GBE

ISSN 1759-6653

www.gbe.oxfordjournals.org

GENOME BIOLOGY AND EVOLUTION

GBE

Genome Biology and Evolution

SMBE

Linking Genomics and Ecology to Investigate the Complex Evolution of an Invasive *Drosophila* Pest

Lino Ometto[1], Alessandro Cestaro[1], Sukanya Ramasamy[1], Alberto Grassi[2], Santosh Revadi[1], Stefanos Siozios[1], Marco Moretto[1], Paolo Fontana[1], Claudio Varotto[1], Davide Pisani[3], Teun Dekker[4], Nicola Wrobel[5], Roberto Viola[1], Ilaria Pertot[1], Duccio Cavalieri[1], Mark Blaxter[5], Gianfranco Anfora[1], and Omar Rota-Stabelli[1,*]

[1]Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Italy

[2]Technological Transfer Centre, Fondazione Edmund Mach, San Michele all'Adige, Italy

[3]School of Biological Sciences and School of Earth Sciences, University of Bristol, Bristol, United Kingdom

[4]Division of Chemical Ecology, Swedish University of Agricultural Sciences, Alnarp, Sweden

[5]Institute of Evolutionary Biology and GenePool Genomics Facility, University of Edinburgh, Edinburgh, United Kingdom

*Corresponding author: E-mail: omar.rota@fmach.it.

Accepted: March 3, 2013

Data deposition: Data for this article have been deposited at the European Nucleotide Archive at EBI under the accessions ID CAKG01000001–CAKG01061569.

Abstract

Drosophilid fruit flies have provided science with striking cases of behavioral adaptation and genetic innovation. A recent example is the invasive pest *Drosophila suzukii*, which, unlike most other *Drosophila*, lays eggs and feeds on undamaged, ripening fruits. This not only poses a serious threat for fruit cultivation but also offers an interesting model to study evolution of behavioral innovation. We developed genome and transcriptome resources for *D. suzukii*. Coupling analyses of these data with field observations, we propose a hypothesis of the origin of its peculiar ecology. Using nuclear and mitochondrial phylogenetic analyses, we confirm its Asian origin and reveal a surprising sister relationship between the *eugracilis* and the *melanogaster* subgroups. Although the *D. suzukii* genome is comparable in size and repeat content to other *Drosophila* species, it has the lowest nucleotide substitution rate among the species analyzed in this study. This finding is compatible with the overwintering diapause of *D. suzukii*, which results in a reduced number of generations per year compared with its sister species. Genome-scale relaxed clock analyses support a late Miocene origin of *D. suzukii*, concomitant with paleogeological and climatic conditions that suggest an adaptation to temperate montane forests, a hypothesis confirmed by field trapping. We propose a causal link between the ecological adaptations of *D. suzukii* in its native habitat and its invasive success in Europe and North America.

Key words: draft genome, genome evolution, population genetics, molecular clocks, *Sophophora* phylogeny.

Introduction

The genus *Drosophila* is one of the most studied in virtually all fields of biology because of an invaluable combination of reproductive (high fecundity and short generation time) and ecological (wide range of niches and fast adaptability) traits. These features have allowed several *Drosophila* species to expand well outside their ancestral range. A classic example is *Drosophila melanogaster*, whose worldwide distribution is the result of an out-of-Africa expansion approximately 15,000 years ago (David and Capy 1988). A more recent example of this invasiveness is *Drosophila suzukii*, which in only a handful of years has invaded several Western countries from its original Asian distribution. The global spread of *D. melanogaster* has little economic consequence, but the spread of *D. suzukii* is of significant concern.

Unlike most of its close relatives, which lay eggs only on decaying or rotten fruits, *D. suzukii* lays eggs and feeds on unripe and undamaged fruits (Dreves 2011; Walsh et al. 2011; Rota-Stabelli, Blaxter, et al. 2013), and consequently, this species is quickly becoming an economically significant

pest of fruit industries. This difference in ecology is reflected in morphological adaptations, such as an enlarged serrated ovipositor (used to break ripening fruits), and must also include additional neurological, lifecycle, and physiological adaptations to finding, and feeding on, unripe food sources. *Drosophila suzukii* is thus a promising model for the study of the origins and bases of behavioral innovation. Understanding the cues by which *D. suzukii* finds its host fruits, and the mechanisms used for invading and feeding thereon, is a key goal in research programs aiming to devise novel control systems (Cini et al. 2012).

To investigate the evolutionary history behind the switch in the reproductive behavior of *D. suzukii* from rotten to fresh fruit, and to better understand how this species established itself in western countries at such an impressive speed, we sequenced and annotated the genome and transcriptome of *D. suzukii* from an Italian Alpine population. On the basis of the combined results of phylogenetic and clock analyses, comparative genomics, and field observations, we propose a paleoecological scenario to explain the peculiar *D. suzukii* ecological behavior.

Materials and Methods

Specimens and Sequencing

Inbred *D. suzukii* lines were established from individuals collected at approximately 500 m above sea level (asl) in Valsugana, Trento, Italy, and subsequently maintained in the laboratory under standard conditions. Genomic DNA was extracted from 10 siblings of an F5 inbred generation (five males and five females), whereas total RNA was extracted from 15 unrelated individuals at various developmental stages (five males and five females adults, three larvae, and two pupae). The pooled cDNA library and two short DNA libraries (180 base pairs [bp] and 300 bp) were sequenced at the GenePool Genomics Facility of the University of Edinburgh, using 100 base paired-end sequencing on the Illumina Hiseq2000 platform (proportions were 0.2, 0.4, 0.4 for the cDNA, 180 bp and 300 bp libraries, respectively). The raw data have been deposited in European Nucleotide Archive (study accession ERP001893) and the assembly in the ENA under accession numbers CAKG01000001–CAKG01061569.

RNAseq Assembly

The RNAseq sequencing generated a total of 35.7 million 100 base paired reads. Data quality was evaluated with fastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, last accessed April 3, 2013) and Tallymer (Kurtz et al. 2008). Low-quality positions were trimmed using fastx (http://hannonlab.cshl.edu/fastx_toolkit/, last accessed April 3, 2013) with a threshold of 0.3. We assembled the resulting 30,951,598 read pairs using two distinct approaches. First, we used Oases (Schulz et al. 2012) with k-mers ranging from 25 to 53, obtaining 24,358 contigs (length 100–15,000 bp). In the second approach, we used ABySS (Simpson et al. 2009) with k-mer 45 and obtained 140,736 contigs. The two sets were merged using cd-hit (Li and Godzik 2006) with an identity threshold of 100% and eventually superassembled using CAP3 (Huang and Madan 1999) using default settings. The final data set consisted of 25,810 putative transcripts with lengths varying from 50 to 16,500 bp.

Nuclear Genome Assembly

Assembly of the nuclear genome was performed using both 180 bp and 300 bp libraries. The 180 bp library generated 67,153,264 100 base read pairs totaling 14.3 gigabases (Gb) and the 300 bp library 51,792,255 100 base read pairs covering 10.4 Gb. The insert sizes of both libraries were close to expectations. We initially partitioned the reads depending on whether they originated from nuclear, mitochondrial, or *Wolbachia* DNA. Nuclear genome assembly was based on reads that were not mappable to a reference database of the genomes of five *Wolbachia* strains (*W. ananassae, W. melanogaster, W. simulans, W. willinstoni,* and *wRi*) or to the *D. melanogaster* mitochondrial DNA (mtDNA). Mapping was performed using Smalt (http://www.sanger.ac.uk/resources/software/smalt, last accessed April 3, 2013; see supplementary table S1, Supplementary Material online). Reads that passed this screening were further cleaned using sickle (https://github.com/najoshi/sickle, last accessed April 3, 2013) with a quality score cutoff of 25 (phred scale) applied to a sliding window of 40 bp. Following this step, reads had an average length of 93 bases (standard deviation [SD] = 14) and 94 bases (SD = 15) for the 180 and 300 bp libraries, respectively, an average quality value of 35, and spanned a total of 20 Gb. Assuming similar genome sizes in *D. suzukii* and *D. melanogaster,* this translates to a coverage of approximately 168-fold. Genome assembly was carried out using ABySS (Simpson et al. 2009) with k-mer size ranging from 48 to 64 (supplementary table S2, Supplementary Material online). After quality assessment of the assemblies, we retained as best assembly the one obtained using a k-mer of 64. All contigs longer than 1 kb have been submitted to the European Nucleotide Archive at EBI web site (http://www.ebi.ac.uk/, last accessed April 3, 2013) under ID CAKG01000001–CAKG01061569.

Assembly of Drosophilid Mitochondrial Genomes

All *D. suzukii* reads that matched the *D. melanogaster* mtDNA were assembled using Geneious (http://www.geneious.com, last accessed April 3, 2013), generating 15 contigs, the longest of which (14,736 bp) was identified as the nearly complete *D. suzukii* mtDNA. This fragment covers all genes but lacks the control region, whose length is unknown. To assist our phylogenetic analyses, we also reconstructed the partial mitochondrial genomes of eight additional *Drosophila* species

(*D. biarmipes*, *D. bipectinata*, *D. elegans*, *D. eugracilis*, *D. ficusphila*, *D. kikkawai*, *D. rhopaloa*, and *D. takahashi*). The draft genomes and transcriptomes of these species were kindly made available by the Baylor College of Medicine and modENCODE Consortium (https://www.hgsc.bcm.edu/content/drosophila-modencode-project, last accessed April 3, 2013). For each species, we separately compared the transcriptome and draft genome against the *D. melanogaster* mtDNA using Basic Local Alignment Search Tool (BLAST) (Camacho et al. 2009). We used Geneious to assemble each set of contigs identified by BLAST, using the *D. melanogaster* mtDNA as a reference. We compared by eye the resulting assemblies to the complete mtDNA genome available for 12 other *Drosophila* species (Drosophila 12 Genomes Consortium et al. 2007), revealing many putative nuclear mitochondrial DNA. Finally, we retained only the transcriptome-based assemblies. These contained a large number of undetermined sites due to the expected intraspecific mtDNA polymorphism in the source *Drosophila* populations.

Repeat Identification

To the genome of *D. suzukii* and the other eight *Drosophila* species mentioned earlier, we added the (draft) genomes of 12 additional *Drosophila* species (*D. melanogaster*, *D. ananassae*, *D. sechellia*, *D. simulans*, *D. yakuba*, *D. erecta*, *D. pseudobscura*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis*, and *D. grimshawi*; downloaded from http://flybase.org, last accessed April 3, 2013). In each genome, we automatically annotated repeats using RepeatMasker (http://www.repeatmasker.org, last accessed April 3, 2013), at default settings, and then used the Repbase database (Jurka et al. 2005) as a reference for de novo identification. We analyzed the entire genome without distinguishing between euchromatin and heterochromatin partitions, as these information are either incomplete or unknown for most of the *Drosophila* species used in this study. We used all fragments irrespective of their length, because the *D. suzukii* genome assembly and some of the other draft genomes contained many contigs shorter than the 200 kb limit recommended (Drosophila 12 Genomes Consortium et al. 2007). We quantified the presence and size of repeats as the percentage of repeated sequences over the draft genome size. This approach has the advantage of reducing biases due to the uncertain draft genome size of the different species, which may vary due to the different assembly strategies and/or genome quality levels, and may not reflect the actual genome size. To account for this inaccuracy, we further calculated the percentage of total repeats using two contrasting and conservative estimates of the putative average *Drosophila* genome size (a minimum at 130 Mb and a maximum at 180 Mb).

Orthologous Gene Set Identification

For comparative genomic analyses, we collated data for 21 *Drosophila* species. We downloaded the latest coding sequences (CDS) data sets available for *D. melanogaster* (release 5.43) and *D. ananassae* (release 1.3) from FlyBase, as well the masked alignments of all single-copy orthologs used in the 12 *Drosophila* project available from ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/alignments/all_species.guide_tree.longest.cds.tar.gz (last accessed April 3, 2013) (Drosophila 12 Genomes Consortium et al. 2007). We also downloaded the assembled RNA-Seq data of eight modENCODE *Drosophila* species (https://www.hgsc.bcm.edu/content/drosophila-modencode-project, last accessed April 3, 2013). We identified best-hit homologous sequences between the nine RNA-Seq and two CDS data sets using pairwise BLASTn (optimized using the parameter "-best_hit_overhang 0.15").

Rates of molecular evolution and tests of positive selection were based on the set of orthologous genes identified in *D. melanogaster*, *D. biarmipes, D. takahashi*, *D. suzukii*, and *D. ananassae*. To minimize the possibility of spurious matches, we filtered matches to exclude any with less than 60% of the length of either sequence aligned. We produced two lists of putative ortholog sets from this five-species set. In the first ("^{STAR}orthologues"), we identified as orthologous genes the reciprocal best hits between *D. melanogaster* and each of *D. biarmipes, D. takahashi*, *D. suzukii*, and *D. ananassae* (see supplementary fig. S3, Supplementary Material online). Using this approach, we identified a total of 2,336 ^{STAR}ortholog quintuplets.

The second, more conservative list, included only those genes found as reciprocal best hits for all pairwise comparisons between the five species (^{WEB}orthologs; see supplementary fig. S3, Supplementary Material online). This data set included 1,021 ^{WEB}ortholog quintets and by definition is a subset of the ^{STAR}orthologs data set. All sequences within each ortholog set were oriented based on the *D. melanogaster* sequence and aligned with MUSCLE (Edgar 2004). We then trimmed partial codons at the 5′- and 3′-ends based on the *D. melanogaster* sequence.

For the ortholog groups used for molecular evolution analyses, we then extracted the portion of the alignments with representation from all taxa. Finally, all alignments were realigned using Prank (Löytynoja and Goldman 2008) as implemented in TranslatorX (Abascal et al. 2010), which aligns protein-coding nucleotide sequences based on their corresponding amino acid translations.

We removed from these two data sets all orthologs sets with alignments shorter than 100 bp. The resulting 2,263 ^{STAR}ortholog quintuplets had a mean length ± standard error (SE) of 1,335.7 ± 29.0 bp (median = 1,092 bp; mode = 942 bp), corresponding to 69.9 ± 29.5% of the *D. melanogaster* gene length. The 1,007 ^{WEB}ortholog quintuplets had a mean length ± SE of 1,575.1 ± 29.8 bp

(median = 1,275 bp; mode = 606 bp), corresponding to 76.4 ± 26.0% of the *D. melanogaster* gene length. We found that the results of our analyses did not change qualitatively when based on ^{STAR}orthologs or ^{WEB}orthologs. Thus, for ease of presentation, and unless specified, we have presented only those obtained using the ^{STAR}orthologs data set.

Analyses of the Rate of DNA and Protein Evolution

Rates of molecular evolution were analyzed for both ^{WEB}orthologs and ^{STAR}orthologs using PAML 4.4 (Yang 2007). We estimated the rate of nonsynonymous substitution, d_N (leading to amino acid changes), and synonymous substitution, d_S (which should accumulate neutrally), over all branches of the phylogenetic tree using the "free-ratio" model (M0′ [Yang 1998]; model = 1 and NSsites = 0). This model allows $\omega = d_N/d_S$, that is, the level of selective pressure experienced by a gene, to vary among branches of the tree. Following the results of the phylogenetic analysis (see later), the input unrooted tree had the structure (D. melanogaster, (D. ananassae,(D. takahashi,(D. biarmipes,D. suzukii))));. We then used PAML to test different models of substitution rates across coding sites (Yang and Nielsen 2000; Yang et al. 2000), with the aim of detecting genes that either evolved at a different rate or underwent positive selection along the *D. suzukii* lineage.

In the first test, we compared models that assumed one or more substitution rates across the phylogeny. The first of such models is the basic "one-ratio" branch model (M0), which assumes a constant ω across the phylogeny (model = 0 and NSsites = 0). Following the manual recommendations, this model was used to get the branch lengths for each gene tree, which were then copied into the tree structure file to be used with the "branch and site" substitution models. The likelihood of the M0 model was compared with that of a branch model that assumed two ω values, one for the *D. suzukii* branch (the so called foreground branch) and one for the rest of the tree (the background branches; model = 2 and NSsites = 0). Subsequently, the value of twice the difference between the two likelihoods ($2\Delta\lambda$) was tested using a χ^2 test with 1 degree of freedom.

The occurrence of positive selection was tested by the branch–site test, which aimed at detecting positive selection affecting a few sites along the *D. suzukii* foreground branch. In this test (branch–site model A, test 2 (Yang et al. 2005), ω can vary both among sites in the protein and across branches on the tree (model = 2, NSsites = 2). As for the branch model, we used tree structures with branch lengths estimated by model M0. The null model fixed $\omega_2 = 1$ (fix_omega = 1, omega = 1), whereas the positive selection model allowed $\omega_2 > 1$ (fix_omega = 0, omega = 1). The likelihood ratio test had 1 degree of freedom. To account for multiple testing, we also estimated the false discovery rate (FDR) of each test using the *q* value approach (Storey 2002) implemented in R (R Development Core Team 2009). We note that the reciprocal best-hit approach is prone to miss genes with high sequence divergence, including those that underwent particularly intense divergent adaptive evolution. Thus, we could have missed targets of positive selection among our sequenced genes.

Codon Usage Analysis

We inferred preferred codons and codon usage bias in *D. melanogaster*, *D. ananassae*, *D. takahashi*, *D. biarmipes*, and *D. suzukii* in the genes of the STARorthologs groups with more than 30 codons. We estimated codon bias using the effective number of codons, *Nc* (Wright 1990), and the frequency of optimal codons, *Fop* (Ikemura 1981): Stronger synonymous codon usage bias is identified by larger *Fop* values and lower *Nc* values. Both indices were calculated using the program CodonW (http://codonw.sourceforge.net, last accessed April 3, 2013). Putative optimal (preferred) codons were identified as those that were significantly overrepresented in the 5% of genes with highest and lowest usage frequencies (supplementary table S3, Supplementary Material online). Base composition affected synonymous codon usage, as shown by the strong correlation between GC and GC3 (GC in the third codon position) content and both *Nc* and *Fop* (Spearman correlation, $P < 10^{-16}$). To remove the potential noise due to this correlation, we estimated a version of the effective number of codons, *Nc*′, which accounts for background nucleotide composition (Novembre 2002). We also used in our analyses the residuals of the regression between GC3 content and *Fop* and *Nc*.

Transcriptome and mtDNA Data Sets for Phylogenetic Analyses

We assembled a data set of 91 orthologous genes from the transcriptomes of 21 *Drosophila* species including *D. suzukii*. Strict orthology within the complete set of *D. melanogaster* genes (Drosophila 12 Genomes Consortium et al. 2007) and the other 20 transcriptomes was assessed using the reciprocal best BLAST hits method. We first identified single copy WEBorthologs between *D. melanogaster*, *D. biarmipes*, *D. bipectinata*, *D. elegans*, *D. eugracislis*, *D. ficusphila*, *D. kikkawai*, *D. rhopaloa*, *D. takahashi*, and *D. suzukii*. We identified the masked alignments of all these WEBorthologs in the 12 *Drosophila* alignments (Drosophila 12 Genomes Consortium et al. 2007), thus selecting 97 groups of putative orthologs. A few of these were removed after manual inspection revealed that they contained incomplete, frame-shifted, and/or dubiously assembled sequences, leaving 91 highly reliable ortholog groups. These were aligned using TranslatorX and concatenated into a superalignment of 200,475 bp, which was further inspected by eye and corrected for the correct frame of codons (inclusion of partial stop codons

that altered the frame) and minor errors that escaped the first manual inspection.

We translated this alignment into amino acids and selected conserved regions using Gblocks (Castresana 2000) with parameters 1:11, 2:17, 3:8, 4:10, and 5:half). We retained 90% of the sites, totaling 60,757 amino acids. The final nucleotide alignment of 182,271 bp, perfectly corresponding to the amino acid alignment, was used for further sequence analyses excluding third codon positions.

A mitochondrial genome alignment was constructed by extracting CDS from available and newly assembled (see earlier) mtDNA. The 12 CDS genes were checked for their correct codon frame and concatenated. We also excluded third codon positions from the mtDNA data set for further sequence analyses.

Phylogenetic Analyses

We performed Bayesian and maximum likelihood (ML) analyses on both the transcriptomic and the mitochondrial genomic data sets. For the Bayesian analyses, we used PhyloBayes3 (Lartillot et al. 2009) setting two independent runs until the maxdiff was less than 0.1. We calculated the 50% majority rule consensus trees by pulling sampled trees after a burn-in that minimized the maxdiff statistic in PhyloBayes3. ML analyses were performed using Phyml (Guindon et al. 2010) on 100 nonparametric bootstrapped replicates. In all cases, a discrete gamma distribution (with four rate categories) was used to model among site rate variation. We performed three main experiments on both data sets using different data set treatments and models of replacement:

1. ML analyses on nucleotide alignments using all the three-codon position and a single nt-general time reversible (GTR) model for all codon positions (nucleotides positions 1 + 2 + 3, GTR + G, ML in fig. 1).
2. Bayesian analyses on nucleotide alignments using the CAT model after exclusion of the third codon position (nucleotide positions 1 + 2, CAT + G, Bayesian).
3. Bayesian analyses on the corresponding amino acid alignments using a six-category Dayhoff recoding and the CAT + GTR model (amino acids-Dayhoff, CAT + GTR + G, Bayesian).

Molecular Clock Analyses

We performed two different molecular clock analyses. We first used PhyloBayes (Lartillot et al. 2009) on both the transcriptomic and mitochondrial genomic data sets at the nucleotide level. We employed a CIR process clock model and a GTR + G model of replacement on both data sets using the fixed tree topology of figure 1*A*. We constrained four nodes as in Prud'homme et al. (2006) using their suggested biogeographical calibrations. To account for uncertainty in biogeographical constrains, we allowed both minima and maxima to be soft, thus allowing the posterior dates to be sampled outside the set bounds (Yang and Rannala 2006). We employed a root prior of 80 Ma with a permissive SD of 40 Myr and assumed a birth–death process along all nodes. We modeled replacement using CAT and the clock using CIR as in Rota-Stabelli, Daley, et al. (2013). In a second approach, we used BEAST (Drummond and Rambaut 2007) without constraining internal nodes and the random local clock but only a normally distributed root prior centered at 80 Ma with SD 20 Myr. We assumed the initial mutation rate of 0.0346 (SD = 0.00281) suggested in Obbard et al. (2012). Because mutation rate refers to unconstrained sites, we used only the 4-fold degenerate sites of the genomic data set for the BEAST analysis.

Field Monitoring and Trapping

Field trapping for tests of distribution by altitude were carried out between 15 April (week 14) and 31 October (week 43) 2011. Forty sites across Trento Province were chosen representing both agricultural and natural ecosystems. Traps were deployed on a large-scale altitudinal gradient and assigned to four altitudinal ranges (<250 m asl [$n = 10$], 250–600 m asl [$n = 10$], 600–1,000 m asl [$n = 10$], >1,000 m asl [$n = 10$]). At each trapping site, we placed, in a shady spot, one plastic transparent bottle with multiple small lateral holes (diameter between 5 and 10 mm) containing 250 ml of apple cider vinegar as bait. Weekly, traps were checked, insects collected, and vinegar replaced. Weekly captures of *D. suzukii* in each trap were averaged per altitudinal range.

Results

Genome, Transcriptome, Mitogenome, and *Wolbachia* Sequencing

We sequenced and assembled a draft genome and transcriptome of *D. suzukii* from an Italian Alpine population. The draft genome was sequenced to high depth (an average of 80× coverage) and comprises 49,558 contigs spanning a total of 160 Mb. The draft transcriptome contains 25,810 unique sequences. Both the size of the genome and its repetitive element content are comparable with that of *D. melanogaster* and other sequenced *Drosophila* (supplementary fig. S1, Supplementary Material online). We also assembled the nearly complete mitochondrial genome for *D. suzukii* (~15 kb), whose size and gene content is similar to that of other sequenced *Drosophila*. Finally, we extracted and assembled the genome of a *Wolbachia* endosymbiont (wSuzi, 1.3 Mb) harbored by the Italian D. *suzukii* population. Preliminary analyses based on several genes identify wSuzi as closely related to *w*Ri from *D. simulans* Riverside (Klasson et al. 2009). A more detailed characterization of wSuzi is presented in Siozios et al. (2013).

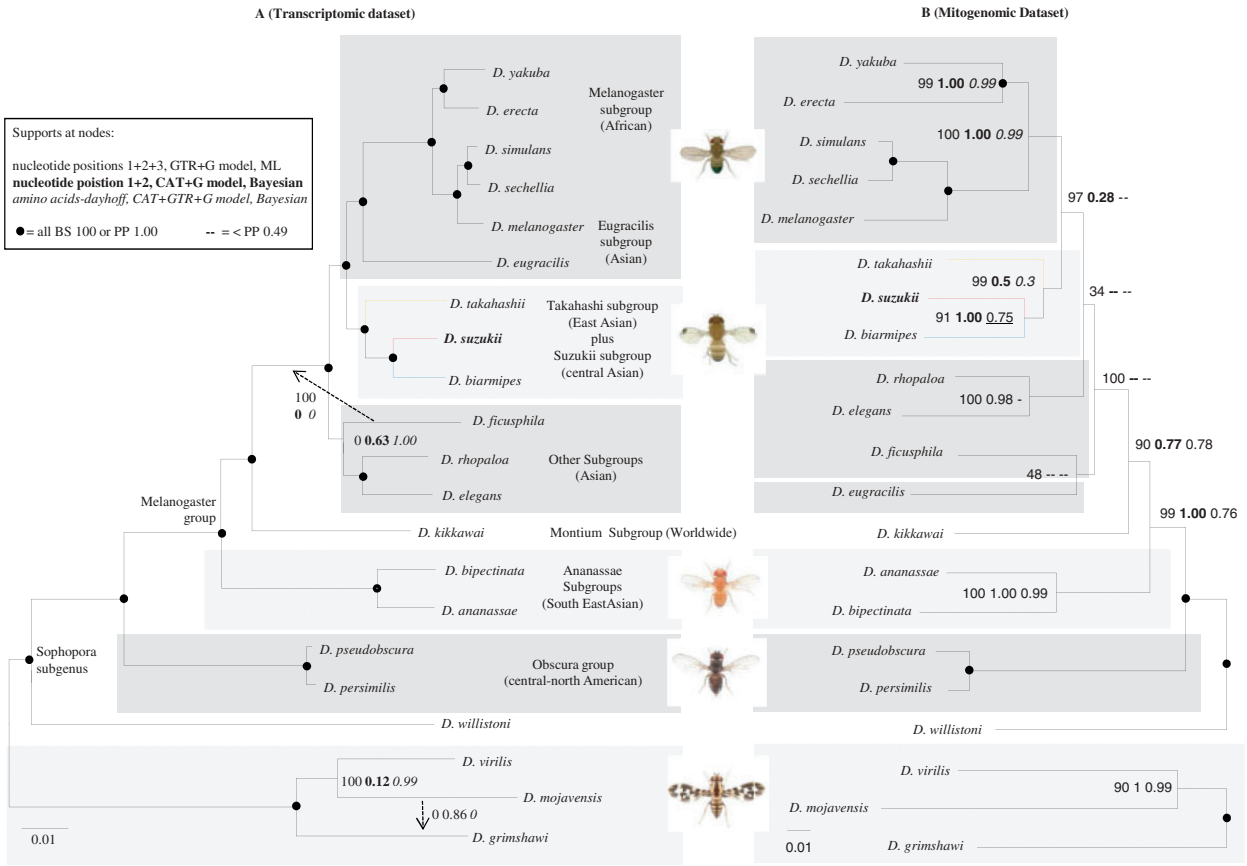


Fig. 1.—The evolutionary affinities of *Drosophila suzukii* and the other *Drosophila* species inferred from phylogenomic and mitogenomic data. (*A*) Phylogenetic analyses of 91 orthologous nuclear genes (200,475 bp). (*B*) Phylogenetic analyses of 12 mitochondrial genes (11,139 bp). Both data sets support an Asian affinity of *D. suzukii*. Drosophila images from Prud'homme and Gompel, used by permission.

Molecular Phylogenetics Using Transcriptomic and mtDNA Data Sets

We used data from the *D. suzukii* genome to conduct a comprehensive multi-locus phylogenetic and dating analysis in the context of genome data from 20 additional *Drosophila* species. We conducted two separate analyses using two distinct data sets.

In the first analysis, we used 91 protein-coding genes extracted from the transcriptomes of the 21 species, covering more than 200,000 nucleotides (fig. 1*A*). We analyzed the aligned data both as nucleotides, excluding third codon positions to exclude likely saturated positions or characters associated with synonymous substitutions, and as amino acid sequences. We also employed different phylogenetic frameworks (Bayesian and ML) and both homogenous and more sophisticated heterogeneous models such as CAT + GTR on a Dayhoff recoded data set (Rota-Stabelli, Lartillot, et al. 2013). All analyses converged on a tree that supported a sister relationship between the *suzukii* and *takahashii* subgroups, and *D. eugracilis* as sister of the *melanogaster* subgroup (fig. 1).

In a second analysis, we reconstructed a phylogeny from the mitochondrial genomes of the 21 *Drosophila* species. We assembled nearly complete mitochondrial genomes for eight additional *Drosophila* species for which whole transcriptome shotgun data were available. Phylogenetic analyses using the same set of experimental procedures used for the transcriptome data set failed to support most of the findings of the genome-derived transcriptome tree (fig. 1*B*).

Molecular Clocks and (Paleo)Ecological Analyses

We performed molecular clock analyses using both the transcriptome and the mtDNA data sets (fig. 2*A*, see Materials and Methods for details). Despite some discrepancies for the ages of nodes closer to the root, the two data sets converged in supporting divergence of *D. suzukii* from *D. biarmipes* in a period between 9 and 6 Ma (i.e., the Tortonian).

To link these clock analyses with the current distribution of *D. suzukii* in Asia, we mapped the current known distribution of *D. suzukii* and their sister species onto a previously compiled climatic model of the Asian Tortonian (fig. 2*B*). The current distribution of *D. suzukii* extends over the Tortonian montane

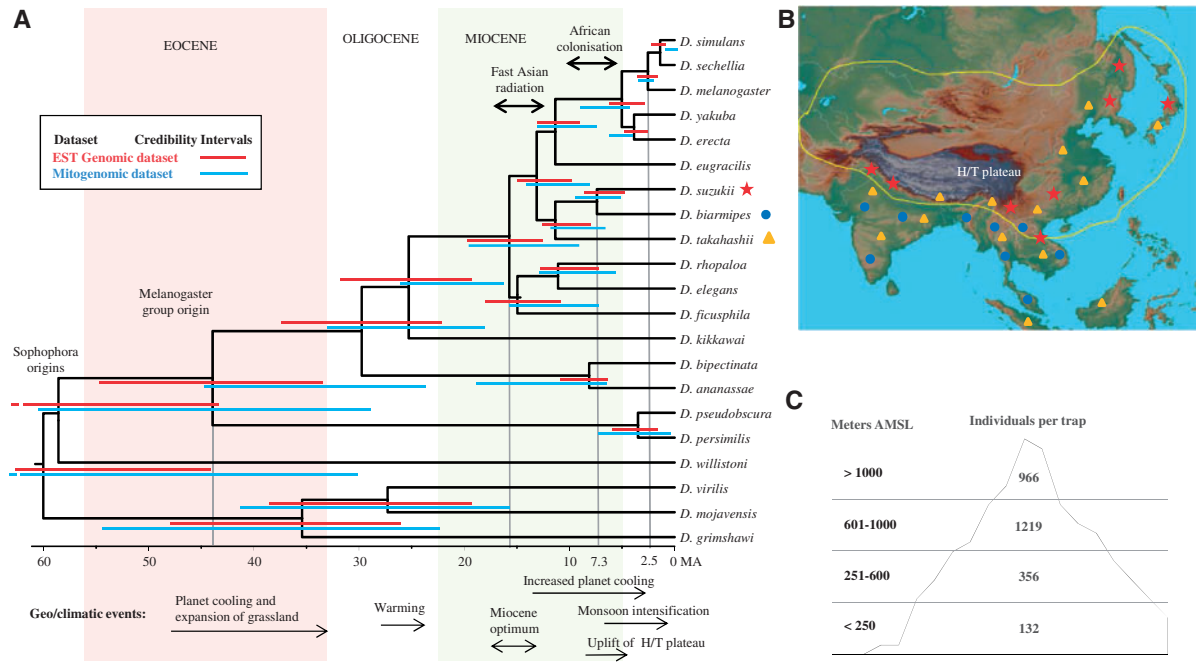


Fig. 2.—Molecular timetrees, paleoclimate, and field trapping suggest a montane-temperate origin of *Drosophila suzukii*. (*A*) Relaxed clock analyses of the *Drosophila* species using both the nuclear and mitochondrial data sets of figure 1. *Drosophila suzukii* is predicted to have diversified toward the late Miocene (Tortonian) simultaneous with an increased uplift of the Himalayan/Tibetan (H/T) plateau and an intensification of the monsoon cycles. Most speciation events (Asian radiation) within the melanogaster group happened just after the mid Miocene climatic optimum in concomitance with further temperature decrease. (*B*) Current endemic geographical distribution of *D. suzukii* (stars) compared with that of *D. biarmipes* (dots) and *D. takahashi* (triangles); yellow line marks the border of temperate (mostly mountainous) forested area during the Tortonian age, the current area being similar but restricted toward the North East. These distributions suggest that *D. suzukii* speciated from *D. biarmipes* by adapting to more temperate mountainous environment. Some species distribution taken from Markow and O'Grady (2005). (*C*) Annual captures per trap at five different altitudes in the Alps confirm a montane/forest optimum for *D. suzukii*, despite greater food resources from fruit production below 600 m asl.

temperate forests, whereas *D. biarmipes* is confined to a more equatorial southern habitat. To investigate a possible preference for temperate climate in *D. suzukii* (see Discussion), we monitored the distribution of *D. suzukii* Italian populations along a gradient of altitude over 1 year (fig. 2*C*). *Drosophila suzukii* preferentially inhabits higher, more temperate altitudes, although the majority of human activity and fruit sources are concentrated at lower altitudes.

Reduced Rate of Molecular Evolution and Reduced Effective Population Size in *D. suzukii*

We explored the patterns of molecular evolution of the *D. suzukii* genome by studying a set of 2,336 orthologous genes from five key species carefully chosen to illuminate key points in its evolutionary history. *Drosophila suzukii* genes are characterized by a slow rate of molecular evolution (fig. 3*A*; supplementary table S4, Supplementary Material online). Both synonymous (d_S) and nonsynonymous (d_N) substitutions rates are significantly lower compared with those of its sister species *D. biarmipes* (fig. 3*B*), consistent with a reduction in substitution rate along the *D. suzukii* branch. This finding is reinforced by a molecular clock analysis that showed that *D. suzukii* has the lowest substitution rate among the *Drosophila* species considered (fig. 3*C*).

We next examined whether in *D. suzukii* the low substitution rate was accompanied by different levels of selective pressure compared with its close relative. The level of overall genomic selective pressure, as measured by the ratio d_N/d_S, is on average significantly lower in *D. suzukii* than in *D. biarmipes* (fig. 3*B*). Interestingly, there is a significantly larger d_N/d_S in autosomal genes of *D. suzukii* compared with those of *D. biarmipes*, whereas the opposite is true for X-linked genes d_N/d_S (fig. 3*B*), consistent with a difference in levels of selective pressure between autosomes and the X chromosome.

To obtain a broader picture of the evolutionary processes, we further analyzed the codon usage in these five species (supplementary tables S2 and S3, Supplementary Material online). In many organisms, synonymous codons are used with different frequencies, leading to codon usage bias. Such bias can be under weak selection ($|N_e s| \approx 1$) and is maintained by the concurrent action of selection, drift, and mutation. Thus, in principle, codon usage bias should be stronger in species with larger effective population size, N_e, compared with species with lower N_e. Both the effective number of codons, *Nc* (Wright 1990), and the frequency of optimal

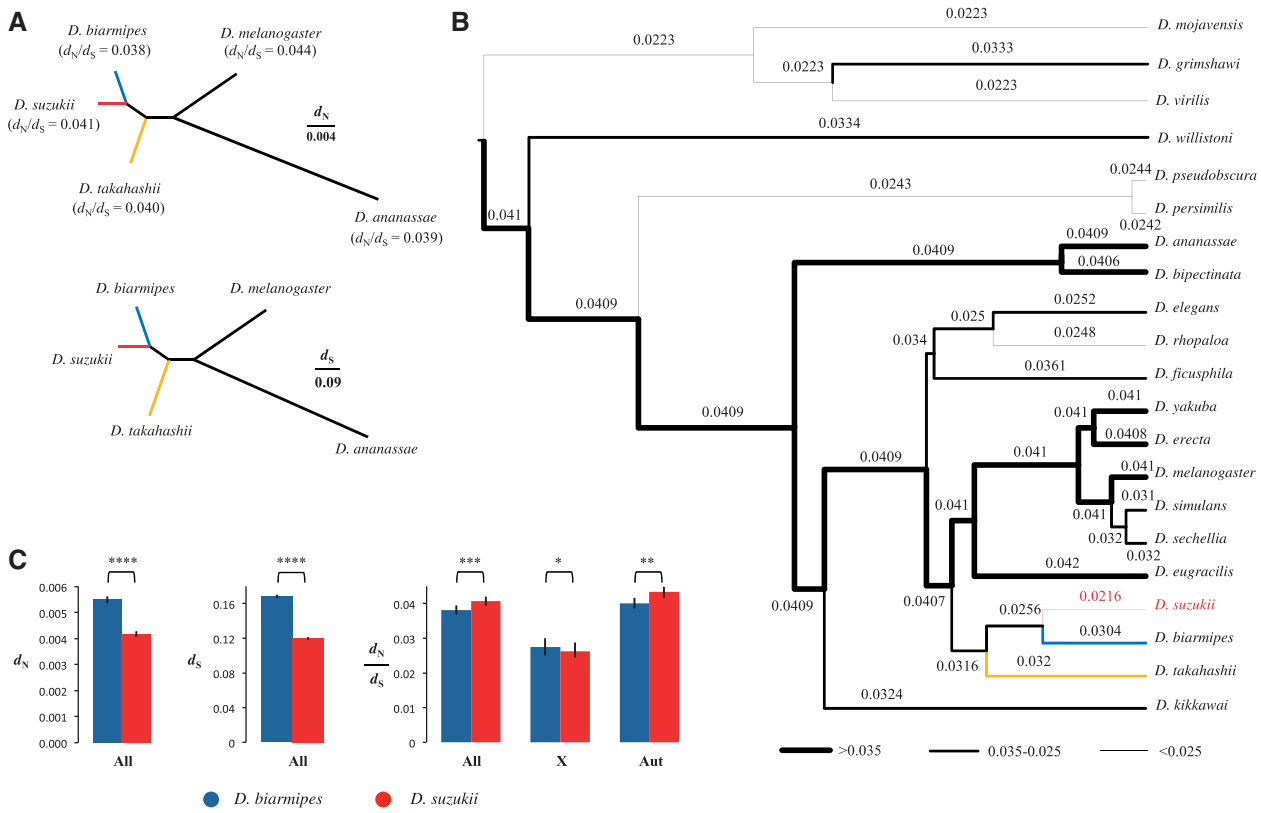


Fig. 3.—The slowly evolving genome of *Drosophila suzukii* can be linked to reduced numbers of generations per year due to winter sexual (female) diapause. (*A*) Consensus evolutionary analysis of 2,336 orthologous genes in five key species. Upper and lower are, respectively, the trees derived from analyses of nonsynonymous (d_N) and synonymous (d_S) substitutions. The d_N/d_S for each species is given in parentheses. (*B*) Branch-specific normally modeled mutation rates as optimized by BEAST using as initial value a mutation rate of 0.0346 neutral substitutions per base pair per million of year (SD = 0.00281). Branch thickness is proportional to the rate. *Drosophila suzukii* is clearly characterized by the lowest rate. Other slower evolving genomes are those of the *virilis–repleta* radiation and of the *pseudobscura* group, which are also preferentially distributed in a temperate/holoartic environment (North American and Central American plateaus). (*C*) A detailed comparison between the rate of molecular evolution in *D. suzukii* and its sister species *D. biarmipes*, for all genes (All) as well for autosomal (Aut) and X-linked genes (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$, Wilcoxon test after controlling for gene length; see also supplementary table S1, Supplementary Material online).

codons, *Fop* (Ikemura 1981), are significantly different between *D. suzukii* and *D. biarmipes* ($P < 10^{-15}$; supplementary table S5, Supplementary Material online) and are consistent with less codon usage bias in the former. Because GC and GC3 (GC in the third codon position) content are significantly correlated to codon usage bias measures in *D. suzukii* and *D. biarmipes* (Spearman's rho > 0.68 for GC and rho < −0.65 for GC3, $P < 10^{-15}$, for both species), we repeated the comparative analyses while correcting for compositional bias. Codon usage bias measures *Nc* and *Fop* do not differ significantly between *D. suzukii* and *D. biarmipes* when correcting for GC or GC3 ($P > 0.139$, for both comparisons). The modified version of *Nc*, which accounts for background nucleotide composition, *Nc′* (Novembre 2002), is significantly larger in *D. suzukii* than in *D. biarmipes* ($P = 6 \times 10^{-11}$), suggesting less codon usage bias in the former.

The analysis of the rate of molecular evolution at the single-gene level revealed only few genes that evolved at different rates in the *D. suzukii* branch compared with the rest of the phylogenetic tree (*D. melanogaster*, *D. ananassae*, [*D. takahashi*, {*D. biarmipes*, *D. suzukii*}]) or where branch-site models detected the occurrence of positive selection specifically affecting sites along the *D. suzukii* branch (tables 1 and 2).

Discussion

The Evolutionary Affinities of *D. suzukii* and the Sister Group of the *Drosophila* Subgroup

Analyses based on 91 nuclear protein-coding genes (fig. 1*A*) confirmed a sister relationship between the *suzukii* and *takahashii* subgroups (Yang et al. 2012). The *melanogaster* subgroup was found to be closely related to *D. eugracilis*, a new hypothesis of *Sophophora* evolution that was extremely robust to various data set treatments (exclusion of third codon positions in nucleotide sequences and translation into

Table 1
Top 10 Genes Identified as Putative Target of Positive Selection along the *Drosophila suzukii* Branch

Drosophila melanogaster Ortholog	P[a]	q Value[b]	p_2[c]	ω_2[d]	Function and Phenotype (Sexual, Neuronal, Thorax)
Cyp4d20	7×10^{-13}	8×10^{-10}	0.0026	15.2	Predicted electron carrier activity. Protein with features of Cytochrome P450. Expression at moderate levels in the following postembryonic organs or tissues: adult head, adult eye, adult heart, adult spermathecae, and adult carcass.
endos	0.00085	0.27948	0.0200	999.0	Predicted sulfonylurea receptor binding activity. Involved in regulation of meiotic cell cycle, mitotic spindle organization, oogenesis, water homeostasis, and response to nutrient. Phenotypically relevant in egg, oocyte, and follicle cell.
Ptp4E	0.00103	0.29554	0.0048	999.0	Predicted transmembrane receptor protein tyrosine phosphatase activity. Involved in motor axon guidance, central nervous system development, and open tracheal system development. Phenotypically relevant in ventral nerve cord.
T48	0.00134	0.33996	0.0084	999.0	Unknown molecular function. Phenotypically relevant in ventral furrow.
CG15626	0.00157	0.36001	0.0030	693.7	Unknown molecular and biological function.
Cyp4aa1	0.00199	0.39060	0.0024	999.0	Predicted electron carrier activity. Involved in insecticide catabolic and hormone metabolic processes.
Osi20	0.00205	0.39060	0.0078	747.8	Unknown molecular and biological function. Phenotypically relevant in trichogen cell.
CG13397	0.00271	0.47676	0.0024	200.4	Predicted alpha-N-acetylglucosaminidase activity. Protein domains suggest involvement in carbohydrate metabolic process.
yemalpha	0.00351	0.52651	0.0022	135.4	DNA binding activity. Involved in female meiosis. Phenotypically relevant in oocyte.
toe	0.00358	0.52651	0.0123	70.3	Predicted molecular function is sequence-specific DNA-binding transcription factor activity. Involved in compound eye development and negative regulation of transcription from RNA polymerase II promoter. Phenotypically relevant in scutum and scutellum (mesothoracic tergum).

[a]Likelihood ratio test probability based on branch–site models of codon evolution, with *D. suzukii* set as foreground branch.
[b]Proportion of false positives (FDR) of the test.
[c]Proportion of sites under positive selection estimated in the foreground branch (*D. suzukii*) by the branch–site model A.
[d]ω estimated for the sites under positive selection in the foreground branch (*D. suzukii*) by the branch–site model A.

the corresponding amino acid sequences) and experimental procedures (use of homogenous and heterogeneous substitution models in both a Bayesian and ML framework; see legend of fig. 1*A*). Not all relationships were well resolved using this large data set. The placement of *D. ficusphila* is data set and model dependent, but the use of the sophisticated CAT + GTR model coupled with Dayhoff recoding of the amino acid data set (performed to reduce possible systematic errors in phylogenomic analyses; Rota-Stabelli, Lartillot, et al. 2013) points toward its grouping with *D. rhopaloa* and *D. elegans*.

To corroborate our phylogenetic results, we further analyzed an mtDNA data set, which failed to support all the findings of the nuclear gene set tree (fig. 2*B*). This is most likely because of a lack of phylogenetic signal in the mtDNA. Thus, the apparently robust bootstrap support (97%) against the sister relationship *D. eugracilis-melanogaster* subgroup vanishes when highly saturated third codon positions are excluded, or when an amino acid data set was employed, indicating that signal contradicting the nuclear phylogeny carried by mitochondrial genomes is concentrated in unreliable third codon positions and/or synonymous substitutions. Overall, our phylogenetic analyses reveal that the African *melanogaster* subgroup evolved from within a rapid Asian radiation, identifying *D. eugracilis* as a key intermediate species to polarize evolutionary traits of the *melanogaster* subgroup. With respect to the placement of *D. suzukii* in the phylogeny, our analyses suggest that this species is the sister taxon of *D. biarmipes*. *Drosophila subpulchrella*, another little-studied fly in the *suzukii* subgroup, has been reported to have feeding behavior similar to that of *D. suzukii* (Mitsui et al. 2010). This species is, however, thought to be most closely related to *D. pulchrella* (hence its name), which is sister to the *suzukii* + *biarmipes* clade (Yang et al. 2012), suggesting independent acquisition of unripe fruit feeding. It will be important to explore the relationships of *D. subpulchrella* using genome-scale data.

Rate of Molecular Evolution Suggests Winter Diapause

Our results indicate that gene sequences evolve at a significantly lower rate in *D. suzukii* than in its sister species *D. biarmpes* (fig. 3 and table 1). We hypothesize that the

Table 2

Top 10 Genes Evolving at a Significantly Different Rate along the *Drosophila suzukii* Branch

Drosophila melanogaster Ortholog	P^a	*q* Value[b]	$\omega_{FB}{}^c$	$\omega_R{}^d$	Function and Phenotype (Sex, Neuron, Thorax)
ran	6×10^{-10}	0.000001	0.0752	0.0001	GTP and protein-binding activity. Involved in regulation of meiotic spindle organization, cell cycle, cell shape, cell adhesion, and actin filament organization. Phenotypically relevant in photoreceptor cell R7, meiotic spindle, karyosome, ommatidium, and pigment cell.
mtm	1×10^{-9}	0.000002	0.1639	0.0190	Phosphatidylinositol-3-phosphatase activity. Involved in mitotic cell cycle, chromosome segregation, and response to wounding. Phenotypically relevant in sessile hemocyte and embryonic/larval hemocyte.
wcd	4×10^{-8}	0.00003	0.3814	0.0782	Unknown molecular function. Involved in ribosome biogenesis, neuroblast proliferation, and female germ-line stem cell division. Phenotypically relevant in trichogen cell and mesothoracic tergum.
l(3)72Dn	9×10^{-8}	0.00005	0.4781	0.1330	Unknown molecular function. Involved in ribosome biogenesis and neurogenesis. Phenotypically relevant in mesothoracic tergum.
CG9135	7×10^{-7}	0.00031	0.2774	0.0150	Predicted guanyl-nucleotide exchange factor activity. Unknown biological function. Phenotypically relevant in mesothoracic tergum.
CG8562	3×10^{-6}	0.00105	0.3100	0.0624	Predicted metallocarboxypeptidase activity. Protein domains suggest involvement in proteolysis.
Oatp33Ea	0.00001	0.00175	0.2131	0.0556	Predicted organic anion transmembrane transporter activity.
lbk	0.00001	0.00175	0.0001	0.0435	Involved in chaeta morphogenesis and oogenesis.
lid	0.00001	0.00175	0.1231	0.0421	Histone acetyltransferase and demethylase activity. Involved in chromatin organization, and histone acetylation and demethylation. Phenotypically relevant in mesothoracic tergum, imaginal disc, and embryonic/larval optic lobe.
dome	0.00001	0.00270	0.0186	0.0886	Transmembrane signaling receptor activity and protein heterodimerization activity. Involved in blastoderm segmentation, hindgut morphogenesis, border follicle cell migration, long-term memory, JAK-STAT cascade, open tracheal system development, and compound eye morphogenesis. Phenotypically relevant in spiracle, integumentary specialization, embryonic hindgut, and compound eye.

[a]Likelihood ratio test probability based on branch models of codon evolution, with *D. suzukii* set as foreground branch.
[b]Proportion of false positives (FDR) of the test.
[c]ω estimated for the focal (*D. suzukii*) branch.
[d]ω estimated for the rest of the phylogenetic tree.

low substitution rate of *D. suzukii* could be due to an idiosyncratic, low mutation rate, and/or because the species has a reduced number of generations per year compared with its relatives. The reproductive ecology of the species supports the second hypothesis, as in its distributional range *D. suzukii* reproduces only during the warm season and is able to overwinter as sexually immature, cold tolerant females (Mitsui et al. 2010). Our genomic evidence supports the hypothesis that *D. suzukii* has a winter sexual diapause and thus had a reduced number of generations since its last common ancestor with *D. biarmipes*.

Reduced Effective Population Size Affects the Efficiency of Positive Selection in *D. suzukii*

The level of overall genomic selective pressures, as measured by the ratio d_N/d_S, is lower in *D. suzukii* than in *D. biarmipes* (fig. 2*B*). This result is consistent with more relaxed selection along the *D. suzukii* lineage, possibly because this species has a smaller effective population size, N_e, than *D. biarmipes*. A reduced N_e would allow the fixation of larger number of slightly deleterious nonsynonymous mutations (Charlesworth 2009), as further supported by the observation that *D. suzukii* has a lower frequency of optimal codons and a lower codon usage bias than *D. biarmipes* (supplementary tables S2 and S3, Supplementary Material online). The alternative possibility that larger d_N/d_S values correspond to pervasive positive selection in the *D. suzukii* genome (i.e., increased fixation of beneficial mutations) is not supported by our data. First, the fixation of favorable alleles in multiple genes would lead to a high dispersion in the distribution of d_N across the genome (Presgraves 2005), whereas the variance in d_N is significantly lower in *D. suzukii* than in *D. biarmipes* (2.9×10^{-5} vs. 4.4×10^{-5}, *F* test $P < 10^{-15}$). Second, only a few genes were detected as significant targets of positive selection in *D. suzukii* (table 1). Thus, the most likely explanation for the low substitution rate of *D. suzukii* is a reduced number of generations per year and a smaller N_e compared with its relatives. It is unlikely that the reduced N_e is a direct consequence of the bottleneck

associated with a colonization of Europe, as the invasion took place only few generations ago (the first record dates back to 2008, Cini et al. 2012), and thus the genome-wide pattern of substitutions should represent that of the ancestral population. We propose instead that the low substitution rate reflects the ecology and evolutionary history of this species. The winter diapause, we suggest, explains both the low substitution rate and the reduced selection efficiency in *D. suzukii* compared with *D. biarmipes*. Overwintering diapause results in recurrent population size bottlenecks, particularly for males, and thus in a lower effective population size N_e, and lower selection efficiency in removing slightly deleterious nonsynonymous mutations, as indicated by its genome-wide higher d_N/d_S.

The hypothesis that males undergo more severe bottlenecks than females is supported by the discrepancy in levels of selective pressure between the autosomes and the X chromosome. As males contain only one copy of X (and two of the autosomes), sex-biased population size changes would alter relative levels of X-linked and autosomal N_e, namely by decreasing N_e of autosomes 2-fold relative to X chromosome in males. We indeed observed a significantly larger d_N/d_S in autosomal genes of *D. suzukii* than of *D. biarmipes*, whereas in X-linked genes, d_N/d_S was lower in *D. suzukii* than in *D. biarmipes* (fig. 3*B*). If we assume that levels of d_N/d_S are a proxy for effective population size, the X/autosome ratio of N_e values, N_{eX}/N_{eA}, seems to be higher in *D. suzukii* than in *D. biarmipes*, possibly leading to differences in the efficiency of selection on the X and autosomes between the two species. One hypothesis to explain this observation is a difference in the efficiency of purifying selection in removing recessive deleterious mutations in hemizygous males, a phenomenon which can often lead to a faster-X effect (Charlesworth et al. 1987; Vicoso and Charlesworth 2009a). Thus, the bottlenecks associated with the winter diapause of *D. suzukii* could be directly responsible for the relative difference in N_{eX}/N_{eA} between the two species (Pool and Nielsen 2007). Other factors that may have affected the differences in the ratio N_{eX}/N_{eA} between *D. suzukii* and *D. biarmipes* include different recombination rates (Vicoso and Charlesworth 2009b) and variance in male reproductive success due to sexual competition among males (Andersson 1994; see Mank et al. 2010 for a review). Additional genetic and behavioral studies will be necessary to disentangle these forces and evaluate their role in the evolution of *D. suzukii*.

Paleobiology and Adaptation to Temperate Ecology

The presence of a winter diapause in *D. suzukii* may be an adaptation that is relevant to the switch in ecology of the species. Relaxed clock studies of both nuclear and mitochondrial genomes (fig. 2*A*) converged on a scenario in which *D. suzukii* diverged from *D. biarmipes* approximately 9–6 Ma, toward the end of the Miocene (Tortonian). Climate modeling has shown that, during the Tortonian, the ecology of region between North India, Indochina, and the Chinese coasts (delineated by the yellow line in fig. 2*B*) was characterized by extended montane temperate forests. Toward the present, forests reduced in extent to the North and East and alternated with scrublands or tropical forests (Pound et al. 2011). The present endemic distribution of *D. suzukii* extends over this region, whereas *D. biarmipes* is endemic to a more equatorial, southern habitat. The distribution of the two species suggests that speciation of *D. suzukii* was accompanied by adaptation to temperate habitats, through the increased uplift of the Tibetean plateau and the concomitant intensification of the monsoon regime in the Tortonian (Zachos et al. 2001).

A strong preference for montane temperate climate in current invasive populations is supported by the results of trapping surveys. Although extensive soft fruit production is concentrated below 600 m asl in the surveyed Trentino Province of North Italy, the majority of the captures we made were at higher altitudes (fig. 2*C*). The proposal that *D. suzukii* originated in a temperate, montane ecology is congruent with its current life habit. In temperate forests, fruit production, and thus the availability of rotting fruit, is highly seasonal, whereas in the tropics fruiting, and thus the production of rotted food sources, is near-continuous (Willson 1991; Ting et al. 2008). For a species occupying a temperate ecosystem, ovipositing in fresh fruit is required to access food. Growing larvae can then accelerate decay and fermentation to provide food for the adult stage. Overwintering diapause bridges the winter months when fruits are scarce if not absent, and low temperatures limit both fermentation and fly activity.

Preadaptations Suggest Invasive Success

An innate predisposition to temperate climates might also explain why *D. suzukii* was able to invade European and North American regions so rapidly. *Drosophila melanogaster* has also invaded temperate climates, but the colonization originated from an ancestral tropical African range and was accompanied by local adaptation (Ometto et al. 2005), whereas invading populations of *D. suzukii* are likely to have already had many traits adaptive in the newly occupied range. Genes under positive selection (table 1) and those with fast evolutionary rates (table 2) are good candidates for loci involved in such adaptations. We found no evidence for a significant overrepresentation of Gene Ontology-defined functional classes in these gene sets, but many of the identified genes are phenotypically linked (through their *D. melanogaster* orthologs) with biology of the genitalia, the neuronal system and (particularly and unexpectedly) with the mesothoracic tergum.

The genetic and neurological bases of the adaptive behavioral and lifecycle traits of *D. suzukii* may hold the keys to

understanding the origin of a novel behavioral repertoire and lead to strategies for control of this pest in western countries. Because the current hypothesis of the phylogeny within the *suzukii* subgroup has not yet been confirmed by whole-genome phylogenetics, we cannot exclude the possibility that *D. subpulchrella* is sister to *D. suzukii* rather than *D. pulchrella*. Under this scenario, some adaptations currently modeled as arising within *D. suzukii* may in fact be shared with *D. subpulchrella*. Genome sequencing of *D. subpulchrella* will clarify this question. Our evolutionary analyses of the *D. suzukii* genome suggest an intriguing causal link between adaptation to temperate environments and its particular biology. The genetic bases of adaptation to temperature could be a key factor to develop new pesticides or containment strategies for this pest.

Supplementary Material

Supplementary figures S1–S3 and tables S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org).

Acknowledgments

The authors acknowledge the Baylor College of Medicine Human Genome Sequencing Center and the NHGRI for prepublication access to their data and staff of the GenePool for their support. They thank Darren Obbard for thorough comments on the manuscript. Computational analyses were performed using the infrastructures of the Foundation Edmund Mach (FEM), Foundation Bruno Kessler (FBK), and the NUI Maynooth High Performance Computing (HPC). O.R.-S. is supported by a Marie Curie/Trento Province Cofound FP7 fellowship. The GenePool is supported by the UK MRC (MR/K001744/1) and NERC (R8/H10/56). This project was funded by Accordo di Porgamma of the *Autonomous* Province of Trento.

Literature Cited

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 38:W7–W13.

Andersson M. 1994. Sexual selection. Princeton (NJ): Princeton University Press.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17: 540–552.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 10:195–205.

Charlesworth B, Coyne JA, Barton NH. 1987. The relative rates of evolution of sex chromosomes and autosomes. Am Nat. 130:113–146.

Cini A, Ioriatti C, Anfora G. 2012. A review of the invasion of *Drosophila suzukii* in Europe and a draft research agenda for integrated pest management. Bull Insectol. 65:149–160.

David JR, Capy P. 1988. Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet. 4:106–111.

Dreves AJ. 2011. IPM program development for an invasive pest: coordination, outreach and evaluation. Pest Manag Sci. 67:1403–1410.

Drosophila 12 Genomes Consortium, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 7:214.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.

Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. Genome Res. 9:868–877.

Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol. 151:389–409.

Jurka J, et al. 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110:462–467.

Klasson L, et al. 2009. The mosaic genome structure of the *Wolbachia wRi* strain infecting *Drosophila simulans*. Proc Natl Acad Sci U S A. 106: 5725–5730.

Kurtz S, Narechania A, Stein JC, Ware D. 2008. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics 9:517.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659.

Löytynoja A, Goldman N. 2008. A model of evolution and structure for multiple sequence alignment. Philos Trans R Soc Lond B Biol Sci. 363: 3913–3919.

Mank JE, Vicoso B, Berlin S, Charlesworth B. 2010. Effective population size and the Faster-X effect: empirical results and their interpretation. Evolution 64:663–674.

Markow TA, O'Grady P. 2005. *Drosophila:* a guide to species identification and use. London: Elsevier.

Mitsui H, Beppu K, Kimura MT. 2010. Seasonal life cycles and resource uses of flower- and fruit-feeding drosophilid flies (Diptera: Drosophilidae) in central Japan. Entomol Sci. 13:60–67.

Novembre JA. 2002. Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol. 19:1390–1394.

Obbard DJ, et al. 2012. Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. Mol Biol Evol. 29:3459–3473.

Ometto L, Glinka S, de Lorenzo D, Stephan W. 2005. Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol Biol Evol. 22: 2119–2130.

Pound MJ, et al. 2011. A Tortonian (Late Miocene, 11.61–7.25 Ma) global vegetation reconstruction. Palaeogeogr Palaeoclimatol Palaeoecol. 300:29–45.

Pool JE, Nielsen R. 2007. Population size changes reshape genomic patterns of diversity. Evolution 61:3001–3006.

Presgraves DC. 2005. Recombination enhances protein adaptation in *Drosophila melanogaster*. Curr Biol. 15:1651–1656.

Prud'homme B, et al. 2006. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature 440:1050–1053.

R Development Core Team. 2009. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Rota-Stabelli O, Blaxter M, Anfora G. 2013. Quick guide: *Drosophila suzukii*. Curr Biol. 23(1):R8–R9.
Rota-Stabelli O, Daley A, Pisani D. 2013. Molecular timetrees reveal a Cambrian colonisation of land and a new scenario for ecdysozoan evolution. Curr Biol. 23(5):329–398.
Rota-Stabelli O, Lartillot N, Philippe H, Pisani D. 2013. Serine codon usage bias in deep phylogenomics: pancrustacean relationships as a case study. Syst Biol. 62(1):121–133.
Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086–1092.
Simpson JT, et al. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117–1123.
Siozios S, et al. Forthcoming 2013. Draft genome of the *Wolbachia* endosymbiont of *Drosophila suzukii*. Genome Announc. 1(1):e00032–13; doi:10.1128/genomeA.00032-13.
Storey J. 2002. A direct approach to false discovery rates. J Roy Stat Soc B. 64:479–498.
Ting S, Hartley S, Burns KC. 2008. Global patterns in fruiting seasons. Global Ecol Biogeogr. 17:648–657.
Vicoso B, Charlesworth B. 2009a. Effective population size and the faster-X effect: an extended model. Evolution 63:2413–2426.
Vicoso B, Charlesworth B. 2009b. Recombination rates may affect the ratio of X to autosomal noncoding polymorphism in African populations of *Drosophila melanogaster*. Genetics 181: 1699–1701.
Walsh DB, et al. 2011. *Drosophila suzukii* (Diptera: Drosophilidae): invasive pest of ripening soft fruit expanding its geographic range and damage potential. J Integr Pest Manag. 2:1–7.
Willson MF. 1991. Dispersal of seeds by frugivorous animals in temperate forests. Rev Chil Hist Nat. 64:537–554.
Wright F. 1990. The "effective number of codons" used in a gene. Gene 87:23–29.
Yang Y, Hou Z-C, Qian Y-H, Kang H, Zeng Q-T. 2012. Increasing the data size to accurately reconstruct the phylogenetic relationships between nine subgroups of the *Drosophila melanogaster* species group (Drosophilidae, Diptera). Mol Phylogenet Evol. 62:214–223.
Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 15:568–573.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17:32–43.
Yang Z, Nielsen R, Goldman N, Pedersen A. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.
Yang Z, Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol Biol Evol. 23:212–226.
Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22: 1107–1118.
Zachos J, Pagani M, Sloan L, Thomas E, Billups K. 2001. Trends, rhythms, and aberrations in global climate 65 Ma to present. Science 292: 686–693.

Associate editor: B. Venkatesh

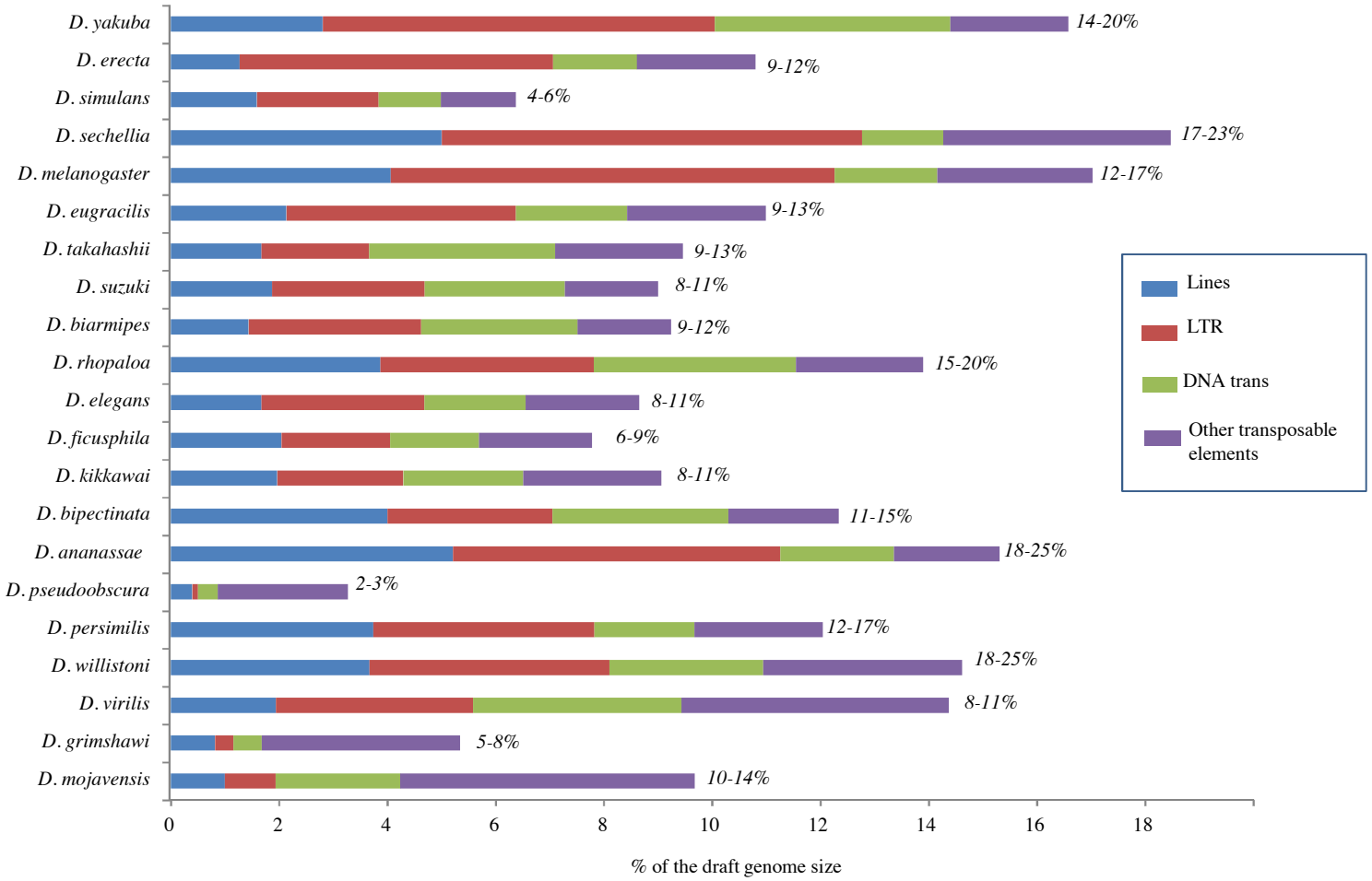


Figure S1: Repeated elements in *D. suzukii* genome. The distribution and number of repeats in *D. suzukii* is similar to that of sister species *D. biarmipes* and *D. takahashii*, thus making sense of their phylogeny. In most other cases there is not a similarity between closely relates species, see for example *D. yakuba* and *D. erecta*.

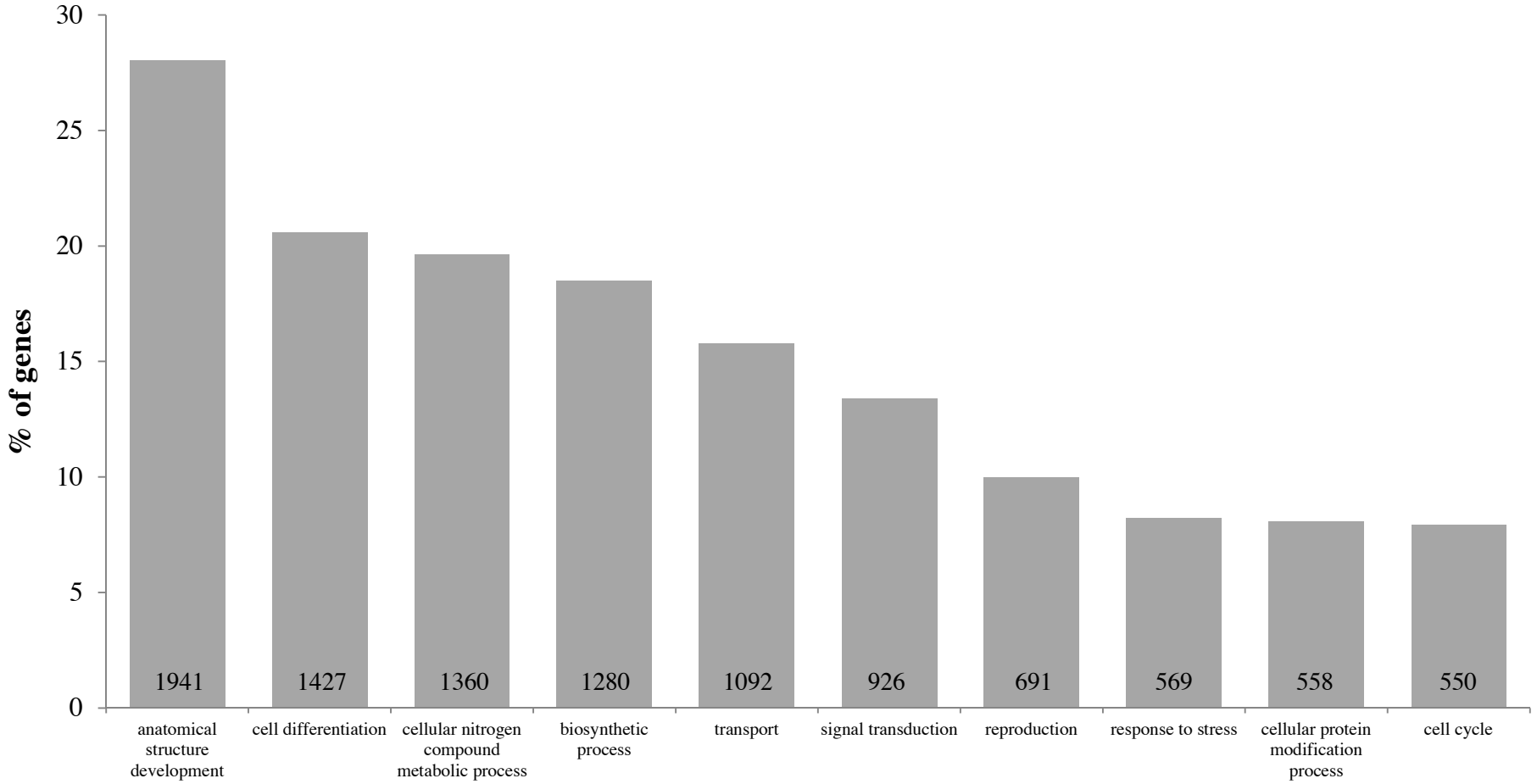


Figure S2. Putative function of the *D. suzukii* genes. *D. suzukii* genes obtained through RNASeq sequencing were blasted against *D. melanogaster* genes. A total of 8,137 reciprocal best hits were retained as putative orthologues, representing 74% of the *D. melanogaster* annotated genes. Assuming conserved synteny and chromosomal organization between *D. suzukii* and *D. melanogaster*, we could verify that chromosomes were evenly covered by our RNA-seq gene sequences (chromosome 2: 72.2%, chr. 3: 76.4%, chr. X: 73.7%; exception is chr. 4: 53.8%). Putative function was assigned based on gene ontology (GO) terms of the *D. melanogaster* genes using the web tool available at http://go.princeton.edu. Only categories representing more than 7% of the total GO terms in the list are shown (the actual number of genes is given within the bars).

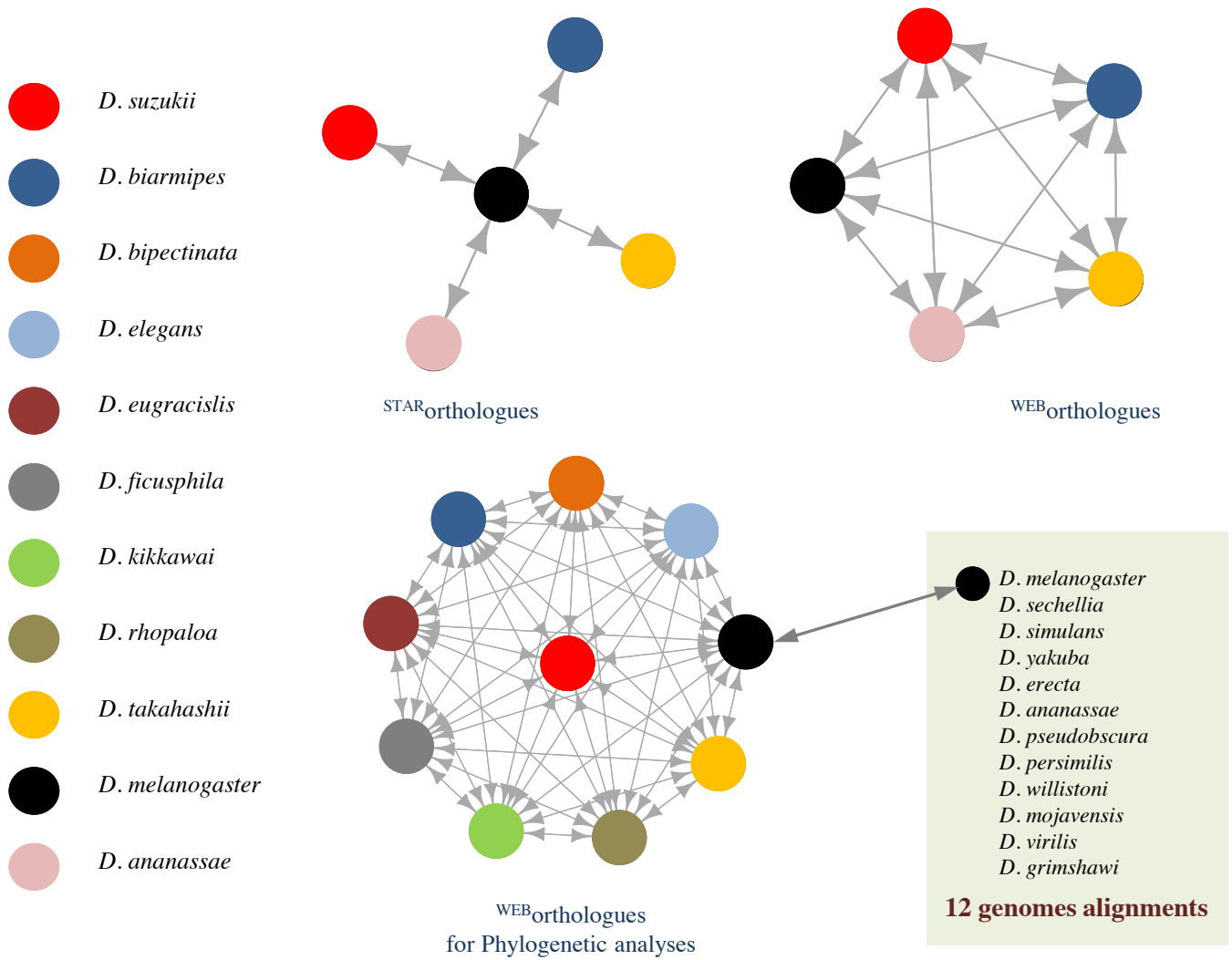


Figure S3. Orthologues search. Putative orthologues were identified using a reciprocal best-hit approach across *Drosophila* species. See Methods for details.

Table S1: *D. suzukii* genome cleaning statistics. Number of reads in libraries and number of reads that blast against different W*olbachia* genomes and *D. melanogaster* mtDNA.

Genome	NCBI accession	*n* reads (180 bp)	*n* reads (300 bp)
***D. suzukii* initial reads**	ERA an	134,306,528	103,584,510
W. simulans	NZ_AAGC00000000.1	126,346	218,494
W. melanogaster	NC_002978.6	161,996	288,266
W. ananasse	NZ_AAGB00000000.1	77,484	259,026
W. willistoni	NZ_AAQP00000000.1	115,601	197,998
W. wRi	NC_012416.1	162,071	288,803
all Wolbachia		399,088	683,606
***D. melanogaster* mtDNA**	NC_001709	59,342	129,646
D. suzukii* reads after cleaning		124,381,960	96,007,805

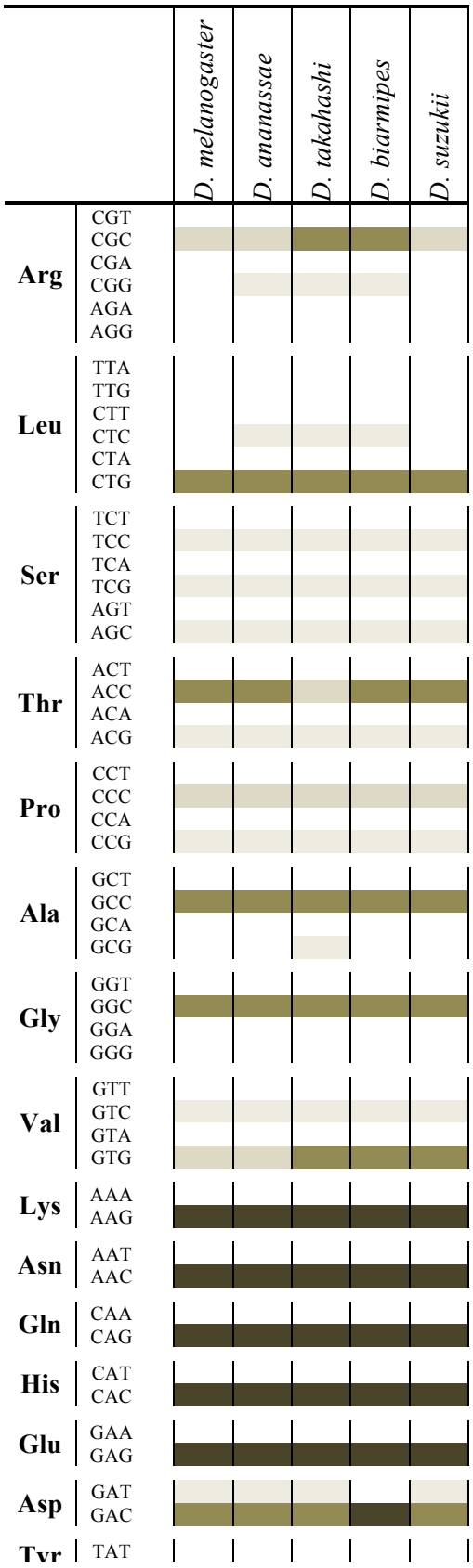
Columns "180bp" and "300bp" indicates the number of matching reads for the tow libraries. * After the quality checks reads had an average length of 93 bp (sd14) for the 180bp library and of 94bp (sd 15) for 300 bp library, and an average quality value of 35.

Table S2: Genome assembly statistics. Abyss trials with different k-mer size.

K-mer size	N contigs	n:200	n:N50	N80	N50	N20	max	Sum (MBp)
48	1,399,155	93,256	7,826	1,369	4,756	18,300	208,969	185.3
54	1,200,237	105,190	9,204	1,283	4,445	15,559	169,965	195.7
64	961,286	131,597	12,820	1,089	3,565	11,309	169,947	209.6

n:200 is the number of contigs shorter than 200 bp, n:N50 is the number of contigs longer than the median, N80 is the size of the 80 percentile, N50 is the median contig size, N20 is the size of the 20 percentile, sum is the overall contigs size in millions of base pairs.

Table S3. Codon usage in Drosophila. Darker colors identify synonymous codons used at higher frequency among the ^{STAR}orthologues.



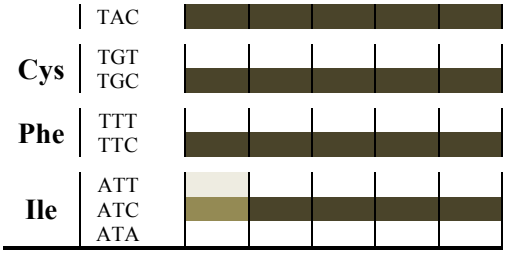

TAC
Cys
TGT
TGC
Phe
TTT
TTC
Ile
ATT
ATC
ATA

Table S4. Total branch length (*tot*) and rate of synonymous (d_S) and non-synonymous substitution (d_N) across [WEB]orthologues along the *Drosophila biarmipes* and *D. suzukii* lineages.

	D. biarmipes			***D. suzukii***			***P*** [b]	
	n [a]	**mean**	**SD**	***n***	**mean**	**SD**	**raw**	**res_L**
tot	1002	0.1092	0.0479	1002	0.0789	0.0372	1×10^{-49}	0.0076
d_N	1002	0.0048	0.0056	1002	0.0037	0.0046	3×10^{-9}	0.0004
d_S	1002	0.1628	0.0765	1002	0.1169	0.0575	6×10^{-51}	0.0376
d_N/d_S	999	0.0352	0.0662	989	0.0365	0.0648	0.5799	0.0058

[a] Exceedingly large d_N/d_S values corresponding to $d_S = 0$ were ignored in the analysis. SD = standard deviation.
[b] Wilcoxon test probability calculated for the raw data and after correcting for gene length (res_L; we used the residuals of the correlation between gene length and the statistic of interest).

Table S5. Mean (Standard deviation, SD) codon usage and GC content in [STAR]orthologues..

	Fop [a]		***Nc*** [b]		***Nc'*** [c]		**GC** [d]		**GC3** [e]	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
D. suzukii	0.668	0.093	43.9	6.7	50.2	4.7	56.6	4.0	72.3	8.9
D. biarmipes	0.745	0.097	41.0	7.1	49.3	4.8	58.3	4.0	76.9	9.1
D. takahashi	0.724	0.094	42.0	6.6	49.5	4.8	57.5	4.0	74.9	8.9
D. melanogaster	0.613	0.086	45.9	6.5	50.9	4.4	55.6	3.7	69.3	8.2
D. ananassae	0.662	0.098	45.6	7.2	50.9	4.4	55.8	4.2	70.0	9.4

[a] Frequency of the optimal codon.
[b] Number of effective codons.
[c] Number of effective codons when accounting for background nucleotide composition.
[d] Percentage of GC content across genes.
[e] Percentage of GC content in the third codon position across gene