universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

"Development of a novel tool to uncover mobile genetic element diversity and trace the invasion of DNA transposons"

verfasst von / submitted by

Lukas Weilguny, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2019 / Vienna 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 220
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Joint-Masterstudium Evolutionary Systems Biology
Betreut von / Supervisor:	Priv.-Doz. Dipl.-Ing. Dr. Robert Kofler

Contents

Abstract

Transposable elements (TEs) are selfish DNA sequences that multiply within host genomes. They are present in most species investigated so far at varying degrees of abundance and sequence diversity. The TE composition may not only vary between but also within species and could have important biological implications. Variation in prevalence among populations may for example indicate a recent TE invasion, whereas sequence variation could indicate the presence of hyperactive or inactive forms. Gaining unbiased estimates of TE composition is thus vital for understanding the evolutionary dynamics of transposons.

To this end we developed DeviaTE, a tool to analyze and visualize TE abundance using Illumina or Sanger reads. Our program only requires sequencing reads and consensus sequences of TEs. Thus, it works in an assembly-free manner, increasing its applicability to non-model organisms for which a high-quality assembly is not available yet. It generates a table and a visual representation of TE composition and provides unbiased estimates of TE abundance. Using published data we demonstrate that DeviaTE can be used to study TE composition within samples, identify clinal variation in TEs or compare TE diversity among species. We also present careful validation with simulated data.

Moreover, we describe a model of DNA transposon invasions and an approach to reconstruct the history of such invasions using our novel tool. We propose that an invasion leaves unique fingerprints within populations, which consist of non-autonomous, internally deleted variants of TEs. Using these TE remnants, we show that the sequence of the P-element invasion in North American and European *Drosophila melanogaster* populations can be retraced. In particular, we find that patterns of internally deleted variants recover the geographic distribution of sampled populations. Additionally, we identify potential origins and routes of the invasion on both continents. With the development of DeviaTE we hope to catalyze future progress in our understanding of TE invasion dynamics and other diverse phenomena, in which TEs play a central role.

1 Introduction

1.1 Transposable elements - potent producers of variation

Transposable elements (TEs) are stretches of DNA that copy themselves within host genomes. They have been found in almost all eukaryotes, and in most bacteria and archaea investigated so far (Biémont and Vieira 2006; Brügger et al. 2002; Wicker et al. 2007). TEs are important mutagens, which generate novel phenotypic variation, e.g. in *Drosophila melanogaster* an estimated 50 - 80 % of observed mutations are due to TEs (Biémont and Vieira 2006). With transposition rates ranging from 10^{-3} to 10^{-5} per element per generation, TEs are highly potent producers of variation compared to the classical nucleotide substitution rate of 10^{-9} per base per generation (Biémont and Vieira 2006).

Casacuberta and González (2013) present a variety of mechanisms by which TEs can cause mutations. The main factor determining the outcome of a mutation is the exact position of the novel TE insertion, which can either lead to the generation of new regulatory elements or disrupt existing ones (Casacuberta and González 2013). A TE can also transpose near exons or introns of genes causing frameshifts, exonization of TE sequences or gene duplication through retrotransposon-mediated sequence transduction (Casacuberta and González 2013; Xing et al. 2006). Moreover, there is evidence of environmental adaptation through biological innovation as a consequence of novel TE insertions. For instance, a high proportion of identified recent TE insertions in *D. melanogaster* are specifically linked to factors influencing adaptation to environmental parameters such as temperature or rainfall (González et al. 2010). Other examples for TE-associated adaptation include resistance to viral infections, insecticides and pesticides (Casacuberta and González 2013; González et al. 2010). It might thus be concluded, that some active TEs can be beneficial for a species in the process of adapting to a new environment despite the possibility of short-term detrimental effects on individuals. This trade-off can be observed for mutations causing various diseases; particularly over 75 human diseases such as different forms of cancer, haemophilia B or Duchenne muscle dystrophy (Burns 2017; Kazazian Jr et al. 1988; Narita et al. 1993). Altogether, deregulation caused by TE insertions in humans is thought to be responsible for 0.5 to 1 % of all illnesses (Biémont and Vieira 2006; McCullers and Steiniger 2017).

To protect against uncontrolled mobilization of TEs, a defense system has been established in the majority of animals. This includes the small non-coding

piRNAs (piwi-interacting RNAs) and their effector proteins of the PIWI clade (Brennecke et al. 2007; Gunawardane et al. 2007). piRNAs bind to complementary TE sequences, upon which their associated PIWI clade proteins silence the TE (Brennecke et al. 2007; Le Thomas et al. 2014). In a genome previously devoid of a TE, the invading element can spread until one of its copies inserts into a piRNA cluster. Such novel cluster insertions can then be used by the host defense system to produce piRNAs for the repression of that TE (Bergman et al. 2006; Goriaux et al. 2014; Malone et al. 2009; Zanni et al. 2013). Although piRNAs and other defense mechanisms against selfish elements have emerged (Brennecke et al. 2007; Yang et al. 2017), TEs have proven to be highly successful invaders. Hence, most genomes contain large fractions of TEs. In maize, for example, TEs account for a striking 85 % of the genome (Schnable et al. 2009). The proportion in *D. melanogaster* is thought to be approximately 20 % (Barrón et al. 2014; Kaminker et al. 2002), with at least 30 % of TEs persisting as full-length and active copies (McCullers and Steiniger 2017).

Overall, transposons have been implicated in diverse phenomena such as human disease (Burns 2017; Kazazian Jr et al. 1988; Narita et al. 1993), environmental adaptation (Casacuberta and González 2013; Schrader and Schmitz 2018), genome evolution (Kazazian Jr 2004), variation of quantitative traits (Mackay et al. 1992), domestication of important crops (Studer et al. 2011) and the generation of new species (Oliver et al. 2013; Werren 2011). Understanding TE biology is thus of vital interest to many different research fields.

1.2 Classification and quantitative variation of TEs

Although some TEs are ancient and present in all kingdoms, generally the diversity among transposons is high (Bargues and Lerat 2017; Du et al. 2010; Sotero-Caio et al. 2017). Therefore, consistent and efficient nomenclature is essential for comparative and evolutionary studies. In an effort to unify the classification and communication of TEs, Wicker et al. (2007) proposed a system mainly based on transposition mechanism, structural elements such as terminal repeats, coding regions and other diagnostic non-coding features, as well as sequence similarity. The top-level of the TE taxonomy distinguishes two classes according to their intermediate state during transposition: RNA transposons (also called 'copy-and-paste') and DNA transposons ('cut-and-paste') (Finnegan 1989; Wicker et al. 2007). Further subdivision into subclasses and orders reflects the number of strands cut in the TE donor site upon transposition and major differences in

transposition mechanism, coding capability and enzymology (Wicker et al. 2007). A lower level of the classification commonly used to describe TEs are families. They are defined by sequence conservation in coding regions following the rule of 80 % similarity in at least 80 % of aligned sequences and a minimum length of 80 bp (Wicker et al. 2007).

Depending on the TE family and the host species, copy numbers can range from a few to hundreds of thousands of insertions (Biémont and Vieira 2006; Pritham and Feschotte 2007). Thus, TE composition varies substantially among and within species (Bargues and Lerat 2017; Bergman et al. 2017b), which can have important biological consequences. Variation in TE abundance among populations may be the hallmark of a recent TE invasion (Anxolabéhère et al. 1988; Kofler et al. 2015) and may even drive speciation (Oliver et al. 2013; Serrato-Capuchina and Matute 2018). Furthermore, some TEs exist as internally deleted variants, which are thereby either inactivated or act as suppressors of the full-length TE (Black et al. 1987). The abundance of such internal deletions may vary among populations, so the strength of TE repression can also differ (Bergman et al. 2017b; Bonnivard and Higuet 1999). Variations of the TE sequence can highlight activity differences among samples, as base substitutions and indels within TEs can lead to elevated or reduced transposition rates (Beall et al. 2002). Finally, terminally deleted variants are likely immobilized (Marin et al. 2000), so variation in the prevalence of such terminal deletions may allow for identification of inactive copies.

1.3 Analyzing and visualizing TE composition - aim 1

Despite this importance of TE variation, few tools exist that allow for the quantification of TE composition within and between species. Some tools for the analysis and visualization of TEs have been published, but they require a reference assembly and notably do not resolve diversity at a nucleotide level (Tempel and Talla 2015; You et al. 2013). However, in eukaryotes for example, a high quality assembly is only available for a mere 25 species so far, of which most are fungi (Lewin et al. 2018). Additionally, even if a reference assembly is available, resulting estimates of TE diversity may be biased because repetitive structures pose a significant challenge to assembly algorithms (Sohn and Nam 2018), such that the variation and abundance of TEs will not be captured well in the resulting contigs.

We therefore aimed to develop a tool to analyze TE diversity, which circumvents the need for a reference assembly. In particular, we reasoned that aligning

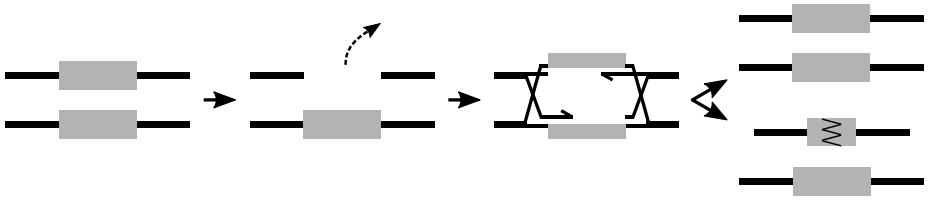

Figure 1: A model for the gap repair following excision of a DNA transposable element (TE). The TE of the top chromatid transposes, upon which the second chromatid is used to repair the resulting double-strand break. A successful repair results in a copy of the original TE. An interruption of the process, however, generates an internally deleted variant of the TE. Modified from Engels et al. (1990).

sequencing reads directly to consensus sequences of TEs will allow for accurate estimates of TE composition. We implemented this approach in our novel program DeviaTE, a tool for an assembly-free analysis of TE diversity. DeviaTE may be used to visualize and quantify TE abundance, single nucleotide poylmorphisms, indels, along with both internal and terminal deletions for multiple TE families and samples. It solely requires consensus sequences of TEs and sequencing reads (Sanger or Illumina) of at least one sample. DeviaTE may be used to study the TE composition of samples, assess TE divergence among species, investigate their activity and autonomy, monitor the progression of TE invasions and study clinal variation of TEs.

1.4 Reconstructing the invasion history of DNA TEs - aim 2

We devised an approach to reconstruct the invasion history of DNA transposons in sampled populations by exploiting traces left behind during the invasion. These traces comprise deletion variants produced by interrupted gap-repair during the transposition of DNA transposons (Engels et al. 1990). Their cut-and-paste transposition mechanism does not lead to an increase in copy number. Rather, a second copy is produced by sister-chromatid mediated gap repair following the excision of the TE (Fig. 1, Engels et al. 1990). Yet, if the gap repair process is interrupted, the newly transcribed copy of the TE can be internally deleted. Such internally deleted variants can still be mobilized but only with the help of a full-length element, which provides the proteins necessary for transposition (Hua-Van et al. 2011; Robillard et al. 2016). Thus, internally deleted copies are commonly described as non-autonomous TEs.

Using DeviaTE's ability to detect internally deleted TE variants and estimate their frequency, we propose that this enables the reconstruction of the invasion

history of a TE family. We hypothesize that using deletion variants as markers allows us to derive the sequence in which populations have been invaded by the TE. This is due to the following properties of internal deletions, which render them suitable markers.

Firstly, a newly invaded population has almost exclusively full length copies. But subsequently, internally deleted insertions are produced by chance during transposition until the TE is silenced by piRNAs of the host defense system. This implies that full-length insertions gradually decrease with each invaded population (Bonnivard and Higuet 1999). After the emergence of repressive piRNAs, no novel deletion variants will be created and their pattern remains a stable fingerprint after the invasion has occurred. Further, internal deletions emerge at a high rate only few generations after an invasion has started (Daniels et al. 1985; Kofler et al. 2018, see also Fig. 8), which is a distinguishing factor to choose internal deletions as markers over other polymorphisms, since TE invasions happen at very short timescales. Additionally, interruption of the gap-repair after transposition is a stochastic event and thus will lead to deletions with arbitrary breakpoints (O'Hare and Rubin 1983). Hence, the deletion variants accumulating in a population are assumed to form a unique, defining fingerprint for that population.

These observations lead us to propose the following model of TE invasions: a transposon spreads in a newly invaded population and internally deleted variants emerge (Fig. 2A). After some time, the invasion is stopped by piRNAs, TE activity ceases and no novel deletion variants will emerge. However, migrating individuals can carry a sample of the TE diversity of the source population to a target population. The introduced TEs will trigger an invasion of the target population with some full-length and some internally deleted insertions from the source population. Similarly, this will lead to the creation of novel deletion variants not found in the source population and to a decrease of full-length elements. The invasion of the target population might again be stopped by piRNAs and other target populations can be invaded by further migration, comparable to a stepping stone model (Kimura and Weiss 1964). As a result, populations will carry unique fingerprints of internally deleted variants that remain stable within that population after the invasion has passed. But the similarity between such patterns contains information on sequential migratory events that lead to the invasion of the TE. Using these signatures, it may thus be possible to infer the sequence and geographic spread of an invasion.

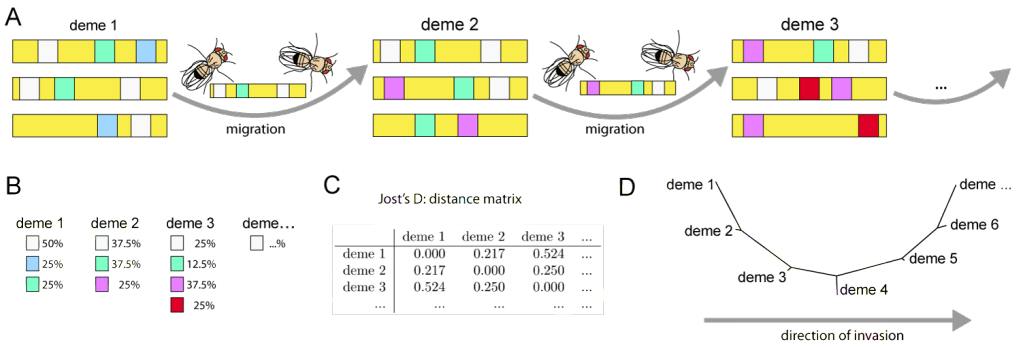

Figure 2: Model for the invasion of a DNA TE and our approach to trace its history using internally deleted TE variants. A) A source population, deme 1, contains a high proportion of full-length elements (white boxes) in their pool of chromosomes and some internally deleted TEs (colored boxes). Each colored box represents a unique deletion variant. Migration from the source population brings a sample of full-length and deleted elements to a recipient population, deme 2. New activity of the TE in the recipient pool ensures the generation of novel deletion variants and a loss of full-length TEs until the TE is repressed. This sequential process creates a unique and stable pattern of internally deleted variants in each population. B) The frequencies of TE variants for each population. Full-length elements decrease in frequency and new variants shape unique deletion patterns. C) Frequencies are used to calculate pair-wise dissimilarities with Jost's D. (see Methods section). D) A dendrogram is constructed from the distance matrix, which allows for the reconstruction of the TE invasion.

1.5 The global spread of the P-element

To test the approach we analyze the global invasion of the P-element in *D. melanogaster*. The DNA transposon P-element is one of the best studied TEs. It has been shown that natural invasions in different Drosophila species have occurred multiple times in a process termed horizontal transfer (HT) (Haring et al. 2000; Kidwell 1983; Kofler et al. 2015; Serrato-Capuchina et al. 2018; Yoshitake et al. 2018). Such horizontal transfer events, in combination with the wide dispersal of Drosophila and consequential admixture of populations, might be crucial factors in avoiding accumulation of deleterious variants and inactivation of DNA transposons (Bartolomé et al. 2009; Schaack et al. 2010). Thus, HT and admixture might be cornerstones of the successful world-wide spread of the P-element. However, repressive regulatory RNAs are generated rapidly and adaptively during the invasion of the P-element (Kelleher 2016; Kelleher et al. 2018). But if migration and ultimately admixture between two populations or species takes place, with one of them carrying the active TE, it might successfully radiate before the piRNA system can stop it. As such, the world-wide spread of the P-element in *D.*

melanogaster is a prominent example (Anxolabéhère et al. 1988; Anxolabéhère et al. 1985), which has been used to unveil details of TE invasion dynamics and the emergence of repressive factors (Kelleher 2016; Quesneville and Anxolabéhère 1997).

The global invasion was hypothesized after lab strains collected in the early 20^{th} century were found to be devoid of the P-element. In contrast, all natural populations sampled after 1980 carried insertions of the TE, with its earliest appearance in the Americas and later observations in Europe and the USSR (Anxolabéhère et al. 1988). This led to the hypothesis of a global P-element invasion in *D. melanogaster*, which had its origin in South America following HT from *D. willistoni* (Daniels et al. 1990). After that, it purportedly proceeded across American populations and later to Europe, Asia and the rest of the world in a period of approximately 60 years (Anxolabéhère et al. 1988; Anxolabéhère et al. 1985).

2 Methods and Implementation

2.1 Implementation of DeviaTE and code availability

DeviaTE is implemented in Python (version 3.6+, Python Software Foundation 2017) and distributed under the GNU GPLv3 License. It is hosted on the Python Package Index (PyPI) and can therefore be installed with the widely-used pip package manager. Additionally, a conda container-type environment is available on the anaconda cloud, which includes all package dependencies. The conda environment ensures an easy and robust installation and makes our tool portable to any Unix-like system. DeviaTE makes use of the Python packages pandas (McKinney 2010, p. v 0.23.4), and pysam (Heger and Jacob 2018, p. v 0.15). For visualization, it uses R and the ggplot2 package (R Core Team 2014; Wickham 2016).

Our program is open-source and hosted on GitHub at: [https://github.com/W-L/deviaTE]. The GitHub repository contains the source code and installation instructions as well as a manual describing the individual scripts, the required arguments and output in more detail. Additionally, a walkthrough with examples using publicly available data and basic use cases is available.

2.2 Demonstrating the features and applicability of DeviaTE

To show some scenarios for the applicability of DeviaTE, we used publicly available data. The TE consensus sequences for the *Drosophila* genus were obtained from a manually curated database, which includes most known families of transposable elements of that species (Bergman et al. 2017a). We used version 9.44 of the database, containing consensus sequences and feature annotations for 179 TE families, most of which from *D. melanogaster*. These 179 families may be further classified into 69 retroviral elements, 58 non-LTR retrotransposons, 41 IR-elements, 4 Foldback elements, 3 MITEs, 2 SINE-like element as well as 1 Helitron and 1 unclassified element. The database was originally compiled by individual

researches, later maintained by the Drosophila Genome Project, and now curated on Github by a team of researches. For our purposes, the sequence collection was parsed from EMBL-format to FASTA.

To demonstrate the basic features of our tool, we separately analyzed TEs in four different datasets. One of which was the TE Burdock in a sample of the Global Diversity Lines, collected in the Netherlands (Grenier et al. 2015). DeviaTE was used with default settings and no normalization was performed. Further, we reexamined the P-element invasion in *D. simulans* from Kofler et al. (2018). In this case, DeviaTE was executed with normalization by million mapped reads and insertion numbers were estimated with the single-copy gene *rpl32*. As another example, we investigated a possible clinal variation of TE composition in a dataset from Bergland et al. (2014). This time, normalization and estimation of TE insertion numbers was performed with the single-copy genes *rpl32, piwi* and *act5C*. Statistical analyses involved a Pearson's product moment correlation for latitude versus estimated number of insertions. For a comparative analysis of TE divergence across three *Drosophila* species, we used high quality Sanger-sequencing data from the Drosophila Genomes Consortium; *D. melanogaster*, *D. sechellia* and *D. simulans* (Drosophila 12 Genomes Consortium et al. 2007). No normalization or copy number estimation were performed. Statistical significance was generally considered at a threshold of $\alpha = 0.05$

2.3 Simulating TE landscapes for validation

To validate our tool and explore its limitations, we simulated TE landscapes with known levels of sequence and structural polymorphism using SimulaTE (Kofler 2018). We simulated sequencing reads after the insertion of artificial TEs in a 100 kb sequence from chromosome 2R of *D. melanogaster*, which is devoid of repetitive sequences. We chose the TE DOC5 because of its sequence length of 4682 bp, which is close to the average of all TEs in *D. melanogaster*. The inserted TEs were simulated with varying levels of polymorphism and with varying read lengths.

To test the accuracy of reported divergence, we simulated reads of a TE insertion deviating in nucleotide composition from 0 to 30 %. These simulations were conducted for short reads of 100 bp and long reads of 1000 bp. Five replicates were used for each level of sequence divergence. Analogous experiments were run for simulating indels. For the validation of allele frequency recovery we simulated populations of 20 genomes to achieve different frequencies of nucleotides at a

biallelic locus at the arbitrarily chosen position 2000. These experiments were repeated for four different levels of background divergence, i.e. 5, 10, 15 and 20 % of the inserted TE in comparison to the consensus sequence, and replicated five times. We chose 20 % as the highest value of divergence because membership of TE families is, among other criteria, commonly defined by sequence homology of at least 80 %. Similarly, TE landscapes containing internally deleted variants were simulated from a population of 20 genomes. Additionally, the size of the deletion was varied (100, 500, 1000, and 2000) to test, whether the size of deletions affects the accuracy of reporting their abundance.

DeviaTE corrects frequencies of internal deletions by raising observed values to the power of a correction factor. The model for this correction factor was derived by minimizing the sum of squares of the residuals in a linear regression of corrected, observed versus simulated frequencies in an optimization procedure. Six different read lengths were used in the linear regression (80, 100, 175, 250, 500, 1000 bp) to cover the range of read lengths commonly achieved by different sequencing platforms. A third-degree polynomial was fitted to the resulting data points to model the relationship between read length and correction factor. This allows for the interpolation of the correction factor for any read length.

All statistical analyses on accuracy were conducted with linear regression of simulated versus observed values in R (R Core Team 2014). Plotted values are means with error bars denoting ±standard error of the mean (Fig. 11).

2.4 Approaches to analyze the P-element invasion history

The datasets for tracing the invasion of the P-element are publicly available and described in Bergland et al. (2014, Accession: SRP044883) and Kapun et al. (2018, Accession: SRP108658). They contain *D. melanogaster* samples, sequenced as populations, covering a broad geographic range along the North American East Coast and Europe, respectively. To avoid oversampling in one geographic location we restricted the samples from Pennsylvania and Ukraine to two randomly chosen ones. An artificial, pseudo-ancestral sample containing only full-length insertions of the P-element was added for selected analyses.

Multiple statistical methods were employed to analyze the P-element in the sampled populations. Firstly, frequencies of internally deleted TE variants were collected from all populations of the analyzed datasets with DeviaTE. A threshold of >1 % frequency was applied to retain variants and deletions with nearly identical breakpoints (±3 bp) were merged. Deletion variant frequencies were

scaled and subsequently used for principal component analysis. In all cases, principal components were selected for further analysis with a cut-off of covering >75 % of total variation. Principal components and geographic coordinates were transformed to pair-wise distance matrices. Multivariate spatial autocorrelation was tested with a Mantel test, implemented in the R package cultevo (Stadler 2018). This test comprises a Monte Carlo procedure including 10000 random permutations of input data to find the distribution of the test statistic under H_0.

Furthermore, deletion frequencies were transformed to pair-wise genetic distances using Jost's D (Jost 2008). Heterozygosities, as part of the calculation of D, were obtained by treating TEs as loci and deletion variants, as well as the full-length TE, as their alleles with respective frequencies. We chose D as our method of differentiation between subpopulations because other more commonly used measures, such as F_{ST} or G_{ST}, converge to zero when gene diversity is very high. In contrast, Jost's D is robust at high levels of diversity and leads to a mathematically more consistent measure for population dissimilarity (Jost 2008).

Dendrograms to infer relationships between samples and the direction of the invasion were calculated from matrices of pair-wise D distances. Statistical significance of a non-random geographic distribution of genetic distances was assessed with a Mantel test, with an MC procedure analogous to above. For unrooted trees, we employed the neighbor-joining algorithm, implemented in bionj (Gascuel 1997). For rooting, we used UPGMA with an artificial, pseudo-ancestral sample containing only full-length TEs as an outgroup.

Full-length frequencies and TE copy numbers in all samples were centered and subsequently tested for univariate spatial autocorrelation using Moran's I, in combination with a weight matrix of inverse geographic distances. Statistical testing involved 10000 Monte Carlo permutations and was performed with the R package ade4 (Chessel et al. 2004). Differences in copy numbers and frequencies between continents were tested with Welch's two sample t-test.

3 Results

3.1 Architecture and usage of DeviaTE

DeviaTE is a novel tool for the analysis and visualization of the abundance as well as the genetic diversity of TE families. As input, our tool requires consensus sequences of TE families and sequencing reads (Sanger or Illumina) from at least one sample, which can be from individuals, pooled populations or tissues. DeviaTE provides quantitative estimates as well as a visual overview of TE diversity, which includes the coverage of ambiguously as well as unambiguously mapped reads, fixed and segregating polymorphisms (SNPs and indels) and internal and terminal deletions. Furthermore, the abundance of TEs is estimated if the sequence of at least one single copy gene is included in the analysis. Although DeviaTE was mainly designed for TEs, we note that it may also be used to analyze the composition of other genomic elements such as genes, gene families, viruses, bacteria and mtDNA.

An analysis of TE composition with DeviaTE proceeds in three steps (Fig. 3, left). For ease of use, a single-command wrapper script called `deviaTE` is also available, which executes all steps automatically (Fig. 3, right). In a first step, reads are trimmed using a modified Mott algorithm, which assigns a score to every sub-string of a read. Each segment has a score equal to the sum of quality scores of all bases in the sub-string. Then, only the highest scoring segment is retained. In other words, the ends of a read are trimmed until only the sub-string with the highest score is left, given some quality threshold. The quality scores can be encoded either in Sanger or Illumina format. Further, it checks whether the resulting sub-string is longer than a specified minimum length. After trimming, reads are aligned to a library of TE consensus sequences (FASTA format) using bwa-sw (Li and Durbin 2010). This is achieved by the executable `deviaTE_prep`. The input to this script are a file containing sequencing reads in FASTQ format and the TE consensus sequences. The output is a sorted and indexed alignment of the sequencing reads against the consensus sequences. To obtain estimates of TE abundance, the sequence of one or more single copy genes may be added to the library of TE consensus sequences.

Next, DeviaTE performs a pileup on the alignment file and generates a table containing the abundance and diversity of TEs (coverage, SNPs, indels, internal and terminal deletions). The second executable called `deviaTE_analyse` is used for this step, and comprises the main task of the workflow. It uses the previously

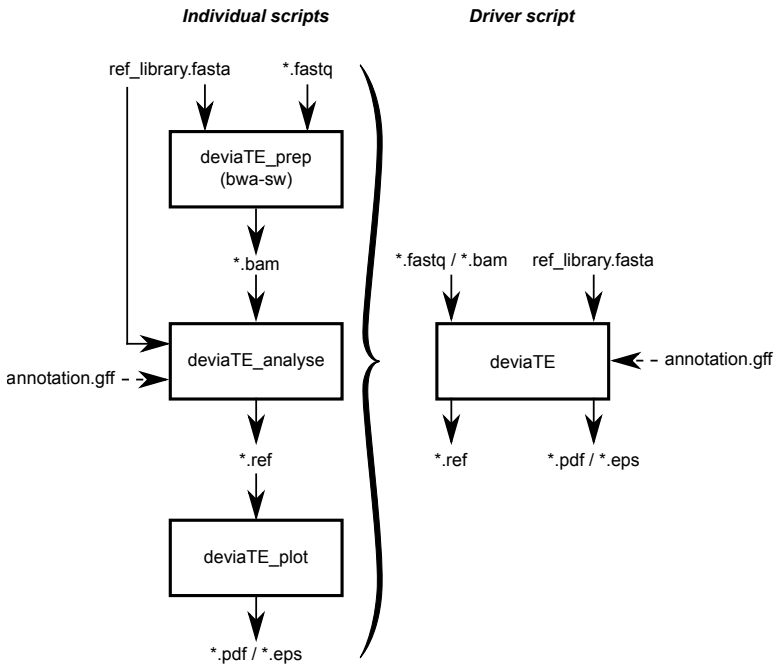

Figure 3: Architecture of DeviaTE. Mandatory input files are shown as solid lines and optional input as dashed lines. Boxes represent executable scripts. DeviaTE consists of three main scripts, to map the sequencing reads (deviaTE_prep), analyze a specific transposon family (deviaTE_analyse), and visualize the results (deviaTE_plot). This design allows for the analysis of several TE families with a single mapping step. A user-friendly driver script is provided, which enables an analysis of multiple input files and transposon families. The input consists of sequencing reads (*.fastq), a library of reference sequences (ref_library.fasta) and an optional annotation of the reference sequences. As output, DeviaTE provides a table containing information on the analyzed TE families (*.ref) and a visualization (*.pdf/*.eps).

prepared alignment file and produces quantitative information at nucleotide resolution for the chosen TE families. Such output tables start with a header section denoted by the pound/hash symbol. The first line contains a timestamp and the command used to generate the file, while the following line states the estimated number of TE insertions, if single gene normalization was selected by the user. The final line of the header section contains the column names of the output table. Each subsequent row corresponds to one position of the TE sequence, containing information about nucleotide counts, presence of polymorphisms and coverage, separated into ambiguous, unambiguous as well as physical coverage.

As the final step in the DeviaTE workflow, the diversity of TEs is visualized with an illustration inspired by Sashimi plots, which are commonly used for quantitative visualization of splicing in RNA-seq data (Katz et al. 2015). In our visualization, internal deletions are shown instead of splicing events. The third script in the pipeline, `deviaTE_plot`, is used for the visualization process, which takes the table of quantitative information from the previous step as input and creates a plot. These illustrations show the coverage of ambiguously and unambiguously mapped reads, the frequency of SNPs, indels, internal deletions and terminal deletions (e.g. Fig. 7). Additionally, a panel with features of the TE is added at the bottom if an annotation of the TE sequences was provided during analysis. The plots can either be created in PDF or EPS format, which enables simple vector graphics processing. DeviaTE can plot an arbitrary number of TE families from one or more samples and automatically align plots by TE (column) and sample (row) (e.g. Fig. 10). In this way, multiple TE families can be visualized by concatenating their output tables and using the merged file as input to the plotting script.

DeviaTE can either be executed by running each step individually or by running the single-command script `deviaTE`, which executes all steps automatically. It can work with raw sequencing reads in FASTQ format or with reads already aligned to TE reference sequences in BAM format. Running DeviaTE on already aligned sequences avoids the time-consuming steps of mapping and detecting internal deletions, such as if a TE family, which was not part of an initial run, should be analyzed. Our tool can also be applied to multiple files at once using simple command-line options. Furthermore, DeviaTE can analyze multiple TE families by providing a comma-separated list of their names.

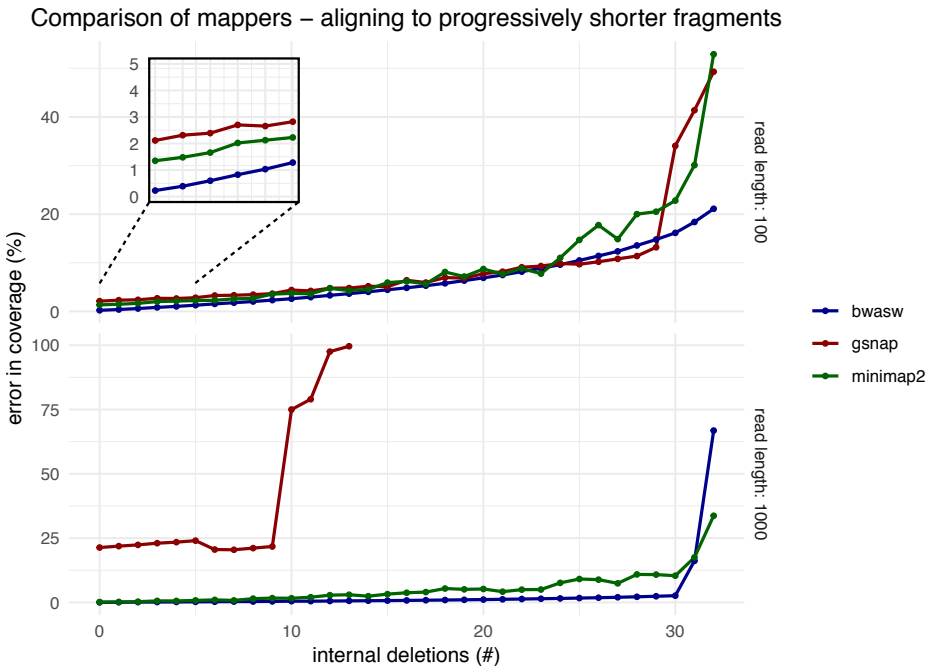

Figure 4: Suitability of different mapping algorithms for aligning reads to internally deleted TEs. The assembly-free nature of our approach necessitates reliable alignments to short reference sequences. We therefore compared three widely-used mapping tools and measured the error in coverage. We introduced a variable amount of internal deletions into the sequence of the TE DOC5, generated reads for these sequences and aligned them back to the consensus sequence of the TE. We found that bwa-sw reproduced the simulated coverage more accurately than gsnap and minimap2 (Li 2018; Li and Durbin 2010; Wu and Nacu 2010) for both short reads of 100 bp (top) and longer reads of 1000 bp (bottom).

3.2 Algorithm to detect internal deletions

Internally deleted TEs are inferred from subsequences of reads mapping to different positions in the reference (i.e. split-reads). For this purpose, we initially evaluated the suitability of different mapping approaches to identify internal deletions. We simulated TEs with varying numbers of internal deletions using SimulaTE (Kofler 2018). Interestingly, the local alignment algorithm bwa-sw performed better than the two split read mappers, gsnap and minimap2 (Fig. 4) (Li 2018; Li and Durbin 2010; Wu and Nacu 2010). Bwa-sw consistently shows a lower error in coverage compared to the other mappers. The difference for short sequencing reads is small (Fig. 4, top), but becomes more apparent for long reads, where edge effects cause reads to be discarded.

Algorithm 1 Find internal deletions from split reads

```
for read in alignment do
  segments ← all mappings of read
  if segments > 1 then
    macs ← Powerset(segments)
    for m in macs do
      overlap ← CheckOverlap(m)
      distance ← CheckDistance(m)
      if overlap ≥ limit then
        Discard(m)
      else if distance ≥ limit then
        Discard(m)
      end if
      scores ← CumulativeQuality(m)
    end for
    HighScoringMac ← max(scores)
    NewMapping ← BuildRead(HighScoringMac)
  end if
end for
```

Figure 5: Algorithm for detecting internal deletions. All subsequences of a single read mapping to a reference sequence are used to construct all possible combinations of the aligned subsequences, here called multiple alignment candidates (macs), to find the best contiguous alignment. After checking for overlaps and inconsistent gaps within reads, the algorithm scores the combinations on the fraction of aligned bases from the total length of the read. Notably, BuildRead generates a novel SAM entry based on the highest scoring mac, which replaces all previous subsequences. This step also constructs a new CIGAR string from the subalignments.

Similarly to BLAST, bwa-sw reports all possible local alignments, i.e. high-scoring-pairs (HSPs) of a read (Li and Durbin 2010). These HSPs may be on different contigs, overlapping or separated by large gaps. To identify internal deletions it is thus necessary to arrange these HSPs into a single best contiguous alignment. Therefore we devised a specialized algorithm to detect internal deletions from such split-read HSPs (Fig. 5). In this procedure, first all possible combinations of HSPs (multiple alignment candidates or macs) from every split-read are constructed, i.e. the powerset of HSPs. Then, combinations with overlapping subsequences and inconsistent alignments (e.g. when large internal regions of reads are not aligned) are removed. Subsequently, solely the combination of HSPs with the largest fraction of the read aligned to one reference is retained. Finally, the algorithm replaces all HSPs of a read by the best combination of HSPs. This involves the creation of a new SAM entry in the alignment file and the construction of its corresponding CIGAR string. To detect terminal deletions, DeviaTE utilizes soft clipped reads, for which a substantial fraction was not mapped to the reference sequence.

3.3 Normalization and copy number estimation

By default, DeviaTE performs no normalization and reports raw abundances of nucleotide counts. This is not suitable for comparing TEs between multiple samples. Therefore, we implemented two different strategies: normalization by million mapped reads and normalization by single-copy genes. The first option normalizes all counts (coverage, SNPs and deletion variants) by million mapped reads and thus accounts for different depths of sequencing, when comparing two or more samples. The second option, the single-copy gene normalization method contrasts all counts with the number of reads mapping to one or more single-copy genes, i.e. genes that are present only once in the genome of the investigated species. Therefore, it accounts for the sequencing depth between samples and additionally obtains an estimate of the insertion copy number of the analyzed TE. Normalization with the coverage of single copy genes may be especially useful when comparing TE abundance between species. In case the genome size varies among samples, normalization to one million mapped reads will result in misleading values, whereas normalization to single copy genes avoids this problem. For this normalization method, the sequence of one or more single-copy genes needs to be present in the library containing the TE consensus sequences. This method can be invoked by providing a list of the single-copy gene names upon execution of the

program. If more than one gene is specified, the average of their coverage is used for normalization. The estimated copy number per haploid genome is provided in the header section of the resulting output table.

Notably, to estimate the abundance of TEs, DeviaTE considers both base and physical coverage, i.e. the coverage of regions spanned by internal deletions. This is crucial as we found that a simple approach, which does not take the physical coverage into account, leads to highly biased results (Fig. 6). To demonstrate this bias, we conducted a simulation experiment involving one *D. melanogaster* chromosome with 50 P-element insertions. Five insertions were full-length and 45 copies had an internal deletion spanning exon 1 to 3, previously described as the KP element (Fig. 6A, Black et al. 1987). DeviaTE was used to estimate the number of insertions with three single-copy genes (*rpl32, piwi* and *act5C*). Estimation was performed without (Fig. 6B, top), and including physical coverage (Fig. 6B, bottom). The estimation ignoring physical coverage leads to a mere 22.79 detected copies, whereas an approach including physical coverage was able to detect 46.06 out of 50 simulated copies and reported a frequency of 88.65 % for the KP element.

3.4 Demonstrating the applicability of DeviaTE

An analysis of TE abundance and diversity may be useful in many different research areas. Thus, we aimed to develop a widely applicable tool that can be used for individuals, pooled populations or tissues of any species. DeviaTE may be used to study TE invasions (Fig. 8), identify clinal variation in TE composition (Fig. 9), estimate TE divergence across species (Fig. 10) or to estimate the proportion of internally deleted (i.e. non-autonomous) and full-length TEs (Fig. 15).

Initially, we demonstrate the basic utility of DeviaTE with a plot illustrating the composition of the long terminal repeat (LTR) retrotransposon burdock in a *D. melanogaster* population from the Netherlands (Fig. 7; data from Grenier et al. 2015). Read depth and sequence polymorphisms are visualized from short sequencing reads mapped to the consensus of the burdock TE family. Overall, the level of divergence is low with some polymorphic sites, fixed differences and few internal deletions. Interestingly, the internal deletions largely coincide with the coding sequence of the TE, as indicated by the annotation at the bottom. Further, ambiguously mapped regions overlap with the long LTRs of burdock (Fig. 7).

Further features of DeviaTE are demonstrated by the analysis of a P-element invasion in experimentally evolving *D. simulans* populations (Fig. 8; data from

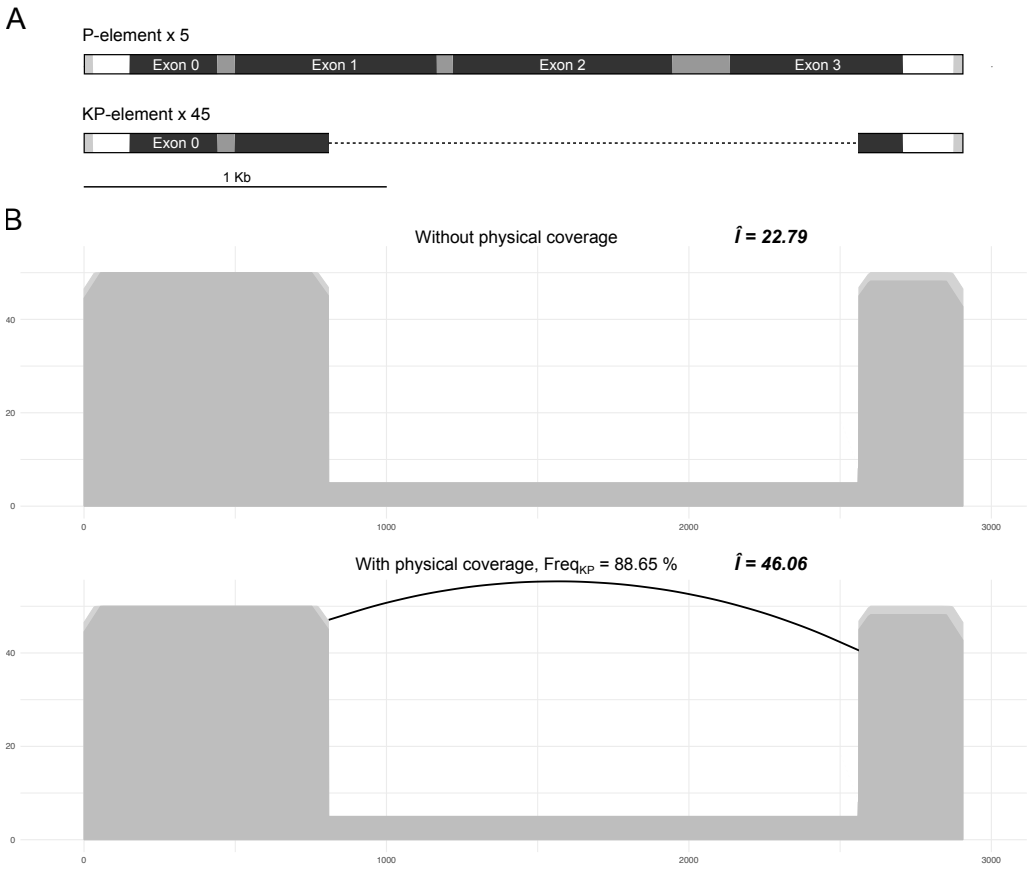

Figure 6: Ignoring the physical coverage of TEs (i.e. regions spanned by internal deletions) leads to biased estimates of TE insertion numbers. A) Structure of the full-length P-element and the internally deleted KP element. We simulated reads from a single *Drosophila melanogaster* chromosome containing fifty insertions of the P-element, five of which were full-length and the remaining 45 were internally deleted KP elements. B) Insertion copy numbers were estimated with DeviaTE using the single copy genes *rpl32, piwi* and *act5C* for normalization. With a naive approach ignoring physical coverage, a mere 22.79 insertions are detected. However, with our approach including physical coverage, DeviaTE identifies 46.06 P-element insertions (out of 50).

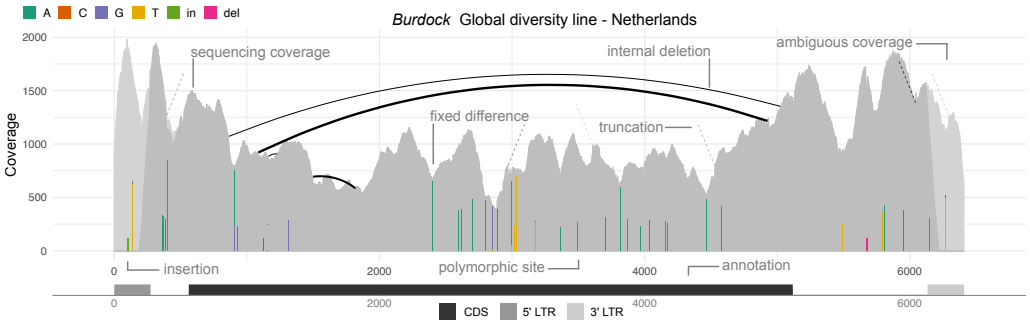

Figure 7: Example for the visualization of TE diversity with DeviaTE using burdock in *D. melanogaster*. Sequencing coverage is shown separately for unambiguously (dark grey) and ambiguously (light grey) mapped reads. Fixed differences and polymorphic sites are shown as colored bars, with the height of the bar corresponding to the frequency of the SNP. The reference allele is not shown in the visualization. Internal deletions are displayed as arcs, where the width of the arcs scales with the abundance of the deletion. Terminal deletions are shown as dashed lines, with opacity indicating their abundance. An annotation of the TE is shown at the bottom.

Kofler et al. 2018). The authors monitored a P-element invasion for 60 generations by sequencing the populations every 10 generations as pools. To allow for a comparison of the TE abundance among samples, we used one of the normalization methods implemented in DeviaTE, namely normalization of the coverage to one million mapped reads. Another feature of DeviaTE is to automatically arrange multiple input files, in this case data from multiple generations, in a vertical grid. A legend for the colors denoting sequence polymorphism is shown at the top and the annotation of the TE is given at the bottom (Fig. 8).

In agreement with Kofler et al. (2018), we observe an increase of P-element copy numbers during the invasion as well as a rapid emergence of internally deleted P-elements, first appearing at generation 10 (Fig. 8). Also note the SNP at position 2040, which is characteristic for the *D. simulans* P-element (Kofler et al. 2015; Yoshitake et al. 2018). Using the coverage of the single copy gene *rpl32* as reference, we estimate that the P-element abundance increased from 0.95 insertions per haploid genome at the base population (G_0) to 15.8 at generation 60 (G_{60}, Fig. 8). This is consistent with estimates of Kofler et al. (2018), who relied on a different approach to estimate P-element abundance, i.e. extrapolating the fraction of reads mapping to the P-element to the estimated genome size of *D. simulans*.

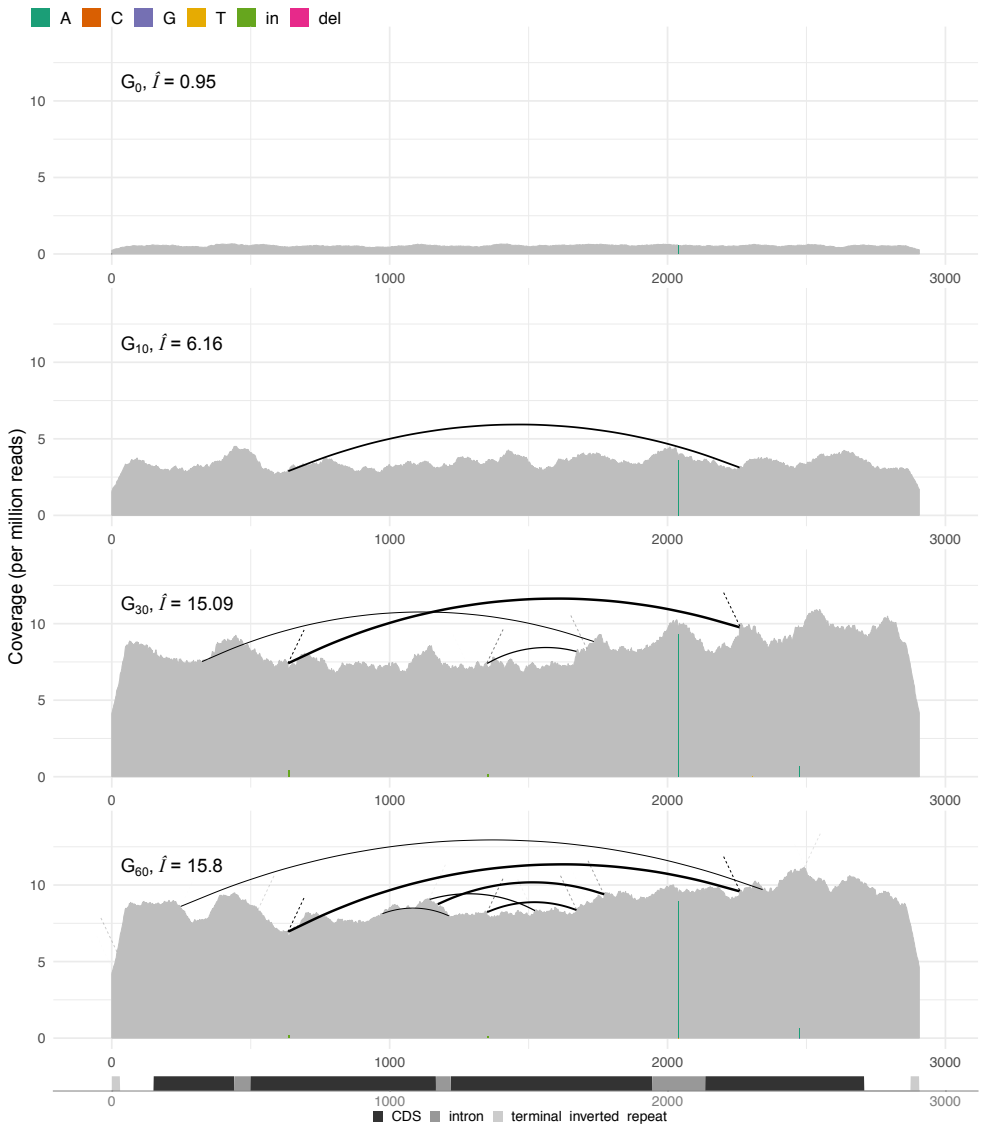

Figure 8: An invasion of the P-element in an experimental *Drosophila simulans* population visualized with DeviaTE. We show the abundance and the diversity of the P-element for four successive time points. The coverage was normalized to one million mapped reads and estimates of insertions per haploid genome ($\hat{I}$) were calculated based on the coverage of a single-copy gene. Note that the abundance of P-elements as well as the number of internally deleted variants increases during the invasion.

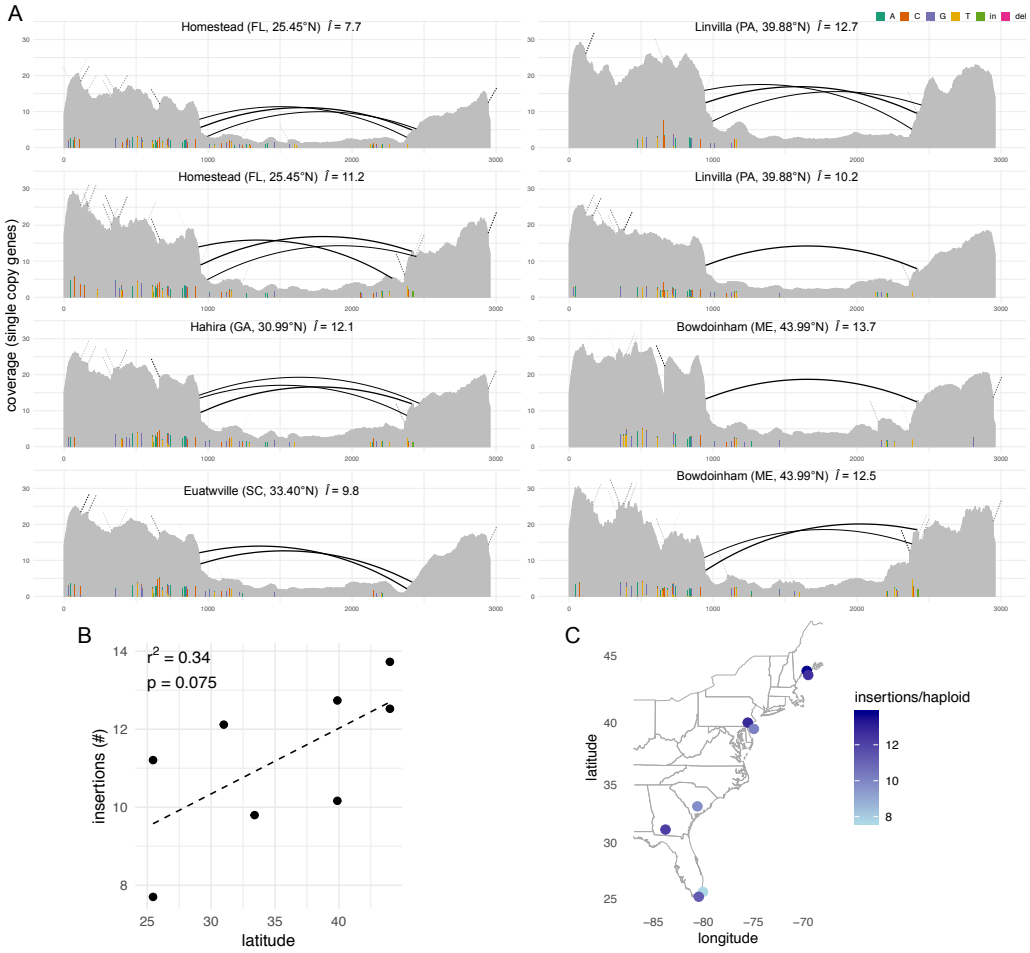

Figure 9: DeviaTE may be used to study clinal variation in TE composition. In this example, we test whether the transposon hobo shows clinal variation in *Drosophila melanogaster* populations sampled along the North American East Coast (data from Bergland et al. 2014). A) Diversity and estimated insertion numbers per haploid genome ($\hat{I}$) of the hobo element in eight populations ranging from Florida (southern-most samples, top left) to Maine (northern-most samples, bottom right). The lowest number of insertions was estimated for a sample from Homestead (Florida, FL) at 7.7 copies, whereas the highest number of insertions, 13.7, was found in a sample from Bowdoinham (Maine, ME). B and C) A linear regression on TE insertion numbers suggests a weak, non-significant relationship between hobo copy numbers and latitude.

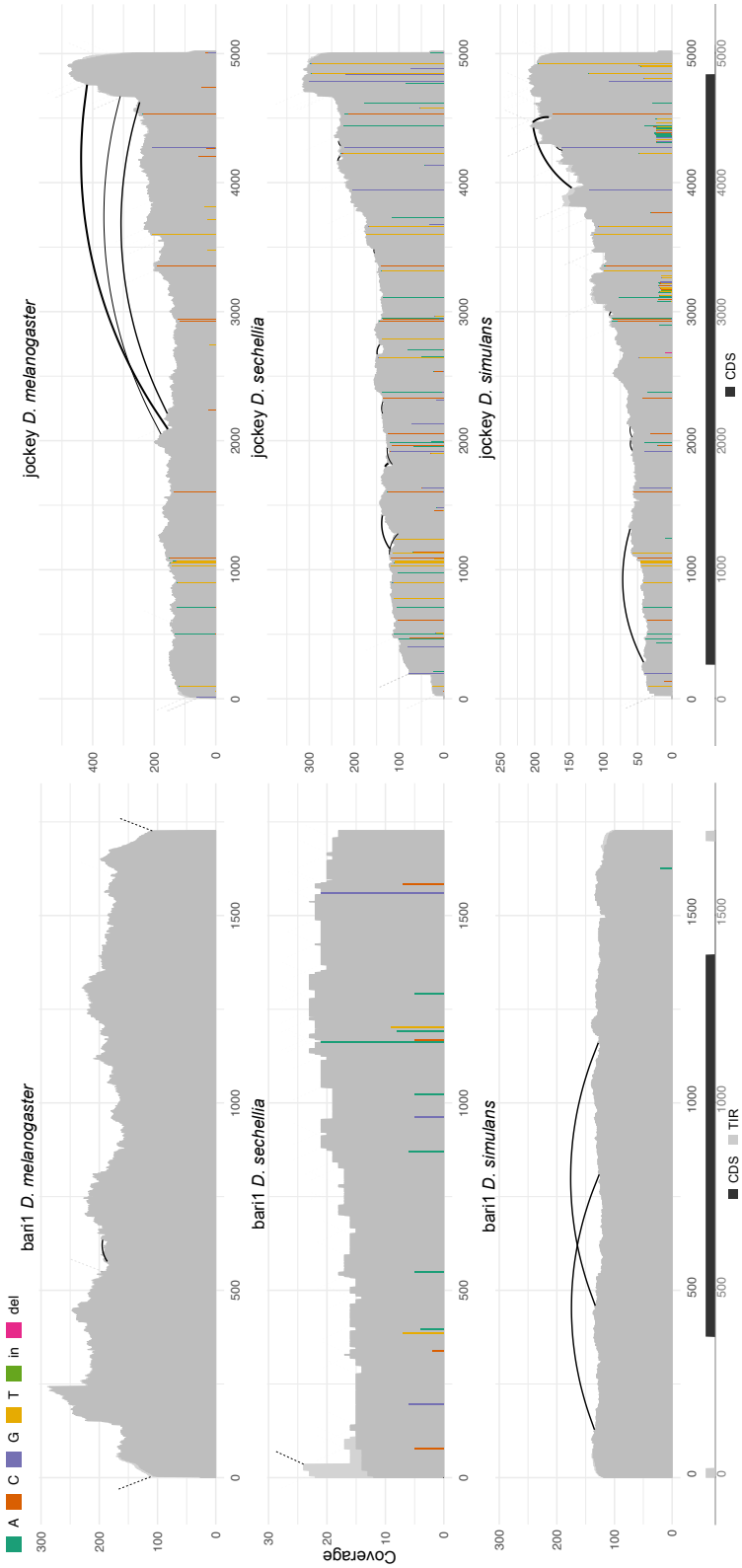

Figure 10: DeviaTE allows for the comparison of TE composition among species. In this example, we show illustrations for bari1 and jockey in the genomes of *Drosophila melanogaster*, *D. sechellia* and *D. simulans*. Bari1 (left column) is a 1.7 kb terminal inverted repeat element that is widespread in the *Drosophila* genus. The copies show high sequence homology across species with few low frequency polymorphisms and internal deletions. The transposon jockey (right column) is a LINE-like non-LTR retrotransposon of 5 kb length. It has a higher level of divergence among species as we find fixed differences and large deletions in all three species. Some fixed differences are shared, while others are exclusive to one species.

DeviaTE also allows for the normalization of the read depth of TEs to the coverage of single copy genes. We demonstrate this feature and an additional usage scenario by applying the tool to data from *D. melanogaster* populations sampled along the North American East Coast (Fig. 9; data from Bergland et al. 2014). We investigated whether copy numbers of the DNA transposon hobo exhibit clinal variation. Using the coverage of multiple single copy genes, we found a weak, non-significant relationship between latitude and hobo copy numbers ($r^2 = 0.34$, $p = 0.075$, Fig. 9B, C).

As an example for the applicability of DeviaTE in comparative studies, we assessed TE divergence of two TE families across three species. With the help of the automatic arrangement functionality, TE families were aligned in columns and species in rows to facilitate comparison. Here we analyzed bari1 and jockey in the genomes of *D. melanogaster*, *D. sechellia* and *D. simulans* (Fig. 10; data from Drosophila 12 Genomes Consortium et al. 2007). While we observe high sequence similarity for bari1 (Fig. 10, left), jockey has an increased level of sequence divergence among species, as indicated by the higher number of SNPs and fixed differences (Fig. 10, right). Additionally, the read depth is less homogeneous along the TE sequence in the case of jockey. Interestingly, this skewed distribution of the sequencing coverage was described before (Kaminker et al. 2002) and may be explained by disrupted replication during the transposition of LINE-like elements (Finnegan 1997). The notable difference in sequence divergence between these transposon families may be indicative of the age of the families. Whereas horizontal transfer is frequent in DNA transposons such as bari1, non-LTR transposons such as jockey are thought to evolve solely vertically and thus show elevated levels of sequence divergence (Bartolomé et al. 2009; Malik et al. 1999).

3.5 Validation - accuracy and limitations

We carefully validated our tool with simulated data generated by SimulaTE (Kofler 2018) and explored up to which levels of divergence DeviaTE accurately reports the expected TE diversity. Firstly, we found that our tool recovers divergence levels of up to 15 % for short reads (100bp) and up to 22 % for longer reads (1000bp) (Fig. 11A). Above these values highly diverged reads fail to align to the reference sequence, which leads to an underestimation of the true divergence levels.

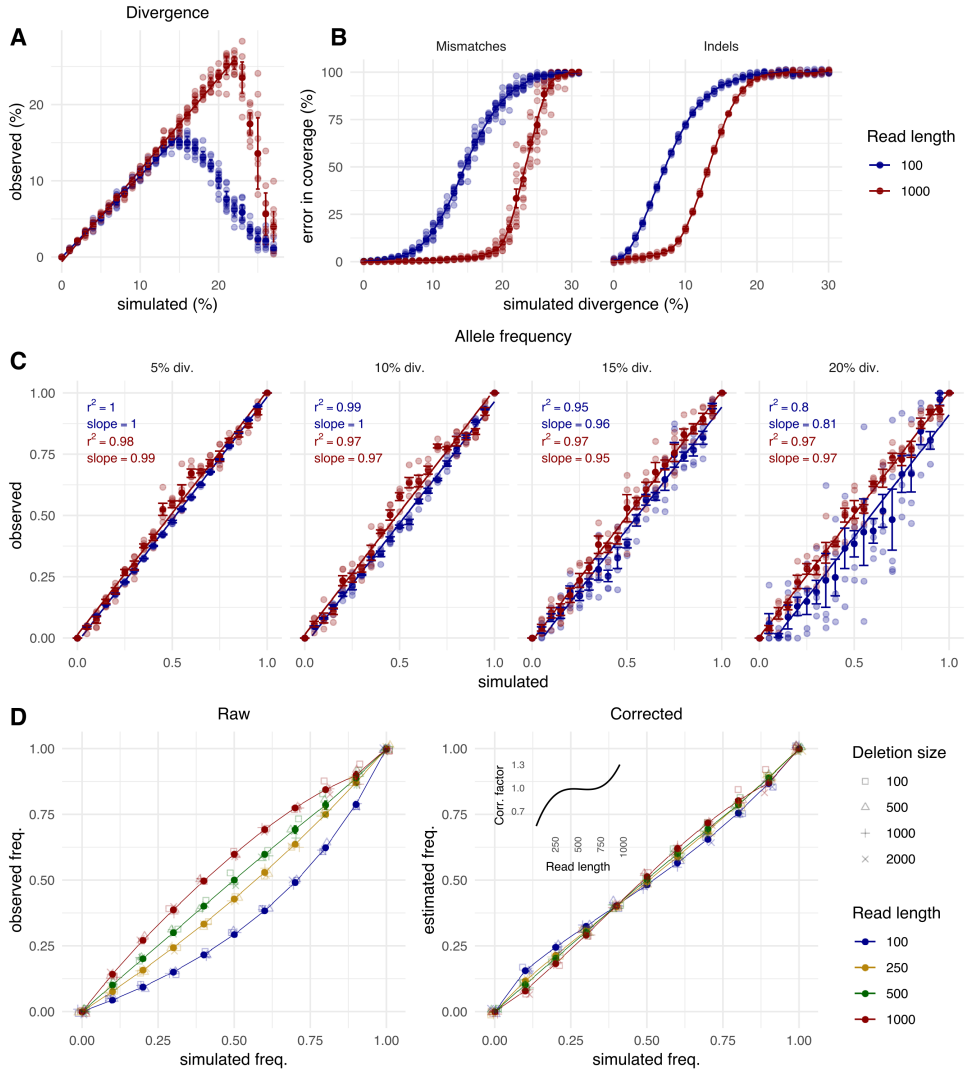

Figure 11: Validation of DeviaTE with simulated data. A) Comparison between simulated and observed sequence divergence. DeviaTE accurately recovers simulated divergence of up to 15 % for short reads (blue) and 22 % for long reads (red). B) Error of the estimated coverage dependent on the simulated divergence of reads. DeviaTE accurately reproduces the simulated coverage if the mismatch rate is smaller than 8 % and 16 % for short and long reads, respectively. Lower divergence levels are tolerated for indels. C) Accuracy of allele frequency estimates dependent on the divergence. DeviaTE accurately reproduces allele frequencies of SNPs up to a divergence of 15 %. D) Accuracy of estimated frequencies of internal deletions. Since raw frequency estimates show a small bias (left), we implemented a read length dependent correction factor (right, inset), which substantially improves the accuracy of frequency estimates (right).

Next, we investigated the impact of diverged or erroneous sequence composition on the accuracy of the estimated sequencing coverage. For short reads, 10 % mismatches lead to a coverage error of 16 %, whereas for long reads 10 % mismatches result in a coverage error of merely 0.7 % (Fig. 11B, left). Less divergence is tolerated to achieve similar results when indels are simulated instead of mismatches (Fig. 11B, right).

To test the accuracy of allele frequency estimates, we simulated a population of 20 haploid genomes with a single SNP of varying frequency (Fig. 11C). At a moderate divergence (< 10 %), the allele frequency is reproduced faithfully (adj. $r^2 = 0.99$ for 100 bp reads, adj. $r^2 = 0.97$ for 1000 bp reads), whereas for higher levels of divergence the accuracy diminishes (Fig. 11C). Thus, for TEs with high sequence homology, the allele frequency of sequence polymorphisms can be determined very accurately, while at low sequence homology the error in coverage leads to erroneous allele frequency estimates.

Lastly, we evaluated the accuracy of DeviaTE in estimating the frequency of internal deletions with varying sizes within TEs (Fig. 11D). The size of the deletion does not have an impact on the results. However, raw frequency estimates show a read length dependent bias, which causes the frequency of internal deletions to be overestimated in long reads and underestimated in short reads (Fig. 11D, left). To overcome this bias, DeviaTE automatically applies a correction factor which results in highly accurate frequency estimates (Fig. 11D, right). The reason for this bias of raw frequency estimates is that the mapping algorithm bwa-sw does not align subsequences of reads that are shorter than 30bp (by default). Hence, internal deletions can only be detected in central regions of reads.

3.6 Tracing the invasion history of the P-element

We aimed to reconstruct the invasion of the P-element in *D. melanogaster* using DeviaTE to detect internal deletions in TEs and estimate the frequency of such deletion variants. It is generally believed that the P-element was introduced in *D. melanogaster* by horizontal transfer from *D. willistoni* in South America (Daniels et al. 1990). Thereafter, it supposedly spread through North America and later reached Europe (Anxolabéhère et al. 1988). We propose a model, in which the signatures of internally deleted variants can be utilized to infer the sequence of the invasion in samples along the North American East Coast with data from Bergland et al. (2014) and across Europe using data described in Kapun et al. (2018). Here, we test the hypothesis that the P-element invaded the North

American continent from South America, spread across populations towards the north and subsequently reached the European continent, where it successively invaded populations in the general direction of west to east.

Patterns of deletion variants recover geographic distribution on both continents To begin with, we analyzed internally deleted TE variants by principal component analysis (Fig. 12, 13). In North American samples, the first two principal components (PCs) account for 64.5 % and 14.5 % of total variance, respectively. The sum of their explained variance is almost 80 % (Fig. 12A, B), thus only the first two PCs were retained for further analysis. Looking at the five top contributing deletion variants, we find that two of them are responsible for more than half of the variation explained by PC1 (Fig. 12C). The well-characterized variant, KP (Black et al. 1987), is also among the top contributing deletions with 18.05 % explained variance.

Notably, populations from the same sampling location have similar values of PC1 (Fig. 12A). We observe that the first principal component generally separates samples by geographic distribution, as the southern-most samples from Florida and the northern-most samples from Maine are at the extremes of the distribution and the remaining samples show intermediate values. To emphasize this finding, we assigned a gradient of colors to the points in Fig. 12 according to the latitude of their sampling location. We confirmed this observation of a latitudinal gradient by a Mantel test, which reports a significant spatial correlation with $r = 0.71$ (Fig. 12D). It finds that PCs are not randomly distributed across sampling locations. Rather, geographically close populations show similarity in their values of PCs, which decreases with the geographic distance of populations to each other.

Furthermore, an artificial sample containing only full-length copies of the P-element is grouped closest to flies from Florida. Assuming that the P-element invasion was initiated by full-length copies, devoid of any internal deletions, then flies from Florida are most similar to this ancestral sample (Fig. 12A). These findings support our hypothesis of a sequential invasion of the P-element with its origin south of the United States of America (USA) and a propagation towards northern latitudes.

In a PCA of samples collected throughout Europe, we observe a separation in four loose clusters, which seem to reflect geographic sampling locations (Fig. 13A). One of the clusters groups samples from Scandinavia, the UK and Germany, another one includes Iberian populations, a third cluster contains eastern Euro-

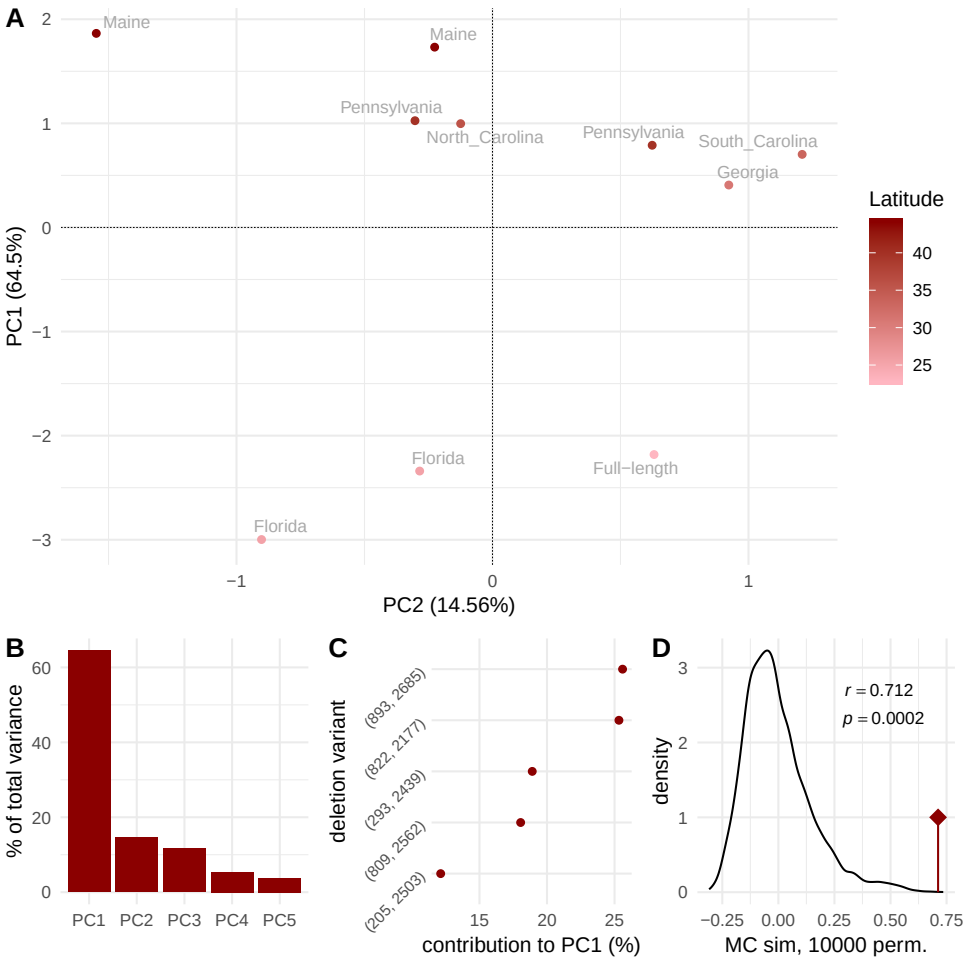

Figure 12: Principal component analysis of deletion variant frequencies in North American samples of *Drosophila melanogaster*. A) Populations cluster according to their geographic sampling location. PC1 separates populations by latitude. B) The percentage of total variance explained by the first five principal components. PC1 and PC2 explain >78 % of variance in deletion frequencies. C) The contribution of internal deletions to PC1. D) Results of a Mantel test on spatial correlation of the first two principal components (red marker). The distribution of the test statistic under H_0 was obtained by Monte-Carlo permutations.

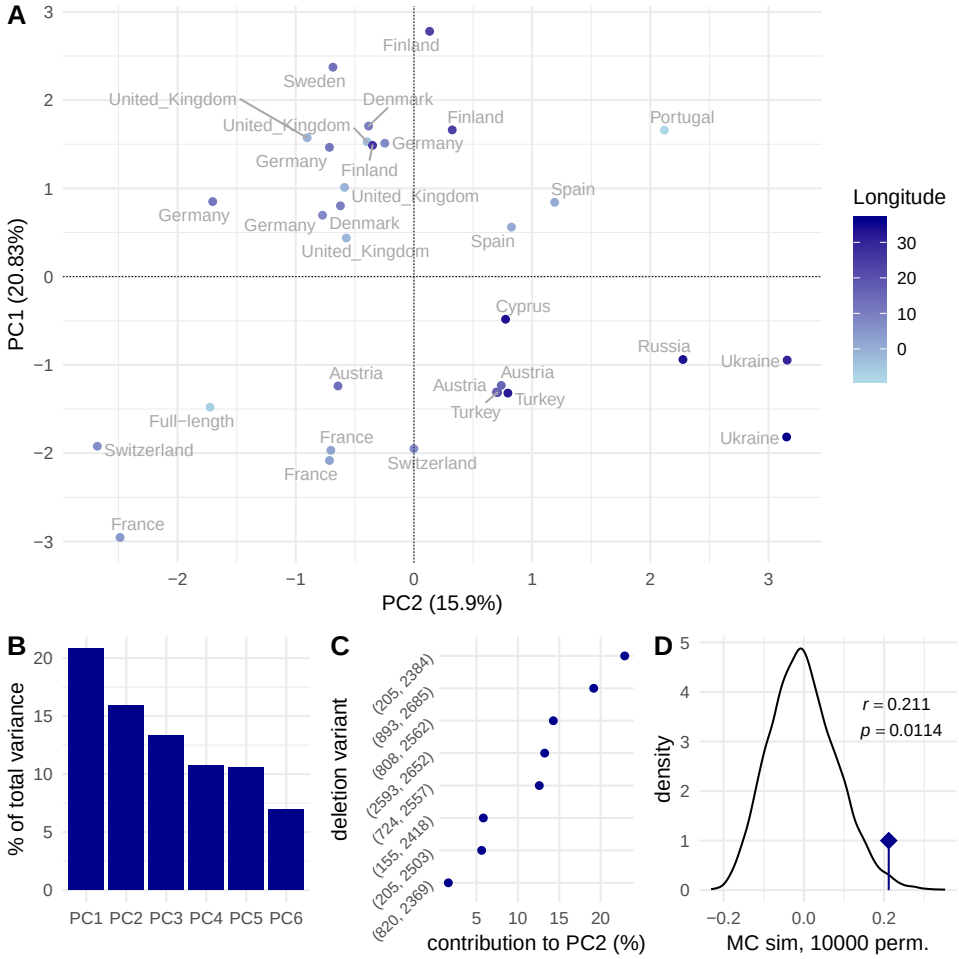

Figure 13: PCA of deletion variant frequencies in European populations of *Drosophila melanogaster*. A) Populations form four loose clusters according to their geographic distribution. Longitude was used to color data points. B) The percentage of total variance explained by the first six principal components. They explain a cumulative variance of 75 % of deletion frequencies. C) The contribution of individual variants. D) A Mantel test shows significant, yet weak spatial correlation of principal components (blue marker). The distribution of the test statistic under H_0 was obtained by Monte-Carlo permutations.

pean samples from the Ukraine, Russia, Cyprus, Turkey and Austria, and the fourth includes French, Swiss and another Austrian population. Interestingly, the artificial, pseudo-ancestral sample containing only full-length TEs clusters close to samples from France and Switzerland. This indicates that the pattern of deletion variants in these locations is most similar to a P-element without any internal deletions. Further, this finding hints at France as a possible starting point for the spread of the P-element throughout Europe.

The percentage of variance accounted for by PC1 is 20.8 % and 15.9 % by PC2 (Fig. 13A, B). Hence, to achieve a cumulative explained variance of 75 %, PCs 1 - 6 are considered for further testing. A Mantel test of pair-wise distances calculated from these components and geographic distances reveals a significant, yet weak spatial correlation, i.e. a non-random geographic distribution of deletion variants (Fig. 13D). To indicate the hypothesized general direction of the invasion, we used longitudinal coordinates to color the points. The contribution of individual variants to PCs ranges from 0.9 % to 22.9 % (Fig. 13C). The KP element also plays an important role on the European continent with a contribution of 14.3 %, which is similar to the contribution of 18.05 % in North America.

Similarity of deletion patterns mostly clusters neighboring populations
To further explore the relationship of the P-element in North American and European samples, we constructed dendrograms for each continent (Fig. 14A, B) and for both datasets combined (Fig. 14C). We computed pair-wise Jost's D from internal deletions as a measure of similarity of two samples to build the trees. Firstly, a dendrogram of the North American P-element variants clusters specimen from the same geographic location (Fig. 14A). Further, it connects the samples seemingly sequential by latitude, similar to results from the PCA (Fig. 12A). We confirmed this non-random distribution of patterns by a Mantel test, showing significant spatial correlation of pair-wise D ($r = 0.834$, $p = 0.0006$). We may conclude from this, that the P-element invasion has left behind traces in the form of deletion variants during its hypothesized spread from south to north.

Next, we investigated the relationship of European samples and found that sampling location is generally reflected in the dendrogram (Fig. 14B). Most samples from northern and some from central Europe are closely related, but some longer, more distant branches also exist. These include a branch of western samples in France and Switzerland as well as a branch of eastern European samples from Austria, Ukraine and Russia, similar to results from the PCA. This indicates

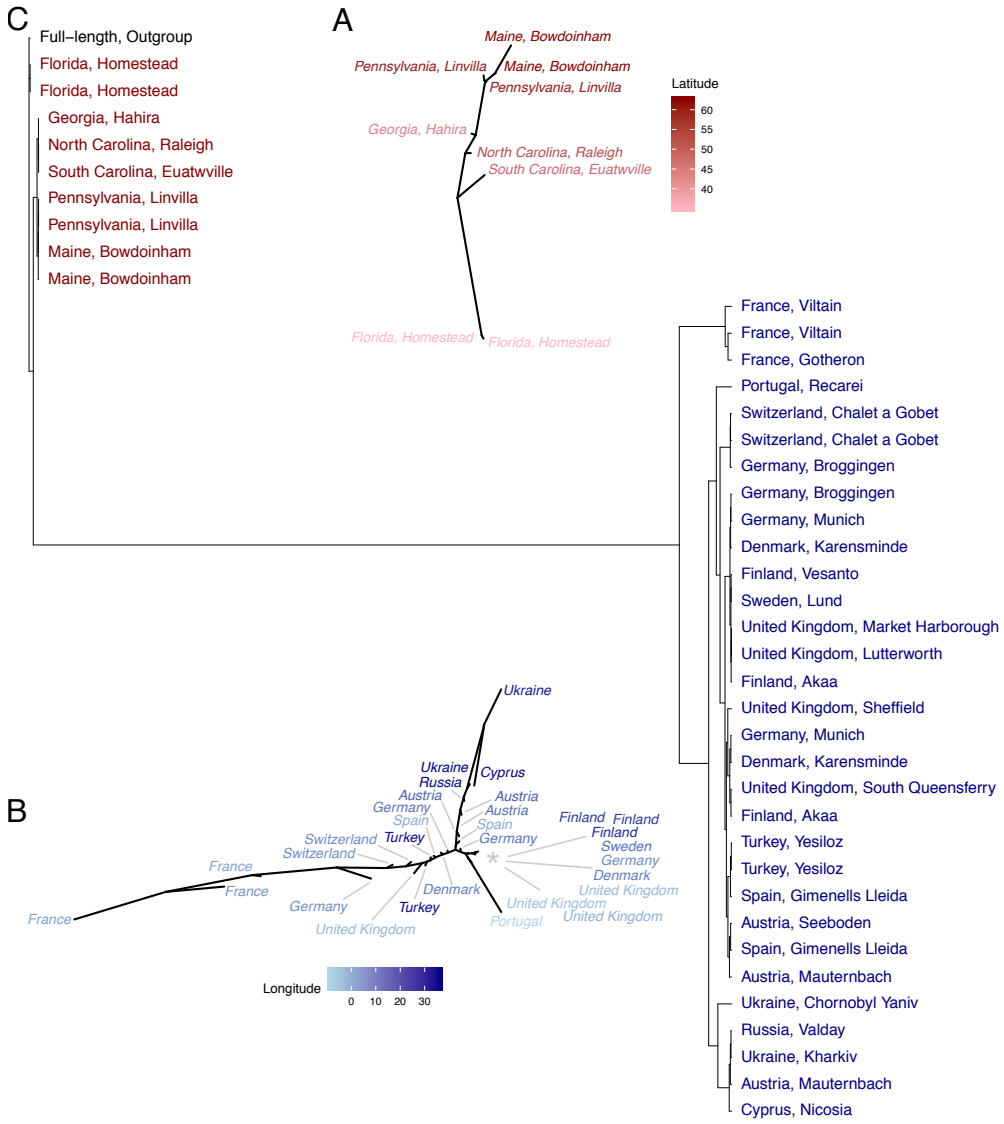

Figure 14: Dendrograms cluster populations by the similarity of their deletion variant patterns. A) An unrooted tree of North American samples clusters populations sequentially by latitude. This supports a hypothesized invasion of the P-element from south to north. Sampling locations are colored by their latitudinal coordinates. B) European samples are clustered less clearly. However, longer branches separate more distant populations from western and eastern Europe. An asterisk denotes populations clustered with small branch lengths. Longitudinal coordinates were used to color samples. C) To infer the direction of the invasion, we included a pseudo-ancestral sample containing only full-length P-elements as an outgroup. Populations cluster well within continents, which are separated by a long branch. Of all populations from both continents, Florida and France are the closest to the root, putting them forward as potential origins of the continent-wide invasions. North American and European populations are colored in red and blue, respectively.

that samples from the extremes of their geographical distribution are least similar. Conversely, a test for spatial auto-correlation does not yield significant results ($r = 0.125$, $p = 0.125$), suggesting that the pattern of deletion variants is not as clearly preserved compared to the pattern in North America. This might be the consequence of more complex geographical structures, which lead to the invasion proceeding in a less clearly defined sequence. Alternatively, ongoing migration and admixture of populations could potentially blur the patterns.

Florida and France are potential origins of continent-wide invasions Besides investigating the relationship of samples, we attempted to infer the direction of the invasion from a dendrogram combining the populations from both continents (Fig. 14C). Additionally to the datasets from North America and Europe, which are denoted in red and blue colors respectively, we introduced an artificial, pseudo-ancestral sample. This population contained solely full-length P-elements and was used as an outgroup to all other populations and consequently for rooting the tree in this analysis. We find that samples cluster very well within continents, which are separated by a long branch. This indicates less similarity between continents than within (Fig. 14C). Interestingly, the outgroup was placed closely to samples from Florida. This confirms, that these populations are most similar to the ancestral form of the P-element and implies that they were invaded first. Moreover, the order of North American samples along their latitudinal coordinates, as observed in the neighbor-joining tree (Fig. 14A), is retained in the rooted tree. Hence, this affirms the proposed sequential invasion along the East Coast of the US towards the north.

The European populations cluster to a clade containing samples from Pennsylvania and Maine. Similar to the unrooted tree in Fig. 14B, we find clearly separated clades for Eastern Europe, for samples from France and for populations from Switzerland and Germany, and less distinct clades of northern and the rest of central European samples. Notably, the most similar populations to North American samples, and thus to the root of the tree, are from France (Fig. 14C). Consequently, France might be regarded as the entry point for the invasion in Europe. From there, it purportedly spread throughout the continent on multiple paths towards eastern Europe, towards the North and onto the Iberian peninsula.

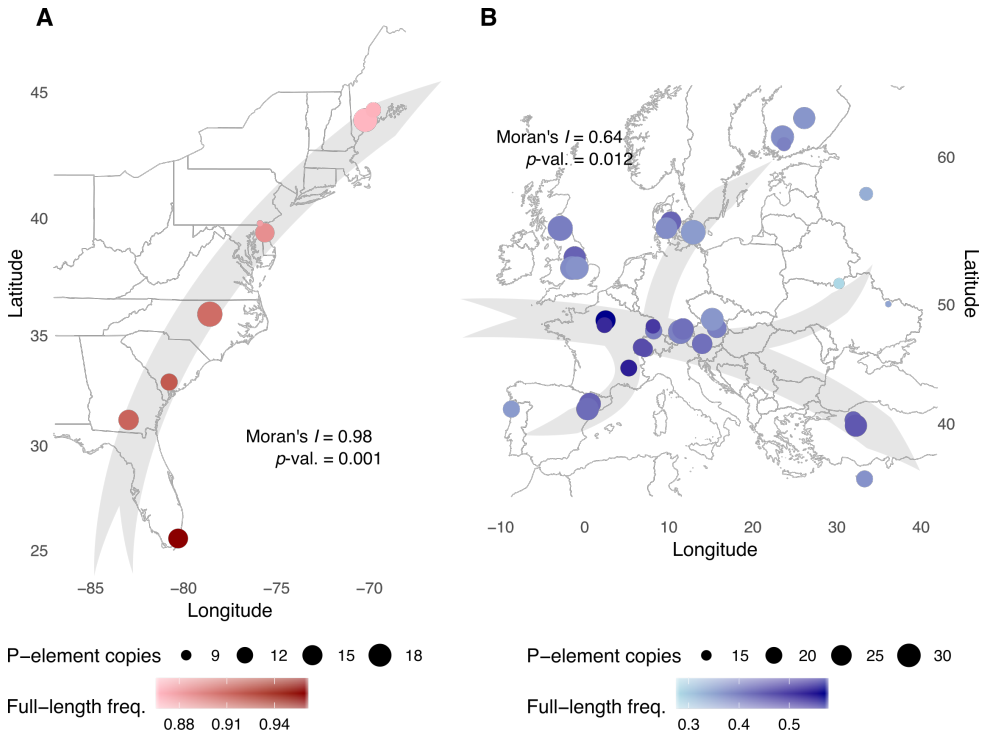

Figure 15: Distribution of the frequency of full-length and total insertions of P-elements. A) North American populations generally have a higher proportion of full-length P-elements. The distribution shows a strong spatial auto-correlation, as percentages gradually decrease from south to north. Hence, the proposed direction of the invasion is indicated with a grey arrow. B) European samples have lower proportions of full-length TEs, with the highest value found in French populations. In contrast, the lowest percentages were found in geographically distant locations. Thus, we propose an invasion originating in France, proceeding to the East and ultimately spreading across the continent through multiple routes.

The frequency of full-length elements gradually decreases during the invasion In a third analysis we investigated the distribution of the frequency of full-length as well as the absolute insertion numbers of P-elements in all samples (Fig. 15). The number of P-element copies ranges from 9 to 18 per haploid genome in North America, and from 15 to 30 copies in Europe, indicated by the size of the circles. While the difference between the continents was found to be significant, with more insertions in Europe on average ($t = 6.2$, $p = 1.04^{-5}$), we did not find any spatial pattern of P-element copy numbers within continents.

With respect to the frequency of full-length TEs, we observed that European samples possess a significantly lower proportion ($t = -31.282$, $p < 2.2^{-16}$), with a range of 28.2 to 57.0 %, as opposed to frequencies of 86.8 to 95.9 % in

North America. Additionally, we found a strong spatial auto-correlation in North America (Moran's $I = 0.98$, $p = 0.001$, Fig. 15A), and a moderate, significant spatial auto-correlation of frequencies in Europe (Moran's $I = 0.64$, $p = 0.012$, Fig. 15B). This implies that populations with a similar frequency of full-length P-elements are geographically close and populations with a dissimilar frequency are more distant to each other.

This observation is consistent with our proposed model of a sequential invasion, in which the proportion of full-length TEs gradually decreases through step-by-step migration events. The results from North America confirm our previous finding, that the P-element invasion had its origin south of Florida, where over 95 % of TEs are full-length, and progressed further north, where only 86 % are full-length (Fig. 15A). As for Europe, the direction of the invasion is not as easily traceable, owing to the more complex geographical structure. However, we find the highest frequency of full-length elements in France, which suggests a role as the starting point of the invasion. Conversely, we found the lowest proportion of full-length elements at geographically more distant locations to France, such as Ukraine, Russia, Portugal and Finland (Fig. 15B).

Finally, we present how the global invasion of the P-element might have proceeded with arrows denoting its direction (Fig. 15A, B). After horizontal transfer of the P-element from *D. willistoni* to *D. melanogaster* in South America, the TE sequentially invaded the North American continent, while the proportion of full-length elements gradually decreased. During this spread, populations acquired patterns of deletion variants, which lead to clustering of samples according to their geographic distribution. After some time, the P-element reached Europe, possibly invading from France and proceeding in different directions. Populations in France consistently show the highest proportion of full-length TEs, while samples in eastern and northern Europe as well as on the Iberian peninsula have increasingly dissimilar deletion variant patterns. Thus, we hypothesize that the invasion of Europe had its origin in France, and progressed generally towards the East. Consequently, the P-element was disseminated throughout the continent on multiple routes, reflecting the more complex geography.

4 Discussion and Conclusion

In this work we presented a novel tool for the analysis of mobile genetic element composition called DeviaTE. It does not require an assembled reference genome; it only needs sequencing reads and consensus sequences of the investigated transposable elements. Thus, it can be employed for model and non-model organisms alike. DeviaTE can be used to study single-nucleotide polymorphisms, indels and structural polymorphisms in the TE sequence, and estimate insertion copy numbers and variant frequencies. The program is open-source, hosted on GitHub and can be easily installed via pip or conda. We presented DeviaTE's versatility and its wide applicability with several examples and performed extensive validation using simulated TE landscapes.

4.1 Highly diverged TEs can reduce accuracy

In the results of the validation, we observed that accuracy and reliability depend on the length of the analyzed sequencing reads. For instance, when reads of 1000 bp length are used, DeviaTE can recover correct information on coverage and polymorphisms for sequences that are 22 % diverged compared to the consensus of their TE family. However, for shorter reads of 100 bp length the accuracy starts to drop at a sequence divergence of 15 % or more (Fig. 11). Although the sequence divergence of TEs with similar features can form a broad spectrum, membership in a TE family is defined as a minimum of 80 % sequence identity (Wicker et al. 2007). This can potentially be an issue, as short sequencing reads of highly diverged TEs fail to get assigned to their TE family and may lead to underestimation of insertion numbers and polymorphism rates.

This implies that in case of evolutionary old TEs, an analysis might be limited to recent insertions as opposed to highly degenerated relics. Such ancient TEs include, for example, the non-LTR retrotransposon clades L1 and CRE (Malik et al. 1999), or LINE and MIR elements (Giordano et al. 2007). Often, old TEs are found in specific locations within the host genome. For instance, pericentromeric and other heterochromatin-rich regions of chromosomes contain the most diverged copies of some TE families (Lerat et al. 2003). Thus, DeviaTE might show reduced accuracy for TEs which reside mainly in these regions, such as the TEs forming the telomeres of *Drosophila* (Casacuberta 2017; Danilevskaya et al. 1994). Another instance potentially causing a bias might be TEs with structures prone to sequence changes. For example, Alu elements, which account for a staggering

11 % of the human genome, contain an unstable A-rich tail that rapidly shrinks during transposition and accumulates mutations (Deininger 2011). However, some authors report the highest sequence divergence for LINE and Alu element subfamilies to be 17.8 and 15.1 %, respectively (Khan et al. 2006; Price et al. 2004). Hence, these elements are suited for an analysis with DeviaTE using reads of short or medium length. Moreover, repetitive sequences showing a sequence divergence to its consensus sequence of 20 % or more, should in any case not be considered a member of that family (Wicker et al. 2007).

4.2 Availability of high-quality, reliable consensus sequences

Our approach of analyzing TEs avoids the need for a genome assembly and thus contributes to the versatility of DeviaTE, since high-quality assemblies are only available for a mere 25 species so far (Lewin et al. 2018). Admittedly, this creates the requirement of reliable consensus sequences of the investigated TEs. This is important in order to correctly assign sequencing reads of TEs to their respective family and could otherwise lead to a bias in estimated copy numbers and polymorphism levels. A single complete, high-quality collection of repeat elements does not exist yet, but multiple efforts are pursued in order to achieve this goal. A widely-used, de facto standard database for prototype sequences of repetitive elements is Repbase Update (Bao et al. 2015). It contains the largest collection of consensus sequences for TEs and other repetitive elements, with currently over 44,000 entries from more than one hundred species and is used in many fields of genome research. Other resources include Dfam with 4150 entries (Hubley et al. 2016) and TREP, which initially contained TEs from *Triticeae* only, but was gradually extended with sequences from other plant and fungal species (Wicker et al. 2002; Wicker et al. 2007). Additionally, manually curated databases for diverse species or clades exist. These include collections for *Drosophila* (Bergman et al. 2017a), conifers (Yi et al. 2018), fish (Shao et al. 2018) and dioecious plants (Li et al. 2016), among others. A comprehensive overview of available repositories is presented in Goerner-Potvin and Bourque (2018).

If, however, the consensus sequence for a specific TE can not be found in a database, multiple tools to generate such sequences are available, e.g. RepARK, REPdenovo or RepeatScout (Chu et al. 2018; Koch et al. 2014; Price et al. 2005). These tools construct prototype sequences of repetitive elements from sequencing reads by assembling high-frequency repeat k-mers. Arguably, the quality of consensus sequences produced by specialized algorithms might be higher than

what can be achieved by general purpose genome assemblers. Thus, if the goal is the analysis of sequence composition of few TEs, rather than the identification of particular insertions of a TE in a genome, using purpose-built, manually curated consensus sequences is more appropriate.

4.3 Using internal deletions as genomic markers to track TE invasions

Another result of this work is a model for the invasion of DNA transposons and an approach to reconstruct the course of such invasions. We show that the invasion of the P-element in North America and Europe can be retraced by exploiting non-autonomous, internally deleted remnants of the TE. The abundance and frequency of deletion variants in each deme can be seen as a fingerprint left behind by the sequential invasion.

The use of internal deletions as markers is based on multiple characteristics, which confer advantages compared to other, more commonly studied markers. For example, SNPs as the canonical form of mutations occur at a rate of approximately 10^{-9} per base per generation (Biémont and Vieira 2006). Thus, during the circa 60 years of the global P-element spread, the generation of unique fingerprints composed of SNPs would not be possible. In fact, only few SNPs have been reported in the sequence of the P-element. These include a species-specific SNP in the P-element of *D. simulans* at position 2040 (Kofler et al. 2015; Yoshitake et al. 2018), another species-specific substitution at position 32 in *D. willistoni* (Daniels et al. 1990) and a SNP at position 2699 of an internally deleted P-element (O'Hare and Rubin 1983). On the contrary, internal deletions of active DNA TEs arise at high rates: internally degenerated variants were found after only tens of generations, in both experimental and natural invasions of *Drosophila* populations (Daniels et al. 1985; Kofler et al. 2015; Kofler et al. 2018).

Other possible markers are specific insertion sites of TEs within the host genome. One would assume that novel insertions in a newly invaded host occur at arbitrary locations. In fact, many TEs show a preference for specific insertion sites and behave in a non-random manner (Wu and Burgess 2004). Therefore, the probability to find the same target sites in multiple, independently invaded populations is not negligible and increases with the number of samples. Conversely, the model for the transposition of DNA TEs by double-strand gap repair of Engels et al. (1990) suggests that internal deletions occur when the repair process is interrupted. The assumption of such interruptions leading to arbitrary deletion

breakpoints seems plausible. Incidentally, it has been observed that internally deleted derivatives of the P-element have heterogeneous sizes and unique deletion breakpoints, implying that there is no bias for their position (O'Hare and Rubin 1983).

Because internal deletions are key to our approach, it follows that only the invasion of TEs which generate deletion variants can be analyzed. Thus, TEs need to cause a double-strand break and subsequent initiation of gap-repair upon excision. In principle, this applies to all DNA, or class II, transposable elements, which mobilize by a 'cut-and-paste' mechanism. In addition to the generation of internal deletions, another prerequisite might be that an invasion happened recently. Depending on the intensity of the traces and the degree of migration between invaded populations, the pattern might otherwise be more difficult to observe. Apart from the P-element, interesting targets possibly generating fingerprints of internally deleted variants include the hobo and pogo elements in *Drosophila*, TEs in other clades such as the classic Ac/Ds system in maize (Rubin and Levy 1997), or the ubiquitous Tc/mariner elements (Auge-Gouillou et al. 1995). Our approach is not applicable to class I TEs, however, since they use an RNA intermediate for transposition and do not cause double-strand breaks.

4.4 Stability of deletion fingerprints

A remaining question is to which degree the fingerprints we observe resemble the outcome of the original invasion. During the time gap between the invasion and the sampling of genomic material from populations, processes might take place, which blur the traces of the invasion. This could hamper clear conclusions about the sequence of a TE invasion, since deciphering the original signal from a long-term outcome is difficult. Major factors that could impede the conservation of the pattern include the reactivation of TEs leading to secondary waves of radiation or migration between demes causing a loss of reproductive isolation.

A TE can be reactivated by the loss of piRNA production. This might be the result of a recombination event leading to the loss of TE insertions within a piRNA cluster or the disruption of a whole cluster by an insertion in its regulatory sequence (Brennecke et al. 2007; Zanni et al. 2013). Recently, it has been observed that environmental factors, such as high temperature, can lead to epigenetic activation of piRNA clusters (Casier et al. 2018). Conversely, this suggests that environmental variables may also lead to the deactivation of piRNA producing loci and thereby to a derepression of a TE. Another mechanism for the onset of a

secondary invasion is the simple deletion of repressive sequences. In *Drosophila* for example, it has indeed been shown that a remarkably high rate of large deletions leads to the elimination of non-essential DNA (Petrov and Hartl 1997), potentially affecting repeat elements in piRNA clusters.

Regarding demographic processes blurring TE variant distributions, some studies investigated the stability of the P and M cytotypes, which express the susceptibility to hybrid dysgenesis by P-elements (Bonnivard and Higuet 1999; Periquet et al. 1989). Interestingly, the distribution of cytotypes shows a longitudinal gradient across Europe (Bonnivard and Higuet 1999), similar to our findings of internally deleted elements (Fig. 15). Periquet et al. (1989) observe no difference in the distribution of either dysgenesis potential, number of P-element copies or frequencies of full-sized and KP elements over a period of five years. Further, they suggest that any changes occur at slow rates or even that a quasi-stable distribution across the continent has been reached.

Likewise, Bonnivard and Higuet (1999) find that the pattern of P and M cytotypes has been stable for 15 years, and hypothesize that buffer zones of certain incompatible cytotypes might reduce the migration between populations (Bonnivard and Higuet 1999). Since the responsiveness to P-element transposition is linked to full-length and internally deleted elements, we expect that such a limitation of migration between demes will also lead to the conservation of deletion fingerprints.

A different TE invasion examined in *D. melanogaster* is the spread of hobo (Bonnivard et al. 2000). Similar to our approach, a special molecular marker solely found in the hobo element was employed to characterize the TE insertions in populations. Interestingly, through a comparison of samples collected in the 1990s and lab strains of the same location from the 1960s, the geographic structure of the marker showed strong stability (Bonnivard et al. 2000). This is further evidence that TE variant patterns in *Drosophila* are conserved for some time after the invasion, and might indicate a minor role of migration in disrupting the geographic distribution of molecular markers.

However, additional insight into the stability of such patterns could be gained by simulation studies with demes showing variable TE composition, and different strengths and modes of migration. Besides simulations, sampling of natural populations at various time points might also elucidate how, and to which extent, demographic processes shape TE composition and influence the distribution of genomic markers.

4.5 Impact of TE composition and invasion histories

With our novel tool DeviaTE, we hope to contribute to the investigation of polymorphisms, abundance variation and dynamics of TEs in various species and groups of species. Its strengths lie in the assembly-free nature and wide applicability to sequencing reads of different technologies and lengths, and from various sources including cells, tissues, individuals or populations. Furthermore, it may be used with any species, as long as consensus sequences of the investigated TEs in a host can be obtained. DeviaTE is aimed to catalyze future progress in the broad spectrum of processes in which TEs play a major role, such as speciation (Oliver et al. 2013; Serrato-Capuchina and Matute 2018), hybrid dysgenesis systems (Bergman et al. 2017b; Black et al. 1987; Bonnivard and Higuet 1999), environmental adaptation (Casacuberta and González 2013; Schrader and Schmitz 2018), as well as disease and aging (Burns 2017; Wood et al. 2016).

Moreover, studies about TE invasions can further our understanding of transposition dynamics and how different genomic components compete analogous to species and employ their selfish properties within the genome ecosystem (Robillard et al. 2016). They could also help investigate the origin and frequency of horizontal transfer events. For example, unravelling TE invasions can solve incongruities between TE and species phylogenies and thus potentially improve our understanding of the evolutionary history of species (Dias and Carareto 2012). Additionally, knowledge of the invasion history of TEs might be important to inhibit the spread of antibiotic resistance and virulence factors, since mobile genetic elements and transposon-like elements are potent means of transfer and conjugation of such genetic information (Malachowa and DeLeo 2010; Waldor et al. 1996). Therefore, our simple method to reconstruct invasion histories is highly applicable and we hope that our approach can help move forward the understanding of TE biology.

List of Figures

Deutsche Zusammenfassung

Transposons (TEs) sind egoistische DNA Sequenzen, die sich in ihrem Wirtsgenom vervielfachen können. Sie wurden in den meisten Spezies, die bisher untersucht wurden, gefunden und weisen einen höchst unterschiedlichen Grad an Häufigkeit und Sequenzverschiedenheit auf. Die Zusammensetzung von TEs kann aber nicht nur zwischen, sondern auch innerhalb von Spezies variieren und wichtige biologische Konsequenzen nach sich ziehen. Unterschiede im Vorkommen innerhalb von Populationen könnten beispielsweise auf eine Invasion eines Transposons hinweisen, wohingegen Variation in der Sequenz das Vorhandensein von hyperaktiven oder inaktiven Varianten bedeuten könnte. Um die evolutionäre Dynamik von Transposons zu verstehen, ist es deshalb wichtig unverzerrte Schätzwerte für die Zusammensetzung von TEs zu erhalten.

Deshalb haben wir DeviaTE entwickelt; ein Programm zur Analyse und Visualisierung von TE Häufigkeit mit Illumina- oder Sanger-sequenzierten DNA-Abschnitten. Unser Werkzeug benötigt lediglich sequenzierte DNA-Abschnitte und Prototypsequenzen von TEs. Damit funktioniert es ohne Gesamtsequenz eines Genoms, was die Anwedung bei Nichtmodellorganismen, für die es bisher keine hoch qualitative Gesamtsequenz gibt, ermöglicht. DeviaTE erstellt eine Tabelle und eine Visualisierung der TE Struktur und liefert unverzerrte Schätzwerte für die TE Häufigkeit. Mit bereits publizierten Daten zeigen wir, dass DeviaTE benutzt werden kann um die Zusammensetzung von Transposons in Stichproben zu untersuchen, geographische Variation in TEs festzustellen oder die Verschiedenartigkeit von TEs zwischen Spezies zu ermitteln. Zusätzlich präsentieren wir eine gründliche Validierung mit simulierten Daten.

Darüber hinaus beschreiben wir eine Modell für Invasionen von DNA TEs und eine Methode um den Ablauf von solchen Invasionen mit unserem neuen Programm zu rekonstruieren. Wir argumentieren, dass eine Invasion einzigartige Fingerabdrücke in Populationen hinterlässt, die aus nicht-autonomen Varianten von TEs mit Deletionen inmitten ihrer DNA Sequenz, besteht. Mithilfe dieser TE Relikte zeigen wir, dass die Abfolge der P-element Invasion in Nordamerikanischen und Europäischen *Drosophila melanogaster* Populationen nachgezeichnet werden kann. Wir stellen fest, dass die Muster von Varianten mit deletierten Sequenzabschnitten die geographische Verteilung der untersuchten Populationen widerspiegeln. Zusätzlich ermitteln wir mögliche Ausgangspunkte und Routen für die Ausbreitung auf beiden Kontinenten. Mit der Entwicklung von DeviaTE hoffen wir, Fortschritte im Verständnis der Dynamik von TE Invasionen und anderer Prozesse, in denen TEs eine wichtige Rolle spielen, zu ermöglichen.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor Robert Kofler. He was incredibly supportive and dedicated throughout my stay in his lab, helped me establish the research project, always had the time to find some motivating words and encouraged me to follow my ideas and goals.

I would also like to thank my office mates Rui, Frank and Odnoo for helpful advice, fruitful conversations and random banter. Discussions with people at the Institute for Population Genetics led to new and exciting ideas and helped me to see my research in a broader perspective. Further, financial aid by the FWF is gratefully acknowledged.

Most crucially, I would like to thank Tania, my parents and my sister for relentlessly supporting me in any possible way and for advice whenever I had to make decisions. Your unceasing encouragement is invaluable!

References

Anxolabéhère, D., Kidwell, M. G., & Periquet, G. (1988). Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of *Drosophila melanogaster* by mobile P elements. *Molecular Biology and Evolution*, *5*(3), 252–269.

Anxolabéhère, D., Nouaud, D., Periquet, G., & Tchen, P. (1985). P-element distribution in Eurasian populations of *Drosophila melanogaster*: A genetic and molecular analysis. *Proceedings of the National Academy of Sciences*, *82*(16), 5418–5422.

Auge-Gouillou, C., Bigot, Y., Pollet, N., Hamelin, M. H., Meunier-Rotival, M., & Periquet, G. (1995). Human and other mammalian genomes contain transposons of the *mariner* family. *FEBS Letters*, *368*(3), 541–546.

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *60*(1), 11.

Bargues, N., & Lerat, E. (2017). Evolutionary history of LTR-retrotransposons among 20 Drosophila species. *Mobile DNA*, *8*(7).

Barrón, M. G., Fiston-Lavier, A.-S., Petrov, D. A., & González, J. (2014). Population genomics of transposable elements in *Drosophila*. *Annual Review of Genetics*, *48*(1), 561–581.

Bartolomé, C., Bello, X., & Maside, X. (2009). Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes. *Genome Biology*, *10*(2), R22.

Beall, E. L., Mahoney, M. B., & Rio, D. C. (2002). Identification and analysis of a hyperactive mutant form of Drosophila *P*-element transposase. *Genetics*, *162*(1), 217–227.

Bergland, A. O., Behrman, E. L., O'Brien, K. R., Schmidt, P. S., & Petrov, D. A. (2014). Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila. *PLOS Genetics*, *10*(11), e1004775.

Bergman, C. M., Quesneville, H., Anxolabéhère, D., & Ashburner, M. (2006). Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome biology*, *7*(11), R112.

Bergman, C. M., Benos, T., Bayraktaroglu, L., Ashburner, M., de Grey, A., Chillemi, J., ... Kaminker, J. (2017a). Drosophila transposable element consensus sequences - v9.44. *https://github.com/cbergman/transposons*, *Last Accessed: 11.01.2019*.

Bergman, C. M., Han, S., Nelson, M. G., Bondarenko, V., & Kozeretska, I. (2017b). Genomic analysis of *P* elements in natural populations of *Drosophila melanogaster*. *PeerJ*, *5*, e3824.

Biémont, C., & Vieira, C. (2006). Junk DNA as an evolutionary force. *Nature*, *443*(7111), 521–524.

Black, D. M., Jackson, M. S., Kidwell, M. G., & Dover, G. A. (1987). KP elements repress P-induced hybrid dysgenesis in *Drosophila melanogaster*. *The EMBO Journal*, *6*(13), 4125–4135.

Bonnivard, E., & Higuet, D. (1999). Stability of European natural populations of *Drosophila melanogaster* with regard to the P-M system: A buffer zone made up of Q populations. *Journal of Evolutionary Biology*, *12*(4), 633–647.

Bonnivard, E., Bazin, C., Denis, B., & Higuet, D. (2000). A scenario for the hobo transposable element invasion, deduced from the structure of natural populations of *Drosophila melanogaster* using tandem TPE repeats. *Genetical Research*, *75*(1), 13–23.

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., & Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell*, *128*(6), 1089–1103.

Brügger, K., Redder, P., She, Q., Confalonieri, F., Zivanovic, Y., & Garrett, R. A. (2002). Mobile elements in archaeal genomes. *FEMS Microbiology Letters*, *206*(2), 131–141.

Burns, K. H. (2017). Transposable elements in cancer. *Nature Reviews Cancer*, *17*(7), 415–424.

Casacuberta, E. (2017). Drosophila: Retrotransposons making up telomeres. *Viruses*, *9*(7), 192.

Casacuberta, E., & González, J. (2013). The impact of transposable elements in environmental adaptation. *Molecular Ecology*, *22*(6), 1503–1517.

Casier, K., Delmarre, V., Gueguen, N., Hermant, C., Viode, E., Vaury, C., ... Boivin, A. (2018). Environmentally-induced epigenetic conversion of a piRNA cluster. *bioRxiv*.

Chessel, D., Dufour, A.-B., & Thioulouse, J. (2004). The ade4 package – I: One-table methods. *R News*, *4*(1), 5–10.

Chu, C., Pei, J., & Wu, Y. (2018). An improved approach for reconstructing consensus repeats from short sequence reads. *BMC Genomics*, *19*(S6), 566.

Daniels, S. B., Peterson, K. R., Strausbaugh, L. D., Kidwell, M. G., & Chovnick, A. (1990). Evidence for horizontal transmission of the P transposable element between *Drosophila* species. *Genetics*, *124*(2), 339–55.

Daniels, S. B., Strausbaugh, L. D., & Armstrong, R. A. (1985). Molecular analysis of P element behavior in *Drosophila simulans* transformants. *Molecular and General Genetics*, *200*(2), 258–265.

Danilevskaya, O., Slot, F., Pavlova, M., & Pardue, M. (1994). Structure of the Drosophila HeT-A transposon: A retrotransposon-like element forming telomeres. *Chromosoma*, *103*(3), 215–224.

Deininger, P. (2011). Alu elements: Know the SINEs. *Genome Biology*, *12*(12), 236.

Dias, E. S., & Carareto, C. M. A. (2012). Ancestral polymorphism and recent invasion of transposable elements in *Drosophila* species. *BMC Evolutionary Biology*, *12*, 119.

Drosophila 12 Genomes Consortium, Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., ... MacCallum, I. (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, *450*(7167), 203–218.

Du, J., Tian, Z., Hans, C. S., Laten, H. M., Cannon, S. B., Jackson, S. A., ... Ma, J. (2010). Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: Insights from genome-wide analysis and multi-specific comparison. *The Plant Journal*, *63*(4), 584–598.

Engels, W. R., Johnson-Schlitz, D. M., Eggleston, W. B., & Sved, J. (1990). High-frequency P element loss in *Drosophila* is homolog dependent. *Cell*, *62*(3), 515–525.

Finnegan, D. (1997). Transposable elements: How non-LTR retrotransposons do it. *Current Biology*, *7*(4), R245–R248.

Finnegan, D. J. (1989). Eukaryotic transposable elements and genome evolution. *Trends in Genetics*, *5*(4), 103–107.

Gascuel, O. (1997). BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, *14*(7), 685–695.

Giordano, J., Ge, Y., Gelfand, Y., Abrusán, G., Benson, G., & Warburton, P. E. (2007). Evolutionary history of mammalian transposons determined by genome-wide defragmentation. *PLOS Computational Biology*, *3*(7), e137.

Goerner-Potvin, P., & Bourque, G. (2018). Computational tools to unmask transposable elements. *Nature Reviews Genetics*, *19*(11), 688–704.

González, J., Karasov, T. L., Messer, P. W., & Petrov, D. A. (2010). Genome-wide patterns of adaptation to temperate environments associated with transposable elements in Drosophila. *PLOS Genetics*, *6*(4), e1000905.

Goriaux, C., Théron, E., Brasset, E., & Vaury, C. (2014). History of the discovery of a master locus producing piRNAs: The *flamenco/COM* locus in *Drosophila melanogaster*. *Frontiers in Genetics*, *5*(257).

Grenier, J. K., Arguello, J. R., Moreira, M. C., Gottipati, S., Mohammed, J., Hackett, S. R., ... Clark, A. G. (2015). Global Diversity Lines – A five-continent reference panel of sequenced *Drosophila melanogaster* strains. *G3: Genes, Genomes, Genetics*, *5*(4), 593–603.

Gunawardane, L. S., Saito, K., Nishida, K. M., Miyoshi, K., Kawamura, Y., Nagami, T., ... Siomi, M. C. (2007). A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in *Drosophila*. *Science*, *315*(5818), 1587–1590.

Haring, E., Hagemann, S., & Pinsker, W. (2000). Ancient and recent horizontal invasions of drosophilids by *P* elements. *Journal of Molecular Evolution*, *51*(6), 577–586.

Heger, A., & Jacob, K. (2018). *pysam: htslib interface for Python*. Retrieved from https://pysam.readthedocs.io/en/latest/

Hua-Van, A., Le Rouzic, A., Boutin, T. S., Filée, J., & Capy, P. (2011). The struggle for life of the genome's selfish architects. *Biology Direct*, *6*(19).

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., ... Wheeler, T. J. (2016). The Dfam database of repetitive DNA families. *Nucleic Acids Research*, *44*(D1), D81–D89.

Jost, L. (2008). GST and its relatives do not measure differentiation. *Molecular Ecology*, *17*(18), 4015–4026.

Kaminker, J. S., Bergman, C. M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., ... Celniker, S. E. (2002). The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biology*, *3*(12), R0084.

Kapun, M., Barrón, M. G., Staubach, F., Vieira, J., Obbard, D. J., Goubert, C., ... González, J. (2018). Genomic analysis of European Drosophila populations reveals longitudinal structure and continent-wide selection. *bioRxiv*.

Katz, Y., Wang, E. T., Silterra, J., Schwartz, S., Wong, B., Thorvaldsdóttir, H., ... Burge, C. B. (2015). Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics*, *31*(14), 2400–2402.

Kazazian Jr, H. H. (2004). Mobile elements: Drivers of genome evolution. *Science*, *303*, 1626–1632.

Kazazian Jr, H. H., Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G., & Antonarakis, S. E. (1988). Haemophilia A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, *332*(6160), 164–166.

Kelleher, E. S. (2016). Reexamining the *P*-element invasion of *Drosophila melanogaster* through the lens of piRNA silencing. *Genetics*, *203*(4), 1513–1531.

Kelleher, E. S., Azevedo, R. B. R., & Zheng, Y. (2018). The evolution of small-RNA-mediated silencing of an invading transposable element. *Genome Biology and Evolution*, *10*(11), 3038–3057.

Khan, H., Smit, A., & Boissinot, S. (2006). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Research*, *16*(1), 78–87.

Kidwell, M. G. (1983). Evolution of hybrid dysgenesis determinants in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences*, *80*(6), 1655–1659.

Kimura, M., & Weiss, G. H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, *49*(4), 561–576.

Koch, P., Platzer, M., & Downie, B. R. (2014). RepARK — *de novo* creation of repeat libraries from whole-genome NGS reads. *Nucleic Acids Research*, *42*(9), e80.

Kofler, R., Hill, T., Nolte, V., Betancourt, A. J., & Schlötterer, C. (2015). The recent invasion of natural *Drosophila simulans* populations by the P-element. *Proceedings of the National Academy of Sciences*, *112*(21), 6659–6663.

Kofler, R. (2018). SimulaTE: simulating complex landscapes of transposable elements of populations. *Bioinformatics*, *34*(8), 1419–1420.

Kofler, R., Senti, K. A., Nolte, V., Tobler, R., & Schlötterer, C. (2018). Molecular dissection of a natural transposable element invasion. *Genome Research*, *28*(6), 824–835.

Le Thomas, A., Stuwe, E., Li, S., Du, J., Marinov, G., Rozhkov, N., ... Aravin, A. A. (2014). Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing the chromatin of piRNA clusters and inducing precursor processing. *Genes and Development*, *28*(15), 1667–1680.

Lerat, E., Rizzon, C., & Biémont, C. (2003). Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. *Genome Research*, *13*(8), 1889–1896.

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., ... Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, *115*(17), 4325–4333.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589–595.

Li, S.-F., Zhang, G.-J., Zhang, X.-J., Yuan, J.-H., Deng, C.-L., Gu, L.-F., & Gao, W.-J. (2016). DPTEdb, an integrative database of transposable elements in dioecious plants. *Database*, *2016*, 1–10.

Mackay, T. F. C., Lyman, R. F., & Jackson, M. S. (1992). Effects of P element insertions on quantitative traits in *Drosophila melanogaster*. *Genetics*, *130*(2), 315–332.

Malachowa, N., & DeLeo, F. R. (2010). Mobile genetic elements of *Staphylococcus aureus*. *Cellular and Molecular Life Sciences*, *67*(18), 3057–3071.

Malik, H. S., Burke, W. D., & Eickbush, T. H. (1999). The age and evolution of non-LTR retrotransposable elements. *Molecular Biology and Evolution*, *16*(6), 793–805.

Malone, C. D., Brennecke, J., Dus, M., Stark, A., McCombie, W. R., Sachidanandam, R., & Hannon, G. J. (2009). Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell*, *137*(3), 522–535.

Marin, L., Lehmann, M., Nouaud, D., Izaabel, H., Anxolabéhère, D., & Ronsseray, S. (2000). *P*-element repression in *Drosophila melanogaster* by a naturally occurring defective telomeric *P* copy. *Genetics*, *155*(4), 1841–1854.

McCullers, T. J., & Steiniger, M. (2017). Transposable elements in *Drosophila*. *Mobile Genetic Elements*, *7*(3), e1318201.

McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 51–56).

Narita, N., Nishio, H., Kitoh, Y., Ishikawa, Y., Ishikawa, Y., Minami, R., ... Matsuo, M. (1993). Insertion of a 5' truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. *Journal of Clinical Investigation*, *91*(5), 1862–1867.

O'Hare, K., & Rubin, G. M. (1983). Structures of P transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. *Cell*, *34*(1), 25–35.

Oliver, K. R., McComb, J. A., & Greene, W. K. (2013). Transposable elements: Powerful contributors to angiosperm evolution and diversity. *Genome Biology and Evolution*, *5*(10), 1886–1901.

Periquet, G., Ronsseray, S., & Hamelin, M. H. (1989). Are *Drosophila melanogaster* populations under a stable geographical differentiation due to the presence of P elements? *Heredity*, *63*, 47–58.

Petrov, D. A., & Hartl, D. L. (1997). Trash DNA is what gets thrown away: high rate of DNA loss in *Drosophila*. *Gene*, *205*(1), 279–289.

Price, A. L., Eskin, E., & Pevzner, P. A. (2004). Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Research*, *14*(11), 2245–2252.

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). *De novo* identification of repeat families in large genomes. *Bioinformatics*, *21*(suppl 1), i351–i358.

Pritham, E. J., & Feschotte, C. (2007). Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. *Proceedings of the National Academy of Sciences*, *104*(6), 1895–1900.

Python Software Foundation. (2017). *Python Language Reference, version 3.6*. Retrieved from http://www.python.org/

Quesneville, H., & Anxolabéhère, D. (1997). A simulation of *P* element horizontal transfer in *Drosophila*. *Genetica*, *100*(1), 295–307.

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Robillard, É., Le Rouzic, A., Zhang, Z., Capy, P., & Hua-Van, A. (2016). Experimental evolution reveals hyperparasitic interactions among transposable elements. *Proceedings of the National Academy of Sciences*, *113*(51), 14763–14768.

Rubin, E., & Levy, A. A. (1997). Abortive gap repair: Underlying mechanism for *Ds* element formation. *Molecular and Cellular Biology*, *17*(11), 6294–6302.

Schaack, S., Gilbert, C., & Feschotte, C. (2010). Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology & Evolution*, *25*(9), 537–46.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., ... Wilson, R. K. (2009). The B73 maize genome: Complexity, diversity, and dynamics. *Science*, *326*(5956), 1112–1115.

Schrader, L., & Schmitz, J. (2018). The impact of transposable elements in adaptive evolution. *Molecular Ecology*, *Epub ahead of print*.

Serrato-Capuchina, A. G., & Matute, D. R. (2018). The role of transposable elements in speciation. *Genes*, *9*(5), E254.

Serrato-Capuchina, A. G., Zhang, S., Martin, W., Peede, D., Earley, E., & Matute, D. R. (2018). Recent invasion of P transposable element into *Drosophila yakuba*. *bioRxiv*.

Shao, F., Wang, J., Xu, H., & Peng, Z. (2018). FishTEDB: A collective database of transposable elements identified in the complete genomes of fish. *Database*, *2018*, 1–9.

Sohn, J. I., & Nam, J. W. (2018). The present and future of *de novo* whole-genome assembly. *Briefings in Bioinformatics*, *19*(1), 23–40.

Sotero-Caio, C. G., Platt, R. N. I., Suh, A., & Ray, D. A. (2017). Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biology and Evolution*, *9*(1), 161–177.

Stadler, K. (2018). *cultevo: Tools, measures and statistical tests for cultural evolution*. R package version 1.0.2. Retrieved from https://kevinstadler.github.io/cultevo/

Studer, A., Zhao, Q., Ross-Ibarra, J., & Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nature Genetics*, *43*(11), 1160–1163.

Tempel, S., & Talla, E. (2015). VisualTE: A graphical interface for transposable element analysis at the genomic scale. *BMC Genomics*, *16*, 139.

Waldor, M. K., Tschäpe, H., & Mekalanos, J. J. (1996). A new type of conjugative transposon encodes resistance to sulfamethoxazole, trimethoprim, and streptomycin in *Vibrio cholerae* O139. *Journal of Bacteriology*, *178*(14), 4157–4165.

Werren, J. H. (2011). Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proceedings of the National Academy of Sciences*, *108*(S2), 10863–10870.

Wicker, T., Matthews, D. E., & Keller, B. (2002). TREP: A database for *Triticeae* repetitive elements. *Trends in Plant Science*, *7*(12), 561–562.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., ... Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*(12), 973–982.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

Wood, J. G., Jones, B. C., Jiang, N., Chang, C., Hosier, S., Wickremesinghe, P., ... Helfand, S. L. (2016). Chromatin-modifying genetic interventions suppress age-associated transposable element activation and extend life span in *Drosophila*. *Proceedings of the National Academy of Sciences*, *113*(40), 11277–11282.

Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, *26*(7), 873–881.

Wu, X., & Burgess, S. M. (2004). Integration target site selection for retroviruses and transposable elements. *Cellular and Molecular Life Sciences*, *61*(19), 2588–2596.

Xing, J., Wang, H., Belancio, V. P., Cordaux, R., Deininger, P. L., & Batzer, M. A. (2006). Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proceedings of the National Academy of Sciences*, *103*(47), 17608–17613.

Yang, P., Wang, Y., & Macfarlan, T. S. (2017). The role of KRAB-ZFPs in transposable element repression and mammalian evolution. *Trends in Genetics*, *33*(11), 871–881.

Yi, F., Ling, J., Xiao, Y., Zhang, H., Ouyang, F., & Wang, J. (2018). ConTEdb: A comprehensive database of transposable elements in conifers. *Database*, *2018*, 1–7.

Yoshitake, Y., Inomata, N., Sano, M., Kato, Y., & Itoh, M. (2018). The *P* element invaded rapidly and caused hybrid dysgenesis in natural populations of *Drosophila simulans* in Japan. *Ecology and Evolution*, *8*(19), 9590–9599.

You, R. N., Kim, W. C., Lee, K. H., Lee, Y. K., Shin, K. S., Cho, K., & Cho, D. H. (2013). REViewer: A tool for linear visualization of repetitive elements within a sequence query. *Genomics*, *102*(4), 209–214.

Zanni, V., Eymery, A., Coiffet, M., Zytnicki, M., & Luyten, I. (2013). Distribution, evolution, and diversity of retrotransposons at the *flamenco* locus reflect the regulatory properties of piRNA clusters. *Proceedings of the National Academy of Sciences*, *110*(49), 19842–19847.