 doi:10.20944/preprints202012.0473.v1

Article

TEfinder: A Bioinformatics Pipeline for Detecting New Transposable Element Insertion Events in Next-Generation Sequencing Data

Vista Sohrab [1], Cristina López-Díaz [2], Antonio Di Pietro [2], Li-Jun Ma [1,3,*], and Dilay Hazal Ayhan [1,3,*]

[1] Department of Biochemistry and Molecular Biology, University of Massachusetts Amherst, Amherst, MA USA
[2] Departmento de Genética, Universidad de Córdoba, Córdoba, Spain
[3] Molecular and Cellular Biology Graduate Program, University of Massachusetts Amherst, Amherst, MA USA
* Correspondence: lijun@biochem.umass.edu (LJM); dayhan@umass.edu (DHA)

Abstract:

Transposable elements (TEs) are mobile genetic elements capable of rapidly altering the genome through their movements. The importance of TE activity has been documented in many biological processes, such as introducing genetic instability, altering patterns of gene expression, and accelerating genome evolution. Increasing appreciation of TEs results in the growing number of bioinformatics software to identify insertion events. However, the application of existing TE finding tools is limited by either narrow-focused design of the package, too many dependencies on other tools, or prior knowledge required as input files that may not be readily available to all users. Here, we report a simple pipeline, TEfinder, developed for the detection of new TE insertions with minimal software and input file dependencies. The external software requirements are BEDTools, SAMtools, and Picard. Necessary input files include TEs present in the reference genome, binary paired-end alignment, reference genome index, and a list of TE names. All inputs can be easily generated with popular variant calling pipelines. We tested TEfinder pipeline among several evolving populations of *Fusarium oxysporum* generated through a short-term adaptation study. Our results demonstrate that this easy-to-use tool can effectively detect new TE insertion events, making it accessible and practical for TE analysis.

Keywords: Transposable elements, Mobile element insertion events, Next Generation Sequencing (NGS), Genome evolution

1. Introduction

Transposable elements (TEs) are DNA sequences that move from one genomic location to another and thus impact genome evolution and organism adaptation [1]. TE transposition can alter the genomic architecture, introduce structural polymorphisms, disrupt coding sequences, and affect transcriptional and translational regulation. Additionally, TEs are capable of changing eukaryotic gene expression by providing cis-regulatory elements such as promoters, transcription factor binding sites, and repressive elements [2,3]. Ultimately, TEs provide a wide array of genomic diversity, functional impact, and evolutionary consequences that can be of notable interest to population genetics, host interaction, and comparative genomics studies.

TEs comprise a significant portion of the genome of humans and of many other organisms, due to their mobilization and accumulation throughout evolution [4]. While some TEs are no longer active, certain TE families remain mobile and their transposition contributes to genetic variation both at the individual and the population level. TEs play important roles in many biological processes,

 doi:10.20944/preprints202012.0473.v1

such as cancer biology [5], neurodegenerative diseases [6], or host-pathogen interactions [7]. Therefore, it is of high interest to identify transposon insertion polymorphisms (TIPs) for the detection of highly active TE families and the understanding of their contribution to genome dynamics and organism adaptation.

Advancements in next-generation sequencing technologies have made *in silico* discovery of transposon insertion events readily accessible. Two features commonly used for the detection of TE insertions are the target site duplication (TSD) and discordantly aligned reads. Upon insertion of certain transposons into a new genomic location, the mechanism of integration results in the duplication of the target sequence at the integration site, which is referred to as the TSD [4]. TSD length varies across superfamilies, and the identification of this structural motif proves useful in determining true transposition events within the genome [8]. Due to the nature of the transposon insertion, the TE sequence should be mapped to existing TE locations within the genome, while the paired-end read partner should be mapped to the unique sequence at the insertion site. This will result in discordant reads, which can be recognized when two paired-end reads are placed to different genomic locations or in a much greater distance that exceeds the expected insert size of the sequencing library. The more discordant reads localized in a genomic region (cluster), the higher the confidence to call it a new insertion site.

Although many bioinformatic tools that detect such events have been developed, the broad application of these tools is limited by heavy external software or file dependencies. For instance, ISMapper [9] can report insertion positions of bacterial insertion sequences (ISs) when provided with paired-end short-read sequences, TE sequences as multi-FASTA queries, and the reference genome. However, this requires specific versions of Python 3 and BioPython, which may not be readily available to all users and lead to difficulties in running the software. Mobile Element Locator Tool (MELT) [10], a Java software package originally developed as a part of the 1000 Human Genomes Project, discovers, annotates, and genotypes mobile element insertions with the only requirement of Bowtie2 [11]. The package is powerful in comprehensive TE analysis in human and chimpanzee genomes [12]. However, the package requires a FASTA file and a repeat masked BED file for every TE in the analysis. Such a high level of external file dependency makes it challenging for users who are interested in TE analysis in non-model organisms that lack well-annotated genomes.

Encountering difficulties in applying available TE insertion detection tools, we developed a simple bash bioinformatics pipeline, TEfinder, to detect new insertion events using tools that are commonly embedded in genomics variant calling workflows, including BEDTools [13], SAMtools [14], and Picard [15]. Required input files include the reference genome FASTA index file, TE annotations of the reference genome, binary alignment of paired-end short-read sequencing data, and a list of TE names of interest. The pipeline reports new insertion events based on the TE annotation of the reference genome. The output files are in BED detail format, which can be integrated into the downstream analysis. Here we report the design of the pipeline and the testing of its performance using short-read sequencing data derived from a short-term evolution experiment in the filamentous fungus *Fusarium oxysporum* f. sp. *lycopersici* 4287 (Fol4287).

2. Materials and Methods

2.1. Requirements

TEfinder is a bash pipeline for detecting TE insertions using paired-end sequencing data. The overall objective is to identify new TE insertion events in a given sample that is different from ones captured in the reference genome. For this, an assembled genome and pair-end sequencing of the sample is required.

Software required to run this tool include BEDTools 2.28.0 or later [13], SAMtools 1.3 or later [14], and Picard 2.0.1 or later [15]. Four user input files are a FASTA index file of the reference genome, a file of paired-end read alignments in binary format (BAM), TEs present in reference genome in GFF format, and a list of TE names to be analyzed in text file format. The reference genome and the read alignments files are essential. The TE GFF file can be produced using any annotation tool (see

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2020 doi:10.20944/preprints202012.0473.v1

Implementation for an example). The list of TE names provides users the option to focus their analysis on selected TE families. Without a specification, the list can be easily derived from the TE GFF file.

Optional arguments have been incorporated for users to customize the tool. The DNA fragment length or insert size of the short-read sequencing library has been set to a default value of 400 base pairs (bp). The maximum distance between reads for merging and forming clusters has been set to 150 bp. The default maximum target site duplication (TSD) length is 20 bp. Modifying this value can be useful if the TSD lengths of the TEs being analyzed are known leading to more targeted TE analysis results. An additional Java argument relating to Picard's maximum memory heap size can be submitted to the pipeline as a fraction of the total memory allocated leading to enhancement of the overall runtime. A working directory name can be provided to the tool which will be created to contain all individual TE family directories and their respective files as well as final outputs generated from the analysis.

The outputs are two BED files reporting insertion events in non-repeat and repeat regions, as well as a BAM file containing the supporting discordant reads for the events from all TE families.

2.2 Implementation

2.2.1 Preprocessing

Preparation of the four input files depicted in the top panel of Figure 1A:

1. **FASTA index file (FAI) of the reference genome.** The reference genome information anchors all downstream analysis in TIPs and enhances efficient access to regions within the FASTA file by incorporating names and lengths of sequences within the FASTA file as well as additional useful information relevant to indexing. Based on the reference genome FASTA file, this index file can be easily generated using SAMtools faidx.

```
samtools faidx reference.fasta -o reference.fai
```

2. **BAM file of aligned reads to the reference genome.** The sample sequencing reads should be aligned to the reference genome using an aligner. In this study, Burrows-Wheeler Aligner (BWA) [16] is used. Users may choose other aligners. The aligned BAM files need to be sorted by coordinates.

```
bwa index reference.fasta
bwa mem reference.fasta sample_R1.fq sample_R2.fq > sample.bam
samtools sort -o sample.sorted.bam sample.bam
```

3. **GFF file of TE annotation in the reference genome.** This file captures genomic locations of all TEs that are present in the reference genome. To generate this file, a library of TE sequences of the reference genome must be provided. For a well-annotated genome, users can obtain this library from existing databases of known repetitive elements such as Repbase [17]. If such a library is not available, users can compile one by running a *de novo* TE family identification software such as RepeatModeler [18] or RepeatScout [19] on the reference genome. An example is shown below to use RepeatScout to discover repetitive sequences and form the TE library of the reference genome.

```
build_lmer_table -sequence reference.fasta -freq reference.freq
RepeatScout -sequence reference.fasta -output TElib.fa -freq reference.freq
```

RepeatMasker [20] can be used to generate the GFF file based on the provided reference genome and the TE library. The output from RepeatMasker should be filtered so that the simple repeats are removed.

```
RepeatMasker -lib TElib.fa -dir workingdir -gff reference.fasta
```

4. **Text file with names of the TEs of interest.** The TE names provided should exactly match the FASTA headers of the TE library in a single column text file.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2020 doi:10.20944/preprints202012.0473.v1

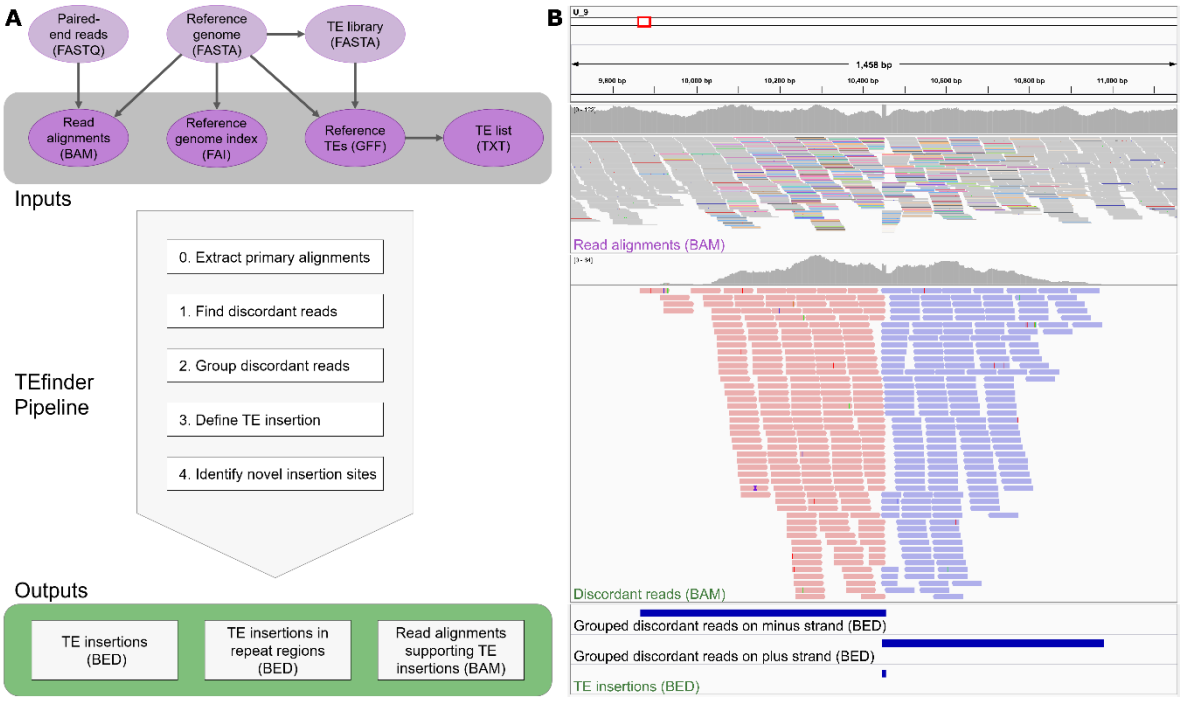


Figure 1. (**A**) Basic workflow of the TEfinder pipeline. The required input files are in dark purple, while the master files which are used to create the input files are in light purple. Arrows show the dependencies of the files. Please check Implementation for details. The steps of the pipeline are in the light gray box and the outputs are in the green box. (**B**) Visualization of a new small Hornet (TIR/hAT) insertion event using IGV. The data were derived from Y3 population after a short-term evolution experiment using Fol4287. (Top) input read alignments file, (middle) output discordant reads alignment file, and (bottom) intermediate and output BED files are displayed. The aligned reads in the top panel are shown in squished mode with reads mapped to different sequences colored non-gray. The aligned reads in the middle panel are shown in collapsed mode and forward reads are in red while reverse reads are in blue. The histograms above the alignments in the top and middle panels show the aligned read counts. The position of the insertion event reported by TEfinder coincides with the duplicated target site (TSD).

2.2.2 Pipeline

TEfinder relies on paired-end sequencing and uses information on discordant reads, which are reads that do not match the expected orientation or insert size. The software required for the package are SAMtools [14], BEDTools [13], and Picard [15].

A typical command to run TEfinder is:

```
TEfinder -bam sample.sorted.bam -fai reference.fai -gff TEs.gff -te List_of_TEs.txt
```

To avoid multi-mapping and chimeric ambiguities from affecting our analysis, the initial step of the tool is the removal of secondary and supplementary alignments from the user-provided sample alignment file (Figure 1A).

For each TE, the package goes through the following logical steps:

1) **Identify discordant reads.** The program extracts all primary reads mapped to known TEs in the reference using BEDTools intersect. Then, alignments of the selected reads and their pairs are extracted from the input BAM file using the FilterSamReads tool of Picard [15]. Among those, reads are selected as discordant if the corresponding mate maps to a different sequence or the read pair has an insert size that exceeds a threshold of 10 times the insert size (Figure 1B, top panel).
2) **Group discordant reads.** Once the discordant read alignments have been filtered, the regions of clustered reads are identified using BEDTools merge so that reads aligned to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2020 doi:10.20944/preprints202012.0473.v1

the plus strand and minus strand are grouped separately (Figure 1B, middle and bottom panels). In this step, the reads must be overlapping or within a given distance to be considered in the same group.

3) **Define a TE insertion site.** Each plus-strand group is coupled with the nearest minus-strand group. The coupled regions go through a filtering step to remove incorrect orientations, considering that TE sequences should be present between forward and reverse groups and not the reverse order. Due to the nature of duplication of the target site upon the transposon insertion, TSD sequences may be present in both forward and reverse strands, resulting in an overlap between forward and reverse clusters. If an overlapping site is smaller than the max TSD length, the location is reported as a possible TE insertion site. If they are not overlapping due to low coverage or other reasons, and the distance between the groups is smaller than the given threshold, the region in between is reported.
4) **Identify new insertion sites.** The most important output file that captures all new insertion events are reported as 'TE_insertions.bed'. As TEs are repetitive sequences, it is no surprise that this pipeline also detects existing TEs in repetitive regions of the reference genome. These events are easily separated from interesting new events and stored under the file name 'TEinsertions_inRepeatRegions.bed' for future considerations. It needs to be noted that some new insertions within repeat regions will be filtered out and reported in this file (Figure 1).

These detected TE insertions output files are in BED detail format with 7 columns: 1) insertion site sequence, 2) start coordinate, 3) end coordinate, 4) TE family, 5) total number of reads supporting the insertion event, 6) unused, 7) additional information regarding the insertion such as the number of forward and reverse reads supporting the event (FR and RR), as well as the insertion region start coordinate (IS) and insertion region end coordinate (IE) obtained from the pipeline.

Additionally, discordant pair alignment files from individual TE families are combined to create 'DiscordantReads.bam' file in the working directory (Figure 1B, middle panel). This file can be used to visualize the events on genome browsers such as Integrative Genomics Viewer (IGV) [21].

2.3. Testing dataset and processing

The performance of TEfinder was tested using a model *Fusarium oxysporum,* a soil-inhabiting ascomycete fungus that causes devastating losses in more than a hundred different crops and disseminated infections in immunocompromised humans [22,23]. One interesting genomic feature of the *F. oxysporum* species complex is the compartmentalization of its genome, where conserved core regions carry essential house-keeping functions while lineage-specific accessory regions are enriched for TEs and associated with host-specific pathogenicity [23,24].

The pipeline was used to identify new TE insertion events among five populations, Y1-Y5, evolved under laboratory conditions. Briefly, the ancestor strain (WT) of *Fol*4287, which was previously used to generate the reference genome assembly [24], was subjected to successive transfers on Yeast Peptone Dextrose Agar plates. After 10 passages with 5 independent biological replicates (Y1-Y5), genomic DNA was extracted from the final populations and sequenced using Illumina HiSeq 2500 platform with 2X71 cycles. The whole-genome shotgun sequencing reads are available at NCBI under project PRJNA682786 and datasets SRR13203443, SRR13203444, SRR13203445, SRR13203446, and SRR13203447.

2.4. Experimental validation

Four detected TE insertion events were validated by PCR. Genomic DNA was extracted from mycelia of the *F. oxysporum* reference strain or of single spore (SS) isolates obtained from the experimentally evolved lines, using the CTAB method [25]. PCR was performed in a thermocycler using the thermostable DNA polymerase of the Expand High Fidelity PCR System (Roche Life Sciences). Each PCR reaction contained 300 nM of each primer, 2.5 mM $MgCl_2$, 0.8 mM dNTP mix,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2020 doi:10.20944/preprints202012.0473.v1

0.05 U/µl polymerase and 5-10 ng/µl genomic DNA. PCR cycling conditions were as follows: an initial step of denaturation (5 minutes, 94°C); 35 cycles of 35 seconds at 94°C, 35 seconds at the calculated primer annealing temperature and 1 minute/1.5 kb extension at 72°C (or 68°C for templates larger than 3 kb); and a final extension step of 10 minutes at 72°C (or 68°C). For each predicted TE insertion event, a pair of specific primers flanking the insertion site was designed. Specific primers designed for validation of TE insertion events are Y1 forward: TCCTCCTGGGTTTCTTGTCAC, Y1 reverse: CTCTTGAAACGTGGTGCAGAC; Y3 and Y4 forward: AGACGGACAAAGAGGTGTGAC, Y3 and Y4 reverse CTCGACACTACCAGGCACTAT; Y5 forward: GAGTATGCTTCCCGATCCTTG, Y5 reverse: GACATCCCTCAATCCGCTGAA.

3. Results

3.1. Data preparation

The *Fol*4287 reference genome is 53.9 MB in 499 scaffolds [26]. The sequencing reads were mapped to the reference using BWA [16] with >99% mapping and median coverages ranging from 67 to 94 (Table 1). RepeatMasker [20] was used to identify the known TEs in the genome with a curated TE library [22,23] that includes 69 TE families. Approximately 4.5% of the entire genome was identified as repetitive sequence with 3.98% of the genome comprised of transposable elements. Accessory sequences include 74% of all TEs in the genome [22].

Table 1. Sample sequencing and mapping summary statistics.

Sample	Total Reads	Discordant Mate Mapping Quality >= 5	Percent Reads Mapping to Reference	Median Coverage
Y1	62,015,365	475,966	99.41	67
Y2	85,688,005	600,267	99.38	94
Y3	83,109,424	578,959	99.44	92
Y4	70,322,907	475,591	99.42	78
Y5	76,034,020	567,823	99.37	83

3.2. Total TE Insertion Events Detected

On a high-performance computing cluster with a memory request of 3 X 50000 MB, the trial runs took an average of 5.2 hours to complete with the minimum time being 3.5 h across the samples. The maximum heap memory for Java was set to 25000 MB to enhance the run time when filtering the alignment to only include discordant reads.

TEfinder detected 576 total new insertion events across whole genome sequencing data from 5 evolved Y1-Y5 populations. Details for each individual sample are summarized in Table 2. As expected, many insertion events are detected among complex repetitive regions that capture most transposons. In fact, the number of new TE insertions reported in each sample population is approximately one third of the insertions detected in known repeat regions.

Table 2. Number of TE insertion events reported by TEfinder in evolved populations of Fol4287.

Sample	New TE Insertions	TE Insertions in Repetitive Regions	Manually inspected Insertions
Y1	105	397	9
Y2	117	449	13
Y3	115	455	11
Y4	118	446	11
Y5	121	423	14

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 December 2020 doi:10.20944/preprints202012.0473.v1

The BAM output file format is intended for visual inspection. Figure 1B captures a new insertion event of the TE small Hornet (TIR/hAT) that reached fixation in the Y3 population visualized in IGV. The confidence level for calling this new insertion event is high with a total of 415 supporting reads spanning a 1114 bp insertion region in a data set with 92X median coverage. Of these supporting discordant reads, 223 reads are grouped in the plus strand cluster and 192 in the minus strand cluster. Since the TEfinder pipeline does not focus on TSD position detection, the reported insertions may not always coincide precisely with the true location. Nevertheless, the pipeline mapped the TSD position of many events precisely. In the example of Figure 1B, two clusters overlap at an 8-base TSD, which coincides with the known TSD of this particular transposon superfamily [8]. Four newly detected TIP by the same TE in Y1, Y3, Y4 and Y5 populations were validated by PCR in two single spore isolation lines (Figure 2). The size difference in the ancestor and the lines with TIP are about 750 bp which coincides with the size of small Hornet transposon.

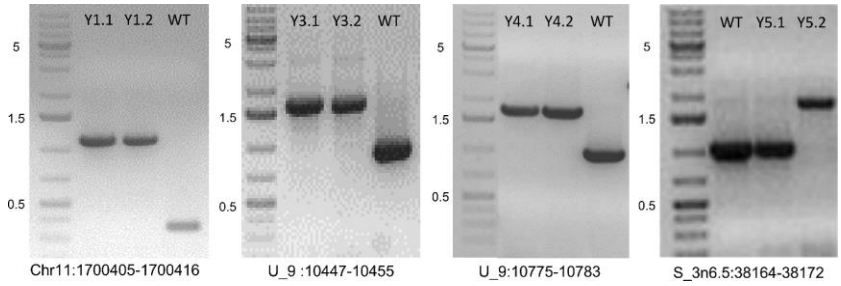
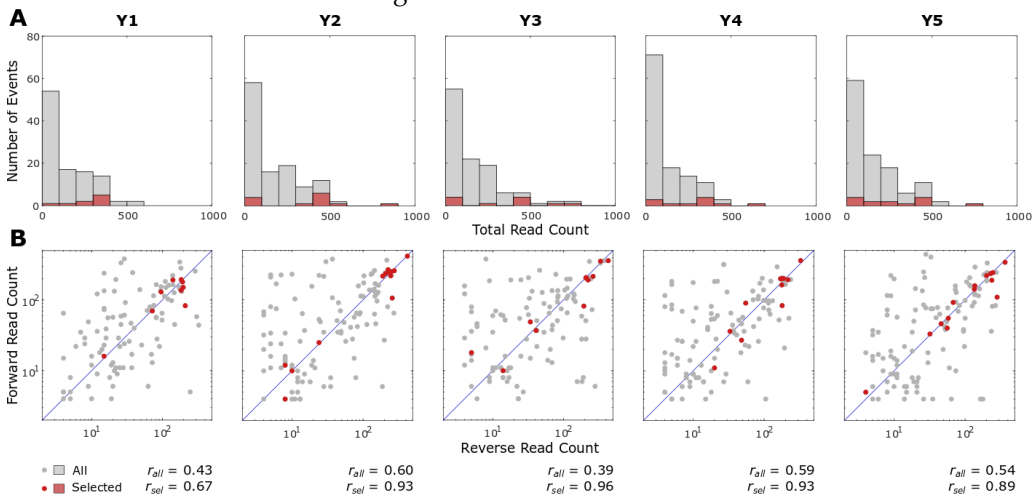


Figure 2. Validation of small Hornet transposon insertion events detected in experimentally evolved populations of Fol4287. PCR was performed with primers flanking the insertion site in the ancestor strain (WT) and two single spore (SS) isolates in populations Y1, Y3, Y4, and Y5 respectively. GeneRuler 1 kb Plus DNA Ladder shown with indicated sizes in kb to the left.

3.3. Read Counts Support New TE Insertion Events

The output BED files of TEfinder report an overview of the total number of reads supporting the insertion event, as well as the number of forward and reverse read counts. In our trial, TEfinder was able to capture low-frequency insertion events with read evidence as little as 9 in the dataset with 83X coverage (Figure 3A, sample Y5). For further functional analysis, we applied a filter to remove cases with a strand bias of 5-fold or larger and visualized them to further filter out the events in locations with noisy alignments and/or in non-uniquely mapped sequences. We selected 9 to 14 such events in each population, which were considered as "true" insertion events (Table 2).

Importantly, the correlation of forward and reverse read counts of these selected insertion events is greatly increased to 0.67-0.96 compared to a correlation of 0.39-0.60 for all reported events (Figure 3B). Therefore, it is highly recommended that users filter the output BED files for read counts and for strong bias of the forward-reverse read counts.



 doi:10.20944/preprints202012.0473.v1

Figure 3. Details of the reported new insertion events in evolved *F. oxysporum* f. sp. *lycopersici* populations Y1 to Y5. (**A**) Distributions of the read counts and (**B**) scatter plots of the forward and reverse read counts of the reported events in non-repeat regions. All events are plotted in gray while selected events are shown in red. Blue lines in (**B**) are x=y lines. The Pearson correlation values of the forward and reverse read counts of all and selected (sel) events are given below the scatter plots.

4. Discussion

Here we report a simple pipeline to detect new TE insertions in related populations when compared to a reference genome. The required input and output file types are commonly used in variant detection workflows and therefore it can be easily implemented in larger pipelines. Testing of the pipeline in evolved populations of the fungus *F. oxysporum* demonstrated that this easy-to-use tool can effectively detect new TE insertion events, making it accessible and practical to address TE activity-related biological questions in population genomics, genome evolution, and other applications.

Paired-end sequence reads are the foundation to read out genetic changes in an individual or a population when comparing to a reference genome. Therefore, sufficient sequence coverage and good quality of the sequence reads are important, as for all variant calling software. However, even with a small number of read evidence, the pipeline can still capture low-frequency events. The optional parameters can be used to adjust the strictness of the algorithm.

Our data suggest that as long as the reference genome is reasonably assembled, the quality of the reference assembly should not affect the performance of the TEfinder pipeline. For instance, the assemblies of the lineage-specific genome regions of the Fol4287 genome used to test TEfinder were fragmented due to high repeat content, while chromosome-level assembly was accomplished for most core regions. Importantly, TEfinder was able to capture the events in both types of regions.

As with other variant calling software, the output files need to be filtered before further analysis. Users can utilize the reported forward, reverse, and total read count values according to their needs. The output BAM file is also useful to do visual confirmations. One feature missing from the output is the allele frequency of the events. However, the read counts and insertion region positions can be used to estimate allele frequencies. We were able to experimentally verify some of the events. Although we are in the early phase in understanding the functional impact of these TIPs, the ability to detect these events with high confidence enables hypothesis generation and establishment of targeted functional studies. Testing of the TEfinder pipeline in evolved fungal populations further confirmed the effectiveness of this tool in identifying TE insertion events in non-reference eukaryotic genomes.

Supplementary Materials: The TEfinder pipeline is available on GitHub at https://github.com/VistaSohrab/TEfinder

Author Contributions:" Conceptualization, V.S., L.-J.M. and D.H.A.; software, V.S. and D.H.A.; data generation, C.L-D., D.H.A, L.-J. M, and A.D.P.; writing—original draft preparation, V.S., L.-J.M. and D.H.A.; review and editing, all authors contributed. All authors have read and agreed to the published version of the manuscript."

Funding: "V.S. is supported by University of Massachusetts Amherst Honors College Research grant. D.H.A and L.-J.M are supported by the National Institute of Health (R01EY030150) and Burroughs Welcome Foundation (1014893). C.L-D is supported by FEDER Junta de Andalucía (27374-R). The sequencing data were generated through the support of the United States Department of Agriculture, National Institute of Food and Agriculture (Grant awards 2011-35600-30379 and MASR-2009-04374) and the National Science Foundation (IOS-1652641), and by the Spanish Ministerio de Ciencia e Innovación MICINN (PID2019-108045RB-I00)."

Acknowledgments: "We thank the MGHPCC for providing high-performance computing capacity for genome assembly process."

Conflicts of Interest: "The authors declare no conflict of interest."

References

1. Chénais, B.; Caruso, A.; Hiard, S.; Casse, N. The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene* **2012**, *509*, 7–15, doi:10.1016/j.gene.2012.07.042.
2. Bourque, G.; Burns, K.H.; Gehring, M.; Gorbunova, V.; Seluanov, A.; Hammell, M.; Imbeault, M.; Izsvák, Z.; Levin, H.L.; Macfarlan, T.S.; et al. Ten things you should know about transposable elements. *Genome Biol.* **2018**, *19*, 1–12, doi:10.1186/s13059-018-1577-z.
3. Huang, C.R.L.; Burns, K.H.; Boeke, J.D. Active transposition in genomes. *Annu. Rev. Genet.* **2012**, *46*, 651–675, doi:10.1146/annurev-genet-110711-155616.
4. Munoz-Lopez, M.; Garcia-Perez, J. DNA Transposons: Nature and Applications in Genomics. *Curr. Genomics* **2010**, *11*, 115–128, doi:10.2174/138920210790886871.
5. Burns, K.H. Transposable elements in cancer. *Nat. Rev. Cancer* **2017**, *17*, 415–424, doi:10.1038/nrc.2017.35.
6. Jönsson, M.E.; Garza, R.; Johansson, P.A.; Jakobsson, J. Transposable Elements: A Common Feature of Neurodevelopmental and Neurodegenerative Disorders. *Trends Genet.* **2020**, *36*, 610–623, doi:10.1016/j.tig.2020.05.004.
7. Seidl, M.F.; Thomma, B.P.H.J. Transposable Elements Direct The Coevolution between Plants and Microbes. *Trends Genet.* **2017**, *33*, 842–851, doi:10.1016/j.tig.2017.07.003.
8. Wicker, T.; Sabot, F.; Hua-Van, A.; Bennetzen, J.L.; Capy, P.; Chalhoub, B.; Flavell, A.; Leroy, P.; Morgante, M.; Panaud, O.; et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **2007**, *8*, 973–982, doi:10.1038/nrg2165.
9. Hawkey, J.; Hamidian, M.; Wick, R.R.; Edwards, D.J.; Billman-Jacobe, H.; Hall, R.M.; Holt, K.E. ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. *BMC Genomics* **2015**, *16*, 667, doi:10.1186/s12864-015-1860-2.
10. Gardner, E.J.; Lam, V.K.; Harris, D.N.; Chuang, N.T.; Scott, E.C.; Pittard, W.S.; Mills, R.E.; Devine, S.E. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* **2017**, *27*, 1916–1929, doi:10.1101/gr.218032.116.
11. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74, doi:10.1038/nature15393.
12. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359, doi:10.1038/nmeth.1923.
13. Quinlan, A.R.; Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842, doi:10.1093/bioinformatics/btq033.
14. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079, doi:10.1093/bioinformatics/btp352.
15. Picard Toolkit. *Broad Institute, GitHub Repos.* **2019**.
16. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–60, doi:10.1093/bioinformatics/btp324.
17. Bao, W.; Kojima, K.K.; Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **2015**, *6*, 11, doi:10.1186/s13100-015-0041-9.
18. Smit, AFA, Hubley, R. RepeatModeler Open-1.0.
19. Price, A.L.; Jones, N.C.; Pevzner, P.A. De novo identification of repeat families in large genomes. *Bioinformatics* **2005**, *21*, i351–i358, doi:10.1093/bioinformatics/bti1018.

20. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0.
21. Thorvaldsdottir, H.; Robinson, J.T.; Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **2013**, *14*, 178–192, doi:10.1093/bib/bbs017.
22. Ma, L.J.; Van Der Does, H.C.; Borkovich, K.A.; Coleman, J.J.; Daboussi, M.J.; Di Pietro, A.; Dufresne, M.; Freitag, M.; Grabherr, M.; Henrissat, B.; et al. Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. *Nature* **2010**, *464*, 367–373, doi:10.1038/nature08850.
23. Zhang, Y.; Yang, H.; Turra, D.; Zhou, S.; Ayhan, D.H.; DeIulio, G.A.; Guo, L.; Broz, K.; Wiederhold, N.; Coleman, J.J.; et al. The genome of opportunistic fungal pathogen Fusarium oxysporum carries a unique set of lineage-specific chromosomes. *Commun. Biol.* **2020**, *3*, 50, doi:10.1038/s42003-020-0770-2.
24. Kistler, H.C.; Rep, M.; Ma, L.-J. Structural dynamics of Fusarium genomes. In *Fusarium, genomics, molecular and cellular biology*; Caister Academic Press, Norfolk, 2013; pp. 31–42.
25. Raeder, U.; Broda, P. Rapid preparation of DNA from filamentous fungi. *Lett. Appl. Microbiol.* **1985**, *1*, 17–20, doi:10.1111/j.1472-765X.1985.tb01479.x.
26. Ayhan, D.H.; López-Díaz, C.; Di Pietro, A.; Ma, L.-J. Improved Assembly of Reference Genome Fusarium oxysporum f. sp. lycopersici Strain Fol4287. *Microbiol. Resour. Announc.* **2018**, *7*, doi:10.1128/MRA.00910-18.