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

Lino Ometto1, Alessandro Cestaro1, Sukanya Ramasamy1, Alberto Grassi2, Santosh Revadi1, Stefanos Siozios1, Marco Moretto1, Paolo Fontana1, Claudio Varotto1, Davide Pisani3, Teun Dekker4, Nicola Wrobel5, Roberto Viola1, Ilaria Pertot1, Duccio Cavaleri1, Mark Blaxter5, Gianfranco Anfora1, and Omar Rota-Stabelli1,*

1Research and Innovation Centre, Fondazione Edmund Mach, San Michele all’Adige, Italy
2Technological Transfer Centre, Fondazione Edmund Mach, San Michele all’Adige, Italy
3School of Biological Sciences and School of Earth Sciences, University of Bristol, Bristol, United Kingdom
4Division of Chemical Ecology, Swedish University of Agricultural Sciences, Alnarp, Sweden
5Institute 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. willistoni*, and *wif*) 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.
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. pseudoobscura*, *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_speciesguide_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* ortholog 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 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, $d_\text{N}$ (leading to amino acid changes), and synonymous substitution, $d_\text{S}$ (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 $\omega = d_\text{N}/d_\text{S}$, 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 $\omega$ 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 $\omega$ 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$\Delta$ln) was tested using a $\chi^2$ 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), $\omega$ 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 STAR Orthologs 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, $P < 10^{-15}$). 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. ficushila, 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 1A. 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 wRI from D. simulans Riverside (Klasson et al. 2009). A more detailed characterization of wSuzi is presented in Siozios et al. (2013).
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. 1A). 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. 1B).

Molecular Clocks and (Paleo)Ecological Analyses

We performed molecular clock analyses using both the transcriptome and the mtDNA data sets (fig. 2A, 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. 2B). The current distribution of *D. suzukii* extends over the Tortonian montane
temperate forests, whereas \textit{D. biarmipes} is confined to a more equatorial southern habitat. To investigate a possible preference for temperate climate in \textit{D. suzukii} (see Discussion), we monitored the distribution of \textit{D. suzukii} Italian populations along a gradient of altitude over 1 year (fig. 2C). \textit{Drosophila suzukii} preferentially inhabits higher, more temperate altitudes, although the majority of human activity and fruit sources are concentrated at lower altitudes.

\textbf{Reduced Rate of Molecular Evolution and Reduced Effective Population Size in \textit{D. suzukii}}

We explored the patterns of molecular evolution of the \textit{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. \textit{Drosophila suzukii} genes are characterized by a slow rate of molecular evolution (fig. 3A; supplementary table S4, Supplementary Material online). Both synonymous ($d_s$) and nonsynonymous ($d_N$) substitution rates are significantly lower compared with those of its sister species \textit{D. biarmipes} (fig. 3B), consistent with a reduction in substitution rate along the \textit{D. suzukii} branch. This finding is reinforced by a molecular clock analysis that showed that \textit{D. suzukii} has the lowest substitution rate among the \textit{Drosophila} species considered (fig. 3C).

We next examined whether in \textit{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 \textit{D. suzukii} than in \textit{D. biarmipes} (fig. 3B). Interestingly, there is a significantly larger $d_N/d_S$ in autosomal genes of \textit{D. suzukii} compared with those of \textit{D. biarmipes}, whereas the opposite is true for X-linked genes $d_{N/X}/d_{S/X}$ (fig. 3B), 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 ($\langle N_e \rangle \approx 1$) 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, $N_c$ (Wright 1990), and the frequency of optimal
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 \( N_c \) 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 \( N_c \), which accounts for background nucleotide composition, \( N_c' \) (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. 1A) confirmed a sister relationship between the suzukii and takahashi 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
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. 1A). Not all relationships were well resolved using this large data set. The placement of \textit{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 \textit{D. rhopaloe} and \textit{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. 2B). 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 \textit{D. eugracilis}-\textit{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 \textit{melanogaster} subgroup evolved from within a rapid Asian radiation, identifying \textit{D. eugracilis} as a key intermediate species to polarize evolutionary traits of the \textit{melanogaster} subgroup. With respect to the placement of \textit{D. suzukii} in the phylogeny, our analyses suggest that this species is the sister taxon of \textit{D. biarmipes}. \textit{Drosophila subpulchrella}, another little-studied fly in the \textit{suzukii} subgroup, has been reported to have feeding behavior similar to that of \textit{D. suzukii} (Mitsui et al. 2010). This species is, however, thought to be most closely related to \textit{D. pulchrellae} (hence its name), which is sister to the \textit{suzukii} + \textit{biarmipes} clade (Yang et al. 2012), suggesting independent acquisition of unripe fruit feeding. It will be important to explore the relationships of \textit{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 \textit{D. suzukii} than in its sister species \textit{D. biarmipes} (fig. 3 and table 1). We hypothesize that the
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. 2B). 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* ($2.9 \times 10^{-5}$ vs. $4.4 \times 10^{-5}$, $F$ 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

### Table 2

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

<table>
<thead>
<tr>
<th>Ortholog</th>
<th>$P^{a}$</th>
<th>q Value$^{b}$</th>
<th>$\omega_{B}^{c}$</th>
<th>$\omega_{N}^{d}$</th>
<th>Function and Phenotype (Sex, Neuron, Thorax)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ran</td>
<td>$6 \times 10^{-10}$</td>
<td>0.000001</td>
<td>0.0752</td>
<td>0.0001</td>
<td>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.</td>
</tr>
<tr>
<td>mtm</td>
<td>$1 \times 10^{-9}$</td>
<td>0.000002</td>
<td>0.1639</td>
<td>0.0190</td>
<td>Phosphatidylinositol-3-phosphatase activity. Involved in mitotic cell cycle, chromosome segregation, and response to wounding. Phenotypically relevant in sessile hemocyte and embryonic/larval hemocyte.</td>
</tr>
<tr>
<td>wcd</td>
<td>$4 \times 10^{-8}$</td>
<td>0.00003</td>
<td>0.3814</td>
<td>0.0782</td>
<td>Unknown molecular function. Involved in ribosome biogenesis, neuroblast proliferation, and female germ-line stem cell division. Phenotypically relevant in trichogen cell and <strong>mesothoracic tergum</strong>.</td>
</tr>
<tr>
<td>I(3)72Dn</td>
<td>$9 \times 10^{-8}$</td>
<td>0.00005</td>
<td>0.4781</td>
<td>0.1330</td>
<td>Unknown molecular function. Involved in ribosome biogenesis and neurogenesis. Phenotypically relevant in <strong>mesothoracic tergum</strong>.</td>
</tr>
<tr>
<td>CG9135</td>
<td>$7 \times 10^{-7}$</td>
<td>0.0031</td>
<td>0.2774</td>
<td>0.0150</td>
<td>Predicted guanyl-nucleotide exchange factor activity. Unknown biological function. Phenotypically relevant in <strong>mesothoracic tergum</strong>.</td>
</tr>
<tr>
<td>CG8562</td>
<td>$3 \times 10^{-6}$</td>
<td>0.00105</td>
<td>0.3100</td>
<td>0.0624</td>
<td>Predicted metalloprotease activity. Protein domains suggest involvement in proteolysis.</td>
</tr>
<tr>
<td>Oatp33Ea</td>
<td>0.00001</td>
<td>0.00175</td>
<td>0.2131</td>
<td>0.0556</td>
<td>Predicted organic anion transmembrane transporter activity.</td>
</tr>
<tr>
<td>lbk</td>
<td>0.00001</td>
<td>0.00175</td>
<td>0.0001</td>
<td>0.0435</td>
<td>Involved in chaeta morphogenesis and <strong>oogenesis</strong>.</td>
</tr>
<tr>
<td>lid</td>
<td>0.00001</td>
<td>0.00175</td>
<td>0.1231</td>
<td>0.0421</td>
<td>Histone acetyltransferase and demethylase activity. Involved in chromatin organization, and histone acetylation and demethylation. Phenotypically relevant in <strong>mesothoracic tergum</strong>, imaginal disc, and embryonic/larval optic lobe.</td>
</tr>
<tr>
<td>dome</td>
<td>0.00001</td>
<td>0.00270</td>
<td>0.0186</td>
<td>0.0886</td>
<td>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 spiralire, integumentary specialization, embryonic hindgut, and compound eye.</td>
</tr>
</tbody>
</table>

---

*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* Estimated for the focal (*D. suzukii*) branch.  
*d* Estimated for the rest of the phylogenetic tree.
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 genomewide pattern of substitutions should represent that of the ancestr al 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 dia pause results in recurrent population size bottlenecks, particularly for males, and thus in a lower effective population size Ne, 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 Ne, namely by decreasing Ne 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. 3B). If we assume that levels of $d_{N}/d_{S}$ are a proxy for effective population size, the X/autosome ratio of Ne values, $N_{X}/N_{\text{autos}}$, 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 $N_{X}/N_{\text{autos}}$ between the two species (Pool and Nielsen 2007). Other factors that may have affected the differences in the ratio $N_{X}/N_{\text{autos}}$ 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. 2A) 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. 2B) 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 Tibetan 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. 2C). 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 rotted fruit, is highly seasonal, whereas in the tropics fruiting, and thus the production of rotten 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 suzuki subgroup has not yet been confirmed by whole-genome phylogenetics, we cannot exclude the possibility that D. subpulchrella is sister to D. suzuki rather than D. pulchrella. Under this scenario, some adaptations currently modeled as arising within D. suzuki may in fact be shared with D. subpulchrella. Genome sequencing of D. subpulchrella will clarify this question. Our evolutionary analyses of the D. suzuki 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 pre-publication 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/G001744/1) and NERC (R8/H10/56). This project was funded by Accordo di Porgamma of the Autonomous Province of Trento.

Literature Cited


Associate editor: B. Venkatesh