Published online 21 December 2010 *Nucleic Acids Research, 2011, Vol. 39, No. 6* ***e36***
doi:10.1093/nar/gkq1291

T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data

Anna-Sophie Fiston-Lavier*, Matthew Carrigan, Dmitri A. Petrov and Josefa González

Department of Biology, Stanford University, 371 Serra St., Stanford, CA 94305-3020, USA

Received September 7, 2010; Accepted November 30, 2010

ABSTRACT

Transposable elements (TEs) are repetitive DNA sequences that are ubiquitous, extremely abundant and dynamic components of practically all genomes. Much effort has gone into annotation of TE copies in reference genomes. The sequencing cost reduction and the newly available next-generation sequencing (NGS) data from multiple strains within a species offer an unprecedented opportunity to study population genomics of TEs in a range of organisms. Here, we present a computational pipeline (T-lex) that uses NGS data to detect the presence/absence of annotated TE copies. T-lex can use data from a large number of strains and returns estimates of population frequencies of individual TE insertions in a reasonable time. We experimentally validated the accuracy of T-lex detecting presence or absence of 768 previously identified TE copies in two resequenced *Drosophila melanogaster* strains. Approximately 95% of the TE insertions were detected with 100% sensitivity and 97% specificity. We show that even at low levels of coverage T-lex produces accurate results for TE copies that it can identify reliably but that the rate of 'no data' calls increases as the coverage falls below 15×. T-lex is a broadly applicable and flexible tool that can be used in any genome provided the availability of the reference genome, individual TE copy annotation and NGS data.

INTRODUCTION

Transposable elements (TEs) are ubiquitous, ancient, diverse, often highly destructive and at times surprisingly constructive members of the genomic community (1). There is virtually no facet of genome biology that has not been affected by the continuous co-evolution of TEs with their host genomes (2–5). In some celebrated cases TE copies have been co-opted to play key functions such as the generation of antibody diversity in the vertebrate immune system (6), maintenance of telomeres in Drosophila (7) and evolution of new and rewiring of old regulatory networks (8–11). It is also quite likely that epigenetic mechanisms such as gene silencing through methylation, RNAi and RiP epigenetic mechanisms evolved as a means of genomic defense against TEs (12).

TEs produce most repetitive DNA in eukaryotic genomes, generating much of the genome bulk and leading to elevated rates of genome rearrangements. In addition, repetitive DNA poses a major challenge to genome sequencing and assembly. It is indisputable that we cannot understand genome structure and function without a thorough understanding of TEs. Beyond that, our very ability to obtain full genome sequences depends crucially on our ability to annotate individual TE copies. The full understanding of TE biology and their effect on the function, variation and evolution of organisms will require not only finding ways to annotate TE copies well in a single genome but also to understand the patterns of presence and absence of individual TE copies in different strains.

The availability of next-generation sequencing (NGS) data is allowing high-throughput detection of various types of copy number variants (CNVs), such as TEs. Sequencing-based approaches enable a more thorough detection of CNVs with a higher resolution than other approaches (e.g. microarray analysis). The annotation of TE copies presents a distinct problem from that of other CNVs because TEs are often highly repetitive with tens, hundreds and at times millions of copies in a genome (13–16). The number of TE insertions in a genome generally [but not always (17)] does not vary significantly among strains while the exact location of individual TEs can be almost entirely non-overlapping in different strains (18–20). The depth coverage of reads mapping to a TE

*To whom correspondence should be addressed. Tel: +1 650 736 2249; Fax: +1 650 723 6132; Email: afiston@stanford.edu sequence is thus not informative about the presence of individual TE insertions. Many TE insertions are also longer than most paired-end inserts making it hard to use the distance between the reads that map to unique places in the genome to detect the presence of TE insertion.

On the other hand TE copies are often well characterized with known sequences of not only functional, full-length copies but also of individual TE copies at least in one or several reference genomes. It is important to use this available information to produce high quality assessment of the presence of individual TE insertions in the sequenced genomes. Here we describe a computational pipeline designed specifically to detect presence/absence of annotated individual TE insertions in sequenced genomes. Designed for Solexa/Illumina data, we called this pipeline T-lex (T for 'Transposable element' and lex for 'soLEXa data') to emphasize that it is designed to detect TEs in NGS data. When using data from multiple strains, T-lex both ascertains presence/absence of individual TE copies for each strain and also returns the frequency estimate for each TE insertion in the tested strains.

T-lex is a broadly applicable and flexible tool that can be used in any genome provided the availability of the reference genome, individual TE copy annotation and NGS data. We demonstrate that T-lex performs well by using *Drosophila melanogaster* individual TE copy annotation and Solexa/Illumina data of two newly sequenced strains. We describe T-lex design and show that provided moderate sequence coverage (15×) it can identify presence/absence for the majority of TE insertions with high levels of accuracy. At lower levels of coverage T-lex does not make more errors but rather returns more 'no data' calls, providing useful feedback about the quality of the data. T-lex is an efficient, accurate and cost-saving tool for TE copies annotation in NGS population genomic data. T-lex is available for download at: http://petrov.stanford.edu/cgi-bin/Tlex_manual.html.

MATERIALS AND METHODS

T-lex pipeline overview

T-lex is a pipeline designed to detect annotated TE copies by specifically looking for two types of reads in NGS data. The presence of a TE insertion in a sequenced genome should generate reads that span the junction of the TE insertion and its flanking sequences. Such reads are generally unique and identify the presence of a TE insertion with high specificity. Similarly, the absence of a TE insertion should be indicated by the presence of reads that span the two flanking sequences and the lack of the sequence of the TE copy. T-lex relies on the availability of the reference genome sequence, the annotation of TE copies specifically identified in the reference genome as well as on having the ability to map reads uniquely to the flanking sequences. Thus, T-lex has limited ability to investigate presence and/or absence of TE insertions embedded in highly repetitive regions, such as TEs inserted inside other TEs or in regions with high repetitive content in general. In practice this means that we are restricted to the analysis of the TE population dynamics in the euchromatic, gene-rich regions that are the regions of the primary interest in most cases. T-lex can be run for a single TE copy in a single NGS data set as well as for several TE copies and/or several NGS data sets sequentially in an automatic fashion.

T-lex is composed of four main modules. The first module, 'TEfilter' identifies and eliminates TE insertions present in highly repetitive regions from the list of TE insertions to be analyzed. The second and third modules are detection modules that assess the 'presence' and the 'absence' of the TE insertions, respectively. The last module, "combine", uses the results from the "presence" and the "absence" modules to ascertain the presence/absence of the TE insertions for each strain and also returns the frequency estimate for each TE insertion in the tested strains. The source script and documentation are available on the T-lex website: http://petrov.stanford.edu/cgi-bin/Tlex_manual.html.

Input data

T-lex requires four input data sets: (i) the list of the TE insertions to be analyzed, (ii) the annotations of these TE insertions, specifically arranged in four tabulated columns: the TE name, the location (i.e. chromosome, BAC name, etc.), the start and the end nucleotide positions of the TE insertion and (iii) the reference genome sequence and (iv) the NGS data in the "official" fastq format (http://en.wikipedia.org/wiki/Fastq). If desired, simple repeats and low-complexity regions in the reference genome can be masked prior to running T-lex using RepeatMasker (21). This step increases the specificity of the mapping of the NGS reads to the reference genome. TE insertion coordinates, i.e. the start and end nucleotide position of each TE insertion, are used to obtain the coordinates of the flanking regions for each TE insertion.

T-lex "TEfilter" module

T-lex extracts 100 bp of the flanking regions for each TE insertion and runs RepeatMasker to mask all the repeats in these regions (22). If at least one flanking region shows repeat density greater than the pre-specified value (50% by default), this TE copy will be eliminated from the list of TE copies to be analyzed. The length of the flanking region analyzed can be changed by the user and should be equal or longer than the maximum read length of the NGS data. RepeatMasker requires specifying the name of the species to be analyzed (22). The user should do this using the option '-S' in the command line of T-lex. The "TEfilter" module both accelerates and ensures an accurate T-lex detection result. By default, the TE copies are filtered. To remove this step the user can set the option "-noFilterTE" (http://petrov.stanford.edu/cgi-bin/Tlex_manual.html).

T-lex "presence" module

The 'presence' module is described in Figure 1a. T-lex uses Maq for the detection of presence of the TE copies (23). It starts by extracting the two flanking sequences (1 kb by default) of each TE copy and 20 bp from the terminal regions of the TE copy. We determined empirically that 20 bp is the length at which only the reads that overlap both the TE copy region and the flanking region are mapped to the TE copy. Using longer terminal regions increases the probability of erroneous mapping of reads generated by related TE copies elsewhere in the genome while using shorter regions leads to high levels of non-specific mapping.

The extracted sequences are then converted into binary fasta (bfa) format, while the NGS fastq files are converted into binary fastq (bfq) format by Maq (23) (Figure 1a). Because fastq to bfq reformatting is one of the most time-consuming steps of the presence detection, this reformatting step can be parallelized using the option '-processes'. This option allows reformatting of several fastq files from the same strain into bfq at the same time. Parallel execution is expected to provide a benefit on some systems, particularly those with high-performance solid-state drives that eliminate the disc-thrashing problem.

The bfq reads are then mapped as single reads on the extracted flanking sequences of the TE copies using Maq with its parameters set to default values. Maq uses these mapped reads to build a contig for each side of the TE insertion (23) (Figure 1a). T-lex uses the alignment of the contigs on the reference sequence to define the presence and/or absence of the TE copy in a strain. If at least five base pairs of one of the contigs overlaps the TE copy sequence, T-lex classifies the TE copy as present (Figure 1a). If less than five base pairs of the TE sequence is present in the overlap, T-lex classifies the TE copy as absent (Figure 1a). Because the lack of reads overlapping the flanking region and the TE can also be due to low read coverage, TE copies that are in fact present in the strain could be erroneously classified as absent. To avoid such erroneous calls we use an additional heuristic step: if in addition to missing reads overlapping the TE copy, we also miss more than five base pairs of the terminal flanking sequence of the contig T-lex returns 'no data' as the call (Figure 1a). This indicates that the sequencing data did not allow T-lex to make solid

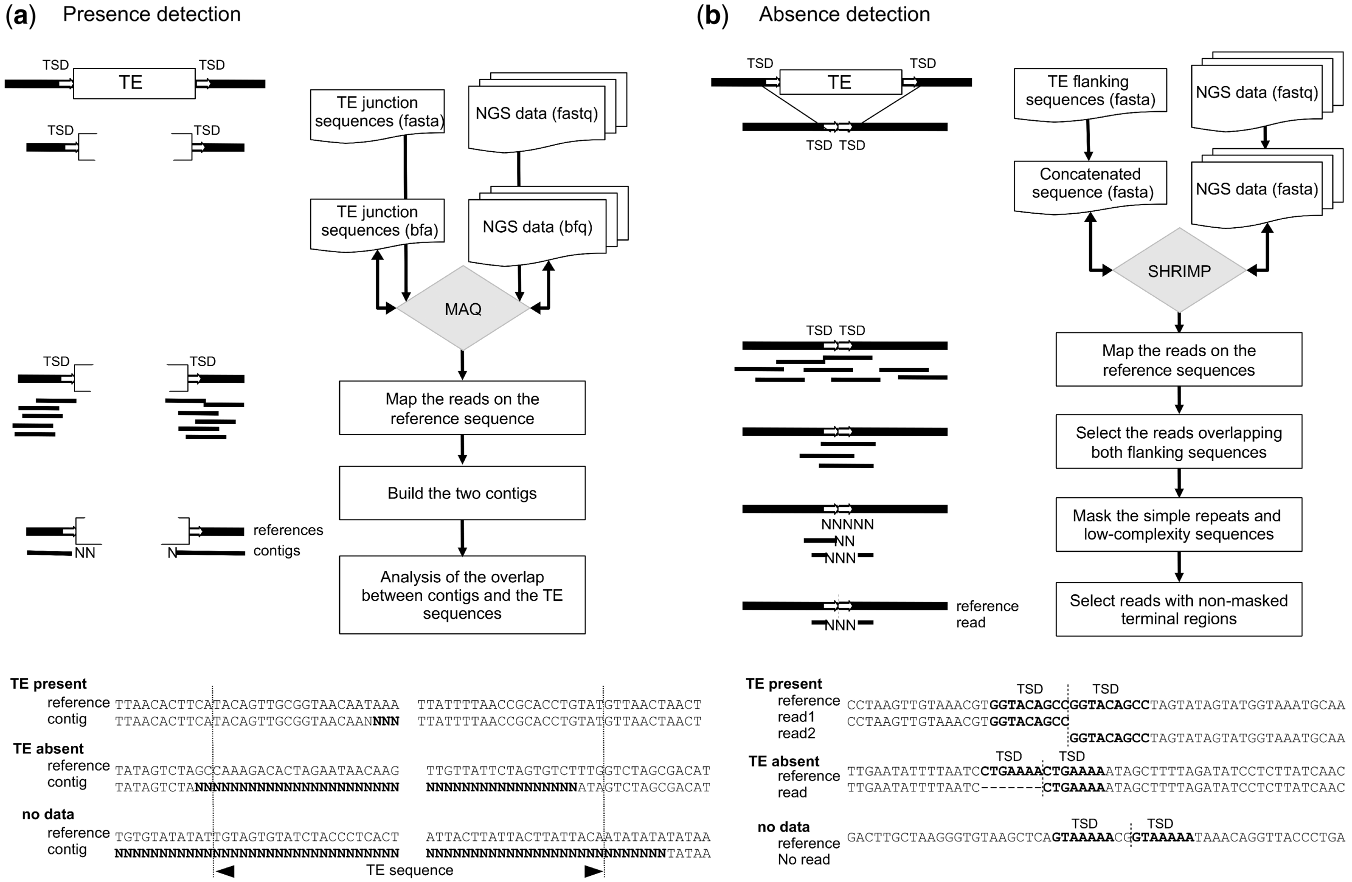
(a) Presence detection
TSD
TE
TSD
TSD
TSD
TE junction sequences (fasta)
NGS data (fastq)
TE junction sequences (bfa)
NGS data (bfq)
MAQ
Map the reads on the reference sequence
Build the two contigs
Analysis of the overlap between contigs and the TE sequences
TSD
TSD
NN
N
references
contigs
(b) Absence detection
TSD
TE
TSD
TSD TSD
TE flanking sequences (fasta)
NGS data (fastq)
Concatenated sequence (fasta)
NGS data (fasta)
SHRIMP
Map the reads on the reference sequences
Select the reads overlapping both flanking sequences
Mask the simple repeats and low-complexity sequences
Select reads with non-masked terminal regions
TSD TSD
NNNNN
NN
NNN
NNN
reference
read
TE present
reference TTAACACTTCATACAGTTGCGGTAACAATAAA TTATTTTAACCGCACCTGTATGTTAACTAACT
contig TTAACACTTCATACAGTTGCGGTAACAANNNN TTATTTTAACCGCACCTGTATGTTAACTAACT
TE absent
reference TATAGTCTAGCCAAAGACACTAGAATAACAAG TTGTTATTCTAGTGTCTTTGGTCTAGCGACAT
contig TATAGTCTANNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNATAGTCTAGCGACAT
no data
reference TGTGTATATATTGTAGTGTATCTACCCTCACT ATTACTTATTACTTATTACAATATATATATAA
contig NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNTATAA
TE sequence
TE present
TSD TSD
reference CCTAAGTTGTAAACGTGGTACAGCCGGTACAGCCTAGTATAGTATGGTAAATGCAA
read1 CCTAAGTTGTAAACGTGGTACAGCC
read2 GGTACAGCCTAGTATAGTATGGTAAATGCAA
TE absent
TSD TSD
reference TTGAATATTTTAATCCTGAAAACTGAAAAATAGCTTTTAGATATCCTCTTATCAAC
read TTGAATATTTTAATC-------CTGAAAAATAGCTTTTAGATATCCTCTTATCAAC
no data
TSD TSD
reference GACTTGCTAAGGGTGTAAGCTCAGTAAAAACGGTAAAAATAAACAGGTTACCCTGA
No read

Figure 1. T-lex presence and absence detection modules. (**a**) The 'presence' detection module is based on the mapping of the NGS reads on the TE insertion junctions. The TE insertion junctions encompass the flanking region of the TE insertion and the terminal TE insertion sequence. After extracting the two TE insertion junctions, the input data are reformatted. T-lex launches Maq to map the reads on these sequences. The reads are then assembled to obtain two contigs (one for each side). Gaps are represented by 'Ns'. The alignments of the contigs are used to define the presence and/or absence of the TE insertion ('Materials and Methods' section). (**b**) The 'absence' detection module is based on the mapping of the reads on the putative ancestral genomic sequence before the TE insertion. The flanking sequences of the TE insertion are extracted and concatenated. NGS data is reformatted. T-lex maps the NGS reads on the putative ancestral sequence using SHRiMP. T-lex masks the simple sequence repeats and low-complexity sequences of the selected NGS reads. Only the non-fully repetitive reads are used to define the absence of the TE insertion (see 'Materials and Methods' section). The repetitive regions are represented here by 'Ns'.

conclusions about the presence of the TE insertion using the presence module (Figure 1a). This procedure allows us to take into account the expected decay in the read number at the extremities of the reference sequence due to the mapping process. In addition, in order to facilitate the manual curation of the TE detection results, T-lex returns for each TE side the alignment of the contigs from all the tested strains in fasta format.

T-lex 'absence' module

The workflow of the 'absence' module is shown in Figure 1b. T-lex starts by formatting the NGS fastq files into fasta. The two 100 bp flanking sequences from each side of the TE insertion are extracted and then concatenated. This should approximate the ancestral sequence prior to the TE insertion except for the expected presence of tandem site duplications (TSDs). TSDs are short (two to ~20 bp) tandem duplications of the target site generated by transposition of practically all TEs (24). In a genome in which a TE insertion is absent, reads spanning the target site of this TE copy are not expected to have a TSD (i.e. only one copy of the target site should be detected). To allow for mapping of reads despite this expected presence of gaps, T-lex uses SHRiMP (25), which was specifically designed to handle long gaps and polymorphisms. This ability of SHRiMP should also reduce the negative effect of possible mis-annotations of TE copies on the performance of the absence module of T-lex.

T-lex looks for reads spanning the two flanking regions with at least 15 bp (by default) of overlap on each side. We recommend that the length of the extracted sequence should be longer than the reads themselves in order to obtain accurate mapping. T-lex launches the RepeatMasker program with the option '-noint' that only masks the simple sequence repeats and low-complexity regions (22). T-lex then selects only those reads that do not fully correspond to repeated sequences in order to overcome the problem posed by the presence of TSDs or poly-A tails (Figure 1b). By default the reads with at least 5 bp of non-repeat terminal sequence on both sides of the read are used to define the absence of the TE. If at least one read maps correctly on both TE sides, T-lex classifies the TE as 'absent' (Figure 1b). If no reads overlap the TE junction, T-lex checks the presence of reads in the flanking regions of the TE. The absence module concludes that the TE is 'present' if at least one read is detected (Figure 1b). In all other cases the absence module returns a 'no data' call. To aid with the manual curation of the results, T-lex returns the number of unique reads overlapping the TE junction and the alignments of these reads to the reference sequence. This information can be used to inform the level of confidence in the absence detection.

T-lex 'combine' module

T-lex then combines the results from the two detection modules (Table 1). The logic is based on the 'asymmetric' detection power of each module: the presence module returns presence results with high confidence whereas its absence calls could often be false negatives especially when the sequence coverage is low; the reverse is true for the absence module, which has much higher confidence in calling absence of a TE copy than its presence. Thus when the presence module returns 'present', the TE copy can be classified as 'present', 'polymorphic' or 'present/polymorphic', while when the absence module returns the 'absent' result, the TE copy can be classified as 'absent', 'polymorphic' or 'absent/polymorphic' (Table 1).

Table 1. Combination of the detection results from the two T-lex detection modules

Absence detection result	Presence detection result	Combination
Absent	Absent	Absent
Absent	No data	Absent/polymorphic
Absent	Present	Polymorphic
Present	Absent	No data
Present	No data	No data
Present	Present	Present
No data	Absent	No data
No data	No data	No data
No data	Present	Present/polymorphic

The 'present' call is the result of both detection modules supporting the presence of the TE copy. Similarly, T-lex returns the 'absent' call when both detection modules support the absence of the TE copy. The 'polymorphic' call is a result of the presence module returning presence and the absence module returning absence, i.e. when we have strong evidence of both presence and absence of the same TE copy. The call of 'present/polymorphic' corresponds to the instance when there is strong evidence of presence (presence call made by the presence module) while the absence module returned 'no data' result; similarly the 'absent/polymorphic' call is the case when there is strong evidence of absence (absence call by the absence module) and 'no data' result from the presence module (Table 1). The flexible architecture of this system easily allows this logic to be modified. It is also straightforward to incorporate additional data in making final calls.

Using the combination of the results, T-lex estimates the frequency for each TE copy in the population adding the number of strains for which the TE copy is 'present' and one-half times the number of 'polymorphic' strains, and dividing by the total number of strains for which T-lex returns data.

T-lex operation and output

The full T-lex process is launched by default. The user can launch only one of the two detection modules adding in the command line the options '-q' for the presence detection and '-p' for the absence detection.

All the T-lex results from all the TE copies and all the strains obtained in a single run of T-lex are stored in one output directory. By default this directory is called 'T-lex_[project]' associated with the name of the project (the user can change it using the option '-O'). Two sub-directories called 'Tfilter' and 'Talign' created, and contain, respectively, the files corresponding to TE filtering and the multiple alignments generated by T-lex. The final results from the two detection approaches are stored in the 'Tresults' file and the TE frequency estimates are stored in the 'Tfreq' file. Using the option '-noclean', the intermediate files such as the reformatted input data will be also returned. This option allows re-running T-lex for the same NGS or the same TEs bypassing the preparation of the reference sequences (using the option '-binref' in the command line) or bypassing the initial formatting of the NGS reads (using the '-binread' option). For additional details please consult the T-lex manual at http://petrov.stanford.edu/cgi-bin/Tlex_manual.html.

T-lex computational requirements

T-lex can be run on all UNIX platforms. We successfully implemented T-lex on a cluster (BioX^2 cluster at Stanford University; http://biox2.stanford.edu/). Computations described in this article were performed on a personal PC [4 Intel(R) Xeon(R) CPU 2.33 GHz with 8 GB of RAM]. To run T-lex, only the RepeatMasker (RepeatMasker Open-3.2.9, www.repeatmasker.org/), Maq (23) and SHRiMP (25) programs need to be installed (http://petrov.stanford.edu/cgi-bin/Tlex_manual.html).

T-lex computational time scales linearly with the number of NGS data sets in which the mapping has to take place. The scaling with the depth of coverage is also approximately linear: the full T-lex pipeline for the detection of 768 TEs in one strain of 15× coverage took approximately 3 h and with 50× coverage took approximately 8 h. T-lex pipeline is easily run in parallel, which can vastly reduce computational time, especially when TEs need to be mapped in a large number of strains.

Data set

We run T-lex on Solexa/Illumina data for two *D. melanogaster* strains (*W1 and Canton-S*) provided by Michael Eisen. The data consists of 100 bp paired-end Solexa/Illumina reads in each strain, with a coverage per strain of ~50× on average. To test the impact of coverage on the performance of T-lex, we subsampled the sequencing data to a lower coverage (10×, 15×, 20×, 30× and 40×). For each coverage level we first estimated the number of reads necessary to achieve this coverage and then randomly extracted the requisite number of reads. We checked the coverage by mapping the selected reads on the reference genomic sequence using the Maq program and repeated the extraction step if the coverage was incorrect (23). We carried out subsampling on each chromosome separately in order to take into account coverage differences among the chromosomes.

Release 5 of the *D. melanogaster* genome sequence was downloaded from the flybase website (http://flybase.org) (21). The TE data set consists of 768 annotated euchromatic and non-nested TE copies in release 5.30 of *D. melanogaster* for which an estimate of the TE insertion frequency in north American populations is available (Supplementary Table S1; 5,26,27). These TEs are broadly distributed across the genome and represent all TE superfamilies present in *D. melanogaster*.

Experimental TE detection

To verify experimentally the TE presence and/or absence in the tested strains, we used the PCR approach described in González *et al* (27) for 34 randomly chosen TE copies. These TE copies span the range of T-lex results in the two strains, e.g. presence in both strains or absence in both strains ('Results' section). All primers were designed using Primer3 (28) and were checked with VirtualPCR (29). One set of primers was intended to assay for the presence of the TE insertion and consists of a 'Left' (L) primer which is designed to be inside the TE sequence and a 'Right' (R) primer that maps to the flanking region to the right of the TE insertion. This PCR gives a band only when the TE copy is present. The other set of primers was intended to assay for the absence of the TE insertion and consisted of a 'Flank' (FL) primer that is located in the left flanking region of the TE insertion and the R primer mentioned above. In this case, the absence of a TE copy in the strain should give a shorter 'absence' band and the presence of a TE copy should give a longer, 'presence' band. We assumed that the 'presence' band is unlikely to be amplified if the TE sequence was longer than 800 bp (27). For each strain, DNA was extracted from 10 adult females.

RESULTS

T-lex validation

We validated the performance of T-lex by detecting the presence/absence for 768 euchromatic and non-nested TE copies whose presence/absence detection was previously investigated using PCR in a large number of *D. melanogaster* strains (Supplementary Table S1; 5,26,27). We ran T-lex for these 768 TE copies on two *D. melagonaster* laboratory strains, *W1 and Canton-S*, for which Solexa/Illumina data (100 bp, paired-end, ~50× coverage in average) is available. The two sequenced strains are inbred but not entirely isogenic—we expect that while most of the TE copies should be detected as fully present or fully absent, some might be polymorphic in these strains and we should detect both the presence and the absence for such TE copies. As input we used the Release 5.30 *D. melanogaster* reference genome (21) in which simple sequence repeats and low-complexity regions have been masked first by RepeatMasker (22).

We ran T-lex using the parameters by default ('Materials and Methods' section). We turned off the 'TEfilter' module of T-lex since we know that the flanking regions of these 768 TE copies do not have repetitive sequences other than *INE-1* elements (Supplementary Table S1; 5,26,27). INE-1 elements are fixed in *D. melanogaster* (24,25) and therefore we do not expect them to interfere with the presence/absence detection of TE copies by T-lex.

T-lex returned 1536 instances of detection, i.e. separate determinations of presence and/or absence. For ~12% (183/1536) of the instances T-lex returned 'no data', with 152 instances coming from 76 TE copies for which T-lex returned 'no data' in both strains (Supplementary Table S2). We investigated whether particular features of these TE copies prevented T-lex from returning a result and determined that in most cases the problem stemmed from mis-annotation of the TE junctions in the reference genome. For the remaining 31 TE copies (i.e. 31 instances), T-lex returns 'no data' in only one strain (Supplementary Table S2). Manual inspection of these 31 results revealed that the 'no data' calls are due to one or a combination of the following cases: low read coverage for at least one of the two TE copy sides, presence of a long low-complexity region at the TE copy junction, or presence of long TSDs. Thus, if we eliminate the 76 mis-annotated TE copies, we are left with 692 TE copies generating 1384 instances of detection. T-lex returns 'no data' result for 31 out of these 1384 instances giving us ~98% detection success rate.

We compared T-lex results for the 661 TE copies for which we had T-lex results for both strains with previously obtained frequency estimates (i.e. 1322 instances of detection; Figure 2 and Supplementary Table S3). These frequency estimates were generated by using PCR with DNA pooled from 64 North America strains and 11 sub-Saharan African strains (Supplementary Table S1; 5,26,27). TE copies had been grouped by these previous studies into four frequency classes: 'fixed', 'common', 'rare' and 'very rare' (Figure 2).

The 105 TE copies classified as 'fixed' were present in all the strains previously analyzed and thus are likely to be present also in the two strains used here (*W1 and Canton-S*). As expected, the vast majority of 'fixed' TE copies (101/105) were found present both in *W1 and Canton-S* strains (Supplementary Table S3). Only two out of the 105 'fixed' TE copies were detected as polymorphic in both strains, one TE copy was present in one strain and polymorphic in the other strain and another TE copy was detected as polymorphic in one strain and absent/polymorphic in the other strain. Manual curation of the results for these four TE copies confirmed T-lex calls. Thus, T-lex does not generate false negative TE presence detections and therefore identifies the presence of TE insertions with 100% sensitivity.

Three hundred fifty-three TE copies had been classified as 'very rare' (Supplementary Table S1; 5,26,27), meaning that they were not detected in any of the previously tested strains and thus should generally be absent from *W1 and Canton-S* as well. As expected, the majority (81%) of the instances of detection (i.e. 574 out of the 706 instances of detection for the 'very rare' TE copies) were true negatives (Supplementary Table S3). The rest (i.e. 132 instances of detection) correspond to 115 TE copies. For only 17 of these TE copies, evidence of presence was provided in both strains: 12 are polymorphic in both strains, three are present in one strain and polymorphic in the other and only two are present in both strains (FBti0020101, FBti0019180). We manually curated all the TEs for which we had evidence of presence in at least one strain (132 instances of detection). We confirmed 114 instances, while in 18 instances the quality of the alignment generated by T-lex was poor and we believe that T-lex determination of presence was false. Thus, after manual curation we end up with 18 false positive therefore we identified 574 out of the 592 instances of 'true' absence detections, giving us an estimate of 97% specificity.

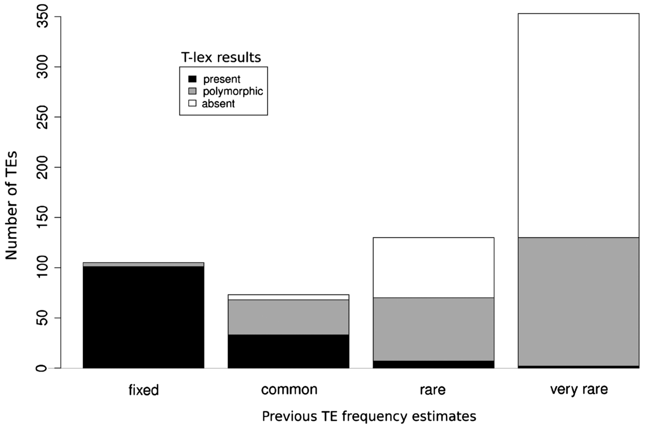

Figure 2. Comparison of T-lex results with previous TE frequency estimates. The 661 TE insertions for which T-lex returns results in both strains are classified as 'present', 'absent' and 'polymorphic'. These results are compared with previous TE insertion frequency estimates that classified the TE insertions as 'fixed', 'common', 'rare' and 'very rare' (Supplementary Table S1; 5,26,27).

'Common' and 'rare' TE copies were previously found to be present at intermediate frequencies, with the 'common' TE copies expected to be present at higher frequencies on average than the 'rare' TE copies. 48% (98/203) of the 'common' and 'rare' TE copies were classified as polymorphic by T-lex. As expected, the 'common' TE copies are detected as present more often than the 'rare' TEs ($\chi^2 = 15.5$, *d.f.* = 3, $P << 0.01$).

We further tested T-lex performance by using PCR to assess the presence/absence of a random subset of 34 TE copies in the *W1 and Canton-S* stocks that were used to prepare the Solexa libraries ('Materials and Methods' section). PCR failed for five TE copies: three TEs in both strains and two TEs in a single strain, while T-lex returns 'no data' for two TEs. At the end we analyzed 56 detection instances, i.e. all the cases in which both PCR and T-lex returned a result in at least one strain (Table 2). Note that because these strains are not entirely isogenic, it is entirely possible and even likely that the flies used to construct the Solexa libraries differed in their pattern of TE presence and absence from the flies used for PCR. T-lex and PCR results agree in 44 instances and disagree in 12 instances. The manual curation of the instances of disagreement resulted in correction of five T-lex results (Table 2). In two cases we have clear misinference of presence by T-lex (FBti0019223, Fbti0019360) such that absent TE copies

Table 2. Comparison of T-lex and PCR results for 56 instances for which T-lex and PCR approaches retuned a result in at least one strain

TE identifier	Strain	T-lex result	PCR result	T-lex result after manual curation
FBti0018879	*Canton-S*	Absent	Absent	Absent
FBti0018879	*W1*	Present	Present	Present
FBti0018880[a]	***Canton-S***	**Present**	**Polymorphic**	**Present**
FBti0018880	*W1*	Present	Present	Present
FBti0018884	*Canton-S*	Polymorphic	Polymorphic	Polymorphic
FBti0018884	*W1*	Absent/polymorphic	Absent	Absent/polymorphic
FBti0018889	*Canton-S*	Present	Present	Present
FBti0018889	*W1*	Present	Present	Present
FBti0018892	*Canton-S*	Polymorphic	Polymorphic	Polymorphic
FBti0018892[a]	***W1***	**Polymorphic**	**Absent**	**Polymorphic**
FBti0018955	*Canton-S*	Present	Present	Present
FBti0018955	*W1*	Present	Present	Present
FBti0018978[b]	***Canton-S***	**Polymorphic**	**Present**	**Present/polymorphic**
FBti0018978	*W1*	Absent/polymorphic	Absent	Absent/polymorphic
FBti0018980	*Canton-S*	Polymorphic	Polymorphic	Polymorphic
FBti0018980	*W1*	Present	Present	Present
FBti0018999	*Canton-S*	Present	Present	Present
FBti0018999	*W1*	Present	Present	Present
FBti0019056	*Canton-S*	Present	Present	Present
FBti0019056	*W1*	Present	Present	Present
FBti0019065[a]	***Canton-S***	**Polymorphic**	**Absent**	**Polymorphic**
FBti0019065	*W1*	Present	Present	Present
FBti0019081	*Canton-S*	Present	Present	Present
FBti0019081	*W1*	Present	Present	Present
FBti0019164	*Canton-S*	Present	Present	Present
FBti0019164	*W1*	Absent	No data	Absent
FBti0019223[b,c]	***Canton-S***	**Polymorphic**	**Absent**	**Absent**
FBti0019223	*W1*	Absent	Absent	Absent
FBti0019294	*Canton-S*	Polymorphic	Polymorphic	Polymorphic
FBti0019294	*W1*	Present	No data	Present
FBti0019296[a]	***Canton-S***	**Absent**	**Polymorphic**	**Absent**
FBti0019296	*W1*	Present	Present	Present
FBti0019344	*Canton-S*	Absent/polymorphic	Absent	Absent/polymorphic
FBti0019344	*W1*	Present	Present	Present
FBti0019360[b,c]	***Canton-S***	**Polymorphic**	**Absent**	**Absent**
FBti0019360	*W1*	Absent	Absent	Absent
FBti0019372	*Canton-S*	Polymorphic	Polymorphic	Polymorphic
FBti0019372	*W1*	Absent	Absent	Absent
FBti0019386	*Canton-S*	Polymorphic	Polymorphic	Polymorphic
FBti0019386	*W1*	Present	Present	Present
FBti0019415	*Canton-S*	Absent	Absent	Absent
FBti0019415[a]	***W1***	**Absent**	**Polymorphic**	**Absent**
FBti0019613[a]	***Canton-S***	**Polymorphic**	**Absent**	**Polymorphic**
FBti0019613	*W1*	Polymorphic	Polymorphic	Polymorphic
FBti0019624	*Canton-S*	Present	Present	Present
FBti0019624	*W1*	Absent	Absent	Absent
FBti0019985[a]	***Canton-S***	**Absent**	**Present**	**Absent/polymorphic**
FBti0019985	*W1*	Absent	Absent	Absent
FBti0020042	*Canton-S*	Absent	Absent	Absent
FBti0020042	*W1*	Absent	Absent	Absent
FBti0020089	*Canton-S*	Present	Present	Present
FBti0020089[b]	***W1***	**Absent**	**Polymorphic**	**Absent/polymorphic**
FBti0020091[a]	***Canton-S***	**Polymorphic**	**Present**	**Polymorphic**
FBti0020091	*W1*	Present	Present	Present
FBti0020125	*Canton-S*	Absent/polymorphic	Polymorphic	Absent/polymorphic
FBti0020125	*W1*	Absent/polymorphic	Absent	Absent/polymorphic
FBti0020190	*Canton-S*	Absent	Absent	Absent
FBti0020190	*W1*	Absent	Absent	Absent

Cases for which T-lex nd PCR results differed are highlighted in bold.
[a] Results do not match after manual curation.
[b] Results match after manual curation.
[c] Cases of misinference by T-lex (FBti0019223, Fbti0019360).

were called polymorphic. In the other three cases manual curation made the calls less definitive (e.g. an automatic call of 'absent' changed to 'absent/polymorphic' after manual curation). This gives 3.6% rate of clear false positives (2/52) and 5% rate of imprecise but not wrong calls (3/52). These error rates are similar to our previous estimates of around 5% rate of misclassification by the pooled-PCR approach (data not shown).

The minimum read coverage requirements

In order to assess the impact of the read coverage on the detection results, we ran T-lex on the same data set of TEs (661 TE copies) in both strains but using the NGS sequenced data subsampled to lower coverage: 10×, 15×, 20×, 30×, and 40× ('Materials and Methods' section). The results are shown in Figure 3. T-lex is designed to arrive at definitive calls only when the data are of sufficient quality to produce unambiguous results. Consequently lower coverage should generally lead to a higher rate of 'no data' results but should rarely lead to incorrect calls. Indeed, the number of cases where the T-lex call for a TE at low coverage changes when the coverage increases was very low in both strains (on average 3% of wrong calls at any level of coverage). However, as expected, lower coverage leads to a higher rate of 'no data' results with the maximum of 30% of 'no data' results returned at 10× coverage (Figure 3). The ability of T-lex to make calls increases monotonically with coverage with the biggest jump between 10× and 15× levels of coverage (30% to 19% of 'no data' results). Increasing coverage also helps to make more definitive calls, i.e. making 'present', 'absent' or 'polymorphic' calls compared to non-definitive calls such as 'absent/polymorphic' that are less specific. With 10× data we could only make 61% of such definitive calls compared to 92% in the 50× data. Note that this problem would be alleviated when either isogenic strains or single individual are used for the generation of NGS data where TEs are either fully absent, fully present, or at worst present at 50% frequency.

DISCUSSION

T-lex is a broadly applicable and flexible tool that can be used in any genome and for any number of TE copies provided the availability of the reference genome, individual TE copy annotation and NGS data. T-lex pipeline identifies NGS reads that indicate the presence and/or absence of individual TEs with high specificity and sensitivity given sufficient coverage of Solexa data (≥10×). The presence of a TE is indicated by reads that overlap the junction between a specific TE and its unique flanking sequences. Similarly, the absence of a TE is indicated by the presence of a read mapping to the junction between the two flanking sequences of a TE.

T-lex is specifically designed to detect TE insertions and performs well despite the presence of TSDs, which are generated by virtually all TEs. It can also handle the presence of simple repeats flanking TE junctions, which is especially important for non-Long Terminal Repeat (non-LTR) TEs that have long terminal poly-A tails. However, note that T-lex could also be used to detect other CNVs, such as segmental duplications, provided that they have been annotated as insertions with a defined start and end nucleotide positions.

T-lex can use both single-end and paired-end data, however it uses only single-end reads even in paired-end data sets. This is done by design. Our intention was to create a program that could be used with all NGS data and be as robust as possible. Even though paired-end reads do allow one to use alternative methods of mapping, such as looking for the presence of one read inside the TE copy and one in the flanking region at the right distance from the TE copy, these alternative approaches would be limited to paired-end data (and not all NGS data is paired-end) as well as would suffer from additional sources of error. For instance, the presence of insertions and deletions in the flanking regions of a TE might affect the distance between the mapped paired reads and this could lead to error by such algorithms.

Similarly we do not use any information about the depth of coverage to differentiate between repetitive and non-repetitive sequences. This is because the numbers of TEs are often large or do not vary among strains

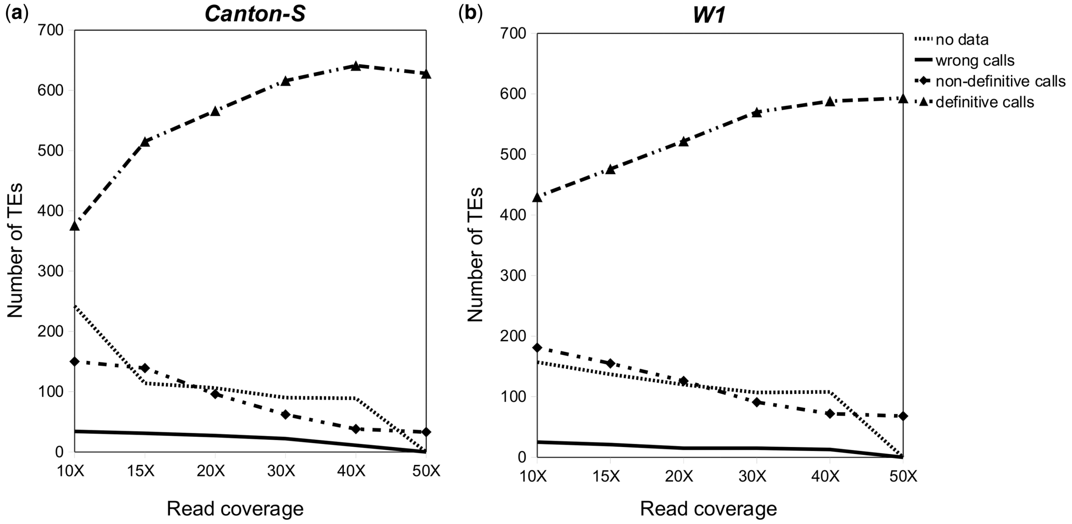

Figure 3. Impact of the read coverage on T-lex results. T-lex results for 661 TE insertions were obtained using NGS data subsampled to different coverages for (**a**) *Canton-S* and (**b**) *W1* strains. The number of 'no data', 'wrong calls' (i.e. non-compatible results compared to the 50× coverage NGS data), 'non-definitive calls' (i.e. 'absent/polymorphic' and 'present/polymorphic') and 'definitive calls' (i.e. 'absent', 'present' and 'polymorphic') results for each coverage are plotted.

significantly making such information not particularly informative or reliable. Nevertheless, T-lex provides the pileups mapping to the flanking regions in its output and this information can be used in manual curation. Specifically, having very few reads mapping to the flanking regions might indicate that mapping was erroneous. Having too many reads could indicate that the flanking regions are themselves repetitive. Indeed, we identified several such cases that upon further manual curation proved to be cases of segmental duplications (data not shown).

T-lex has been designed to be modular both in terms of the computational tools and in terms of the data used for mapping. As a result we provide the user with a high degree of flexibility. For instance, one of the longest steps in T-lex is the reformatting of data from fastq to bfq that can take from 12 min (10× data) to 1 h (50× data) with our data sets. However, this step can be skipped by providing T-lex with the reformatted data and adding the '-binreads' option ('Materials and Methods' section). Similarly, the preparation of the reference genome for the absence mapping (extraction of TEs and concatenating of the flanking sequences) can be done only once and then reused subsequently ('-binref' option). Note, that this also means that one can provide more precise ancestral sequences if one chooses to do so by using the ancestral sequence from the strains where the TE insertion was found to be absent. In the future, additional modules could be added to the T-lex pipeline. At present we are developing a 'T-*denovo*' module that will search for new TE copies from known TE families. We are also working on a module that will evaluate TE frequency using pooled data from multiple strains. The flexibility of the T-lex design should allow these and other extensions to be easily added.

Another exceptional feature of T-lex is that it allows for an easy manual curation of the presence/absence calls, which might be necessary when the quality of the TE annotations is low. As part of the output files, T-lex returns (i) for each TE side the alignments of the contigs from all the tested strains, (ii) the number of unique reads overlapping the TE junctions and (iii) the alignments of these reads to the reference genome. Inspection of these alignments might lead to redefining the TE sequence coordinates and as mentioned above T-lex can be run again using the corrected TE coordinates if needed (using the –binref option). The user can also modify the parameters used by T-lex both for the presence and the absence detection (for a detailed list of the parameters used by T-lex consult T-lex manual at http://petrov.stanford.edu/cgi-bin/Tlex_manual.html). For example, if the user suspects that the TE copy is longer than annotated, the user can ask T-lex to retrieve longer reference sequences both for the presence and the absence detection. Finally, note that the ability of SHRiMP to handle long gaps and polymorphisms should also reduce the negative effect of possible miss-annotations of TEs on the performance of the absence module.

We tested T-lex on a set of *D. melanogaster* TEs whose presence/absence detection was previously investigated using PCR in a large number of strains (Supplementary Table S1; 5,26,27). We tested T-lex performance with data of different coverage by subsampling our data from 50× down to 10×, 15×, 20×, 30× and 40×. We found that even at 10× coverage T-lex performs well in that it generates very few (on the order of 3%) erroneous results. The low coverage does lead to a fairly high rate of 'no data' calls, as expected, such that with the 10× coverage ~30% of TEs could not be mapped. This proportion went down to 19% when the coverage increased to 15× and decreases slowly with increasing coverage (Figure 3). In addition, the sequencing data used for validation in this study comes from not entirely isogenic (although fairly highly inbred) strains with the library made from multiple individuals. This means that polymorphic TE copies can be present at different frequency in the Solexa data complicating the detection process. As expected, the data of higher coverage allowed T-lex to ascertain a higher proportion of TEs as polymorphic, by finding evidence of both presence and absence. It appears that 15× coverage is roughly appropriate for detecting a TE copy as present and absent in isogenic strains, while higher coverage is required for libraries made from multiple outbred individuals. Note that running T-lex on multiple strains sequenced to even low coverage should generate reasonable estimates because even though a polymorphic TE copy in any one strain might be misinterpreted as either present or absent (or not mapped at all) the overall estimated frequency of this TE insertion in multiple strains would be affected much less. Indeed, we find that increasing the coverage shifts the presence and absence calls to polymorphic calls with roughly equal probabilities suggesting that the misinference is random and thus would not systematically bias the estimate of overall frequencies of TE copies in a sample.

Although tested with Drosophila data, T-lex can be used for any species for which the TE copy annotations and NGS data are available. Currently, there are individual TE copy annotations available for other organisms: *Arabidopsis Thaliana* (30), *Oryza sativa* (31), soybean *Glycine max* (32) the plant-parasitic nematode *Meloidogyne incognita* (33) and the pea aphid *Acyrthosiphon pisum* (The International Aphid Genomics Consortium) and several more are in progress (http://urgi.versailles.inra.fr/). Moreover, with the decline of sequencing costs, we may expect in the very near future an exponential increase of the amount of sequencing data for many model and even non-model organisms. Large-scale, population-level NGS projects are moving forward for numerous organisms such as vertebrates (34), *Drosophila* (DGRP http://service004.hpc.ncsu.edu/mackay/Good_Mackay_site/DBRP.html and DPGP http://www.dpgp.org/) and Arabidopsis (35). T-lex should therefore allow investigating particular TE insertions that might be of interest to the user as well as to perform genome-wide population dynamics analyses of TEs. Given the abundance, ubiquity and the role of TEs in generating chromosomal rearrangements, regulating gene expression and altering gene function, a thorough understanding of the population dynamics of TEs is essential for the understanding of the eukaryotic genome evolution and function.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Michael Eisen for the Illumina data for the two strains analyzed in this article and Ruth Hershberg, Scott W. Roy and Diamantis Sellis for their comments on the article. The authors also thank all members of the Petrov Lab for the discussions that significantly improved T-lex and this article.

FUNDING

Funding for open access charge: National Institutes of Health (grant number GM089926 to D.A.P.).

Conflict of interest statement. None declared.

REFERENCES

1. Craig,N.L., Craigie,R., Gellert,M. and Lambowitz,A.M. (eds), (2002) *Mobile DNA II*. ASM Press, Washington, DC.
2. Kidwell,M.G. and Lisch,D.R. (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution*, **55**, 1–24.
3. Biemont,C. and Vieira,C. (2006) Genetics: junk DNA as an evolutionary force. *Nature*, **443**, 521–524.
4. Gonzalez,J., Macpherson,J.M. and Petrov,D.A. (2009) A recent adaptive transposable element insertion near highly conserved developmental loci in Drosophila melanogaster. *Mol. Biol. Evol.*, **26**, 1949–1961.
5. Gonzalez,J., Karasov,T.L., Messer,P.W. and Petrov,D.A. (2010) Genome-wide patterns of adaptation to temperate environments associated with transposable elements in Drosophila. *PLoS Genet.*, **6**, e1000905.
6. Agrawal,A., Eastman,Q.M. and Schatz,D.G. (1998) Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature*, **394**, 744–751.
7. Levis,R.W., Ganesan,R., Houtchens,K., Tolar,L.A. and Sheen,F.M. (1993) Transposons in place of telomeric repeats at a Drosophila telomere. *Cell*, **75**, 1083–1093.
8. Wang,J., Keightley,P.D. and Halligan,D.L. (2007) Effect of divergence time and recombination rate on molecular evolution of Drosophila INE-1 transposable elements and other candidates for neutrally evolving sites. *J. Mol. Evol.*, **65**, 627–639.
9. Cordaux,R., Udit,S., Batzer,M.A. and Feschotte,C. (2006) Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl Acad. Sci. USA*, **103**, 8101–8106.
10. Ackerman,H., Udalova,I., Hull,J. and Kwiatkowski,D. (2002) Evolution of a polymorphic regulatory element in interferon-gamma through transposition and mutation. *Mol. Biol. Evol.*, **19**, 884–890.
11. Naito,K., Zhang,F., Tsukiyama,T., Saito,H., Hancock,C.N., Richardson,A.O., Okumoto,Y., Tanisaka,T. and Wessler,S.R. (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature*, **461**, 1130–1134.
12. Slotkin,R.K. and Martienssen,R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.*, **8**, 272–285.
13. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
14. Kaminker,J.S., Bergman,C.M., Kronmiller,B., Carlson,J., Svirskas,R., Patel,S., Frise,E., Wheeler,D.A., Lewis,S.E., Rubin,G.M. *et al.* (2002) The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol.*, **3**, RESEARCH0084.
15. Wicker,T., Robertson,J.S., Schulze,S.R., Feltus,F.A., Magrini,V., Morrison,J.A., Mardis,E.R., Wilson,R.K., Peterson,D.G., Paterson,A.H. *et al.* (2005) The repetitive landscape of the chicken genome. *Genome Res.*, **15**, 126–136.
16. Kordis,D. (2009) Transposable elements in reptilian and avian (sauropsida) genomes. *Cytogenet. Genome Res.*, **127**, 94–111.
17. Yang,G., Zhang,F., Hancock,C.N. and Wessler,S.R. (2007) Transposition of the rice miniature inverted repeat transposable element mPing in Arabidopsis thaliana. *Proc. Natl Acad. Sci. USA*, **104**, 10962–10967.
18. Charlesworth,B., Jarne,P. and Assimacopoulos,S. (1994) The distribution of transposable elements within and between chromosomes in a population of Drosophila melanogaster. III. Element abundances in heterochromatin. *Genet. Res.*, **64**, 183–197.
19. Naito,K., Cho,E., Yang,G., Campbell,M.A., Yano,K., Okumoto,Y., Tanisaka,T. and Wessler,S.R. (2006) Dramatic amplification of a rice transposable element during recent domestication. *Proc. Natl Acad. Sci. USA*, **103**, 17620–17625.
20. Lockton,S. and Gaut,B.S. (2010) The evolution of transposable elements in natural populations of self-fertilizing Arabidopsis thaliana and its outcrossing relative Arabidopsis lyrata. *BMC Evol. Biol.*, **10**, 10.
21. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of Drosophila melanogaster. *Science*, **287**, 2185–2195.
22. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–467.
23. Li,H., Ruan,J. and Durbin,R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.
24. Wicker,T., Sabot,F., Hua-Van,A., Bennetzen,J.L., Capy,P., Chalhoub,B., Flavell,A., Leroy,P., Morgante,M., Panaud,O. *et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, **8**, 973–982.
25. Rumble,S.M., Lacroute,P., Dalca,A.V., Fiume,M., Sidow,A. and Brudno,M. (2009) SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.*, **5**, e1000386.
26. Petrov,D.A., Fiston-Lavier,A.S., Lipatov,M., Lenkov,K. and González,J. (2010) Population genomics of transposable elements in Drosophila melanogaster. *Mol. Biol. Evol.*, doi:10.1093/molbev/msq337.
27. Gonzalez,J., Lenkov,K., Lipatov,M., Macpherson,J.M. and Petrov,D.A. (2008) High rate of recent transposable element-induced adaptation in Drosophila melanogaster. *PLoS Biol.*, **6**, e251.
28. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
29. Lexa,M., Horak,J. and Brzobohaty,B. (2001) Virtual PCR. *Bioinformatics*, **17**, 192–193.
30. Buisine,N., Quesneville,H. and Colot,V. (2008) Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics*, **91**, 467–475.
31. Juretic,N., Bureau,T.E. and Bruskiewich,R.M. (2004) Transposable element annotation of the rice genome. *Bioinformatics*, **20**, 155–160.
32. Du,J., Grant,D., Tian,Z., Nelson,R.T., Zhu,L., Shoemaker,R.C. and Ma,J. (2010) SoyTEdb: a comprehensive database of transposable elements in the soybean genome. *BMC Genomics*, **11**, 113.
33. Abad,P., Gouzy,J., Aury,J.M., Castagnone-Sereno,P., Danchin,E.G., Deleury,E., Perfus-Barbeoch,L., Anthouard,V., Artiguenave,F., Blok,V.C. *et al.* (2008) Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. *Nat. Biotechnol.*, **26**, 909–915.
34. Genome 10K Community of Scientists: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.*, **100**, 659–674.
35. Weigel,D. and Mott,R. (2009) The 1001 genomes project for Arabidopsis thaliana. *Genome Biol.*, **10**, 107.