REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
28-08-2018	Article	-

4. TITLE AND SUBTITLE

A maximum-likelihood approach to estimating the insertion frequencies of transposable elements from population sequencing data

5a. CONTRACT NUMBER
W911NF-10-1-0444

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611103

6. AUTHORS

Xiaoqian Jiang, Haixu Tang, Wazim Mohammed Ismail, Michael Lynch, Rebekah Rogers

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

Indiana University at Bloomington
509 East 3rd ST

Bloomington, IN 47401 -3654

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
58126-LS-MUR.53

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

Transposable elements (TEs) contribute to a large fraction of the expansion of many eukaryotic genomes due to the capability of TEs duplicating themselves through transposition. A first step to understanding the roles of TEs in a eukaryotic genome is to characterize the population-wide variation of TE insertions in the species. Here, we present a maximum-likelihood (ML) method for estimating allele frequencies and detecting selection on TE insertions in a diploid population, based on the genotypes at TE insertion sites detected in multiple individuals sampled from the population using paired-end (PE) sequencing reads. Tests of the method on simulated data show that it can

15. SUBJECT TERMS

transposable elements; insertion polymorphism; purifying selection; maximum-likelihood; population genomics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Patricia Foster
UU	UU	UU	UU		19b. TELEPHONE NUMBER 812-855-4084

REPORT DOCUMENTATION PAGE (SF298)
(Continuation Sheet)

Continuation for Block 13

Proposal/Report Number: 58126.53-LS-MUR

Report Title: A maximum-likelihood approach to estimating the insertion frequencies of transposable elements from population sequencing data
Report Type: Article

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 2-Awaiting Publicat
Journal: Molecular Biology and Evolution
Publication Identifier Type: DOI Publication Identifier: 10.1093/molbev/msy152
Volume: Issue: First Page #:
Date Submitted: 8/28/18 12:00AM Date Published: 8/1/18 12:00AM
Publication Location:
Article Title: A maximum-likelihood approach to estimating the insertion frequencies of transposable elements from population sequencing data
Authors: Xiaoqian Jiang, Haixu Tang, Wazim Mohammed Ismail, Michael Lynch, Rebekah Rogers
Keywords: transposable elements; insertion polymorphism; purifying selection; maximum-likelihood; population genomics
Abstract: Transposable elements (TEs) contribute to a large fraction of the expansion of many eukaryotic genomes due to the capability of TEs duplicating themselves through transposition. A first step to understanding the roles of TEs in a eukaryotic genome is to characterize the population-wide variation of TE insertions in the species. Here, we present a maximum-likelihood (ML) method for estimating allele frequencies and detecting selection on TE insertions in a diploid population, based on the genotypes at TE insertion sites detected in multiple individuals sampled from the population using paired-end (PE) sequencing reads. Tests of the method on simulated data show that it can accurately estimate the allele frequencies of TE insertions even when the PE sequencing is conducted at a relatively low coverage (=5X). The method can also detect TE insertions under strong selection, and the detection ability increases with sample size in a population, although a substantial fraction of actual TE...
Distribution Statement: 1
Acknowledged Federal Support: **Y**

A maximum-likelihood approach to estimating the insertion frequencies of transposable elements from population sequencing data

Authors: Xiaoqian Jiang[1*], Haixu Tang[2*], Wazim Mohammed Ismail[2], Michael Lynch[3]

Affiliations:

1, Department of Biology, Indiana University, Bloomington, IN 47405

2, School of Informatics and Computing, Indiana University, Bloomington, IN 47405

3, Center for Mechanisms of Evolution, Arizona State University, Temple, AZ 85287

*Corresponding authors:

Xiaoqian Jiang, Email: jxq198409@hotmail.com

Haixu Tang, Email: hatang@indiana.edu

Abstract

Transposable elements (TEs) contribute to a large fraction of the expansion of many eukaryotic genomes due to the capability of TEs duplicating themselves through transposition. A first step to understanding the roles of TEs in a eukaryotic genome is to characterize the population-wide variation of TE insertions in the species. Here, we present a maximum-likelihood (ML) method for estimating allele frequencies and detecting selection on TE insertions in a diploid population, based on the genotypes at TE insertion sites detected in multiple individuals sampled from the population using paired-end (PE) sequencing reads. Tests of the method on simulated data show that it can accurately estimate the allele frequencies of TE insertions even when the PE sequencing is conducted at a relatively low coverage ($= 5\mathrm{X}$). The method can also detect TE insertions under strong selection, and the detection ability increases with sample size in a population, although a substantial fraction of actual TE insertions under selection may be undetected. Application of the ML method to genomic sequencing data collected from a natural *Daphnia pulex* population shows that, on the one hand, most ($> 90\%$) TE insertions present in the reference *D. pulex* genome are either fixed or nearly fixed (with allele frequencies > 0.95); on the other hand, among the non-reference TE insertions (i.e., those detected in some individuals in the population but absent from the reference genome), the majority ($> 70\%$) are still at low frequencies (< 0.1). Finally, we detected a substantial fraction (~9%) of non-reference TE insertions under selection.

Keywords: transposable elements; insertion polymorphism; purifying selection; maximum-likelihood; population genomics

Short title: ML method to estimate allele frequencies of TE insertions

Introduction

Transposable elements (TEs) are a class of DNA components found in most eukaryotic genomes. Due to their selfish spread within host genomes, TEs play critical roles in shaping the host's genomic architecture. As an extreme example, TEs constitute ~85% of the maize genome (SanMiguel et al. 1996). While in rare cases, TE insertions may contribute to the creation of regulatory sequence (Chuong et al. 2017), in most cases, mobilizations of TEs have negative impacts on gene functions, and thus may result in deleterious mutations (Hurst and Werren 2001). Therefore, studying the insertion polymorphisms of TEs is necessary for understanding the demographic mechanisms by which these aggressive elements spread through host populations.

The spread and maintenance of TEs in a population are determined by the relative rate of gained insertions (i.e., transposition activity) and of TEs loss by selection (i.e., related to fitness effects) (Charlesworth and Langley 1989; Charlesworth and Lapid 1992; Sniegowski and Charlesworth 1994). Given the generally negative effects of TE insertions, host individuals containing excessive TEs often experience selective disadvantages (Pasyukova et al. 2004). TE insertions with strongly deleterious effects will be removed by purifying selection before they have a chance to produce new insertions with less deleterious effects (Blumenstiel et al. 2002; Hazzouri et al. 2008). Only TE insertions with sufficiently mild deleterious effects on the host genome can avoid being eliminated, and thus have opportunities to spread in a population. It was reported that most TEs in *Drosophila* are present at low frequencies, supporting the hypothesis that most TE insertions are selected against (Franchini et al. 2004; Cridland et al. 2013; Barrón et al. 2014). The accumulation of mildly deleterious mutations in TE sequences leads to eventual transposition inactivation of the inserted element (Maside et al. 2005). Therefore, among thousands of TEs identified in various species, only a few of them have been formally demonstrated to be autonomous (Lynch 2007).

In principle, the number of TEs can reach equilibrium in a population with the number of TEs gained by transposition equaling the number lost due to selection and/or excision (Charlesworth and Charlesworth 1983). Estimates of TE insertion frequencies in a population can reveal important information about the genomic structure and the strength of natural selection on the genome, and thus are fundamental in population genomics (Biémont 1992; Yang and Nuzhdin 2003). Previous studies using PCR to survey TE insertions were biased towards insertions with high population frequencies (Petrov et al. 2003; Petrov et al. 2011), while unbiased estimates of frequencies of TE insertions obtained by DNA display in previous studies were limited to only a few TE families (Maside et al. 2001).

With the development and rapidly declining cost of next-generation sequencing (NGS) technologies, new bioinformatics tools have been proposed to characterize TE insertions at the whole-genome scale (Hormozdiari et al. 2010; Ewing 2015; Rishishwar et al. 2017). All of the designed approaches first map NGS reads onto the reference genome, and then search for TE insertions based on paired-end reads or by splitting reads that span the breakpoint of the inserted TE. Specifically, most methods based on paired-end reads take as inputs the paired-end reads, the available reference genome sequence, and a TE sequence library. If one of the two paired-end reads is mapped onto a TE sequence in the library, while the other one is mapped onto a non-repetitive region in the reference genome (i.e., the flanking region of the inserted TE), these paired-end reads support the presence of a TE insertion. This approach is capable of identifying novel TE insertions that are not contained in the reference genome, although it is limited by the availability of element sequences collected in the TE library. Despite the limitation of currently available methods, researchers have moved from the analysis of small datasets to genome-wide data for all TEs in a host genome. Such approaches were summarized in a recent review of TEs in *Drosophila melanogaster* (Barrón et al. 2014). For instance, in a pooled-sequencing sample, it is possible to estimate the frequencies of TE insertions in a population represented by the pooled samples. One previous study developed a software tool PoPoolation to identify almost twice as many "novel" (non-reference) TE insertion sites in a *Drosophila melanogaster* population as the "known" (reference) sites in the reference genome (Kofler et al. 2012).

Although TE insertions are an important type of genomic structural variation, estimates of TE insertion frequencies are not as straightforward as the estimates of frequencies of single nucleotide variations (SNVs) from NGS data. Even though the computational methods mentioned above can detect TE insertions, the high sequence similarity among TE elements in the same family may introduce bias in TE detection. Specifically, current TE detection algorithms may miss some TE insertions due to relatively low read coverage and incompleteness of the reference assembly with repetitive DNA, and thus may underestimate the frequencies of these TE insertions. When the frequencies are estimated from individual genome sequencing data, the detection bias may also affect the genotyping results. For example, a heterozygous insertion site may be interpreted as a homozygote if one of the alleles (present or absent with respect to the novel TE insertion) is not called from the sequencing data. Developing appropriate statistical methods is crucial to address this issue for genome-wide surveys of TE insertions in populations.

Here, we present a maximum-likelihood (ML) method for estimating the frequencies of TE insertions from population genome sequencing data. Combined with bioinformatics pipelines to detect the presence/absence of TE insertions using NGS data, the ML method can estimate the allele frequencies at polymorphic sites of TE insertions after correcting for the bias in TE detection based on the observed genotypes of the sequenced individual genomes. Like other programs for identifying TE insertions using paired-end (PE) reads (Adrion et al. 2017), our TE detection algorithm is unable to discover nested/overlapped TEs or TEs in complex genomic regions (e.g., containing a large number of tandem duplications). Therefore, the method presented here is suitable for genotyping TE insertions identified in well-assembled, non-repetitive genomic regions using whole genome re-sequencing data. Evaluation of the method on simulated data shows that unbiased estimates of TE insertion frequencies can be obtained with low read coverage. Our method can also identify putative TE insertions under strong purifying selection based on the correct estimates of allele frequencies of TE insertions. Application of this method to a natural microcrustacean *Daphnia pulex* population reveals that the allele-frequency distributions of reference and non-reference TE insertions follow different patterns, and that a subset of TE insertions is under negative selection.

Results

Estimating the performance of TE detection algorithms

We developed a bioinformatics pipeline, including two algorithms, to detect the presence and absence of TE insertions at each site in an individual genome (see Methods). To test the performance of our TE detection algorithms, we selected a chromosome representative of scaffold_1 (with length of 1.6 MB nucleotides) in the *D. pulex* assembly *PA42*, randomly inserted TE sequences from the *D. pulex* TE library into the chromosome, simulated Illumina paired-end sequencing reads, and then ran our TE detection algorithms on the simulated sequencing data (see methods). Our results on the simulation study suggested that, under a reasonably high coverage ($20\mathrm{X}$), both the TE-presence and TE-absence detection algorithms showed high accuracy to detect TE insertions, with low false-negative rates (FNR=0 for both algorithms) and low false-positive rates (FNR, 0.03 and 0 for the TE-presence and TE-absence detection algorithms, respectively). Under a relatively low coverage ($10\mathrm{X}$), the false-negative rates increased for both the TE-presence (FNR=0.11) and the TE-absence (FNR=0.12) detection algorithms, while the false-positive rates decreased to FPR=0.005 for the TE-presence detection algorithm (the false-positive rate of the TE-absence detection algorithm remains at 0). These results indicated that our TE

detection algorithms perform well on sequencing data with a high coverage, but may miss some TE insertions when the sequencing coverage is relatively low.

Estimating TE-insertion frequencies on simulated data

We present a maximum-likelihood (ML) method for estimating the allele frequencies of TE insertions in a population based on the number of reads supporting the presence and absence of the TE insertion at each site in each sequenced individual genome (see Methods). We used p_{--}, p_{+-}, and p_{++} to denote the frequencies of genotypes: -- (homozygous, where neither chromosome contains the TE insertion), +- (heterozygous, where only one chromosome contains the TE insertion), and ++ (homozygous, where both chromosomes contain the TE insertion) in the population, respectively, and thus, the allele frequency of the TE insertion (+) is $q = p_{+-}/2 + p_{++}$.

To estimate the performance of the ML method for estimating TE insertion frequencies q, we introduced an important parameter θ in the analysis to account for the relative detection abilities of the presence and absence of TE insertions, with $\theta = 0.5$ when no detection bias exists (see Methods). We first need to estimate the parameter θ in advance. With the genotypes derived from the simulated reads in 100 clones (with a preset TE allele frequency $q = 0.5$), the bias parameter θ is estimated to be 0.51, which is consistent with the observation that our TE detection method has no bias in detecting TEs based on the simulated paired-end reads from a well-assembled genomic region. Consequently, the estimate of allele frequency $q = 0.53$ by the ML method is close to the preset 0.5.

The purpose of this study is to develop an ML method for estimating the TE insertion frequencies under different conditions. Therefore, we simulated different genotypes to further evaluate the performance of our ML method when detection bias indeed exists (see Methods). For each simulation, a total of 100 replicates was conducted for the estimates of p_{--} and p_{+-}, with different preset values for the parameter θ, coverage μ, sample size *N*, inbreeding coefficient *f*, as well as the allele frequency q_0 of the TE insertion under Hardy-Weinberg equilibrium (HWE), selective disadvantages of TE insertions s_1 (for genotype p_{+-}) and s_2 (for p_{++}). To demonstrate the advantage of the ML method, we also estimated the allele frequency by using a naïve method, which simply computes the frequency based on observations from the TE insertion detection algorithms (i.e., the numbers of reads supporting the presence or absence of TE insertions) without correction for detection bias.

We observed that the ML method can yield nearly unbiased estimates of TE-insertion allele frequencies q (FIG. S1). Changes of sample size *N* and inbreeding coefficient *f* do not seem to bias

estimates of q, although the standard deviation of estimated q decreases with the increasing sample size *N*. We note that unbiased estimates of q can be reached by even the naïve method if the depth of coverage is moderate or high ($\mu \geq 10\text{X}$) or no strong bias in TE-insertion detection algorithms ($0.2 \leq \theta \leq 0.8$).

In contrast, for sites with relatively low read coverage μ ($\leq 10\text{X}$) and/or strong bias in detecting TE insertions ($\theta \geq 0.9$ or $\theta \leq 0.1$), where some reads supporting the presence or absence of TE insertions may not be detected, the estimates of q are inaccurate by the naïve method (FIG. 2). For $\theta < 0.5$, where the TE-presence detection algorithm is more likely to miss reads than the TE-absence detection algorithm, i.e., a heterozygote (+-) may be interpreted as a homozygote (--), the naïve method underestimates q (FIG. 2A,C,D), whereas for $\theta > 0.5$, where the TE-absence detection algorithm is more likely to miss reads than the TE-presence detection algorithm, i.e., a heterozygote (+-) may be interpreted as homozygote (++), the naïve method overestimates q (FIG. 2B). In these cases, the parameter θ introduced in the ML method is crucial to correct this bias, resulting in more accurate estimates of q than those of the naïve method. Even in the cases with a very low depth of coverage ($\mu = 3\text{X}$ in FIG. 2C) or a strong bias ($\theta = 0.9$ in FIG. 2B), although the ML method cannot give very accurate estimates of q due to insufficient numbers of supporting reads, it still significantly improves the accuracy of estimates of q compared to the naïve method.

The accuracies of the observed genotypes (++, +-, --) depend on the product of μ and θ: if $1/\mu \leq \theta \leq 1 - 1/\mu$, the observations are consistent with the true genotypes of the individuals, and thus q can be estimated accurately by the naïve method. Notably, although θ is known in the simulation experiment, the ML method assumes it is unknown and derives $\hat{\theta}$ from inferred heterozygous individuals, which is very accurate when the sequencing coverage is high ($\mu \geq 20\text{X}$) but leads to slight deviations from θ when the coverage is moderate or low ($\mu \leq 10\text{X}$). For instance, when $\mu = 10\text{X}$, we estimated $\hat{\theta} = 0.28$ when $\theta = 0.3$ is simulated. However, it seems that such slight deviation will not influence the accuracy of estimating q using the ML method.

Finally, as expected, increasing the sample size (*N*) can improve the accuracies of estimates of $\hat{\theta}$ and allele frequency $\hat{q}$. However, for a very low coverage $\mu = 3\text{X}$, even a large sample size $N = 500$ cannot further improve the accuracy of the estimates of q (data not shown), indicating that a modest sequencing coverage is required for the effective application of the ML method.

Detecting selection on TE insertions in simulated data

The ML method can also detect putative selection on TE insertions (see Methods). To evaluate the power of the ML method for detecting selection, we simulated data with different values of q_0, s_1, and s_2 under three selection models with inbreeding coefficient *f* known in advance: the recessive-effect model ($s_1 = 0$ and $s_2 \neq 0$), the additive-effect model ($s_2 = 2s_1$), and the dominant-effect model ($s_2 = s_1$). For each model, a series of q_0 ranging from 0.1 to 0.9 were simulated with sample sizes $N = 100$ or $N = 500$, in an attempt to test the impact of *N* on the statistical power of the ML method. The other parameters are set as coverage $\mu = 20\text{X}$, parameter $\theta = 0.3$, and inbreeding coefficient $f = 0.02$, which are comparable with those of the *D. pulex* KAP population sequencing data. We repeated the simulation 100 times for each parameter setting. The log-likelihood ratio test (LRT) was used to test if the maximum likelihood computed using the ML method is significantly better ($\text{p} - \text{value} \leq 0.01$) than that under the neutral model ($s_1 = s_2 = 0$). If the neutral (null) model was rejected, we concluded that selection is present.

Overall, the ML method has a low false-positive rate on detecting selection, generally ≤ 0.01 regardless of q_0, and is essentially zero in most cases (FIG. 3, for the recessive-effects model). On the other hand, the statistical power of the method is relatively low, but increases with the larger sample size N, as indicated by true-positive rate (TPR) in FIG.3. Although the statistical power does not linearly increase with the strength of selection (s_1 or s_2), sites under strong selection ($s_2 > 0.8$) can be detected in most cases where the TE insertion frequency is relatively high ($q_0 > 0.5$). For instance, the TPR reached 1.0 when $0.3 < q_0 < 0.9$ and $s_2 > 0.65$ for the sample size $N = 500$. In contrast, weak selection ($s_2 < 0.2$) is difficult to detect using the ML method regardless of q_0, indicating the limitation of the ML method. However, TE-insertion sites under strong negative selections generally have little chance to reach very high allele frequencies; therefore, the TE-insertion with high allele frequency in a natural population is unlikely under strong selection.

Among the three selection models, the additive-effect model is not as easily detected compared with the other two models (see FIG. S2). For the TE insertion sites detected under selection, we examined which model fits the data better by using the LRT statistics. We found that selection can be detected by the LRT when any of the three selection models is assumed. However, the selection model used in the simulation cannot be distinguished from the other two models in all examined cases, as the likelihood value computed under the true model is not significantly greater than the likelihood computed under the other two models. Moreover, the ML method cannot report accurate estimates of the selection strengths (i.e., s_1 or s_2) because different combination of q_0, s_1 or s_2 may produce similar observed

genotype frequencies (see simulation results in FIG. S3). For instance, we cannot distinguish the genotype frequencies under strong selection but with a high allele frequency ($s_2 > 0.8$ and $q_0 = 0.9$) from the frequencies under a neutral model but with a low allele frequency $q_0 < 0.9$ (FIG. S3. A2).

Genome-wide identification of TE insertion sites in *D. pulex* KAP population

We applied our method to a population of *D. pulex*, a microcrustacen commonly found in ephemeral ponds and a well-documented model system in ecological and evolutionary genomics. The *D. pulex* population studied here reproduces by parthenogenesis for 3 to 5 generations each year but also undergoes obligate sexual reproduction as the pond dries up annually. An early season sample with 96 clones was collected; sequencing reads were obtained on these clones and then were used to identify TE insertions in each clone (see method). The *D. pulex* TE library as input of our TE detection algorithms have been established in a previous study, which is a comprehensive *D. pulex* TE library annotated from two available *D. pulex* reference genome assemblies (*TCO* and *PA42*), including 1,461 full length and 27,849 fragmented TE sequences (Jiang et al. 2017).

On average, reference (present in the *PA42* assembly) and non-reference (absent in the *PA42* assembly) TE insertion sites were detected in 65 and 64 clones with the sequencing coverage of ~20X, respectively. The mean reads count supporting the reference and non-reference TE insertion sites among these clones are ~30X and ~18X, respectively. We initially identified on average ~7,000 TE insertions in each clone using the TE-presence detection algorithm. Under the condition that the presence and/or absence of TE insertion have to be supported by at least 5 reads in >=50 clones, a total of 17,658 polymorphic TE insertion sites were identified in the KAP population, including 2,263 (~13%) reference TE insertion sites and 15,395 (~87%) non-reference TE insertion sites. Compared with the previous comprehensive comparison of TE insertion difference between *D. pulex* asexual and sexual groups, we identified fewer TE insertions (17,658 in this study vs 19,301 in 8 sexual *D. pulex* clones, Jiang et al. 2017). For the purpose of estimating allele frequencies, we used a strict filtering process in this study that only TE insertion (present or absent) sites supported by at least 5 reads in >=50 clones were retained, which is the main reason for this discrepancy. Under such strict conditions, our TE detection algorithm may have abandoned a substantial number of TE insertion sites detected only in a small number of clones. Moreover, as stated above, like many other TE detection algorithms based on paired-end reads, our method is unable to detect nested TEs or TEs in complex genomic regions. Thus, our estimates should be interpreted as a lower bound of the actual numbers of TEs in the population.

In order to estimate the TE allele frequencies, we first estimated the parameter θ. We treated the reference and non-reference TE insertions separately in our analysis as reads mapping may be differential for the regions with or without a TE in the reference genome. Using the number of reads supporting the presence and absence of TE insertions in the heterozygous clones in a population, we estimated that the average θ across TE insertion sites for reference and non-reference TE insertions in the KAP population are 0.78 and 0.35, respectively. Note that the parameter θ deviates from 0.5 in both cases, indicating that our TE-insertion detection algorithms would likely overestimate the allele frequencies of reference TE insertions and underestimate the allele frequencies of non-reference TE insertions if the bias was not corrected. Hence, we used different θ in the analyses for reference and non-reference TE insertions. The average inbreeding coefficient of KAP population is estimated as $f = 0.02$ in a previous study (Lynch et al. 2016), implying that the KAP population is close to HWE but is slightly inbred. All estimates of parameters used in the ML analysis are summarized in Table 2.

Our results show that about two-thirds of the 17,658 TE insertions have low allele frequencies (63.4% insertions with allele frequencies < 0.10), some are fixed or nearly fixed (14.2% insertions with frequencies > 0.95), with the remaining sites segregating at intermediate frequencies (22.4% insertions with frequencies between 0.10 and 0.95). The distributions of allele frequencies of reference and non-reference TE insertions are quite different: most reference TE insertions are nearly fixed (93.9% with allele frequencies > 0.95) and few have low frequencies (0.3% with allele frequencies < 0.10), whereas more than two-thirds of non-reference TE insertions have low frequencies (72.7% with frequencies < 0.10) and few have high frequencies (2.6% with frequencies > 0.95). The distributions of TE insertion frequencies are shown in FIG. 4.

As shown in the simulation results section, although the ML method may detect selection on the TE insertions with very high allele frequencies (> 0.9), TE insertions under strong selection in a natural population are unlikely to have arisen to such frequencies. Moreover, the false-positive rates in detecting selection are relatively high ($\mathrm{FPR} \approx 0.05$) in some cases with low sequencing coverage $\mu = 5\mathrm{X}$ (see FIG. S4). Therefore, we only tried to search for potential selection on the TE insertion sites with allele frequencies < 0.9 and in clones with high sequencing coverage (10X). In total, 18 reference and 8,189 non-reference TE insertions were retained for this analysis. The ML method revealed that 4 of 18 reference, and 734 of 8,189 (~7.9%) non-reference TE insertion sites are under potential negative selection in the KAP population. We noted that all TE insertion sites inferred to be under selection have allele frequencies q between 0.3 and 0.7. Only TE insertions with mild deleterious effects on host fitness

may reach a relatively high allele frequency. Therefore, it is not surprising that we did not detect any TE insertion sites with high allele frequencies (> 0.7) under selection.

The ML method is relatively conservative in detecting selection and cannot detect weak purifying selection on TE insertions, especially for those with low frequencies, as demonstrated in our simulation experiments. This is one possible reason for why only a small fraction of novel TE insertions is detected to be under selection in the KAP population. On the other hand, the clones sequenced in the KAP population were collected from hatchlings, and thus the selection effect on these clones may not be significant, which may also explain our observation.

Discussion

High-throughput genome sequencing has become an important approach to characterizing genome-wide structural variations, including the mapping of TE insertion sites and insertion polymorphisms. Several tools have been developed to identify novel TE insertions from paired-end reads (Keane et al. 2013; Lee et al. 2014; Zhuang et al. 2014). These methods can be exploited on individual sequencing data to characterize genome-wide TE insertion polymorphisms in a population from which the genomes are sampled. However, to accurately estimate the allele frequencies of TE insertions at population-levels for a diploid organism, we need to address two additional issues.

For diploid organisms, we need to detect the genotypes of TE insertions based on the number of reads supporting both the presence and absence of TE insertions, Some existing algorithms using paired-end reads can report such information (Cridland et al. 2013; Adrion et al. 2017). However, the heterozygous individuals may not be distinguished from the homozygous individuals in the condition of low coverage. Second, some TE detecting algorithms were designed for using pooled-sequencing data, which assume that the pooled sample is a good representative of a population (Kofler et al. 2012). However, this is not always the case. Furthermore, the assumption that the number of reads supporting presence of a TE insertion is proportional to the frequency of TE insertion in the pooled-sequencing data may not hold either, due to the potential bias in detecting the presence and absence of a TE insertion, respectively. As shown in our results, such bias existed in the natural *D. pulex* population data: the parameter θ in our method deviated from 0.5 (no bias), and θ was distinct for reference and non-reference TE insertions. We emphasized that, although similar algorithms have been applied to the characterization of TE insertion polymorphisms from pooled-sequencing data in previous studies, the detection bias was also observed in other algorithms for direct estimates of allele frequencies of TE insertions (Kofler et al.

2016). Therefore, even though individual genome sequencing is costly compared to the widely employed pooled-sequencing approach (Schlotterer et al. 2014), it may provide more accurate estimates of allele frequencies of TE insertions.

In this paper, we developed a ML method and utilized simulated data to demonstrate the accuracy of the ML method to estimate allele frequencies of TE insertions when detection bias is present. As shown in our simulation results, estimates of allele frequencies deviate from the true data when sequencing coverage is not sufficiently high (FIG. 1). The parameter θ introduced in the ML process is necessary to correct potential detection bias in estimating allele-frequency of TE insertions. The ML method performs well even when the coverage is moderately low ($\mu = 5\mathrm{X}$), indicating that it is ready to be applied to genome-wide characterization of TE polymorphisms in well-assembled genomic regions in population sequencing projects.

An advantage of the ML framework is that it allows the assessment of significance of a putative selection model through a LRT statistic. A correctly specified likelihood function can be used to evaluate if selection models are statistically significantly better than neutral models for explaining population data, as long as the inbreeding coefficient f is known. Purifying selection potentially affects the allele frequencies of TE insertions, as only TE insertions with no significant selective disadvantage are unconstrained to drift to high frequencies. Nevertheless, detection of the most likely selection model is a critical challenge in the analysis of TE polymorphism. Although a large sample size improves the statistical power for detecting selection (FIG. 3), it is generally difficult to determine the most likely selection model, especially for the TE insertions with low allele frequencies (< 0.2). This is because the frequencies of TE insertions that subjected to purifying selection in the population will not change much in one generation. For instance, for the recessive-effect model case ($s_1 = 0$), as long as the TE frequency in a population is low, selection will not be efficient in reducing the frequencies of deleterious TE insertions because deleterious insertions may be hidden in heterozygotes. As a result, it is almost impossible to rid a population of recessive deleterious TE insertions, even when such TE insertions are lethal in a homozygous state (i.e., $s_2 = 1$).

Although some examples have shown that TE insertions may regulate expression levels of the neighboring genes and bring some benefits to the host genomes (Daborn et al. 2002; Aminetzach et al. 2005), most TE insertions are considered to be deleterious or neutral to the host (Hurst and Werren 2001). Therefore, we only considered the negative selection of TE insertions in this study. Despite the ML method is conservative to detect selection, the observations that most TEs are present at low

population frequencies in *D. pulex* and a fraction of novel TE insertion sites are under selection support the concept that TE insertions are mostly deleterious to the host genome.

A previous study proposed a method to infer the probability distribution of the TE insertion allele frequencies in a population, based on the age of TE insertions (Blumenstiel et al. 2014). Both the previous and our method provide conservative estimates of the selection on TE insertions and are suitable for analyzing re-sequencing data in which TEs are identified in a well-assembled genome, and both assumed TE insertion alleles are independent. Without the assumptions of a constant transposition rate, the previous study also provided the evidence of negative selection against most TE insertions, which is consistent with our results. The previous method also detected a small subset of TEs under positive selection. But it cannot be used for DNA transposons due to the challenge of inferring their age, indicating that our method may be complementary to the previous method. Even with its own limitations, the ML approach developed in our study represents a new method for detecting TE insertions under selection only using paired-end sequencing data.

Recently, in an effort to establish *D. pulex* as a model system for evolutionary genomics, a project to sequence ~100 clone isolates from each of ~30 *D. pulex* populations throughout the geographic range of the species has been initiated (Lynch et al. 2016). Inspired by these available paired-end sequencing data, we have developed here a bioinformatics pipeline for the identification of TE insertions, estimating allele frequencies of TE insertions, and detecting selection on TE insertion sites in a population. By applying our pipeline to a natural *D. pulex* population data, we computed the distribution of population-level allele frequencies of TE insertions. We observed a distribution of TE insertion frequencies in *D. pulex* similar to that reported in *Drosophila melanogaster* (Kofler et al. 2012), i.e., most TEs identified in the KAP population were not present in the *PA42* reference genome, while a high fraction of TE insertions are segregating at very low frequencies, e.g., a high fraction (73%) of novel TE insertions in the KAP population occur at frequencies < 0.1.

Much like other algorithms for identifying TE insertions using paired-end reads, our TE detection algorithms are unable to discover TEs located in complex genomic regions, and the detection power also depends on the completeness of the TE library and the quality of the reference genome assembly. Even for the well-assembled *Drosophila melanogaster* genome, long repeats are common in heterochromatin and it is difficult to detect TEs in heterochromatic regions using paired-end sequencing data (Chakraborty et al. 2018). TE abundance and diversity are highly variable in different species. Therefore, new methods for accurately mapping the reads are still desirable. In this work, we focused on our ML

method to correct the bias in estimating allele frequencies once we obtained the genotypes of TE insertions, although we may miss some variants of TE insertions in the population (and thus cannot estimate their allele frequencies). However, our ML method performs better than the naïve method in estimating TE allele frequencies in all simulated conditions.

The framework outlined here is the first of its kind to provide accurate estimates of allele frequencies of TE insertions using individual sequencing data in a population, as well as attempting to detect potential selection on TE insertion sites without additional genomic information. It is worth noting that the method can be applied to other type of insertion polymorphism, and is not limited to TE insertions.

Materials and methods

Identification of TE insertion sites using paired-end NGS data

In a diploid genome, two alleles can be observed at each TE insertion site: *presence* of a TE insertion (denoted as +) or *absence* of a TE insertion (denoted as -). To infer the allele frequencies of TE insertions (+), we first need to obtain the frequencies of genotypes --, +-, and ++ in a population. We developed a bioinformatics pipeline to detect the presence and absence of TE insertions at each site in an individual genome from its paired-end whole genome sequencing data.

First, a previously published graph-based algorithm was used to identify potential TE insertions from the paired-end (PE) reads (Lee et al. 2014). This algorithm takes as input a reference genome, a TE library, and PE sequencing data, and reports a list of read-pair clusters to identify TE insertions (i.e., the allele +). Second, we implemented an additional python script to detect the absence of a TE (i.e., the allele -) as follows. PE reads were first mapped to a reference genome with BWA (Li 2013) and were then located onto unambiguous positions using Samtools (Li et al. 2009). Here, we classified the *putative* TE insertion sites into two types: *reference* if the TE insertion is present in the reference genome, and *non-reference* if the TE insertion is absent from the reference genome (FIG. 1). Two different criteria were applied for determining the absence of reference and non-reference TE insertions at a site, respectively: 1) for a reference TE insertion site, if one or more PE reads are mapped across the flanking regions of the TE on the reference genome; 2) for a non-reference TE insertion site, if the distance between the read pairs mapped on the reference genome is consistent with the insert size of the PE read library (FIG. 1A). This algorithm also reports the numbers of PE reads supporting the absence of TE insertions at each site (FIG. 1B). We note that our TE detection methods are not designed to detect

nested TEs or TEs in complex regions, and thus our estimates may provide a lower bound for the actual numbers of TEs in each clone.

Estimating false-negative/positive rates (FNRs/FPRs) for TE detection algorithms

To estimate FNRs and FPRs of the two TE detection algorithms, we simulated paired-end sequencing reads using a recent published algorithm simulaTE (Kofler 2017) and analyzed the simulated data. We generated one TE-free clean chromosome with RepeatMasker (Smit et al. 2004), using scaffold_1 of *D. pulex* assembly *PA42* (~1.6 MB). Next, we randomly inserted a set of 100 TEs chosen from *D. pulex* TE library with the length about 2 KB (Jiang et al. 2017). To detect the TE insertions, we assigned the simulated TE insertions with two different frequencies, 1.0 (presence) and 0 (absence). Finally, we simulated Illumina PE reads with a read length of 100 bp, insert size of 300 ± 50 bp, and two coverages 20X and 10X, respectively (see supplemental materials for parameter setting). We then identified TE insertions using our TE detection algorithms with the simulated PE reads, the TE library, and the clean chromosome. This process was conducted in 100 simulated clones.

False-positive rates were estimated for TE insertions as FPR=FP/TN, where FP is the number of discovered TEs falsely inferred to be insertions, and SN is the number of the simulated TE insertions. False-negative rates were estimated as FNR=FN/SN, where FN is the number of simulated TE insertions that were not identified. The average FPR and FNR across 100 simulated individuals was reported, providing the estimates of accuracies for the detection algorithm used in this study.

Estimating the allele frequencies of TE insertions by a ML approach

The ML method for estimating the allele frequencies of TE insertions in a population takes as input the numbers of reads supporting the presence and absence of the TE insertion at each site in each sequenced individual genome (as obtained using the algorithms described above).

Letting n_i^+ and n_i^- be the numbers of reads from an individual genome *i* that support alleles + and - at a specific TE insertion site, respectively, the likelihood of observing n_i^+ and n_i^- supporting reads in an individual *i*, based on the genotype frequencies, is

$$P(n_i^+, n_i^-) = p_{--}\Phi_{--} + p_{+-}\Phi_{+-} + p_{++}\Phi_{++} \ , \qquad (1)$$

where p_{--}, p_{+-}, and p_{++} denote the frequencies of genotypes --, +-, and ++ in the population, and the Φ_j denote the conditional probabilities of observing the supporting reads provided the individual *i* is of genotype *j* = ++, +-, or --.

The basic assumption in our model is that while all detected read pairs supporting the presence or absence of a TE insertion are true, some reads may not be detected by the TE detection algorithm due to low coverage. Given the inputs of n_i^+ and n_i^- from a number of individuals, there are three types of observations: if $n_i^+ > 0$ and $n_i^- = 0$, the individual *i* may have the genotype +- or ++; if $n_i^- > 0$ and $n_i^+ = 0$, the individual *i* may have the genotype +- or --; and if $n_i^+ > 0$ and $n_i^- > 0$, the individual *i* must have the heterozygous genotype +-. The general expressions for Φ_j and the corresponding likelihood conditional on different observations (i.e., n_i^+ and n_i^-) are summarized in Table 1.

Note that we introduce a parameter θ in the analysis (see Table 1) to account for the relative detect abilities of the two alleles + and - by using paired-end (PE) reads. We define θ and $(1-\theta)$ as the average probabilities of sampling TE-presence and TE-absence allele, respectively, in heterozygotes (genotype +-). θ is assumed to be different for reference and non-reference TE insertion sites in a population, respectively, but is constant across the genome for both cases, and thus can be estimated independently prior to the ML analysis, by setting $\theta/(1-\theta)$ to be the ratio of the numbers of reads supporting alleles - and + in all definitive heterozygous individuals across either the reference or non-reference TE insertion sites in the population, i.e., $\theta = [\sum_{i=1}^{N_h}(n_i^+/n_i^-)/n_h]/[1+\sum_{i=1}^{N_h}(n_i^+/n_i^-)/n_h]$, where n_h is the total number of reference or non-reference heterozygotes.

Given the supporting reads in *N* individuals sampled in a population, the log likelihood of the full set of data in a single locus is

$$L = \sum_{i=1}^{N} \ln P(n_i^+, n_i^-) \qquad (2)$$

The ML solution of the actual genotype frequencies ($\hat{p}_{--}$, $\hat{p}_{+-}$, and $\hat{p}_{++} = 1 - \hat{p}_{--} - \hat{p}_{+-}$) is obtained by maximizing L with respect to n_i^+ and n_i^-. In this paper, we used a grid search to solve the ML maximization problem, allowing potential solutions of p_{--} and p_{+-} as m/N, where m is an integer between 0 and *N*. The estimated allele frequency of TE insertions in the population is then

$$\hat{q} = \hat{p}_{+-}/2 + \hat{p}_{++} \qquad (3)$$

Detecting purifying selection on TE insertions

The ML method can be used to detect putative selection on TE insertions. When no selection and no inbreeding occurs, the genotype frequencies are expected to follow the Hardy-Weinberg equilibrium (HWE), i.e., the expected frequencies of --, +-, and ++ are $(1-q_0)^2$, $2q_0(1-q_0)$, and ${q_0}^2$, respectively, where q_0 is the allele frequency of the TE insertion. When selection occurs, the relative contributions of

the three genotypes to the next generation become $(1-q_0)^2 w_{--}$, $2q_0(1-q_0)w_{+-}$, and ${q_0}^2 w_{++}$, respectively (note that these values should be normalized to obtain the actual genotype frequencies), where w_{--}, w_{+-}, and w_{++} are the fitness of the genotypes --, +-, and ++, respectively. Letting s_1 and s_2 denote the *selective disadvantages* of the TE insertions in the genotypes +- and ++, respectively, the fitness of the three genotypes --, +-, and ++ can be written as $w_{--} = 1$, $w_{+-} = 1 - s_1$, and $w_{++} = 1 - s_2$, respectively.

Finally, note that the expected genotype frequencies will further differ from the above if the inbreeding coefficient f measuring the deviation from HWE prior to selection is nonzero. Here, we assume f can be estimated from the genotype frequencies across single nucleotide polymorphic (SNP) sites unassociated with the TE insertions from population genomic data. Thus, the expected frequencies of genotypes p_{--}, p_{+-}, and p_{++} of a TE insertion are:

$$E(p_{--}) = [(1-q_0)^2 + f q_0(1-q_0)]/\bar{W} \quad (4a)$$

$$E(p_{+-}) = 2q_0(1-q_0)(1-f)(1-s_1)/\bar{W} \quad (4b)$$

$$E(p_{++}) = [{q_0}^2 + f q_0(1-q_0)](1-s_2)/\bar{W} \quad (4c)$$

, where $\bar{W} = 1 - 2q_0(1-q_0)(1-f)s_1 - [{q_0}^2 + f q_0(1-q_0)]s_2$ is a normalization factor. Note that q_0 represents the frequency of the allele + (containing TE) prior to selection, which is distinct from the observed allele frequency in the population (denoted as q in the previous section).

The parameters in equations (4a-4c) to be solved for by ML become q_0, s_1, and s_2 (instead of p_{--} and p_{+-}). In the simplest case of no selection ($s_1 = s_2 = 0$), the likelihood equations need only be solved for one unknown parameter $\hat{q}_0$. By contrasting the selection and neutral models, one can test the hypothesis that no selection occurs at each TE insertion site using a log-likelihood ratio test (LRT) statistic,

$$LRT = 2(LL_1 - LL_0)\,, \quad (5)$$

where LL_0 denotes the log-likelihood under the null hypothesis, i.e., the neutral (no selection) model ($s_1 = s_2 = 0$), while LL_1 denotes the log-likelihood under the alternative hypothesis in which the likelihood function (Equation 2) is maximized subject to p_{--} and p_{+-}. The LRT statistic approximately follows a χ^2 distribution with one or two degree of freedom (s_1 and/or s_2). If the neutral model ($s_1 = s_2 = 0$) is rejected by the LRT statistic, we conclude that the TE insertion site is under selection. We

developed several perl and python scripts to evaluate the ML parameters and the LRT statistic for detecting selection.

Evaluation of ML approach by simulation

We used computer simulation to evaluate two aspects of the performance of the ML approach based on genotyping data: the accuracy of allele-frequency estimation, and the power to detect selection on TE insertions. We started our simulation from the expected genotype frequencies in the population computed using the different combinations of preset parameters *f* and θ to mimic the potential bias in detecting the presence and absence of TE insertions from paired-end sequencing data. The parameters q_0, s_1, and s_2 were set based on the neutral model or three biological selection models. Any alternative models allowing for selection ($s_2 \neq 0$ and/or $s_1 \neq 0$) can be written as $s_1 = \lambda s_2$, where each different λ represents a different selection model. The three selection models considered here include: $\lambda = 0$, representing the recessive-effect model ($s_1 = 0$ and $s_2 \neq 0$); $\lambda = 0.5$, representing the additive-effect model ($s_2 = 2s_1$); and $\lambda = 1.0$, representing the dominant-effect model ($s_2 = s_1$).

The *expected* genotype (++, +-, or --) at each TE insertion site in a simulated individual is then sampled based on the genotype frequencies according to equations (4a-4c). Finally, we simulated the numbers of reads supporting the presence or absence of TE insertions (i.e., n_i^+ or n_i^-) in the individual based on its expected genotype and the parameters of θ and coverage μ.

In the simulations, the allele frequency q_0 was assigned a value between 0 and 1, and the selection coefficients s_1 and s_2 were set to be between 0 and 1 as TE insertions are considered to be generally deleterious (under purifying selection). The remaining parameters were also preset in line with some population sequencing projects, including the *Daphnia* project data to be analyzed here (see the next section): sample size *N* was set to be between 30 and 500; the inbreeding coefficient f was set to be between 0.0 and 0.30; the expected sequencing coverage μ (see below) was set to be between 3 and 30; and the bias parameter θ was assigned a value between 0.1 and 0.9.

For each simulation experiment, the parameters *N*, θ, and f were kept constant, and the actual sequencing coverage of an individual was drawn from a Poisson distribution with the mean of μ. We implemented this simulation process in Python, and the simulation results were used to evaluate the accuracy of the estimated TE insertion frequency q and the detected TE insertion sites under selection using the ML method. All code and data used in this study are released in Github (https://github.com/xiaoqian1984/TE_detection).

Application to *D. pulex* population sequencing data

In this study, the *D. pulex* population sequencing data includes 96 clones, and was collected in the spring of 2013 from Kicka Pond (KAP), located in Illinois, immediately after resting-egg hatchling to minimize the chance of sampling genetically identical individuals. These clones were maintained in benign laboratory conditions to keep animals reproducing parthenogenetically. DNA was extracted from each clone using a cetyltrimethylammounium bromide method (Doyle 1987). Paired-end sequencing performed on the Illumina MiSeq platform generated 100 or 150 bp reads. The raw reads of this study have been previously deposited in NCBI Sequence Read Archive (SRA) with accession number SRS1785808 (Lynch et al. 2016).

Raw reads from all clones were first preprocessed for quality control by using Trimmomatic and Fast-Toolkit (Gordon and Hannon 2010; Bolger et al. 2014). Several subsequent steps were used to further filter the data: 1) we first abandoned clones with low genome coverage ($\leq 5\mathrm{X}$); 2) we then applied a published tool MAPGD, which uses population-genomic data to estimate pairwise relatedness, to prevent the inadvertent inclusion of clone mates in the original clone collection (Ackerman et al. 2017). 3) we abandoned the sequencing data of clones with average insert size < 250 bp of PE reads because our TE insertion identification algorithm cannot find TE insertions if PE reads have short insert size. After these filtering steps, a total of 73 clones were retained in the KAP population for further analysis, with average coverage ~20X and average insert size ~340 bp of PE read library.

The input of TE identification algorithms include: a *D. pulex PA42* assembly as the reference genome (Ye et al. 2017); a previously established *D. pulex* TE library consisting of 1,461 full-length and 27,849 fragmented TEs (Jiang et al. 2017); and the whole-genome sequencing data from each of the 73 *D. pulex* clones from the KAP population. We abandoned TE insertion sites with < 5 supporting PE reads in each clone, and we also excluded the TE insertion sites covered in < 50 clones in the population.

Acknowledgements

This project was supported by National Science Foundation (DBI-1262588 to H. Tang and DEB-1257806 to M. Lynch), National Institutes of Health (1R01-GM101672 and R35GM122566-01 to M. Lynch).

Reference

Ackerman MS, Johri P, Spitze K, Xu S, Doak TG, Young K, Lynch M. 2017. Estimating seven coefficients of pairwise relatedness using population genomic data. *Genetics*:genetics. 116.190660.

Adrion JR, Song MJ, Schrider DR, Hahn MW, Schaack S. 2017. Genome-wide estimates of transposable element insertion and deletion rates in *Drosophila melanogaster*. *Genome Biol Evol* 9:1329-1340.

Aminetzach YT, Macpherson JM, Petrov DA. 2005. Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*. *Science* 309:764-767.

Barrón MG, Fiston-Lavier A-S, Petrov DA, González J. 2014. Population genomics of transposable elements in *Drosophila*. *Annu Rev Genet* 48:561-581.

Biémont C. 1992. Population genetics of transposable DNA elements. *Genetica* 86:67-84.

Blumenstiel JP, Chen X, He M, Bergman CM. 2014. An age-of-allele test of neutrality for transposable element insertions. *Genetics* 196:523-538.

Blumenstiel JP, Hartl DL, Lozovsky ER. 2002. Patterns of insertion and deletion in contrasting chromatin domains. *Mol Biol Evol* 19:2211-2225.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*:btu170.

Chakraborty M, VanKuren NW, Zhao R, Zhang X, Kalsow S, Emerson J. 2018. Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nat Genet* 50:20.

Charlesworth B, Charlesworth D. 1983. The population dynamics of transposable elements. *Genet Res* 42:1-27.

Charlesworth B, Langley CH. 1989. The population genetics of *Drosophila* transposable elements. *Annu Rev Genet* 23:251-287.

Charlesworth B, Lapid A. 1992. The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster.* I. Element frequencies and distribution. *Genet Res* 60:103-114.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* 18:71.

Cridland JM, Macdonald SJ, Long AD, Thornton KR. 2013. Abundance and distribution of transposable elements in two *Drosophila* QTL mapping resources. *Mol Biol Evol* 30:2311-2327.

Daborn P, Yen J, Bogwitz M, Le Goff G, Feil E, Jeffers S, Tijet N, Perry T, Heckel D, Batterham P. 2002. A single P450 allele associated with insecticide resistance in *Drosophila*. *Science* 297:2253-2256.

Doyle JJ. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull Bot Soc Am* 19:11-15.

Ewing AD. 2015. Transposable element detection from whole genome sequence data. *Mob DNA* 6:1.

Franchini LF, Ganko EW, McDonald JF. 2004. Retrotransposon-gene associations are widespread among *D. melanogaster* populations. *Mol Biol Evol* 21:1323-1331.

Gordon A, Hannon G. 2010. Fastx-toolkit. *Computer program distributed by the author, website http://hannonlab. cshl. edu/fastx_toolkit/index. html [accessed 2014–2015]*.

Hazzouri KM, Mohajer A, Dejak SI, Otto SP, Wright SI. 2008. Contrasting patterns of transposable-element insertion polymorphism and nucleotide diversity in autotetraploid and allotetraploid *Arabidopsis* species. *Genetics* 179:581-592.

Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, Sahinalp SC. 2010. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 26:i350-i357.

Hurst GD, Werren JH. 2001. The role of selfish genetic elements in eukaryotic evolution. *Nat Rev Genet* 2:597-606.

Jiang X, Tang H, Ye Z, Lynch M. 2017. Insertion Polymorphisms of Mobile Genetic Elements in Sexual and Asexual Populations of *Daphnia pulex*. *Genome Biol Evol* 9:362-374.

Keane TM, Wong K, Adams DJ. 2013. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* 29:389-390.

Kofler R. 2017. SimulaTE: simulating complex landscapes of transposable elements of populations. *Bioinformatics* 1:2.

Kofler R, Betancourt AJ, Schlotterer C. 2012. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. *PLoS Genet* 8:e1002487.

Kofler R, Gómez-Sánchez D, Schlötterer C. 2016. PoPoolationTE2: comparative population genomics of transposable elements using Pool-Seq. *Mol Biol Evol* 33:2759-2764.

Lee H, Popodi E, Foster PL, Tang H. 2014. Detection of Structural Variants Involving Repetitive Regions in the Reference Genome. *J Comput Biol* 21:219-233.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078-2079.

Lynch M. 2007. The origins of genome architecture: Sunderland: Sinauer Associates.
Lynch M, Gutenkunst R, Ackerman M, Spitze K, Ye Z, Maruki T, Jia Z. 2016. Population Genomics of *Daphnia pulex*. *Genetics*:genetics. 116.190611.
Maside X, Assimacopoulos S, Charlesworth B. 2005. Fixation of transposable elements in the *Drosophila melanogaster* genome. *Genet Res* 85:195-203.
Maside X, Bartolome C, Assimacopoulos S, Charlesworth B. 2001. Rates of movement and distribution of transposable elements in *Drosophila melanogaster*: in situ hybridization vs Southern blotting data. *Genet Res* 78:121-136.
Pasyukova E, Nuzhdin S, Morozova T, Mackay T. 2004. Accumulation of transposable elements in the genome of *Drosophila melanogaster* is associated with a decrease in fitness. *J Hered* 95:284-290.
Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. *Mol Biol Evol* 20:880-892.
Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, González J. 2011. Population genomics of transposable elements in *Drosophila melanogaster*. *Mol Biol Evol* 28:1633-1644.
Rishishwar L, Mariño-Ramírez L, Jordan IK. 2017. Benchmarking computational tools for polymorphic transposable element detection. *Brief Bioinform* 18:908-918.
SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, et al. 1996. Nested retrotransposons in the intergenic regions of the *maize* genome. *Science* 274:765-768.
Schlotterer C, Tobler R, Kofler R, Nolte V. 2014. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nat Rev Genet* 15:749-763.
Smit A, Hubley R, Green P. 2004. RepeatMasker Open-3.0. 2004. *Seattle (WA): Institute for Systems Biology*.
Sniegowski PD, Charlesworth B. 1994. Transposable element numbers in cosmopolitan inversions from a natural population of *Drosophila melanogaster*. *Genetics* 137:815-827.
Yang H-P, Nuzhdin SV. 2003. Fitness costs of Doc expression are insufficient to stabilize its copy number in *Drosophila melanogaster*. *Mol Biol Evol* 20:800-804.
Ye Z, Xu S, Spitze K, Asselman J, Jiang X, Ackerman MS, Lopez J, Harker B, Raborn RT, Thomas WK. 2017. A new reference genome assembly for the microcrustacean *Daphnia pulex*. *G3: Genes, Genom, Genet* 7:1405-1416.
Zhuang J, Wang J, Theurkauf W, Weng Z. 2014. TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Res* 42:6826-6838.

Figures and Figure legends

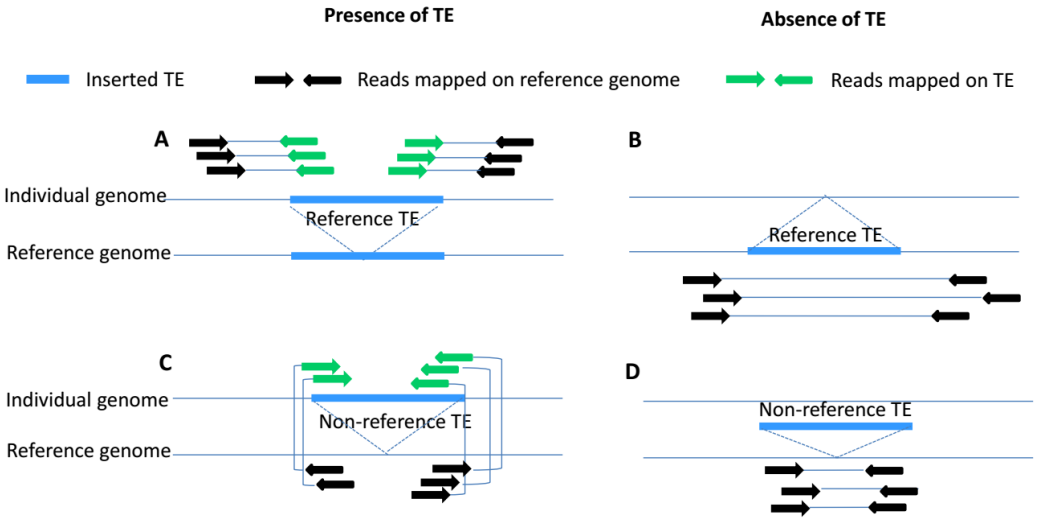

FIG. 1. Outline of the method used to identify the presence and absence of TE insertions in an individual genome. "Reference" TE insertions (A, B) and "Non-reference" TE insertions (C, D) are depicted separately, representing the cases where an TE insertion is or not in the reference genome, respectively. Two different algorithms are used to detect the presence (A, C) and absence (B, D) of TE insertions (see Methods). Paired-end reads are depicted with arrowed lines in the colors of black and green, representing reads mapped to TEs and unique regions in the reference genome, respectively.

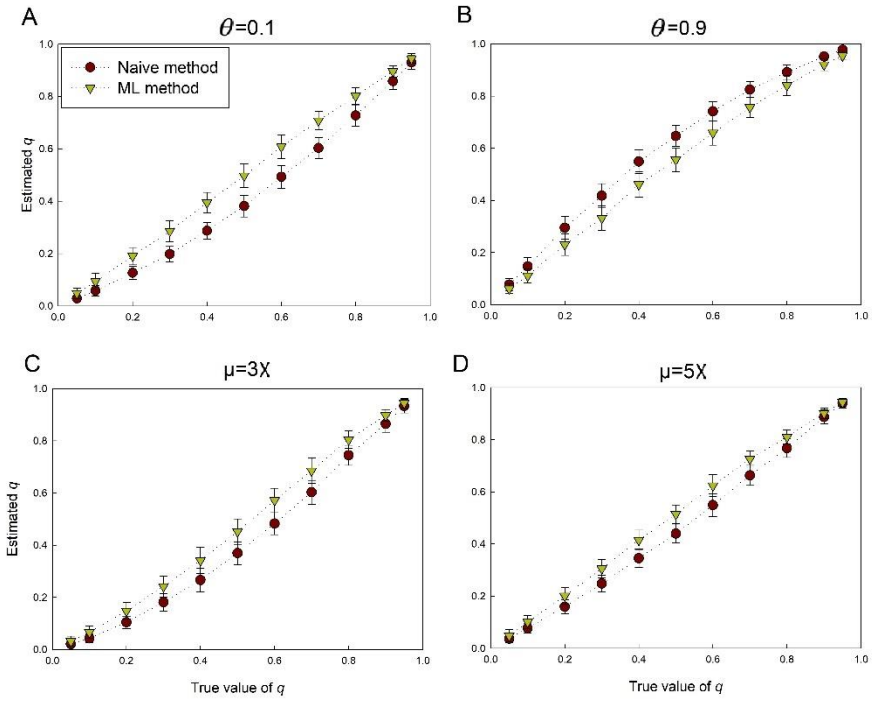

FIG. 2. Estimates of TE insertion frequencies q with simulated data. The mean and standard deviation of q estimated by using the naïve method and the ML method are shown in comparison with the true values of q, under the condition of low coverage (μ) or strong bias of TE insertion detection (θ) in the simulation. (A) $\theta = 0.1$, (B) $\theta = 0.9$, (C) $\mu = 3\text{X}$, (D) $\mu = 5\text{X}$. Except for the parameters shown in the subtitles, the settings of parameters are: $N = 100$, $\mu = 10\text{X}$, $\theta = 0.3$, $f = 0$, $s_1 = s_2 = 0$. The parameter θ was assumed unknown and estimated in the ML method. A total of 100 replicates were conducted in the simulation for each set of parameter values. The mean estimates of $\hat{q}$ under additional parameter settings are shown in FIG. S1.

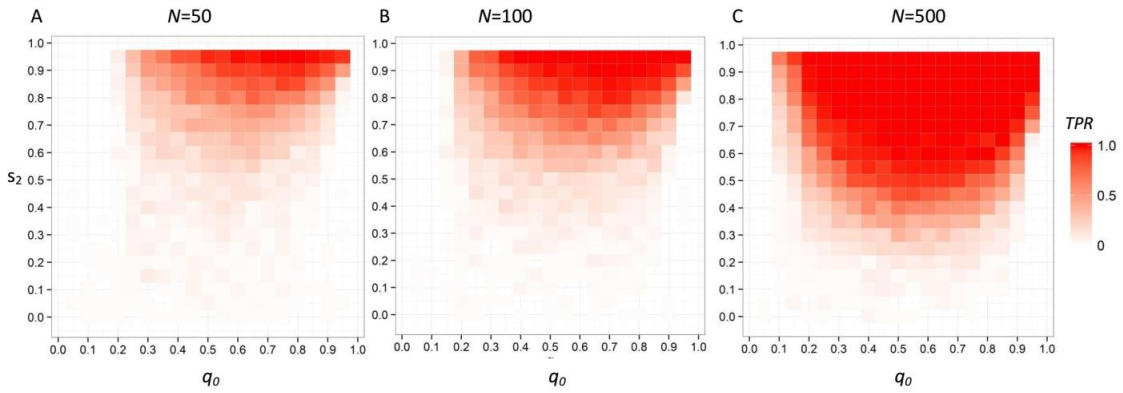

FIG. 3. **The power of detecting selection on TE insertions with the ML method.** The recessive-effects model ($s_1 = 0$ and $s_2 \neq 0$) with sample sizes of *N*=50 (A), *N*=100 (B), and *N*=500 (C) were used here. The true-positive rate (TPR) is depicted. The setting of other parameters are: $\mu = 20\text{X}$, $\theta = 0.3$, $f = 0$. A total of 100 replicates were simulated for each set of parameter values. The power analyses of selection detection using the ML method with two other selection models are shown in FIG. S2.

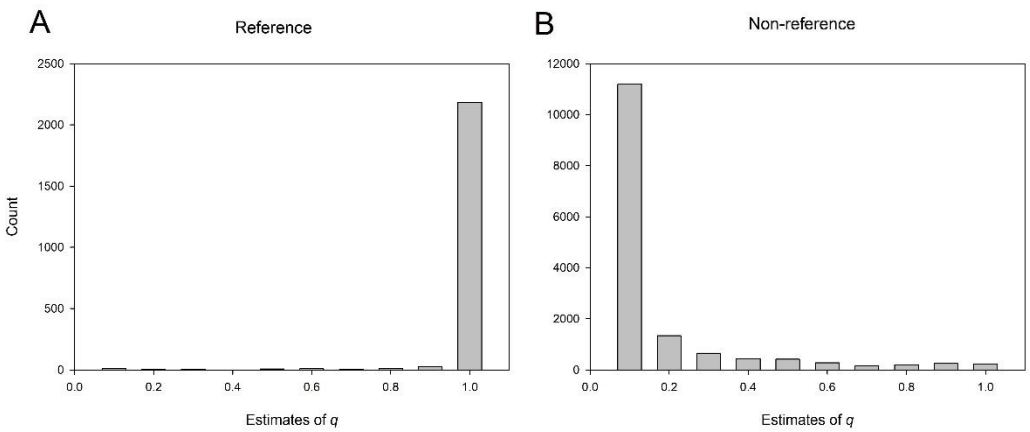

FIG. 4. Distribution of the estimates of *q* for reference (A) and non-reference (B) TE insertions in the *D. pulex* KAP population.

Tables

Table 1. The probability of the genotype Φ and the corresponding likelihood given each of the three types of observations in an individual genome *i*

Observation	Probability of the observation given the genotype			$P(n_i^+, n_i^-)$
	$\Phi_{-/-}$	$\Phi_{+/-}$	$\Phi_{+/+}$	
$n_i^+ = 0\ , n_i^- > 0$	1	$[1-\theta]^{n_i^-}$	0	$p_{--} + p_{+-}[1-\theta]^{n_i^-}$
$n_i^+ > 0\ , n_i^- > 0$	0	1	0	p_{+-}
$n_i^+ > 0\ , n_i^- = 0$	0	$\theta^{n_i^+}$	1	$p_{+-}\theta^{n_i^+} + p_{++}$

Φ_j denotes the conditional probability of observing the supporting reads provided the individual *i* is of genotype p_j = ++, +-, or –
n_i^+ and n_i^- denote the number of reads supportting the presence and absence of TE insertions, respectively.
The parameter θ denotes the detection bias.

Table 2. Parameter estimates for TE insertions in the KAP population.

Type	All[a]	Fixed[b]	θ	*N*	μ	*f*
Reference	2,263	2,124	0.78	65	30X	0.02
Non-reference	15,395	471	0.35	64	18X	0.02

a, Number of total TE insertions summarized in KAP population
b, Number of fixed TE insertions (>0.95 frequency)
The bias parameters θ, coverage parameter μ, and inbreeding coefficient *f* denote the average values across all individuals.
N is the average number of sequenced clones a TE insertion is found in the KAP population.