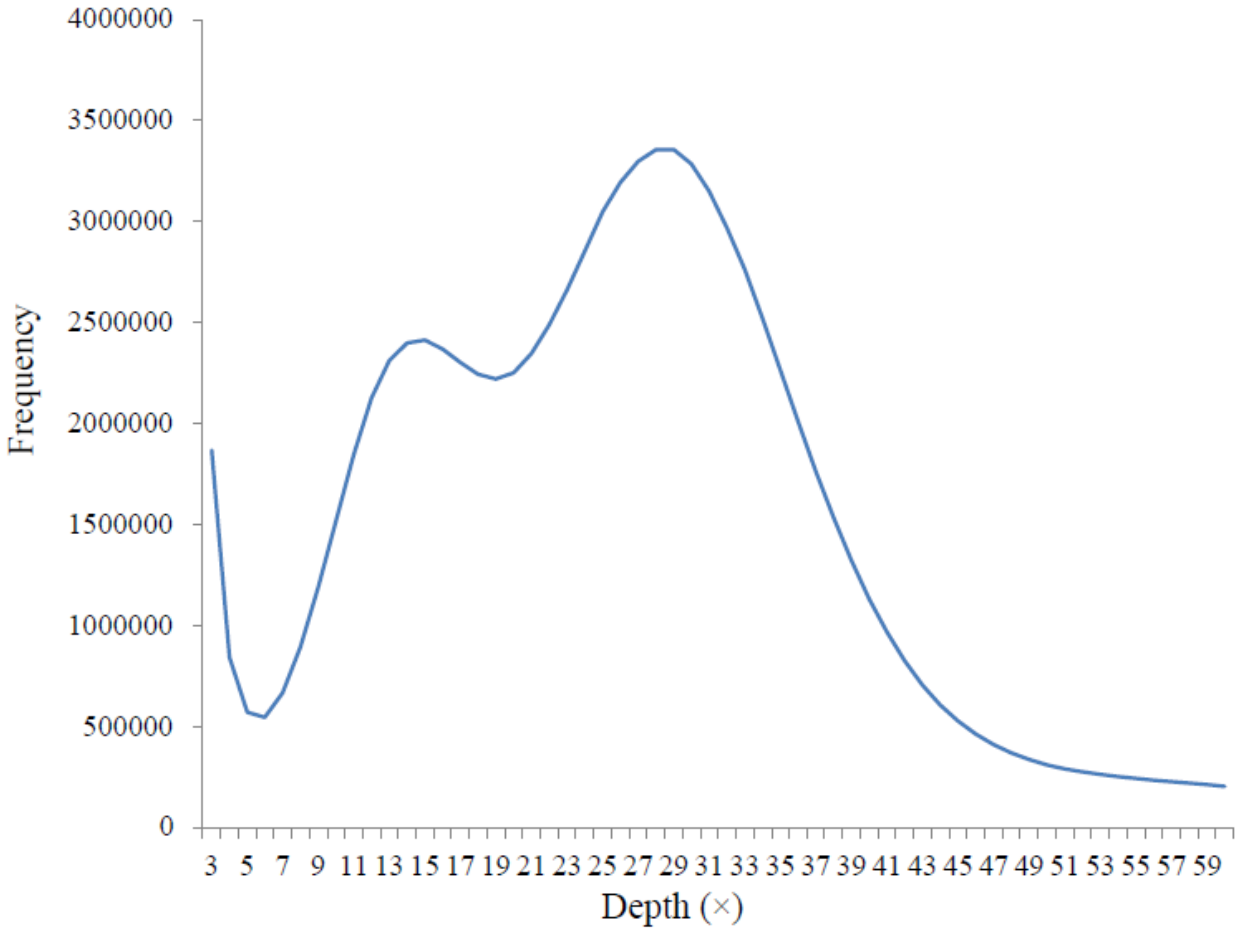

Figure S2. 17-K-mer distribution. The genome size of the grain beetle was estimated at 138Mb by short insert-size libraries. Additional peak at the half of the K-mer depth suggested relatively high heterozygosity.

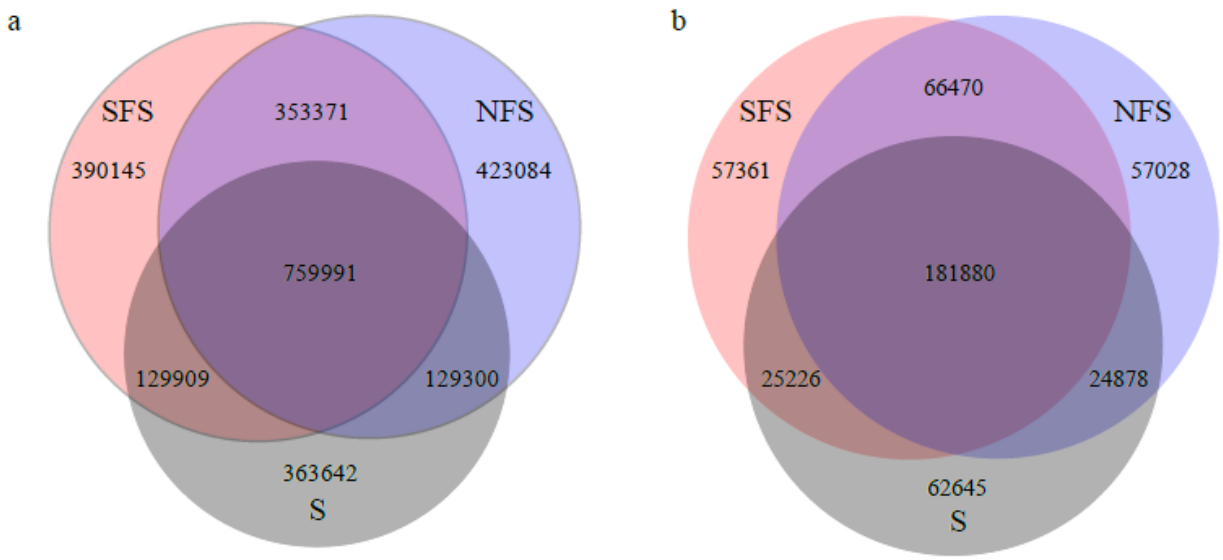

Figure S3. Venn diagrams of the genome-wide variations among the three populations. (a) SNPs from the south-facing slope (SFS), north-facing slope (NFS), and silo (S) populations. (b) Insertions and deletions (Indels) from the SFS, NFS, and S populations.

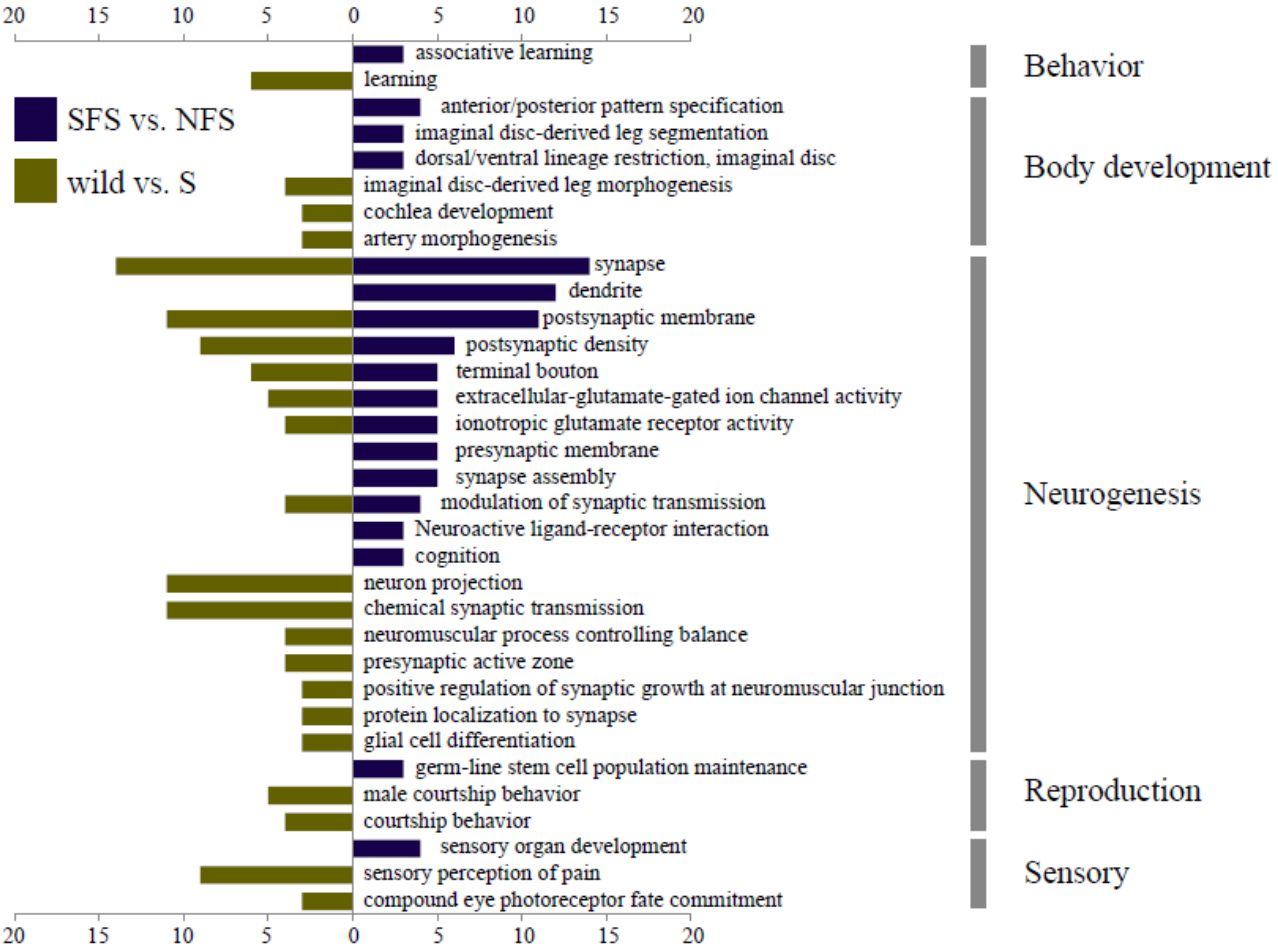

Figure S4. Functional enrichment analysis of population-specific transposable elements (TEs) related genes from the SFS vs. NFS pair and the wild vs. S pair.

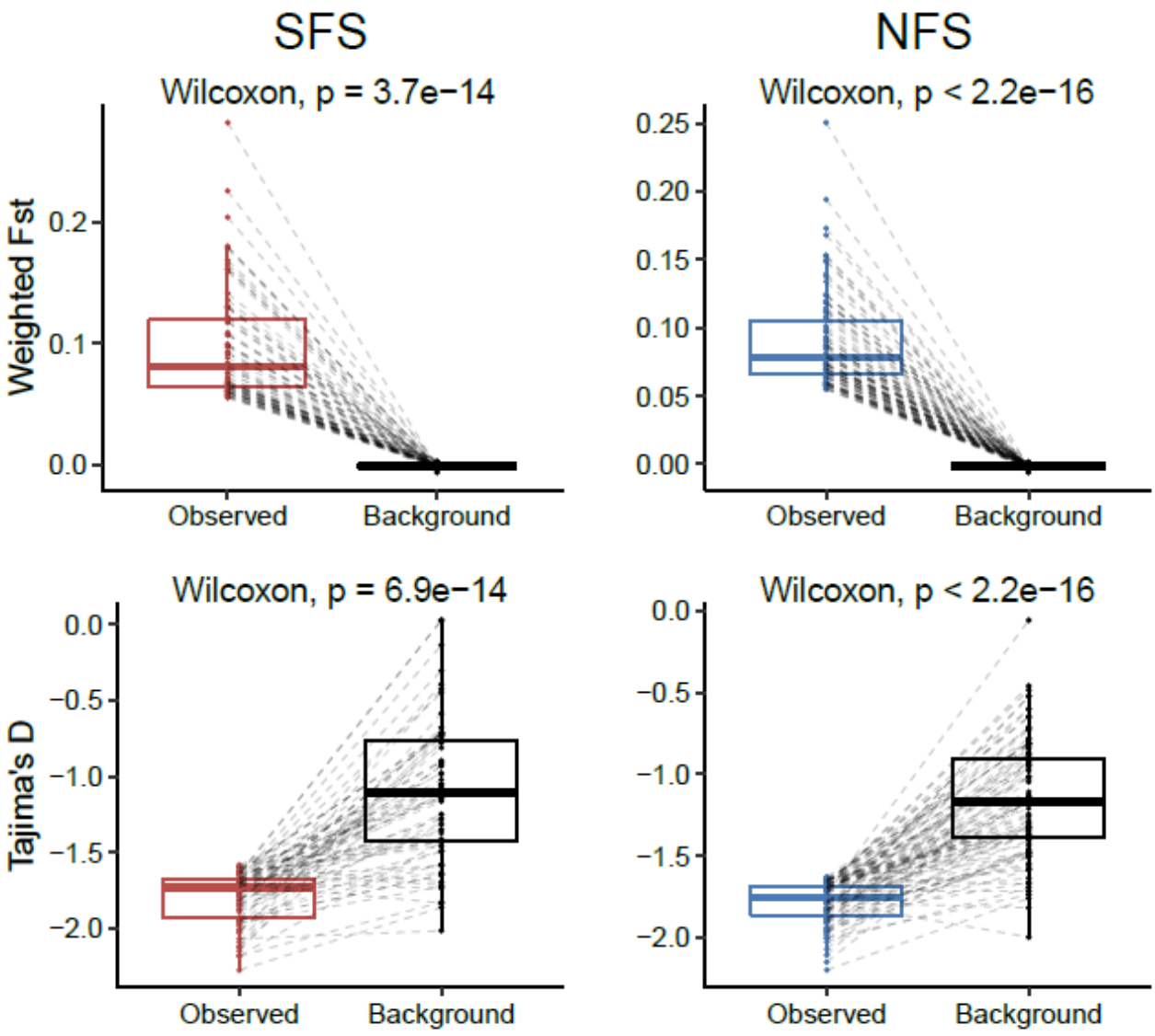

Figure S5. Per-mutation test of selected regions between SFS and NFS. Samples from SFS and NFS were randomly separated into two groups equally 1000 times. The “Observed” box showed observed Fst and Tajima’s D values from selected region in SFS and NFS, while the “Background” box showed the mean value of 1000 bootstraps in corresponding regions. The significance was tested by paired Wilcox test.