A new genome-wide method to track horizontally transferred sequences: application to Drosophila

L. Modolo, Franck Picard, E. Lerat

A New Genome-Wide Method to Track Horizontally Transferred Sequences: Application to *Drosophila*

Laurent Modolo[1], Franck Picard[1], and Emmanuelle Lerat[1,*]

[1]Université de Lyon, France, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France

*Corresponding author: E-mail: emmanuelle.lerat@univ-lyon1.fr.

Accepted: January 28, 2014

Abstract

Because of methodological breakthroughs and the availability of an increasing amount of whole-genome sequence data, horizontal transfers (HTs) in eukaryotes have received much attention recently. Contrary to similar analyses in prokaryotes, most studies in eukaryotes usually investigate particular sequences corresponding to transposable elements (TEs), neglecting the other components of the genome. We present a new methodological framework for the genome-wide detection of all putative horizontally transferred sequences between two species that requires no prior knowledge of the transferred sequences. This method provides a broader picture of HTs in eukaryotes by fully exploiting complete-genome sequence data. In contrast to previous genome-wide approaches, we used a well-defined statistical framework to control for the number of false positives in the results, and we propose two new validation procedures to control for confounding factors. The first validation procedure relies on a comparative analysis with other species of the phylogeny to validate HTs for the nonrepeated sequences detected, whereas the second one built upon the study of the dynamics of the detected TEs. We applied our method to two closely related *Drosophila* species, *Drosophila melanogaster* and *D. simulans*, in which we discovered 10 new HTs in addition to all the HTs previously detected in different studies, which underscores our method's high sensitivity and specificity. Our results favor the hypothesis of multiple independent HTs of TEs while unraveling a small portion of the network of HTs in the *Drosophila* phylogeny.

Key words: horizontal transfer, genome-wide method, *Drosophila*, transposable elements, FDR.

Introduction

Thanks to next-generation sequencing (NGS) technologies and to recent advances in de novo genome-assembly algorithms, we now have access to an increasing number of complete eukaryotic genomes. This methodological shift toward deep sequencing has changed the scale of investigation for many genomic studies and now allows the study of horizontal transfers (HTs) between eukaryotic species (Gilbert et al. 2010, 2013; Gilbert and Cordaux 2013).

HTs are defined by an exchange of genetic material between two reproductively isolated organisms (Gilbert et al. 2009) or by a movement of genetic information across normal mating barriers between more or less distantly related organisms (Keeling and Palmer 2008). Contrary to prokaryotes, for which HTs are common and well described (Fall et al. 2007; Juhas et al. 2009; Weinert et al. 2009), HTs are thought to be rare in eukaryotes, and their underlying mechanisms remain unknown (Andersson 2005). Proposed hypotheses to explain HTs in eukaryotes range from virus-mediated HTs using direct transfer of episomes (O'Brochta et al. 2009), viral particles, or infection (Kim et al. 1994; Dupuy et al. 2011) to parasite-mediated transfers (Gilbert et al. 2010). Overall, the main difference between eukaryotes and prokaryotes regarding HTs resides in the type of DNA material that is transferred: HTs usually involve genes in prokaryotes (Ochman et al. 2000), whereas in eukaryotes, HTs usually involve noncoding DNA and transposable elements (TEs) (Schaack et al. 2010). Following the availability of complete assembled genomes, more attention has been directed to the detection of HTs in eukaryotes, but most studies rely on similar approaches to the ones used for the detection of HTs in prokaryotes (Doyon et al. 2011). However, the differences in the type of horizontally transferred sequences between prokaryotes and eukaryotes and in the quantity of DNA to be investigated raise specific methodological challenges that need to be addressed to obtain a broader picture of genome-wide HT dynamics in eukaryotes (de Carvalho and Loreto 2012).

A particularity of the detection of HTs in eukaryotes is that it first requires the genome-wide identification of candidate

pairs of sequences that are not necessarily predefined. This point has motivated the development of several approaches, such as the surrogate method, which relies on differences in nucleotide patterns consistent with foreign DNA (Ragan 2001; Putonti et al. 2006). Nevertheless, this type of approach displays such a high rate of false detections that it is not efficient for real case studies (Azad and Lawrence 2011). Other genome-wide approaches start with all-to-all Blast searches between genomes of many species and detect HTs using an arbitrary cutoff using *e*-values (Shi et al. 2005) or a lineage probability index (Podell and Gaasterland 2007). However, although these strategies have given better results than the surrogate method, the lack of statistical framework for the detection of HTs in both methods has limited the interpretation of their results and the precise assessment of their specificity and sensitivity (de Carvalho and Loreto 2012). Overall, the promises of genome-wide approaches have been tempered by these common drawbacks, which explain the prevalence of sequence-specific approaches in HT studies even for genome-wide data sets.

When focusing on sequence-specific approaches to study HTs, we can discriminate between tree-topology-based approaches and sequence-divergence-based approaches. In prokaryotes, the gold standard for detecting HTs relies on the study of incongruences between the phylogeny of the sequences undergoing HTs and the phylogeny of the species. Because the pairwise identity of a horizontally transferred sequence is higher than expected according to the divergence time of the two species (Silva et al. 2004; Loreto et al. 2008), a phylogenetic tree based on this sequence will be discordant from the species tree. Unfortunately, phylogenetic approaches require a large taxonomic sampling of genes to have sufficient power of detection, which is often lacking in eukaryotes. Moreover, this method poorly differentiates HT genes from ancestral gene duplication(s) followed by gene loss(es) (Roger 1999). Phylogenic incongruences can also be produced when two or more variants of the ancestral lineage sequence have been stochastically inherited by the derived lineages (Dias and Carareto 2012). Finally, another pitfall of these approaches is the possibility of phylogenetic reconstruction artifacts, which can lead to strongly supported but false trees and thus to false positives for HT detection.

Studying pairwise sequence divergences constitutes an alternative that is commonly used when working with eukaryotes. It can rely on different divergence metrics, such as the synonymous substitution rates (d_S or K_s), to test the consistency of the number of synonymous differences accumulated between two sequences with the divergence time between the two species. Confounding factors can also decrease the power of d_S-based approaches. Codon usage bias, for instance, can result in a reduced d_S for the reference genes, which can decrease the sensitivity of detection of sequences with low d_S (Wallau et al. 2012). Purifying selection and variable rates of sequence evolution can also lead to spurious HT detections or a lack of power for identity-based methods (Capy et al. 1994; Pace et al. 2008). Finally, a third line of evidence for the detection of HTs is a patchy distribution of the sequences within a group of taxa (as they are not vertically transmitted). However, because of stochastic losses, the lack of coverage of some parts of the genomes and the random sampling of the population alleles in the sequenced strains, this third line of evidence is hardly self-sufficient to infer an HT event (Keeling and Palmer 2008; Schaack et al. 2010).

One strategy to control for spurious HT detections has been to focus on one line of evidence for the detection of HTs and to rely on the two others for validation purposes (Loreto et al. 2008; Gilbert et al. 2010). However, when dealing with eukaryotes, the absence of evidence for phylogenetic incongruences and the absence of a patchy distribution are likely to be poor validating arguments, as they do not constitute strict evidence against the possibility of an HT (Wallau et al. 2012). Another weakness of current sequence-specific approaches is that both tree-topology- and sequence-divergence-based approaches are restricted to coding sequences (CDSs). This represents only a small part of most eukaryote genomes and introduces an important detection bias for the analysis of horizontally transferred sequences.

In eukaryotes, for which HT events involve noncoding DNA and TEs, only 330 cases of horizontally transferred TEs have been described to date (Wallau et al. 2012) compared with rates as high as 30% of lateral gene transfers per phylogenetic branches for prokaryotes (Abby et al. 2012). TEs are DNA segments that are able to replicate and insert themselves into the genome using different mechanisms (Finnegan 1997; Wicker et al. 2007; Jurka et al. 2011). One of the outstanding features of TEs is their ability to cross species boundaries and invade new genomes (Daniels et al. 1990; Pinsker et al. 2001; Ludwig et al. 2008). These elements can represent the most abundant part of large eukaryotic genomes, as is the case of the maize genome (85%) (Schnable et al. 2009) and of the human genome (between 45% and 78% according to the detection method [Lander et al. 2001; de Koning et al. 2011]).

Notably, among the 330 horizontally transferred TEs detected, 178 concern drosophilid species, and from the 101 putative HT events proposed in *Drosophilae* in 2008, only 15% were confirmed by the three lines of evidence we have mentioned (Loreto et al. 2008). Regardless of this overrepresentation of drosophilids, the majority of these 330 HT detections relied on sequence-specific studies of candidate sequences. With this approach, only a small part of the genomes is exploited, which leads to an underestimation of the number of HTs. Our proposed genome-wide approach aims to solve this bias by requiring no prior knowledge concerning the sequences of interest and evaluating all the identifiable pairs of sequences between two genomes with an identity-based approach. Our method addresses the detection of all HTs genome wide as a multiple-testing problem to

handle this large number of identity-based detections and to control the proportion of false positives in the results (Wei et al. 2009). We also propose two new filtering methods to sort out spurious HT detections corresponding to conserved sequences in the results.

We applied our method to the genome-wide detection of all putative HT sequences between two *Drosophila* species: *Drosophila melanogaster* and *D. simulans*. These two cosmopolitan *Drosophila* species have a divergence time estimated between 4.3 and 6.5 Myr (Tamura et al. 2004) and are highly similar on many points, except in their TE content. TEs in *D. melanogaster* represent a large amount of the genome (15% [Dowsett and Young 1982]), with mainly young and active (highly similar) copies (Bowen and McDonald 2001; Kaminker et al. 2002; Lerat et al. 2003). In contrast, the TEs in *D. simulans* are represented mainly by old and degraded copies (Lerat et al. 2011) and only account for 6.85% of the genome (Hu et al. 2013). To explain the differences in the TE landscape between these two species, previous studies based on a restricted number of TEs have shown that numerous HTs were likely to be involved (Bartolomé et al. 2009; Lerat et al. 2011) (see Carareto [2011] for a review). To obtain a broader picture of HT between these two genomes, we performed a whole-genome comparison study between *D. melanogaster* and *D. simulans* assuming that undefined fragments of DNA may have been transferred from one species to the other. These undefined fragments of DNA can contain any types of sequences, such as TEs, nuclear genes, or intergenic DNA, thus removing any detection bias toward CDSs. As a result, we detected 10 new putative horizontally transferred TEs in addition to all the horizontally transferred TEs described by different studies between *D. melanogaster* and *D. simulans*, bringing to light a portion of the rich network of HTs that seems to link together the *Drosophila* species.

Materials and Methods

Our method can be divided into two main parts. For the first part, it relies on a multiple-testing framework to identify with a high sensitivity all the sequences that may have been horizontally transferred between two species at the genome scale. This approach is divided into three different steps described later. Then, we developed a multiple-testing framework to evaluate the output of multiple identity-based detections of HTs while controlling for the expected proportion of false positives in the results. A novelty of our approach is the modeling of the data throughout the genome as candidate sequences that are structured spatially, accounting for their dependency structure with a nonhomogeneous Markov model (NHMM) to increase the power of the multiple-testing correction (Kuan and Chiang 2012). For the second part of our method, we discriminate between putative HTs and other mechanisms, leading to a high pairwise identity to increase our specificity. For this purpose, we propose two novel validation procedures that can be applied for genome-wide studies to control for the numerous sources of spurious detections inherent to the detection of HT.

We will thereafter introduce the software, the algorithms, and the statistical models that we used for the different parts of this approach (supplementary fig. S1, Supplementary Material online). In our application, genome A corresponds to the genome of *D. melanogaster* and genome B to the genome of *D. simulans*.

Description of the Tree Steps for the Detection of Putative Horizontally Transferred Sequences between Two Genomes A and B

Step 1: Selection of the Sequences of Interest

To identify HT events, we define a sequence of interest as part of a pair of sequences with a higher pairwise nucleotidic identity than expected between the two species A and B. This first part of the pipeline aims to delimit such sequences in the two genomes. To achieve this goal, we start by retrieving the list of all the identifiable pairs of sequences between the two species A and B. For this step, we performed a nucleotidic all-to-all `Blast` (version 2.2.26) of one genome against the other (Altschul et al. 1990). The output of such a Blast defines a many-to-many cardinality between sequences from the two species, meaning that a given sequence from one species can be linked to many sequences in the other species, and vice versa. These types of links are complex and represent a large quantity of data to address. Moreover, as we cannot observe two different horizontally transferred sequences at the same locus in the species A, we filter the resulting pairs of sequences to only retain the best match for each position of the genome of A. For the task at hand, we only need the best local alignments of sequences for each position along the genomes because the other alignments would have a lower identity and thus a lower probability to correspond to an HT event.

To parse the Blast output and obtain a one-to-one cardinality from a many-to-many cardinality, we developed in python the program `htdetect.py` (available from the online resources). This program uses the fact that when working on two different genomes, there is always a genome of better quality (genome A) than the other genome (genome B). Our algorithm can be divided into the four following stages (fig. 1):

1. Compute the identity between each pair of sequences and the corresponding *P*-values to account for the identity and the size of the pair of sequences (see unilateral binomial test later) (fig. 1*A*).
2. Order all the pairs of sequences according to their position in genome A (fig. 1*B*).
3. Merge all the overlapping pairs of sequences in genome A to obtain a one-to-many cardinality from the many-to-many cardinality (fig. 1*C*).

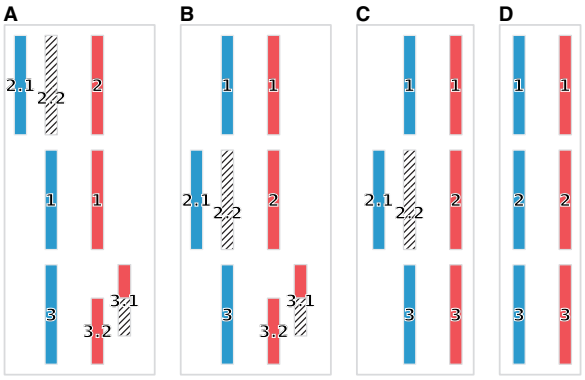

Fig. 1.—Algorithm to reduce the many-to-many cardinality in the results of an all-to-all nucleotidic Blast to a one-to-one cardinality between a genome A (red) and a genome B (blue). (*A*) Compute the identity between each pair of sequences and the corresponding *P* values (see Materials and methods, stage 3), and order all the pairs of sequences according to their position on genome A (the sequence order is 1-2-3). (*B*) Merge all the overlapping pairs of sequences in the genome A to go from a many-to-many cardinality to a one-to-many cardinality (remove the dashed part of the sequence 3.1). (*C*) Keep the sequences with the lowest *P* values from genome B for each pair of sequences that were merged in stage 3 to obtain a one-to-one cardinality (remove the dashed sequence 2.2). (*D*) One-to-one cardinality between the two genomes.

4. Keep the sequences with the lowest *P* values from genome B for each pair of sequences that have been merged in step 3 to obtain a one-to-one cardinality (fig. 1*C* and *D*).

In the hypothetical case where both genomes are of equivalent quality, the above steps will be strictly symmetrical to obtain a one-to-one cardinality.

Step 2: Computation of the Expected Pairwise Identity between the Compared Species

To test H_0 : "the number of differences is greater than or equal to the expected number of differences," for each of the filtered sequences, we compute the expected pairwise nucleotide identity between the species A and B, given their time of divergence. For this purpose, we used a global pairwise alignment of the genome of the species B against the genome of species A. We compute the number of identical nucleotides for each nonoverlapping window of size 1 kb along each chromosome arm of the species A. The size of 1 kb was empirically chosen as a trade-off between the resolution for the identity computation (of 0.01%) and information about the identity variation (a large window size only gives access to the average identity). To compute the nucleotide identity percentage between the species A and B for each of these windows, we removed the unknown nucleotides and the gaps from the computation.

We then used a Gaussian kernel smoothing function of these nonoverlapping window identity scores to obtain the distribution of the nucleotide identity between the two species. As this identity distribution is skewed to the right in the case of our application to *Drosophila* species (fig. 2), we chose to use the highest mode of this distribution as the expected pairwise identity between the two genomes, instead of the mean or a given quantile.

Step 3: Test of the Sequence Pairwise Identity

To model the pairwise identity, for every pair of sequences n, we denote by W_n the number of different nucleotides between the two sequences. The distribution of W_n is $\mathcal{B}(L_n, p_n)$, where L_n is the length of the pair of sequences of interest and p_n is the probability of having a nucleotidic dissimilarity. Our aim is to test $H_0 : \{p_n \leq p_0\}$ accounting for L_n, in which $1 - \widetilde{p_0}$ is the expected identity calibrated using the reference distribution constructed from the global alignment of the two genomes (=95.62% for our application). Thus, we compute for each pair of sequences n the probability $P(w_n^{\text{obs}}) = P\{W_n \geq w_n^{\text{obs}}\}$ of having a number of different nucleotides lower than expected, or unilateral *P*-value.

The number of tests N equals the number of candidate pairs of sequences for each chromosome arm and for the whole genome. Thus, for a given level of type I error (e.g., $\alpha = 0.05$), with a crude estimate under independence of the tests, the number of false positives ($N \times \alpha$) can be larger than the number of positives.

At each position along the genome of species A, we have a *P*-value denoted by $P(w_n)$ that is distributed according to a uniform distribution in [0,1] under H_0. From this *P* value, we want to infer an indicator variable denoted by S_n, such that $S_n = 1$ if H_0 is rejected at position n and $S_n = 0$ otherwise. To proceed, we use the local false discovery rate (ℓFDR) strategy, which consists in assessing the posterior probability that S_n is under H_0 (Efron et al. 2001). Instead of using raw *P*-values, a standard strategy consists in using the inverse probit transform, such that $z_n = \Phi^{-1}(P(w_n))$, which results in centered standard Gaussian variables for the z under H_0, whereas the others follow an unknown density distribution f_1. Then, the posterior probability of being under H_0 is $\ell FDR_n = P(S_n = 1|z_n)$. The decision rule consists in selecting positions $n = 1, \ldots, \ell$, such that $\ell = \max\left\{i : (1/i)\sum_{j=1}^{i} \ell FDR_i \leq \alpha\right\}$, where $\ell FDR_1, \ldots, \ell FDR_N$ is ordered and α is the false discovery rate (*FDR*) level (Benjamini and Yekutieli 2001).

By mapping candidate sequences along the genome of species A, we expect the probability for one locus to have a higher pairwise nucleotide identity than expected to depend on its neighbors. Moreover, with the fragmentation of the candidate sequences due to the nucleotidic Blast, we also could detect small adjacent pieces of this locus instead of a unique DNA fragment, and because of their small sizes, each of these pieces of alignment could be statistically nonsignificant on its own. In the case of dependency, all the

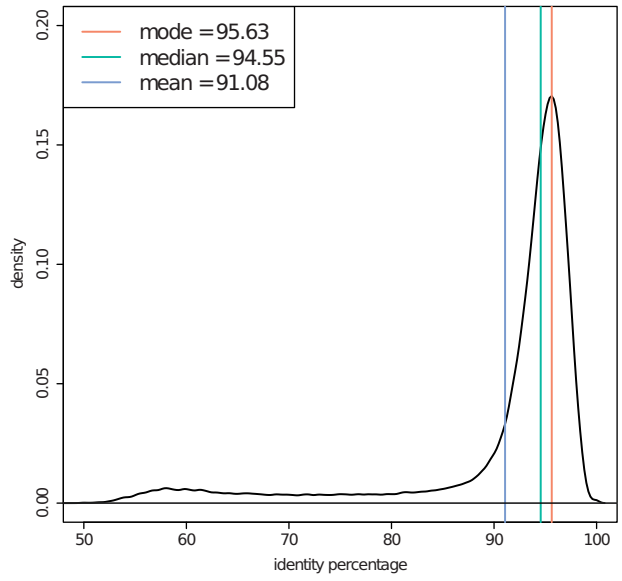

Fig. 2.—Distribution of the pairwise nucleotide identity, genome wide with nonoverlapping windows of size 1 kb, of the *Drosophila simulans* genome alignment on the *D. melanogaster* genome. The vertical bars represent the values of the mean, the median, and the mode of this distribution.

multiple-testing procedures not accounting for the dependency structure are suboptimal (Wei et al. 2009), meaning that if the procedure controls for the *FDR* for a given level, it does not minimize the false nondiscovery rate. Because decision at position n may depend on neighbor tests, we used the local index of significance (*LIS*) to compute $P(S_n = 0|z_1, \ldots, z_N)$ (Sun and Tony Cai 2009).

To proceed, we considered a homogeneous hidden Markov model in which S_n is the hidden states ($S_n \in \{0, 1\}$), which is governed by transition probabilities $P(S_{n+1}|S_n)$. Moreover, we also accounted for the genomic context of each sequences, like GC content or the distance between the sequences that can influence the transition and emission probability of the model, as we do not expect the dependency of a given sequence to its neighbors to be the same between every sequences. We considered a logistic regression to account for covariates $X_1, \ldots, X_N$ characterizing the sequences, such that:

$$P(S_1 = j|X_1 = x) = \frac{\exp\left(\lambda_j + \rho_j^n \times x\right)}{\sum_{k=0}^{1} \exp\left(\lambda_k + \rho_k^n \times x\right)}$$

$$P(S_n = j|S_{n-1} = i, X_n = x) = \frac{\exp\left(\sigma_{ij} + \rho_j^n \times x\right)}{\sum_{k=0}^{1} \exp\left(\sigma_{ik} + \rho_k^n \times x\right)}$$

with $i, j = \{0, 1\}$ (Kuan and Chiang 2012). The model parameters $\Psi = (\kappa, f_1, \lambda_i, \sigma_{ij}, \rho_j)$, with κ being the proportion of *P*-values equal to 1, can be estimated using the EM algorithm. We developed a zero-inflated Gaussian distribution to handle unilateral tests with the appropriate *z*-values transformed. This model is implemented in the `R` package `fdrDEP` available on the `CRAN` for multiple unilateral hypothesis testing.

The *LIS* statistics are computed for each chromosome arm of the species A and concatenated to control for the *FDR* at a level of 10% for the whole genome of A with the Benjamini, Hochberg, and Yekutieli procedure (Wei et al. 2009).

Filtering for True Putative HT Events

With steps 1–3, we could have detected highly similar fragments of sequence alignments that would not have been significant for the whole corresponding sequences, so we first recovered the full length of each annotated DNA fragment detected in the species A. To reconstruct the complete sequences for these results, we used the `bedtools` suite (version 2.17.0, options `intersectBed -a annotations.gff -b results.bed -wa`) (Quinlan and Hall 2010) to extract the annotated sequences corresponding to results with positions intersecting the ones from the species A. Then, we applied the two following filters to sort out conserved sequences from our results for nonrepeated and repeated sequences.

For Nonrepeated Sequences

For CDSs, we expect to observe an effect of selection because nonsynonymous mutations can be deleterious, neutral, or advantageous. Thus, for the CDSs identified with our approach, we can compute their d_S values using orthologous genes. We then performed the same unilateral binomial test as for the nucleotidic identity to determine whether the d_S of a given CDS is significantly lower than the expected identity between the two species considered while controlling for the *FDR* at a level of 10% (Benjamini and Yekutieli 2001).

In addition, to take into account non-CDSs that cannot be used in d_S approaches, we developed a new validation procedure based on sequence conservation, which can be applied to both coding and non-CDSs. In the set of detected sequences, a sequence identified with the same level of significance, both between *D. melanogaster* (the species A) and *D. simulans* (the species B) and between *D. melanogaster* and other *Drosophila* species, would illustrate a conserved sequence across the phylogeny rather than multiple HTs at the same position in *D. melanogaster*. Thus, we performed the same analysis with four other species from the 12 *Drosophila* genomes project: *D. sechellia*, *D. yakuba*, *D. pseudoobscura,* and *D. virilis*, as a gradient of phylogenetically divergent species, before subtracting these results from those of the *D. melanogaster–D. simulans* analysis. We used the `bedtools` suite to subtract the `.bed` tracks of the results of each species along the *D. melanogaster* genome. Figure 3 describes the decision rule used in this subtraction according

to the corresponding phylogenetic tree. This step provided us with a landscape of all the sequences with a pairwise identity higher than expected between *D. simulans* and *D. melanogaster* and not conserved in the other *Drosophila* species. This last filter relies on the strong hypothesis that a pair of sequences absent between a given pair of species is not missing due to random sampling of the population alleles in the sequenced individuals, a lack of genome coverage, or a misassembly.

As TEs and other repeated sequences are present at multiple loci between a pair of genomes, they were excluded from this filtering step and were validated separately, considering that we could not discriminate which TE copy identified between two genomes corresponded to a specific locus in the genome of the species A.

For Repeated Sequences

With our genome-wide approach, the set of TEs detected was not restricted to elements with a coding capacity, preventing us from relying on the d_S metric for their validation. Moreover, for TE family with a large number of copies, we can expect one or more of these copies to be more identical than expected between the two genomes just by chance. To account for the full set of detected TEs and analyze each detected TE family, we developed a new validating procedure based on the recent dynamics of the detected TEs in the genomes of species A and B. We worked under the hypothesis that, after an HT, a TE escapes the host defense mechanisms for a time and quickly replicates itself in the new host genome (Anxolabéhère et al. 1988; Le Rouzic and Capy 2005; Granzotto et al. 2011). Thus, in the case of an identifiable horizontally transferred TE, we expected to observe many highly similar copies of the TE corresponding to this burst of transposition in one or both genomes, in contrast to few conserved TE insertions (Lerat et al. 2011; Dias and Carareto 2012).

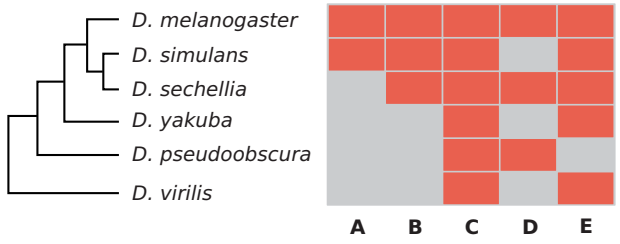

Fig. 3.—Decision rule for the filtering step about selective pressure, with the presence (red) or absence (gray) of a pair of sequences between the corresponding species. (*A*) Putative HT between *Drosophila melanogaster* and *D. simulans*. (*B*) Putative HT between *D. melanogaster* and *D. simulans* prior to the *D. sechellia* speciation event or conserved sequences between *D. melanogaster*, *D. simulans,* and *D. sechellia*. (*C*) Conserved sequences in the *melanogaster* subgroup. (*D–E*) Conserved sequences with stochastic loss or ancestral polymorphisms.

To help in the synthetic interpretation of the recent history of each TE in our results, we start by defining the most identical pair of copies between the two genomes, as the last putative horizontally transferred copy in the case of an HT. For each TE family, this most identical pair of copies between the two genomes is defined as the pair of copies with the lowest *P*-values from all the detected copies using the 80-80-80 rule (Wicker et al. 2007). Then, we Blast each of these most identical copies on the genome of A using a nucleotidic `Blast` (version 2.2.26) (Altschul et al. 1990). We built an index of the similarity of each copy of these elements compared with the most identical pair of copies between the two genomes, normalized by the size of the copies. We called this index the activity track. These activity tracks are used to rank between 0 and 1 all the copies of each identified TE according to their divergence from the corresponding most identical pair of copies between the two genomes, with 1 corresponding to a low degree of divergence and a recent activity of this TE and 0 corresponding to old and divergent copies. The activity track corresponds to the probability of having a pairwise nucleotidic identity with the most identical pair of copies less than or equal to the expected identity $1-\widetilde{p_0}$, estimated using the reference distribution constructed from the global alignment of the two genomes. For every pair of TE copies n, we denote by W_n the number of different nucleotides between the two copies. The distribution of W_n is $\mathcal{B}(L_n, p_n)$, where L_n is the length of the alignment between the copies and p_n is the nucleotidic dissimilarity. Our aim was to compute for each pair of sequences n the probability $P\{W_n \leq w_n^{\text{obs}}\}$ corresponding to the activity track. The same analysis is performed with the genome of species B to get an overview of the TE activity in both genomes. We developed in python the program `activity_tracks.py` (available from the online resources) to compute this index.

Finally, we manually inspected the results in `.bed` format on each chromosome arm of *D. melanogaster* to look for cluster of sequences with a higher identity than expected using the `integrative genome viewer` software (Thorvaldsdóttir et al. 2013).

All the statistical analyses in this article were performed using the software `R` (version 3.0.0) (R Core Team 2013).

Data Acquisition

We used the last available versions of the genomes of *D. melanogaster* (species A) (version r5.49), *D. sechellia* (version r1.3), *D. yakuba* (version r1.3), *D. pseudoobscura* (version r2.30), and *D. virilis* (version r1.2) and the corresponding annotation tracks from flybase (http://flybase.org [Marygold et al. 2013]). For *D. simulans* (species B), we did not work at first on the genome sequenced by the 12 *Drosophila* genomes project (Drosophila 12 Genomes Consortium 2007). Indeed,

this genome is a patchwork of six independently derived strains with the assembly of six major chromosome arms representing only 101.3 Mb of the 137.8 Mb expected. Moreover, this genome presents several major misassemblies and has the worst read quality of the 12 *Drosophila* genomes (Hu et al. 2013). This is why we used the *D. simulans* genome that was resequenced in 2012 and assembled from the *w501* strain of the original Sanger data, in addition to a high-coverage Illumina sequencing of iso-females of this same strain (Hu et al. 2013). However, to be able to compare our approach with previous studies, we also conducted a second analysis with the genome of *D. simulans* (version 1.3) available from flybase (http://flybase.org).

The genome pairwise alignments were retrieved from the UCSC website (http://hgdownload.cse.ucsc.edu).

The sequence annotation tracks used to obtain the full length of the corresponding TEs and CDSs and annotate the noncoding DNA were downloaded from flybase (http://flybase.org) in `.gff` format (Marygold et al. 2013).

Instead of computing the d_S of the detected CDSs, we used the d_S data of the 11,000 orthologs from the 12 *Drosophila* genomes, available from the study of Heger and Ponting (2007).

Quality of the TE Content in the Genome of *D. simulans*

We used the software `SeqGrapheR` (Novăk et al. 2010) to analyze the TE content of the *D. simulans* genome directly from a uniform random sample of 900 k reads obtained from the 2012 genome project (`SRA:SRX159034`) (Hu et al. 2013). The assembled repetitions were annotated using `RepeatMasker` (version 3.3.0) (Smit AF, Hubley R, Green P, unpublished data).

Data Access

All the scripts used for our pipeline are available in a `git` repository at: git://dev.prabi.fr/modolo2013.

Results

Genome-Wide Detection of Sequences with a Higher Nucleotidic Identity than Expected

Defining the Set of Candidates for HT Detection

We kept the best local alignments obtained by the Blast search of *D. simulans* against *D. melanogaster* for each position in the genome of *D. melanogaster*, thereby taking into account the repeated content that is often removed from genome-wide alignment (i.e., best global alignment). The cumulative size of the filtered sequences decreased with the divergence time between a given species and *D. melanogaster*, which is consistent with the nucleotidic Blast algorithm (table 1). For example, we retrieved approximately 112 Mb of sequences between *D. melanogaster* and *D. simulans* (divergence time of 5.4 ± 1.1 Myr), compared with only 13 Mb between *D. melanogaster* and *D. virilis* (divergence time of 42.9 ± 8.7 Myr). However, such a trend was not observed for the number of filtered sequences, which can be explained by the fragmentation of the retrieved sequences, which increased with the phylogenetic distance (table 1). With this set of candidates, we used our method to determine whether the observed pairwise nucleotidic identity for each of these pairs of sequences was higher than expected between the considered species and *D. melanogaster*.

Assessing the Reference Distribution for Nucleotidic Identity

We computed a reference nucleotidic identity distribution with the analysis of the global alignment of the genome of *D. simulans* along the genome of *D. melanogaster* (fig. 2). This distribution accounted for the variations in nucleotidic identity along the two genomes, in contrast to the common mutation rate of $1.1 \pm 0.2 \times 10^{-8}$ mutations per site per year per lineage for the *Drosophila* phylogeny that has been computed on a limited number of nuclear genes (Tamura et al. 2004). Consequently, this mutation rate based on the molecular clock hypothesis (Weir and Schluter 2008) may not be representative of the pairwise nucleotidic identity between the whole genomes of *D. melanogaster* and *D. simulans* (Drosophila 12 Genomes Consortium 2007) and is not suitable for a genome-wide analysis. For the detection of HTs between *D. melanogaster* and *D. simulans*, we were only interested in the expected nucleotide identity corresponding to the accumulation of mutations between these two species since their time of divergence. Thus, we choose the highest mode of identity distribution as a reference to compute the unilateral *P*-values of our tests, which quantified the probability of each candidate to have a nucleotidic identity exceeding 95.63%, while accounting for the size of the alignment (fig. 2).

Controlling for False Positives in the Context of Genomic Dependencies

As in many genomic studies, the number of statistical tests to perform was large (168,325 pairs of sequences for the comparison *D. melanogaster* vs. *D. simulans*). If no multiple-testing procedure is applied, we can roughly expect to declare an average of 10% of the tests (16,832) to be false positives by retrieving all the *P*-values below 0.1, which can be higher than the number of true positives (Finner and Roters 2002). By applying the standard Benjamini–Hochberg multiple-testing correction with an *FDR* level of 10% (Benjamini and Hochberg 1995), without taking into account the dependency structure between the tests, we only retrieved 605 CDSs, 934 TE insertions, and 2,345 intergenic DNA fragments. Thus, we used our method to assess the probability that each pair of sequences has a higher pairwise identity than expected while accounting for its dependency to its neighbors, adjusted to

Table 1
Results of the Filter of the All-to-All Nucleotidic Blast between *Drosophila melanogaster* and the Corresponding Species

Species	Sequence Size (kb)			Number of Sequences			Divergence Time to *D. melanogaster* (Myr)
	Row	Filtered	Significant[a]	Row	Filtered	Significant[a]	
D. simulans	550,226	112,748	9,012	4,468,121	168,325	11,927	5.4
D. sechellia	1,219,599	111,909	5,452	7,947,377	170,394	7,025	5.4
D. yakuba	1,972,352	91,584	977	23,960,790	239,011	3,185	12.8
D. pseudoobscura	102,146	22,241	593	1,431,447	213,790	11,323	30.0
D. virilis	184,640	13,463	298	2,186,411	117,831	6,305	42.0

NOTE.—Row, results corresponding to a many-to-many cardinality; filtered, results corresponding to a one-to-one cardinality.
[a]Results corresponding to the significative identity-based tests after multiple-testing correction.

other genomic covariates to increase our sensitivity. Indeed, the GC content of the sequences as well as the distance between a pair of sequences and the next on a chromosome arm and the presence of TEs are likely to be proxies of the similarity of a pair of sequence to its neighbors. We also expected the recombination rate to be an important factor, but no significant correlation was found between the recombination data available for the genome of *D. melanogaster* and the *P*-values of our tests. With this correction applied to the 168,325 tests, we retrieved 7.3 Mb of sequences, including 2,651 fragments from CDSs (2.46 Mb), 3,967 fragments from insertions of 28 different TE families (201 kb), and a large number of intergenic DNA fragments (13,806 sequences corresponding to 4.68 Mb), between *D. melanogaster* and *D. simulans*.

Distinction between "True" HT Events and Biological False Positives

HT Sequences in the Light of Other Drosophila *Species*

We detected a set of sequences with an identity higher than expected between the genomes of *D. simulans* and *D. melanogaster* that was not reduced to HT sequences, thus we started by retrieving the full-length sequences of each annotated fragment from the genome of *D. melanogaster*. Then, we discriminated putative HT sequences from the sequences displaying a signature of functional constraints. We tested whether the d_S of the 2,651 detected CDSs was significantly lower than expected in the *D. melanogaster-D. simulans* analysis, and we finally retained 26 CDSs.

To discriminate between conserved and horizontally transferred sequences for the full set of detected nonrepeated sequence (i.e., both coding and non-CDSs), we used the comparative analysis between *D. melanogaster* and *D. simulans*, and between *D. melanogaster* and other *Drosophila* species. This subtraction allowed us to remove approximately 40% of the base pairs for the intergenic DNA (thus keeping 2.79 Mb of the 4.68 Mb), with a consistently high pairwise identity with *D. melanogaster* in this phylogeny. This result is consistent with the results from Casillas et al. (2007) where 38.6% of the noncoding DNA in *D. melanogaster* display the signature of functional constraints. We also retained 28 of the 2,651 CDSs with this second approach.

The intersection of the results from the d_S analysis with the ones from this subtraction led to the detection of 11 CDSs annotated from RNA-Seq data but of unknown function (Marygold et al. 2013). These 11 CDSs were sparsely distributed along all the major chromosome arms of *D. melanogaster* and found in clusters of CDSs with significant pairwise nucleotide identity but nonsignificant d_S. Thus, these 11 CDS in our results could be biological false positives caused by the dependency model used in the multiple-testing correction, lowering their probability of being under the null hypothesis due to their conserved neighbors, which does not support the hypothesis of their HT. For the detected noncoding DNA, we were not able to use the *D. melanogaster* annotations to retrieve the full-length sequences of the DNA fragments. This class of fragmented DNA, representing 63.91% of the detected DNA in our results, was annotated based on the *D. melanogaster* annotation tracks (supplementary table S1, Supplementary Material online) but was only analyzed as neighboring sequences of the detected CDSs and TE sequences.

Horizontally Transferred TEs

For the repeated sequence, we used the activity track to study the recent activity of the detected TE family. According to the activity track distributions, most of the detected TE families in our results presented a recent period of activity in *D. melanogaster* (supplementary figs. S2 and S3, Supplementary Material online), with a large number of copies highly similar to the most identical pair of copies between the two genomes (see e.g., the diver element, fig. 4*A*). However, for some elements such as the ancient element *INE-1*, described as having invaded the ancestor lineage of *D. melanogaster* and *D. simulans* (Kapitonov and Jurka 2003), the activity track showed a majority of divergent copies with only few ones close to the most identical copy between the two genomes (fig. 4*B*), as expected by chance for a large number of

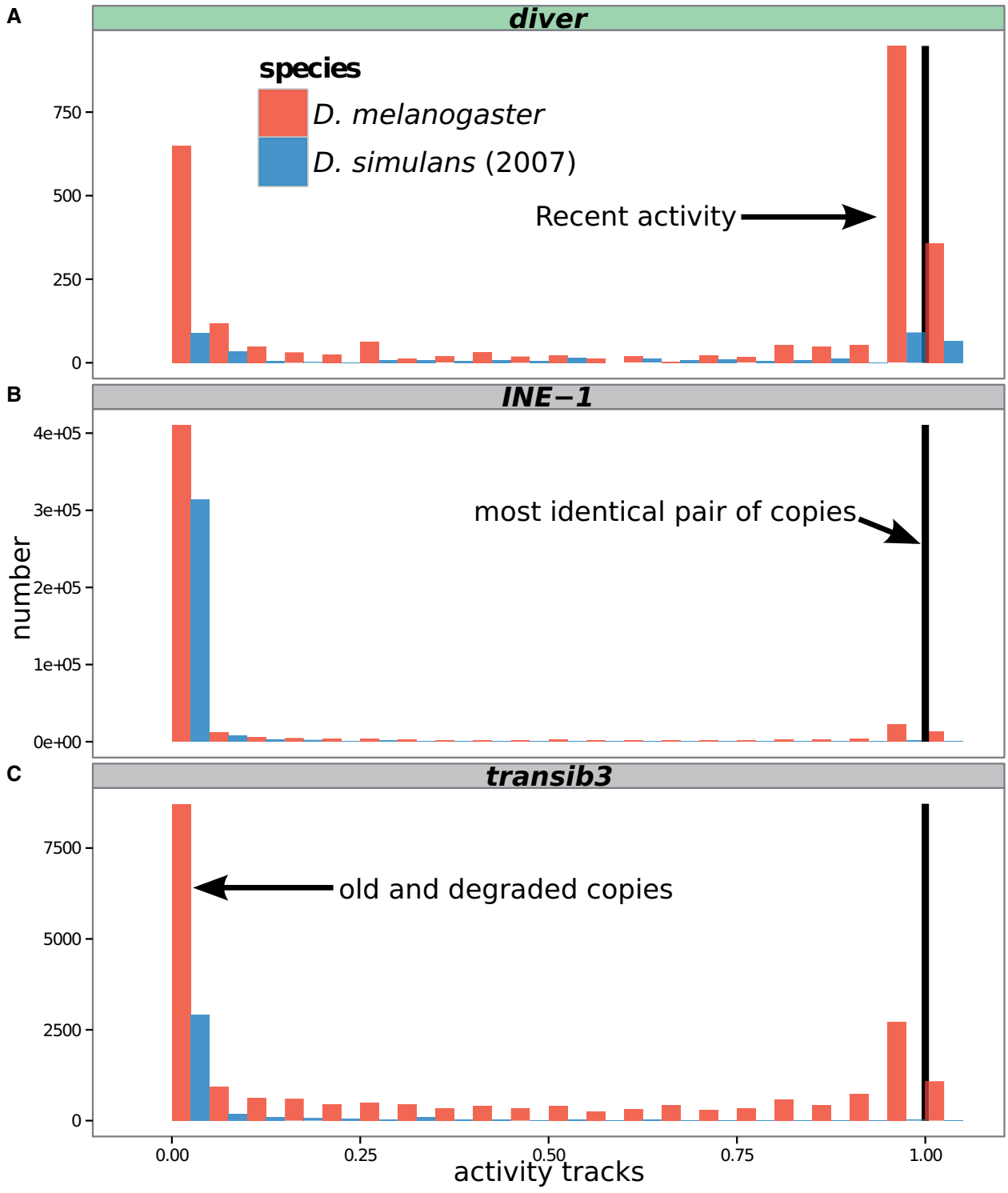

Fig. 4.—Density distributions of the activity tracks computed with the 2007 version of the genome of *Drosophila simulans*. The rightmost black bar corresponds to the most identical pair of copies between the two genomes, whereas the colored bars represent the number of copies ranked according to their similarity to this most identical pair of copies for a given TE. The red bars represent the activity tracks in *D. melanogaster,* whereas the blue bars represent the activity tracks in *D. simulans* 2007. (*A*) Example of a TE family presenting a recent period of activity corresponding to a putative HT from *D. simulans* toward *D. melanogaster*. (*B*) Example of a TE family with an activity not consistent with an HT between *D. simulans* and *D. melanogaster*. (*C*) Example of a TE family with different waves of activity.

old and degraded copies. With the activity track ranking the TE copies according to the most identical pair of copies between the two genomes which can be seen as the last putative horizontally transferred copy in the case of an HT, we were able to balance the direction of the transfers, which is crucial to understand the horizontally transferred TE history and dynamics. In the case of a horizontally transferred TE from a first species toward a second species, we can expect the TE to be present in a small number of highly similar and potentially active copies in the first species and in a large number of highly similar copies in the second species (supplementary fig. S3, Supplementary Material online). We observed this pattern for elements such as *diver*, which had a large number of copies with an activity track close to 1 in *D. melanogaster* and few copies in *D. simulans* (fig. 4*A*). This pattern was consistent with its HT from *D. simulans* toward *D. melanogaster*, even with few copies that were statistically significant in the two genomes. In the genome of *D. simulans*, most of the activity track distributions were bimodal, with few TE copies close to 1 and a large number of copies close to 0 corresponding to old and degraded copies (supplementary fig. S3, Supplementary Material online), which was consistent with the observations made for some of these TE families in the genome of *D. simulans* (Lerat et al. 2011). In contrast, in *D. melanogaster*, most of the TE copies had an activity track close to 1, which was representative of young and active TE populations. These differences of TE landscape between these two species support the hypothesis of multiple horizontally transferred TEs from *D. simulans* (fig. 4*A*) and from other species (fig. 4*C*) toward *D. melanogaster*.

With the 2012 version of the genome of *D. simulans*, we were able to identify 21 TE families with 10 new cases of horizontally transferred TEs, which were not previously identified as horizontally transferred between these two species (supplementary fig. S2, Supplementary Material online). However, 11 TEs were missing from the 24 horizontally transferred TEs previously described by different studies between *D. simulans* and *D. melanogaster* (de la Chaux and Wagner 2009; Bartolomé et al. 2009; Carareto 2011; Lerat et al. 2011). From these 11 TEs, the elements were only present in a few noncomplete copies in the 2012 version of the genome of *D. simulans*, which explain their absence from our results. For the elements *F*, *copia*, *gypsy5,* and *gypsy10*, the TE copies were highly divergent from those present in the genome of *D. melanogaster* and displayed a nonsignificant nucleotidic identity. To confirm the absence of the *412* element, known to be active in some populations of *D. simulans* (Vieira and Biémont 1997), we performed a de novo assembly of the TEs directly from the reads of the 2012 *D. simulans* genome project. The reads corresponding to this 7,566 bp element represented 50 kb of the 137.8 Mb genome of *D. simulans*, with the majority of the reads matching the long terminal repeats and few reads mapping within the element, which was concordant with the 2012 assembly. Therefore, the absence of these 11 horizontally transferred TEs from our results was likely the result of their absence from the assembled strain in the 2012 version of the genome of *D. simulans* rather than a lack of sensitivity of our method. Using the genome of *D. simulans* from the 12 *Drosophila* genomes project (Drosophila 12 Genomes Consortium 2007) (supplementary fig. S3, Supplementary Material online), we were able to recover in one analysis the 24 HTs previously described in the literature, including the 11 families missing from our analysis with the *D. simulans* genome of 2012. The 10 new and the 24 previously described TEs all presented activity track distributions consistent with the after effect of a horizontally transferred TE from *D. simulans* toward *D. melanogaster* (supplementary figs. S2 and S3 and table S2, Supplementary Material online). Thus, given the number of horizontally transferred TEs detected between *D. melanogaster* and *D. simulans* in the short time since their divergence, a parsimonious hypothesis could be the introgression of one or more fragments of DNA containing different TEs instead of multiple independent HTs.

Introgression versus Multiple HT Events

To obtain a broader view of the HTs between *D. melanogaster* and *D. simulans*, and discriminate between introgression and multiple independent HTs, we manually inspected the 11 CDSs, the TE insertions from the 21 families left and the 10,232 fragments of noncoding DNA in the final results along each chromosome arm of *D. melanogaster* and with the 2012 genome of *D. simulans*. In the case of introgression, we expected to observe the simultaneous transfer of these three types of sequences in one large DNA fragment. However, we found no sequence containing three or even two of these different types of sequences in the final results. This absence of completely introgressed fragments could be a consequence of the fragmentation of the detected sequences between the two genomes. However, we also did not find any obvious clusters of these different types of DNA along the chromosome arms of *D. melanogaster*. Overall, the types of detected sequences in our study support the prevalence of TEs and noncoding DNA in HTs between these two species. However, the informations contained in the genome of the sequenced individuals are not sufficient to support the hypothesis of multiple independent horizontally transferred TEs toward *D. melanogaster* rather than introgression events.

To better understand the horizontally transferred TEs involving *D. melanogaster*, we performed the same horizontally transferred TE analysis with the data from our comparison with the four other *Drosophila* species (*D. sechellia*, *D. yakuba*, *D. pseudoobscura,* and *D. virilis*) (supplementary figs. S4–S7 and table S2, Supplementary Material online). We detected numerous horizontally transferred TEs in this restricted window of time starting 5.4 ± 1.1 Ma, corresponding to the expected identity threshold between *D. melanogaster*

and *D. simulans*. The comparison between *D. melanogaster* and *D. sechellia* provided evidence for HTs of 21 elements between these two species (supplementary fig. S4, Supplementary Material online). *Drosophila sechellia* is the only species with a divergence time to the ancestor that it shares with *D. simulans* within the time window of our analysis, so for this species we can discriminate between horizontally transferred TEs involving *D. melanogaster* and the ancestor of *D. simulans* and *D. sechellia*, and horizontally transferred TEs involving *D. melanogaster* and *D. sechellia* (fig. 3). For the element (with an activity track nonconsistent with an HT in the 2007 version of the genome of *D. simulans* and absent from the 2012 version), the results rather indicate recent activity in both species, which suggest the existence of a third donor species, such as another *D. simulans* strain than those sequenced in 2007. We also observed the same pattern for 17 elements between *D. melanogaster* and *D. yakuba* (supplementary fig. S5, Supplementary Material online). However, we did not find any clear evidence of recent horizontally transferred TEs or a burst of transposition in the analysis of the TEs detected between *D. melanogaster* and *D. pseudoobscura*, or between *D. melanogaster* and *D. virilis*, which could be explained by the degree of fragmentation of the corresponding sets of sequences (supplementary figs. S6 and S7, Supplementary Material online). Finally, recent activity of the *Roo* element was detected in *D. melanogaster*, *D. simulans*, *D. sechellia,* and *D. yakuba*, which could be the result of independent HTs of this element toward these four species after their divergence (de la Chaux and Wagner 2009).

Discussion

In eukaryotes, the study of horizontally transferred sequences is confined to CDSs and often focuses on specific TEs. Thus, there are two systematic biases in the detection of HTs in eukaryotes: the candidate HT must be known and must have a coding capacity (Keeling and Palmer 2008). We propose a new genome-wide approach that aims to bypass these biases inherent to sequence-specific approaches by considering all the best local alignments of one genome to another as possible horizontally transferred sequences. Then, we test each of these sequences to retrieve those with a higher nucleotidic identity than expected between the two species while accounting for the multiplicity of the tests and their dependency structure throughout the target genome. We detected 2,651 CDSs, 3,967 insertions from 28 different TE families, and a large number of intergenic DNA fragments (13,806) more identical than expected from the 4,468,121 pairs of sequences identifiable between *D. melanogaster* and the 2012 version of the genome of *D. simulans*. Finally, we discriminated between spurious HT detection and putative HTs in our results with two novel validation procedures for genome-wide HT detection. And after manual inspection of the results, we retained 21 TE families as horizontally transferred between these two species, validating the prevalence of TE sequences in HTs between these two species without detection bias toward this type of sequence.

Genome-Wide Identification of Putative Horizontally Transferred Sequences

Previous genome-wide approaches used a wide range of procedures to infer sequences more identical than expected given the phylogeny of the species to detect HTs in eukaryotes (Loreto et al. 2008; Wallau et al. 2012). However, none of these procedures relied on a statistical testing framework to validate their sensitivity and specificity. This explains why sequence-specific approaches are still used: their particular reliability despite the limited set of sequences considered (Wallau et al. 2012). The collegial tests for the identity-based detection of horizontal transferred sequences in eukaryotes rely on the synonymous substitution rate, often in the form for a codon-based *Z*-test (Pace et al. 2008; Gilbert and Cordaux 2013). In our study, the set of candidates sequences was not restricted to the small coding portion of eukaryote genomes, and this justified the use of a binomial test to retrieve the sequences with a higher pairwise nucleotidic identity than expected between two species without any codon information while accounting for the size of each candidate. This simple model for codon substitution is sensitive enough to detect recent HTs for which we can expect a small saturation between sequences. The saturation corresponds to the occurrences of multiple mutations at a single nucleotide (or site), which leads to an underestimation of the nucleotidic divergence between two sequences because we can only observe the last mutation in the case of multiple mutations per site. A pair of sequences with saturation is expected to have more single mutations per site than multiple mutations per site. Thus, the complex cases, ill-defined by the model, will also correspond to the sequences in the "uninteresting" side of our unilateral hypothesis and will be correctly assigned to the set of nonsignificant sequences.

In genome-wide analyses, we often face multiple-testing issues, and our results underscore the importance of working with a well-defined statistical framework to control the number of incorrect detections and increase the power of the study. We also took advantage of the fact that when comparing two genomes, we always have a genome of better quality to map the detected candidate sequences, to greatly reduce the dimensionality of the data to be analyzed, thus increasing the power of our study (Storey and Tibshirani 2003). For our analysis, the now standard Benjamini–Hochberg *FDR* (Benjamini and Yekutieli 2001) procedure had a too low specificity to produce relevant results. This was caused by the dependency between the tests in our analysis, which was taken into account with the *LIS* framework to increase the specificity of our approach (Sun and Tony Cai 2009). Modeling this dependency between each pair of

candidates along the chromosome arms of *D. melanogaster* with a homogeneous Markov model (HMM) was not sufficient to retrieve all the horizontally transferred TEs described in the literature with the genome of *D. simulans* from the 12 *Drosophila* project (Drosophila 12 Genomes Consortium 2007). For our purposes, an HMM would have tended to homogenize the *LIS* statistics according to the information provided by the adjacent without taking into account any information about the type of sequences or the distance between them (i.e., the nonsignificant pairs of sequences surrounding a TE copy), which can explain these missing horizontally transferred TEs. With an NHMM, we were able to enrich the standard Markovian dependency according to covariates, such as the distance between the statistical tests along the genome and the presence of TEs, and to detect the 24 horizontally transferred TEs described in the literature (Carareto 2011) with an *FDR* level of 10% and the 2007 version of the genome of *D. simulans*, in addition to 10 new ones with the 2012 version of the genome (supplementary table S2, Supplementary Material online).

Our approach can be applied to any pair of sequenced species in which one species has an assembled genome, into which candidate sequences will be placed, to model the dependency structure between the tests with a NHMM. The specificity of this method is high enough to detect sequences more identical than expected between closely related species while controlling for the *FDR* in the results. This procedure could also be used with any other unilateral tests for different biological problems or to model the nucleotidic differences in ancient HTs or between more divergent species with a greater prevalence of saturation.

Two New Methods to Confirm HTs in Genome-Wide Studies

Most of the methods for HT detection in eukaryotes use sequence-specific approaches and rely on strong d_S evidence to infer putative HTs (Bartolomé et al. 2009; Lerat et al. 2011). In the remaining studies, the candidate sequences generally involve distantly related species and recent HT events where the identity line of evidence can be self-sufficient (Loreto et al. 2008), for example, the case of the TEs *SPIN* and *OC1* (Gilbert et al. 2010, 2013). This can also be the case for recent HTs, such as the well-known example of the *P* element transferred from *D. willistoni* to *D. melanogaster* less than 100 years ago, for which the nucleotidic identity is almost of 100% between the two species (Daniels et al. 1990). We were able to retrieve sequences with an identity percentage higher than 99% between *D. melanogaster* and *D. yakuba* for the elements *Doc*, *jockey,* and *transib3*, which was unexpected and could be sufficient to infer their HT. However, the number of obvious cases was small, and we needed to confirm the other HTs by other lines of evidence.

When studying the pattern of sequence divergence between genomes to infer HTs, we can encounter a large number of confounding factors that need to be checked (Siepel et al. 2005; Pollard et al. 2010). These factors range from natural turnover (gain or loss of functional elements) to the effect of purifying and positive selection, which can act on entire sequences or on parts of sequences, canceling the effect of divergence. For the study of HTs, we can add to this list the effect of different evolutionary rates for the sequence under consideration or the effect of stochastic losses in the phylogeny of the candidate sequence(s) (Loreto et al. 2008). Moreover, we have to rely on orthologs and sequence identification, which is nontrivial and can lead to numerous false positives (Gronau et al. 2013). The possibility of misplaced DNA sequences in the genomic database, polymerase chain reaction mispriming, contamination, incomplete sequence data, and poorly rooted trees can also be technical sources of errors for HT detection (Lisch 2008). Therefore, to differentiate between putative HT events and the possibility of vertical transmission, we need to investigate other lines of evidence (Loreto et al. 2008; Gilbert et al. 2010). In genome-wide studies of HT, in contrast to sequence-specific approaches, all the candidate sequences are not assumed to have been horizontally transferred from one species to the other, and the procedures need to include validation steps to produce sound results.

Validation of the Nonrepeated Content

Our approach follows an identity-based line of evidence to detect HTs, so we would need phylogenetic clues to validate them. In the case of *D. melanogaster* and *D. simulans*, which are almost at a terminal node of the *Drosophila* phylogenic tree, phylogenetic incongruences would mostly consist of nonsignificant differences in branch lengths compared with those expected. Even if incomplete lineage sorting could remain a problem, for a sequence-specific identity-based approach, the validation procedures mainly consist of showing evidence that the high observed nucleotidic identity is not the result of other mechanisms than HT, such as purifying selection or a mutational cold spot (Pace et al. 2008; Casillas et al. 2007). When dealing with genome-wide data, tools such as SCONE or the ones from the PHAST package can produce conservation tracks from multiple genome-alignment between different species (Asthana et al. 2007; Hubisz et al. 2011). However, these conservation tracks consist of quantitative scores to measure the departure from neutrality for each nucleotide, and these scores are difficult to incorporate into a statistical test to determine whether a given detected fragment is conserved or horizontally transferred.

We thus developed a more conservative approach that also accounted for non–CDSs, by subtracting the results of the *D. melanogaster–D. simulans* comparison from those retrieved

for the comparison between *D. melanogaster* and other *Drosophila* species (*D. sechellia*, *D. yakuba*, *D. pseudoobscura,* and *D. virilis*). With this comparative analysis, we discriminated between putative horizontally transferred sequences and sequences under purifying selection that are expected to be conserved across the phylogeny genome-wide for coding and noncoding DNA. The use of those four other species strengthened our results by preventing the detection of stochastic loss or ancestral polymorphisms. These two mechanisms could lead to the detection of conserved sequences between *D. melanogaster* and *D. simulans* that are absent from a third species, which is unlikely to occur simultaneously in the five analyzed species.

Finally, as this line of evidence is easily accessible and should always be considered for the study of the HT of CDSs, we also checked for CDS d_S values lower than expected given the time of divergence of the considered species. Overall, our results show an absence of HTs involving CDS between *Drosophila* species (Schaack et al. 2010), which is not caused by detection bias toward these types of sequences. Because this validation procedure is restricted to the nonrepeated content of our results, we also developed a second validation procedure to assess the TEs identified.

Validation of Horizontally Transferred TEs

To identify horizontally transferred TEs among the set of TEs with an identity higher than expected between *D. melanogaster* and *D. simulans*, we analyzed their recent dynamics since their last putative HT (Dias and Carareto 2012). The TE dynamics and maintenance in the host genome can be described as a birth-and-death processes (Schaack et al. 2010; Le Rouzic et al. 2013). The death of a TE corresponds to the inactivation of all its copies by the host defense mechanisms or the accumulation of disabling mutations (Jurka et al. 2012). On the other hand, the birth of a TE corresponds to an active copy colonizing a novel host devoid of specific transposition controls against this TE, which immediately leads to the burst of transposition of the founder copy in the new genome (Le Rouzic and Capy 2005; Naito et al. 2009). Bursts of transposition have been recorded for different TEs in numerous *Drosophila* species (García Guerreiro 2012) and can be easily identifiable because all the resulting TE copies are almost identical to each other. Afterward, most of the copies accumulate stochastic mutations and are lost over time by attrition, except for a minority of copies that can become exapted and can be identified as DNA segments conserved across species (Margulies et al. 2003; Siepel et al. 2005; Pace et al. 2008). Because TEs are likely to evolve neutrally after their insertion, we could use the neutral rate of substitution to compute the timing of a burst of transposition by calculating the pairwise divergence between all the TE copies and their consensus as an approximation of the founder copy as described in the literature (Pace and Feschotte 2007; Schaack et al. 2010; Le Rouzic et al. 2013). However, the consensus is not always a good approximation of the ancestral copy. Thus, instead of studying the complete history of a TE family with a consensus approach, our method focuses on the period of time surrounding its last putative HT between the considered species and ignores the events older than the divergence time between *D. melanogaster* and *D. simulans*. Thus, this change in the time scale provided us with a better temporal resolution for the study of the last bursts of transposition. In *D. melanogaster*, where the TE activity was recent (Bowen and McDonald 2001; Lerat et al. 2003), we were not able to clearly discriminate between the different activity periods of the TE families with an approach based on an estimated neutral mutation rate between all the copies of a TE family and their consensus sequences (Ray et al. 2008). Moreover, for the TEs with different waves of activity, such as the element *transib3* (fig. 4*C*) in the studied species, a consensus would correspond to a hypothetical copy dated in the middle of the waves of activity rather than to the ancestral copy. Our approach solves these drawbacks of consensus-based TE analysis and accounts for highly variable lengths of copies between TE families.

Another important point concerning HTs is to determine the direction of these transfers. In the cases of horizontally transferred TEs, we could expect a species with a high number of TE copies to have a higher probability to horizontally transmit one of its copies to another species, resulting in numerous identical copies in one species and few in the other. For this scenario to be valid, the transferred TEs would need to be almost instantly regulated in the receiver species to stay at a low copy number or for the receiver species to be sequenced before their burst of transposition. For both cases, these TE insertions would not have a high frequency in the species and would most likely not be observed in the sequenced strains. In an opposite scenario (a horizontally transferred TE from a species with few putatively active but controlled TE copies), a TE is transferred to a species where this TE is unknown for the host regulation system, which would lead to a burst of transposition and a quick fixation of this TE in the receiver species. Consequently, we are more likely to observe the results of this second scenario in the sequenced individuals, and we can use it to decipher the direction of detected horizontally transferred TEs (Dias and Carareto 2012).

Overall, our results show that different waves of activity seem to have occurred for different TE families and that their dynamics can be used to describe the numerous horizontally transferred TEs between *D. melanogaster* and *D. simulans*. After a horizontally transferred TE and a burst of transposition, we expect to observe a unique wave of activity before the control of the element, so the presence of other waves seems to be indicative of a complex history of the TE dynamics in *Drosophila*.

Drosophila melanogaster as a Target of Multiple Independent Horizontally Transferred TE Events

Exchange of TEs with D. simulans

In regard to the number of horizontally transferred TEs that have been detected, a parsimonious hypothesis would be their simultaneous transfer by introgression instead of independent HTs. Thus, we can wonder why no traces of introgression were detected between *D. melanogaster* and *D. simulans*, when hybrids are known to have been possible between these two species (Sawamura et al. 2004; Barbash 2010). A first genomic explanation could be that due to mutations and recombinations, the signal of an introgressed fragment of DNA has been lost over time. In this case, we can wonder why this DNA fragment would have undergone such high recombination and mutation rates, when most of the DNA is still identifiable between *D. melanogaster* and *D. simulans*.

As "nothing in evolution makes sense except in light of population genetics" (Lynch 2007), we can try to answer this question at a population level. For TEs, many steps are necessary for an HT to be successful (Le Rouzic and Capy 2005). After passing through the new host barriers, the TE must transfer itself into the germ line to be transmitted to the descendants. Then, the TE needs to have a sufficient transposition rate to propagate into the host genome and to increase in frequency in the species by vertical transmission. For TEs, which are able to actively colonize genomes, most of the TE insertions in natural populations are absent from the sequenced genome, as shown by the study of 113 *D. melanogaster* strains isolated from natural population (Kofler et al. 2012).

In the case of an introgression, all the cells in the progeny of the backcross with the hybrid will have a copy of the introgressed DNA fragment, so the first step of contaminating the germ line is always successful. Afterward, this introgressed DNA fragment has to increase in frequency in the species to be likely to be observed in the individuals actually sequenced. In contrast to the active TE copies, the introgressed fragment cannot actively replicate itself in the new genome, and its probability of fixation is simply its frequency in the population, at least in diploid organisms with a large effective population size, such as *D. melanogaster* (Nolte and Schlötterer 2008). As a result, the frequency of this introgressed fragment would be almost null in comparison to the effective population size of *D. melanogaster*, and even with the carrier subpopulation hypothesis (Jurka et al. 2011), where the population is divided into demes in each of which we can observe an effect of genetic drift that favors the fixation of low-frequency alleles, the introgressed fragment would have a low probability to be transmitted to the other demes and to be fixed in the species (Ghosh et al. 2012). Therefore, we would need to use *D. melanogaster* and *D. simulans* population-genetic data to be able to detect any traces of introgression events, as in the recent study of introgression events between *D. simulans* and *D. sechellia* from Brand et al. (2013).

This population-genetic aspect of the genomic data needs to be taken into account, as it can explain other aspects of our TE-based results. For example, the differences in the detection of horizontally transferred TEs between *D. melanogaster* and *D. simulans* found between the 2012 version of the *D. simulans* genomes sequenced from one strain (Hu et al. 2013) and the version from the 12 *Drosophila* genomes project sequenced from five different strains (Drosophila 12 Genomes Consortium 2007) can be explained by the variability of TE insertions between the populations of *D. simulans* (Vieira and Biémont 2004).

With Other Drosophila *Species*

Overall, the extensive evidence of horizontally transferred TEs detected in *D. melanogaster* seems to indicate that the fixation of new TEs could be facilitated in this genome. The timing of most HTs involving *D. melanogaster* was estimated between 1.4 and 2.3 Ma, before the worldwide expansion of *D. melanogaster* and *D. simulans* that happened 15,000 years ago (Stephan and Li 2007; Cararetto 2011). The *melanogaster* subgroup is endemic to Afrotropical regions, with the proto-melanogaster founder dated between 17 and 20 Ma from the oriental region of Africa (Lachaise et al. 1988). Thus, a parsimonious hypothesis for the numerous horizontally transferred TEs detected among *D. melanogaster*, *D. simulans*, *D. sechellia,* and *D. yakuba* would place them at a time when these species were all living in Africa, before the worldwide expansion of *D. melanogaster* and *D. simulans*. In this scenario, there would have been fewer geographical barriers to hamper the fixation of horizontally transferred TEs into sympatric populations with a smaller repartition area than the worldwide populations of *D. melanogaster*. The arrival of these new TE copies in the genome *D. melanogaster* may have been a springboard for the worldwide expansion of this species, as the load of TEs can be correlated with the colonization of new territory (Vieira et al. 1999, 2002). In contrast, a stronger population structure in *D. simulans* (Mousset and Derome 2004) could explain the polymorphisms of TE insertion that have resulted in different TE loads between populations (Vieira and Biémont 2004) and that may have independently favored the worldwide expansion of this species, even if in both cases the cause of such a mechanism is not yet understood.

Conclusions

We have developed a novel approach for the genome-wide detection of all putative HT sequences independently of their coding capability between two genomes. Our method relies on a well-defined testing framework to approach this genome-wide problem as a multiple-testing problem. We successfully applied this method between the genomes of *D. melanogaster* and *D. simulans*, underscoring the sensitivity

of our approach to detect HTs between closely related species. Like previous studies of HTs in eukaryotes, we validated these results with other lines of evidence. We also proposed two novel approaches to remove bias due to the detection of conserved sequences, by a comparative analysis with phylogenetically related species in the case of CDS and non–CDSs and by an analysis of their recent activity in the case of TEs. After these validation steps, we retrieved all the horizontally transferred TEs previously described in different studies (see Carareto [2011] for a review) and very few spurious CDS, attesting to the sensitivity and the specificity of our approach.

By a manual analysis of our results along each chromosome arm of *D. melanogaster*, we did not detect any trace of introgression between *D. melanogaster* and *D. simulans*, even if this does not completely rule out this hypothesis. We also detected a large number of horizontally transferred TEs involving *D. melanogaster* and other *Drosophila* species with our assessment steps, bringing to light a small portion of the network of horizontally transferred TEs in this phylogeny. This large number of HTs for different TE families also supports the model of birth and death, where HT events are a vital part of the TE life cycle that prevents their extinction (Schaack et al. 2010). We are just beginning to understand the complex horizontally transferred TE network in eukaryotes, and our approach could be applied to any pair of sequenced species to increase our knowledge of the dynamics of these sequences, which seem to jump both within and between species.

Supplementary Material

Supplementary figure S1–S7 and tables S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank C. Carareto for her critical reading of the manuscript. The English of the manuscript has been edited by the American Journal Experts company. This work was supported by the ANR-09-BLANC-0103-01.

Literature Cited

Abby SS, Tannier E, Gouy M, Daubin V. 2012. Lateral gene transfer as a support for the tree of life. Proc Natl Acad Sci U S A. 109:4962–4967.
Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.
Andersson JO. 2005. Lateral gene transfer in eukaryotes. Cell Mol Life Sci. 62:1182–1197.
Anxolabéhère D, Kidwell MG, Periquet G. 1988. Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of *Drosophila melanogaster* by mobile *P* elements. Mol Biol Evol. 5:252–269.
Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S. 2007. Analysis of sequence conservation at nucleotide resolution. PLoS Comput Biol. 3:e254.
Azad RK, Lawrence JG. 2011. Towards more robust methods of alien gene detection. Nucleic Acids Res. 39:e56.
Barbash DA. 2010. Ninety years of *Drosophila melanogaster* hybrids. Genetics 186:1–8.
Bartolomé C, Bello X, Maside X. 2009. Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes. Genome Biol. 10:R22.
Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc. 57:289–300.
Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. Ann Statist. 10:467–498.
Bowen NJ, McDonald JF. 2001. *Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res. 11:1527–1540.
Brand CL, Kingan SB, Wu L, Garrigan D. 2013. A selective sweep across species boundaries in *Drosophila*. Mol Biol Evol. 30:2177–2186.
Capy P, Anxolabéhère D, Langin T. 1994. The strange phylogenies of transposable elements: are horizontal transfers the only explanation? Trends Genet. 10:7–12.
Carareto CM. 2011. Tropical Africa as a cradle for horizontal transfers of transposable elements between species of the genera *Drosophila* and *Zaprionus*. Mob Genet Elem. 1:179–186.
Casillas S, Barbadilla A, Bergman CM. 2007. Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. Mol Biol Evol. 24:2222–2234.
Daniels SB, et al. 1990. Evidence for horizontal transmission of the *P* transposable element between *Drosophila* species. Genetics 124: 339–355.
de Carvalho MO, Loreto ELS. 2012. Methods for detection of horizontal transfer of transposable elements in complete genomes. Genet Mol Biol. 35:1078–1084.
de Koning AP, et al. 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 7:e1002384.
de la Chaux N, Wagner A. 2009. Evolutionary dynamics of the LTR retrotransposons roo and rooA inferred from twelve complete *Drosophila* genomes. BMC Evol Biol. 9:205.
Dias ES, Carareto CMA. 2012. Ancestral polymorphism and recent invasion of transposable elements in *Drosophila* species. BMC Evol Biol. 12: 119.
Dowsett AE, Young MW. 1982. Differing levels of dispersed repetitive DNA among closely related species of *Drosophila*. Proc Natl Acad Sci U S A. 79:4570–4574.
Doyon JP, Ranwez V, Daubin V, Berry V. 2011. Models, algorithms and programs for phylogeny reconciliation. Brief Bioinform. 12:392–400.
Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218.
Dupuy C, et al. 2011. Transfer of a chromosomal Maverick to endogenous bracovirus in a parasitoid wasp. Genetica 139:489–496.
Efron B, Tibshirani R, Storey JD, Tusher V. 2001. Empirical Bayes analysis of a microarray experiment. J Am Statist Assoc. 96:1151–1160.
Fall S, et al. 2007. Horizontal gene transfer regulation in bacteria as a "spandrel" of DNA repair mechanisms. PLoS One 2:e1055.
Finnegan DJ. 1997. Transposable elements: how non-LTR retrotransposons do it. Curr Biol. 7:R245–R248.
Finner H, Roters M. 2002. Multiple hypotheses testing and expected number of type I. errors. Ann Statist. 30:220–238.
García Guerreiro MP. 2012. What makes transposable elements move in the *Drosophila* genome? Heredity 108:461–468.
Ghosh A, Meirmans PG, Haccou P. 2012. Quantifying introgression risk with realistic population genetics. Proc Biol Sci. 279:4747–4754.
Gilbert C, Cordaux R. 2013. Horizontal transfer and evolution of prokaryote transposable elements in eukaryotes. Genome Biol Evol. 5: 822–832.

Gilbert C, Pace JK, Feschotte C. 2009. Horizontal spinning of transposons. Commun Integr Biol. 2:117–119.
Gilbert C, Waters P, Feschotte C, Schaack S. 2013. Horizontal transfer of OC1 transposons in the Tasmanian devil. BMC Genomics 14:134.
Gilbert C, et al. 2010. A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature 464:1347–1350.
Granzotto A, Lopes FR, Vieira C, Carareto CMA. 2011. Vertical inheritance and bursts of transposition have shaped the evolution of the BS non-LTR retrotransposon in *Drosophila*. Mol Genet Genomics. 286:57–66.
Gronau I, Arbiza L, Mohammed J, Siepel A. 2013. Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. Mol Biol Evol. 30:1159–1171.
Heger A, Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes. Genome Res. 17:1837–1849.
Hu TT, Eisen MB, Thornton KR, Andolfatto P. 2013. A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. Genome Res. 23:89–98.
Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinform. 12:41–51.
Juhas M, et al. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev. 33:376–393.
Jurka J, Bao W, Kojima KK. 2011. Families of transposable elements, population structure and the origin of species. Biol Direct. 6:44.
Jurka J, et al. 2012. Distinct groups of repetitive families preserved in mammals correspond to different periods of regulatory innovations in vertebrates. Biol Direct. 7:36.
Kaminker JS, et al. 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol. 3: RESEARCH0084.
Kapitonov VV, Jurka J. 2003. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc Natl Acad Sci U S A. 100:6569–74.
Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet. 9:605–618.
Kim J, et al. 1994. Rodent *BC1* RNA gene as a master gene for *ID* element amplification. Proc Natl Acad Sci U S A. 91:3607–3611.
Kofler R, Betancourt AJ, Schlötterer C, Schlo C. 2012. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. PLoS Genet. 8: e1002487.
Kuan PF, Chiang DY. 2012. Integrating prior knowledge in multiple testing under dependence with applications to detecting differential DNA methylation. Biometrics 68:774–783.
Lachaise D, et al.. 1988. Historical biogeography of the *Drosophila melanogaster* species subgroup. In: Hecht MK, Wallace B, Prance GT, editors. Evolutionary biology. Springer. p. 159–225.
Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
Le Rouzic A, Capy P. 2005. The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. Genetics 169:1033–1043.
Le Rouzic A, Payen T, Hua-Van A. 2013. Reconstructing the evolutionary history of transposable elements. Genome Biol Evol. 5:77–86.
Lerat E, Burlet N, Biémont C, Vieira C. 2011. Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. Gene 473:100–109.
Lerat E, Rizzon C, Biémont C. 2003. Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. Genome Res. 13:1889–1896.
Lisch D. 2008. A new *SPIN* on horizontal transfer. Proc Natl Acad Sci U S A. 105:16827–16828.
Loreto ELS, Carareto CMA, Capy P. 2008. Revisiting horizontal transfer of transposable elements in *Drosophila*. Heredity 100:545–554.
Ludwig A, Valente VL, Loreto EL. 2008. Multiple invasions of Errantivirus in the genus *Drosophila*. Insect Mol Biol. 17:113–124.
Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A. 104:8597–8604.
Margulies EH, Blanchette M, Haussler D, Green ED. 2003. Identification and characterization of multi-species conserved sequences. Genome Res. 13:2507–2518.
Marygold SJ, et al. 2013. Flybase: improvements to the bibliography. Nucleic Acids Res. 41:751–757.
Mousset S, Derome N. 2004. Molecular polymorphism in *Drosophila melanogaster* and d. simulans: what have we learned from recent studies? Genetica 120:79–86.
Naito K, et al. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature 461: 1130–1134.
Nolte V, Schlötterer C. 2008. African *Drosophila melanogaster* and *D. simulans* populations have similar levels of sequence variability, suggesting comparable effective population sizes. Genetics 178:405–412.
Novāk E, Neumann P, Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11:378.
O'Brochta DA, et al. 2009. Transpositionally active episomal *hAT* elements. BMC Mol Biol. 10:108.
Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.
Pace JK, Feschotte C. 2007. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. Genome Res. 17:422–432.
Pace JK, Gilbert C, Clark MS, Feschotte C. 2008. Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc Natl Acad Sci U S A. 105:17023–17028.
Pinsker W, Haring E, Hagemann S, Miller WJ. 2001. The evolutionary life history of *P* transposons: from horizontal invaders to domesticated neogenes. Chromosoma 110:148–158.
Podell S, Gaasterland T. 2007. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. Genome Biol. 8:R16.
Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20:110–121.
Putonti C, et al. 2006. A computational tool for the genomic identification of regions of unusual compositional properties and its utilization in the detection of horizontally transferred sequences. Mol Biol Evol. 23: 1863–1868.
Quinlan AR, Hall IM. 2010. Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.
R Core Team. 2013. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Ragan MA. 2001. Detection of lateral gene transfer among microbial genomes. Curr Opin Genet Dev. 11:620–626.
Ray DA, et al. 2008. Multiple waves of recent DNA transposon activity in the bat, *Myotis lucifugus*. Genome Res. 18:717–728.
Roger A. 1999. Reconstructing early events in eukaryotic evolution. Am Nat. 154:S146–S163.
Sawamura K, Karr TL, Yamamoto MT. 2004. Genetics of hybrid inviability and sterility in *Drosophila*: dissection of introgression of *D. simulans* genes in *D. melanogaster* genome. Genetica 120:253–260.
Schaack S, Gilbert C, Feschotte C. 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol. 25:537–546.
Schnable PS, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115.
Shi SY, Cai XH, Ding DF. 2005. Identification and categorization of horizontally transferred genes in prokaryotic genomes. Acta Biochim Biophys Sin. 37:561–566.
Siepel A, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.

Silva JC, Loreto EL, Clark JB. 2004. Factors that affect the horizontal transfer of transposable elements. Curr Issues Mol Biol. 6: 57–71.
Stephan W, Li H. 2007. The recent demographic and adaptive history of *Drosophila melanogaster*. Heredity 98:65–68.
Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 100:9440–9445.
Sun W, Tony Cai T. 2009. Large-scale multiple testing under dependence. J Roy Statist Soc. 71:393–424.
Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol Biol Evol. 21: 36–44.
Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. Brief Bioinform. 14:178–192.
Vieira C, Biémont C. 1997. Transposition rate of the *412* retrotransposable element is independent of copy number in natural populations of *Drosophila simulans*. Mol Biol Evol. 14:185–188.
Vieira C, Biémont C. 2004. Transposable element dynamics in two sibling species: *Drosophila melanogaster* and *Drosophila simulans*. Genetica 120:115–123.
Vieira C, Lepetit D, Dumont S, Biémont C. 1999. Wake up of transposable elements following *Drosophila simulans* worldwide colonization. Mol Biol Evol. 16:1251–1255.
Vieira C, et al. 2002. Evolution of genome size in *Drosophila*.. Is the invader's genome being invaded by transposable elements? Mol Biol Evol. 19:1154–1161.
Wallau GL, Ortiz MF, Loreto ELS. 2012. Horizontal transposon transfer in eukarya: detection, bias, and perspectives. Genome Biol Evol. 4: 689–699.
Wei Z, Sun W, Wang K, Hakonarson H. 2009. Multiple testing in genome-wide association studies via hidden Markov models. Bioinformatics 25: 2802–2808.
Weinert LA, Welch JJ, Jiggins FM. 2009. Conjugation genes are common throughout the genus *Rickettsia* and are transmitted horizontally. Proc R Soc B. 276:3619–3627.
Weir JT, Schluter D. 2008. Calibrating the avian molecular clock. Mol Ecol. 17:2321–2328.
Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.

Associate editor: Josefa Gonzalez