ELITE: Efficiently Locating Insertions of Transposable Elements

Anwica Kashfeen
UNC Chapel Hill
Department of Computer Science
anwica@cs.unc.edu

Harper B. Fauni
UNC Chapel Hill
Department of Genetics
hfauni@live.unc.edu

Timothy A. Bell
UNC Chapel Hill
Department of Genetics
timothy_a_bell@med.unc.edu

Fernando Pardo-Manuel de Villena
UNC Chapel Hill
Department of Genetics
fernando@med.unc.edu

Leonard McMillan
UNC Chapel Hill
Department of Computer Science
mcmillan@cs.unc.edu

ABSTRACT

A large fraction of mammalian genome consists of transposable elements (TEs). These elements are segments of DNA that either move or are copied from one place in the genome to another. Such movements can cause deleterious mutations and drive chromosome evolution. Existing approaches search for TE insertion (TEi) by aligning millions of mostly irrelevant short reads to either a reference genome or a TE sequence library. Here we present a new *local genome assembly* based pipeline, called ELITE, for identifying and characterizing TEi. ELITE uses an msBWT-based data structure to store and index all the reads from a high-throughput sequencing dataset and leverages a sampled FM-index to detect TEi efficiently. In comparison with two existing tools, ELITE is faster and has a higher precision and recall rate in predicting TEi. ELITE also works on real data, which we validated using PCR assays of surrounding genomic context. Additional features of ELITE include finding zygosity status of a predicted TEi, discovering unannotated TEs that are distantly related to the target one, and providing a summary of TEi sharing pattern within a population.

CCS CONCEPTS

• **Applied computing → Computational biology**.

KEYWORDS

repeats, local genome assembly, transposable element, msBWT, FM-index, high throughput short reads.

ACM Reference Format:
Anwica Kashfeen, Harper B. Fauni, Timothy A. Bell, Fernando Pardo-Manuel de Villena, and Leonard McMillan. 2019. ELITE: Efficiently Locating Insertions of Transposable Elements. In *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19), September 7–10, 2019, Niagara Falls, NY, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3307339.3342182

1 INTRODUCTION

In 1940 Barbara McClintock discovered a type of genomic rearrangements where segments of DNA either *jump* or are *copied* from one place to another. These mobile segments are called *Transposable Elements* (TEs). It has subsequently been discovered that a significant fraction of eukaryotic genomes sequences are composed of TEs and their vestiges. About 45% of human [6], 37% of mouse [7], and 85% of maize genomes [24] consist of TE-derived sequence.

TEs also play vital roles in the biology and evolution of organisms [21][19]. Many TE classes tend to relocate and/or copy themselves into and around genes. Such insertions can interrupt, modify, or sometimes even completely disable the associated gene's function. Many diseases such as hemophilia A, neurofibromatosis, choroideremia, cholinesterase deficiency, Apert syndrome, and β-thalassemia are reported to be consequences of TE translocations [27][26][20][5][3]. Due to TEs important role in genome biology, various TE localization tools have been developed.

One class of tool such as RetroSeq [13], TEMP [30], MELT [9] relies on an initial alignment step to a reference genome to identify discordant read pairs. Detecting TEis by resolving these pairs depend heavily on the quality and efficiency of the alignment method and fails to report the exact insertion site without any additional steps. As reported by [22] the best-performing algorithms have precision rates ranging from to 63% to 95% for simulated data, and between 25% to 73% for real data. Moreover, they can only locate the class of TE that is present in the reference genome. However human genome, which is a consensus of several individuals, may not include all classes of TE.

A second TEi detection approach aligns short reads to a catalog of consensus TE sequences. These tools such as ITIS [12], RelocaTE2 [4] attempt to find all the *split reads* containing both TE and non-TE segment. The non-TE parts of these reads are then clustered, combined, and mapped to a reference genome to identify the insertion site. These methods can give the precise location of a TEi but typically employ lower throughput aligners (i.e., BLAST[1] or BLAT[14]). Moreover, for any new template of TE, they must repeat all the steps of the alignment pipeline.

We present a new TE identification, mapping, and characterization tool for Efficiently Locating Insertions of Transposable Elements, (ELITE). Unlike previous alignment-based approaches, ELITE is a targeted *local-genome-assembly-based* method. ELITE uses a branch-and-bound Depth-First-Search (DFS) algorithm for

ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA

ACM ISBN 978-1-4503-6666-3/19/09…$15.00
https://doi.org/10.1145/3307339.3342182

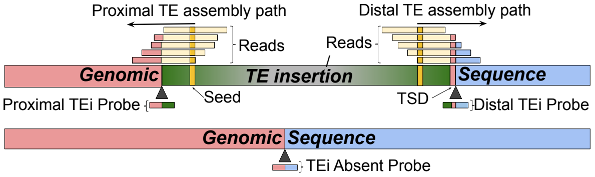


Figure 1: An overview of ELITE: ELITE first finds two highly conserved seeds from any annotated TE sequence. One near the proximal boundary, and the other near the distal boundary (shown in gold). It assembles sequence paths around the seeds from unaligned HTS read data using an FM-index. TE annotation is used to guide the path traversal in a depth-first search to estimate the TEi boundaries. The overhangs of the assembled sequences are used to map the TEi location. When possible, nearby insertions are merged informed by the sharing of a target-site duplication (TSD) sequence. Finally, up to three probes are constructed for use in testing other samples for a similar polymorphic TEi.

the assembly, which efficiently searches the entire unaligned read set of a high-throughput sequencing dataset represented as a multi-string Burrows-Wheeler Transform (msBWT) [11], using an FM-index [8]. Our runtime comparison shows that ELITE is faster than two existing TE detection tools, TEMP[30] and MELT[9]. ELITE allows more divergence in the TE sequence while keeping a conserved TE kmer called seed. This divergence enables ELITE to discover new TE families, that may not be annotated in any standard TE database [2] [29]. ELITE also reports a summary of TE sharing within a given set of samples, which includes TEis that are present in all samples, TEis that are shared by only a subset of samples, and TEis that are unique to a specific sample.

2 METHODS

Given a template TE, and a sample, we aim to find all the locations of that and similar TE classes in that sample. As a preprocess, we construct an msBWT of a whole-genome high-throughput sequencing dataset. An msBWT sorts all the reads alphabetically and then assigns an index to each read. An auxiliary data structure called FM-index is built on the fly. Searching of *kmer* is done incrementally in reverse or suffix order. For example: to search ACT, it will first find all the read indices containing Ts, then CTs, and finally ACTs. We use this backward search approach to assemble TE sequences from an interior seed towards a boundary. Searching is done for the two sides of a TE, which we call *proximal* and *distal*. The assembled sequence beyond the TE boundary called the *context* is then mapped to a reference genome to find the location of insertion. This TEi discovery phase for the proximal side is illustrated in the figure2. The execution order of the steps goes from right to left. After the discovery phase, several additional steps are taken which involves merging proximal and distal TEis, identifying their zygosity and creating different probes to assess TEi pattern in a population. All the steps are described in detail in the following subsections.

2.1 Choosing a Seed

We use a seed-based search approach to perform local assembly where the seed is a highly repeated substring from the given TE. Multiple seeds can be used to allow some mutations in the seed itself. All the seeds, however, need to be within a certain distance from the TE's boundary because ELITE needs to have enough bases beyond the boundary to infer the context of insertion. To ensure this, we only consider the *kmer* as seed whose distance from TE boundary is no more than half of the read length. Among all these, we recommend a seed that occurs most in the sample's genome. We provide two recommended seeds, i.e., proximal and distal seed, where each one is closer to its respective boundary.

2.2 Finding TE-like Sequences

The first step of our TEi discovery phase finds all mutated versions of TE in a given sample. To allow for mutations as well as capture related, and perhaps unannotated, TEs, we find all the sequences that are less than a given edit distance away from the original TE. We start with a seed as an input, which is our initial assembled sequence and use a depth-first-search algorithm1 to extend it further towards the boundary of the TE (figure 2a).

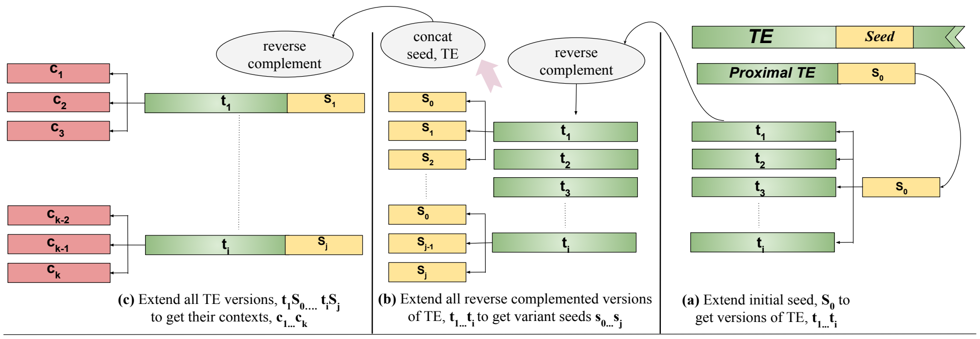
Algorithm 1 Extending seed k-mer

```
procedure ExtendKmer($range$, $seed$, $newTE$)
    if $newTE$ Reaches TE Boundary then
        Add $newTE$ to the $TE\ list$
    for $base\ in\ A, C, G, T$ do
        $newTE \leftarrow base + newTE$
        $newRange \leftarrow findIndicesOfStr(base,(range))$
        $dist \leftarrow editDistance(erv,newTE)$
        if $newRange > t1$ and $dist < t2$ then
            ExtendKmer($newRange$, $seed$, $newTE$)
                                        ▷ t1 and t2 are threshold parameters
```

The algorithm at first finds the range of indices for the seed k-mer using a sampled FM-index of the compressed msBWT [11][8], where the range represents the number of occurrences of seed substring in the set of sequenced reads (specifically their interval in an implicit suffix array). This range, along with the seed k-mer, is used to initialize the recursive DFS used by the local assembly. At each step in the recursion, the algorithm adds a new possible base before the seed and updates the suffix array range (*newRange*) concerning the newly added base and then continues along this child's path in the recursion tree.

If at any place the value of *dist* is greater than a certain threshold $t2$ we prune that path. This prunes from our search sequences that differ too much from the given TE template. Recursion is also terminated if the newly added base causes the value of *newRange* to drop below a threshold $t1$.

At this point, all the versions have the same bases as seed at the end. To allow for variants in the seed, we then remove it from all the TE versions and assemble each version in the reverse direction towards the seed (figure 2b). Since searching in an msBWT using an FM-index is done by extending suffixes, finding a prefix is more straightforward than finding a suffix. Thus, when extending TE sequences in this second DFS pass, where the seed is removed, we conduct the search using the TE's reverse complement sequence, thus searching for alternative seeds that prefix. It works because of DNA's double-stranded structure where one strand contains



(c) Extend all TE versions, $\mathbf{t_1S_{0....}\ t_iS_j}$ to get their contexts, $\mathbf{c_{1...}c_k}$

(b) Extend all reverse complemented versions of TE, $\mathbf{t_{1...}t_i}$ to get variant seeds $\mathbf{s_{0...}s_j}$

(a) Extend initial seed, $\mathbf{S_0}$ to get versions of TE, $\mathbf{t_{1...}t_i}$

Figure 2: TEi Discovery phase: We take the proximal sequence of a given TE and select a kmer located distal to the TE's start as a seed. A seed whose distance from the TE start is less than half of the read length, is of suitable length and has the highly repeated in the data set is chosen. The seed is then extended to find all the potential versions of TE (shown in green). To incorporate any variants that may appear in the seed, we then extend all of the potential TEs in the opposite direction. This is accomplished in the BWT by searching for the extension's reverse complement (shown in yellow). Finally all the versions are further extended to get all the possible prefixes, which establishes their genomic context (shown in pink).

the reverse complemented sequence of the other to form a double helix. Now to get back to the original strand, we again reverse complement the TE versions. We repeat the same assembly process for the distal side of a TE in the opposite direction.

2.3 Mapping TE Insertion Sites

After finding all the versions of a TE, we then attempt to locate them in the genome. We assemble the genomic context by extending the prefix and suffix of each discovered proximal and TE-like sequence, respectively, to find the corresponding genomic context (figure 2c). We use the same algorithm described in the previous subsection, but initialized with the suffix array range of each TE version found by the *ExtendKmer* routine. We set the edit distance threshold $t2$ to a large value and the support threshold, $t1$ to zero to allow the DFS to extend until it runs out of reads containing the given TE version. ELITE then maps all of the contexts found against the reference genome using bowtie2 [15][16] to get the chromosome and position. In this phase, ELITE only keeps the contexts that are uniquely mapped. Finally, it examines the reference sequence adjacent to the alignment position to assess whether it differs or is a close match (as determined by $t2$) to the TE sequence used to find the context originally. If the reference sequence adjacent to the mapped TE differs, we consider the mapping a non-reference TEi.

2.4 Determining Zygosity

As a third step, for any non-reference TEi, ELITE determines if it is present in the homozygous or heterozygous state. To find the zygosity, we extend the mapped context sequence found in step 2 in the direction of TE, which differs depending on whether it was discovered from the proximal or distal side of the TE template. If we find sequence paths, one matching the expected TE, and a second similar to the reference sequence (as determined by the edit-distance parameter, $t2$) then we report the TEi as heterozygous. But, if the extended sequence only leads to the expected TE-like sequence, then we report it as homozygous. Usually, heterozygous TEis are more likely to be active as they show segregation.

2.5 Merging Proximal and Distal TEis

ELITE independently considers the two sides of a TE during the first three assembly steps. It merges any non-reference TEi if there exists a pair of proximal and distal context which map to the same position after adjusting the TSD. ELITE also attempts to find another side of non-reference TE when a mapped TEi is discovered only from one side. ELITE uses genomic sequence adjacent to the context on the side of the detected TEi from the reference to once again guide a DFS to find the TE-like context adjacent to it. If ELITE finds a reference-like context flanked by the expected TSD followed by any sequence that is a significant edit distance from the originating context, the TEi is considered merged. For merging TEs in the reference, we find all the proximal-distal context pairs that are near the length of the TE plus or minus a gap parameter. If a consistent TSD is observed, we merge the proximal-distal context pair. ELITE keeps all mapped, but unresolved one-sided TEis if there is enough reads to support the insertion. Because TEis can be onesided if the context on the other side is unmappable, or if some insertion or deletion modified the missing context following the TEi.

2.6 Assessing TEi pattern in a population

A final optional feature of ELITE is that, once a TEi is identified and mapped, several targeted sequence probes (up to three) are generated to accelerate the testing of subsequent samples. When a TEi is discovered, one or more of three probe types are created1. The first is a TEi specific *proximal* probe for finding split-reads that contain the normal genomic context and the adjacent TE sequence. The second *distal* probe includes genomic sequence at the other end of the insertion preceded by a TSD sequence and the distal TE sequence. The third *absence* probe represents the expected genomic sequence without the TEi. Ideally, corresponding proximal and distal probes are derived from the TE-like sequence found during the discovery phase. Absence probes are derived from all the samples, which has some nonTE-like sequence in the same chromosome and position. The presence or absence of a TEi is confirmed by querying the TEprobe (*proximal* and *distal*) and *absence* probe, respectively. However, having both makes a sample heterozygous in that site.

3 ALGORITHM COMPLEXITY ANALYSIS

Our algorithm to find TE-like sequence and the corresponding genomic contexts is essentially an exponential DFS algorithm. Given that a DNA sequence is consists of 4 bases, i.e., A, C, G, and T, in the worst case scenario, it will traverse all 4^r possible path to extend r bases before a seed. Fortunately, the genome is finite, and it's not possible for a genome to have all 4^r sequences when r is too large. In addition, msBWT allows us to look for any prefix of length r in $O(r)$ time. Finding seed k-mer and extending it ($O(k) + O(r)$) are the only two operations that we use in our discovery phase.

4 RESULT AND DISCUSSION

We have applied ELITE to six short-read sequencing data sets and examined its error rate in the context of expected sharing based on their origin/pedigree. We also report on the performance of ELITE applied to a synthetic dataset where the truth is known to compare ELITE to two TE detection tools. To validate several of ELITE's TEi prediction on real data, we performed standard PCR methods.

4.1 TE Discovery in Real Data

Laboratory inbred mouse strains are widely used in biomedical research as their genomes are assumed to be fixed and reproducible. We examined six such mouse strains using ELITE to determine the variability of TEis between them. The most commonly used mouse strain, C57BL/6J, is the basis for the mouse reference genome (GRCm38.68), which is the primary source of existing TEi annotations. Thus, we ran ELITE on a C57BL/6J sample and a second related strain, B6N-Tyr^{c-Brd}/BrdCrCrl, to assess the degree of TE activity relative to the reference. We consider this pair the B6 type.

Table 1: Total number of TEis found by ELITE in each sample for the six TE templates. A large fraction of these are also in reference. Total 8585 TEis are shared by all samples. A small fractions are private to only one sample indicating potentially recent TE movement.

Sample	TEis	Reference	Shared	Polymorphic	Private
C57BL/6J(m03636A)	11661	11519	8585	3072	4
B6N-Tyr^{c-Brd}/BrdCrCrl	11407	11264	8585	2818	4
A/JCr(f001)	11119	9086	8585	2529	5
A/JOlaHsd(m001)	10639	8673	8585	2047	7
A/JOlaHsd(f015)	11148	9099	8585	2550	13
A/J(f321)	11190	9148	8585	2591	14

We also ran ELITE on four additional samples from a second widely used lab strain, A/J. Two of the A/J samples are independent samples from the same vendor, which allows us to examine TEi pattern relative to that vendor. C57BL/6J(m03636A) incorporates multiple sequencing runs and is a mix of read lengths, 125-bp to 150-bp, all others resulted from a single run (with multiple lanes) and used a uniform 150-bp read length. The Illumina sequencing data from each sample was used to construct an msBWT for each sample as described by Holt [11].

In each sequenced sample, ELITE looked for TEis using six different templates i.e., ERVB7_1-LTR_MM, ERVB4_2B-LTR_MM, RLTRETN_MM, RLTR1IAP_MM, MERVL_LTR, and IAPEY3C_LTR obtained from Repbase. Separate seeds were found for each TE template as described previously. For each template, we selected two conserved kmers as seeds each of length 25 where one is from the proximal side, and the other is from the distal side. The seeds are located within 60-80 bases from the terminal end of the TE sequences. The edit distance threshold, $t2$ was set according to this offset to allow for no more than 1 edit per TE 8 bases, and the minimal read support threshold used was 4 for approximately 25× - 30× genome coverage. The number of mapped TEis per data set is shown in Table 1. These are broken down according to the number of TEis that are included in the reference sequence, those that are common to all six samples, or *shared*, and those that are shared by two to five samples, which we call *polymorphic*, and finally those that appear in only a single sample, which we call *private*.

Table 2: Non-reference TEi sharing patterns. Expected sharing patterns are highlighted in green based on the population structure. There are two primary groups in our population i.e., *B6*, and *AJ*. The *AJ* samples can be further broken down into subpopulations *A/JOla* according to vendor. Sharing patterns that do not match these expectations are likely indirect indicators of ELITE's false-positive and false negative rates indicated in blue and pink respectively. A group of three unclustered and unexpected sharing patterns fill out the table shown in yellow. Combinations of multiple false-negatives and/or false-positives create these patterns.

C57BL/6J	B6N-Tyr^{c-Brd}	A/JCr	A/JOla(m001)	A/JOla(f015)	A/J(f321)	count
0	0	1	1	1	1	1789
0	0	1	0	1	1	74
0	0	0	1	1	1	1
1	1	1	1	1	1	113
1	1	1	0	1	1	8
1	0	1	1	1	1	3
0	1	1	1	1	1	2
1	1	0	1	1	1	1
0	0	0	1	1	0	45
0	0	1	0	0	1	35
1	1	0	0	0	0	10
1	1	0	1	0	0	2
1	1	0	0	0	1	1
0	0	1	1	0	0	2
0	1	1	1	0	0	2
0	0	0	0	1	1	1

Since the mouse reference genome is based on B6, we see a large fraction of TEi found in C57BL/6J and B6N-Tyr^{c-Brd}/BrdCrCrl are also present in reference. Table 2 breaks down the patterns of sharing detected between samples focusing only on those TEis absent from the reference genome. As expected, the single largest pattern of TEi sharing is between the four A/J samples. The second most common pattern of sharing is TEis that appear in all six samples, but do not appear in the reference. There is also significant sharing in subpopulations. In particular, between the two A/JOlaHsd samples, from a common vendor, and the A/JCr and A/J samples distributed from two separate vendors.

By analyzing the sharing patterns of TEis we also gain some insight into ELITE's error rate. The expected patterns of sharing are highlighted as green rows in Table 2. Many of the unexpected sharing patterns are shown clustered with their closest expected pattern and include a highlighted cell which we hypothesize is due to a specific error types. We indicate presumed false negatives in pink, and presumed false positives in blue. ELITE's false-negative error rate brings into question the validity of private TEi which is present only in one sample and has no sharing pattern. Thus, those TEis are best validated by external means, including PCR based experiments which we report on later.

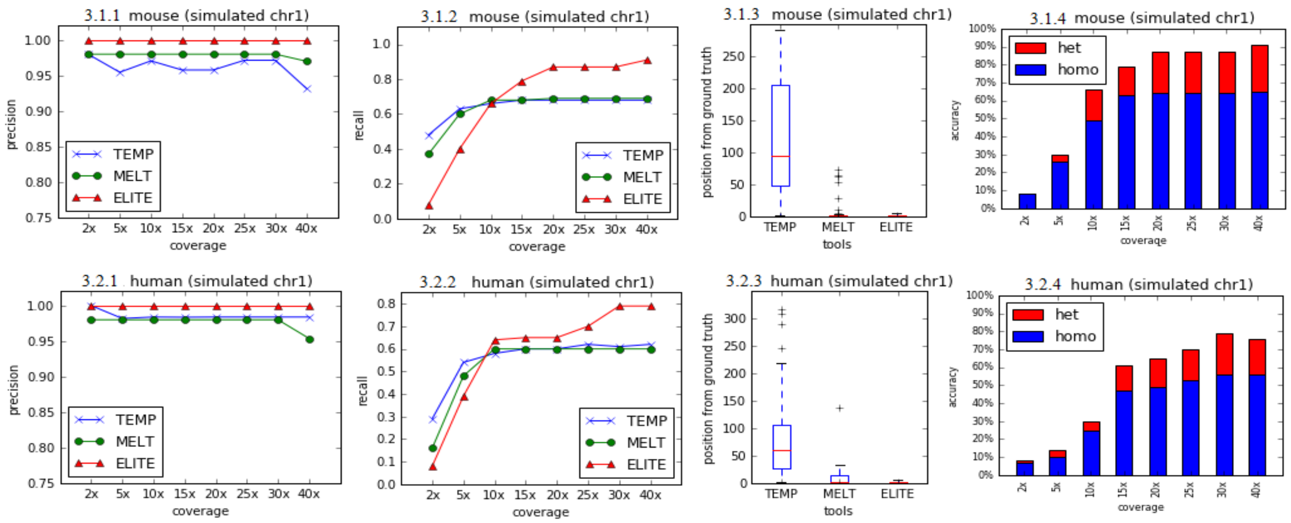


Figure 3: 3.1.1 and 3.2.1: The precision rate of ELITE, MELT, and TEMP for TEi discovery in simulated mouse and human genome as a function of genomic coverage is shown in red, green and blue respectively. As we can see, ELITE did not produce a single false positive in either mouse or human genome resulting in a precision rate of 1. However, TEMP (in both human and mouse) and MELT (in mouse only) produced some false negatives when the coverage is around 40x. 3.1.2 and 3.2.2: The recall rate of three tools. For each tool, it increased with the increase in coverage Although, at low coverage, ELITE performed poorly compared to others, but caught up when the coverage is around 10x. At coverage higher than 10x, MELT and TEMP had many false negatives resulting in a reduced recall rate compared to ELITE. 3.1.3 and 3.2.3: The absolute distance between predicted and actual position for each TEi at coverage 40x. On average, this distance for TEMP was around 100 and 50 bp in mouse (3.1.3) and human (3.2.3) respectively. MELT's predicted position was very close to the ground truth with few exceptions. However, ELITE outperformed both by finding all the TEis within 6bp resolution. 3.1.4 and 3.2.4: The accuracy of zygosity prediction by ELITE in mouse and human respectively. Blue bars stand for homozygous TEi, and red bars stand for heterozygous TEi. Similar to recall, zygosity prediction accuracy also increased with the increase of coverage. At coverage 2x, ELITE failed to locate any heterozygous insertion. Other two tools do not have the zygosity prediction feature.

The last three sharing patterns (total 5 TEis, shown in yellow) are likely a result of multiple errors. There are only two presumed false-negative TEis in the sample with the highest coverage, C57BL/6J. However it is the only sample with a mix of read lengths (125-bp, 151-bp), and the shorter reads may play a role in introducing that error. The single sample with the highest predicted error rate based on these anomalous sharing patterns is A/JOlaHsd(m001), and the type of error is dominated by an access of false negatives (74+8 false negatives vs 2 false positives). This is consistent with the fact that A/JOlaHsd(m001) has lower than usual coverage, which is what we believe drives the false-negative error rate of ELITE.

4.2 Evaluation of ELITE on Simulated Data

To estimate the precision and recall of TEi discovery, we ran ELITE and two additional state-of-the-art TE detection tools, i.e., MELT and TEMP on simulated data. We inserted 100 AluY and 100 ERVB7_1-LTR_MM into chromosome 1 of human and mouse reference genome respectively. Location of these insertions is chosen randomly. Among the 100 TEis in mouse and human, 80 are inserted as homozygous and 30 are heterozygous to estimate ELITE's zygosity prediction rate. We used Samtools [17] to simulate 150-bp paired-end reads at different coverages ranging from 2x to 40x. For each set of simulated data, we built an msBWT index from paired-end reads as required by ELITE. Additionally, as a preprocessing step for both MELT and TEMP, we aligned all the short reads to the corresponding reference genome using BWA. We considered a TEi is found if it's within 500 bases of the ground truth location. Recall and precision rate for all the three tools are shown in figure 3. This figure also includes the distance from ground truth to prediction position and accuracy of ELITE's zygosity prediction.

4.3 Validation via PCR

We validated the presence, absence, and zygosity of the nine predicted TEis found in three of the sequenced samples for which we had available DNA (B6N-Tyr^{c-Brd}/BrdCrCrl(m001), A/JCr(f001), and A/JOla(f015)). We selected at least one TEi from the following categories: shared by everyone (both AJ and B6), polymorphic either in AJ or B6, and private in only one sample. To validate the zygosity predictions we chose four homozygous and five heterozygous TEis. Private TEis are biologically the most interesting ones indicating recent activity as those are absent in other closely related samples. Thus we selected more TEis that are in the private category, and prioritize the ones that are within a gene. In addition to these, we chose two one-sided TEis where our algorithm failed to find both sides of a TE. We created 27 PCR assays (3 assays per TEi) 4. All of those PCR results suggest genotypes that are consistent with ELITE's TEi findings. Thus, via independent methods we have confirmed 100% of 9 of ELITE predicted TEis.

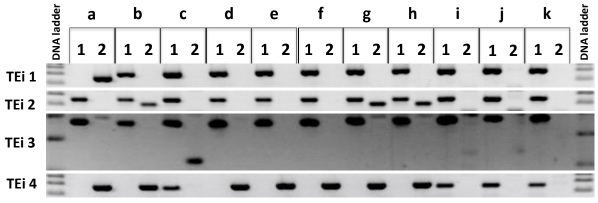


Figure 4: The PCR products from 1) Forward-Reverse and 2) Forward-LTR-Reverse primers reactions for each sample are shown side by side. Reaction 1 and 2 detect the absence and presence of the TEi respectively. Sequenced samples: a) A/JCr(f001), b) A/JOlaHsd(f015), c) B6N-Tyr^{c-Brd}/BrdCrCrl(m001). Additional test samples: d) A/J(m93), e) A/JCrl(f002), f) A/JOlaHsd(m001), g) A/JOlaHsd(m002), h) A/JOlaHsd(f003), i) B6N-Tyr^{c-Brd}/BrdCrCrl(m001), j) C57BL/6J(m001), k) C57BL/6J(f002). TEi 1 is private and found only in A/JCr(f001) at Chromosome 2:58948253. Its presence is confirmed in that sample (a), but it does not appear in any other test sample including a second A/JCr (e). TEi 2 was found in A/JOlaHsd(f015), is also private, and was found in a heterozygous state at Chromosome 4:98626789. The assay (b) confirms this, as well as it being present and heterozygous in 2 of 3 test samples from the same vendor (g,h). TEi 3 is private and was found in B6N-Tyr^{c-Brd}/BrdCrCrl(m001) at Chromosome 1:125336107, where it is confirmed (c). It does not appear in the test sample from the same strain (i). TEi 4 was predicted to be shared among all A/J samples at Chromosome 1:28791918. It was confirmed in the two sequenced samples (a,b) and it also appears in all A/J test samples (d,e,f,g, and h). All four TEis were identified as non-reference, and they are not found in any of the reference-like test samples (j,k).

4.4 Validation via an A/J genome assembly

We used an assembled A/J genome[18] as a means to test if ELITE's predicted TEis are also included in it. At first, we first considered those TEis that are shared by all of our A/J samples. For each TEi, we queried for the context discovered by ELITE in the assembled A/J genome and verified whether it was followed by the expected TE sequence. More than 90% TEis reported in all A/J samples by ELITE were also found in the A/J reference genome. Then we considered only the new predicted TEis (i.e., those not already in the reference) that appear in *any* A/J sample, 1544 of 2113 (73%) appear in the A/J reference genome. Of these, 1440 of the 1789 (80%) that ELITE found in *every* A/J sample also appear in the A/J assembled genome. Next, we considered the predicted false negative cases from Table2. Of the 75 predicted TEis that are shared by all but one A/J sample, 53 (71%) appear in the A/J reference, and all of the 53 were missing from the low coverage A/JOlaHsd(m001) sample. Overall, there is a substantial agreement between the TEi's found by ELITE and those incorporated into the A/J genome assembly. Moreover, it appears that most of ELITE's errors are due to false-negatives in samples with low-coverage. For any missing TEis, the context itself was absent from the new genome, thus not allowing us to verify the presence or absence of the predicted TEi.

4.5 Runtime Comparison

We measured the total time spent on each of the four AJ samples to discover TEis by ELITE, MELT, and TEMP (Table3). At first, we constructed msBWT of each sample for running ELITE. On the other hand, for running MELT and TEMP, we created bam files by aligning all the short reads of each sample to the mouse reference genome using bwa-mem. Each bam file is then sorted and indexed using samtools. We used 6 threads to run our msBWT construction pipeline whereas 30 threads for running all the preprocessing steps of MELT and TEMP due to its computational demand. These preprocessing steps for each tool were run on a machine with the following specification: Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz, 6 cores, 256 GB memory. Running the actual TE searching tool after the preprocessing step does not require heavy resource. Hence we ran the rest of the steps on an Intel(R) Xeon(R) E5420 CPU, 4 cores, 2.50 GHz with 32 GB RAM. ELITE, MELT, and TEMP are written in Python, Java, and Perl, respectively.

Table 3: Total time required for each tool to locate six classes of TE in four A/J samples (in minutes). Data preprocessing time of each tool is shown in the last two columns, where the BWT index corresponds to ELITE, and Alignment timing corresponds to both MELT and TEMP.

Sample	ELITE	MELT	TEMP	BWT index	Alignment
A/JCr(f001)	55	95	312	581	507
A/JOlaHsd(m001)	42	72	238	178	267
A/JOlaHsd(f015)	32	55	200	403	407
A/J(f321)	45	77	360	408	514

As we can see from table 3, ELITE is about 1.7 times faster than MELT. TEMP is significantly slower than both ELITE and MELT. It is also apparent that the preprocessing step for each tool is the most computationally heavy step. However, even in this case, except for sample A/JCr(f001), creating a BWT index is always faster than typical alignment. Other than providing a faster way to detect TEi, msBWT is not biased to any reference genome and has many uses, including data compression, fast kmer query, local assembly, local alignment, error correction, etc [25][23][10][28].

5 CONCLUSIONS

We have developed a tool ELITE, that uses a novel *local-genome-assembly-based* algorithm to efficiently discover TEis and ran it effectively on real large-scale data. In addition to several independent validation methods, we also proved the legitimacy of ELITE's findings by showing its presence in real DNA. We showed different pattern of TEi discovered by ELITE within a population is highly consistent with their origin. ELITE's TEi presence and absence probe are useful for the biologists to create primer for PCR validation. Results produced by ELITE also led to interesting biological finding i,e, many private TEis segregating in a closely related population are in the heterozygous state. They are thus providing stronger evidence for recent activity. Overall, the large number of TEis found by ELITE indicate that TEs are a significant source of genetic variation that must be taken into consideration.

Funding: *This research was supported by NIH NHGRI grant U24HG010100 Genomic Resources for the Collaborative Cross.*

REFERENCES

[1] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. 1990. Basic local alignment search tool. *Journal of molecular biology* 215, 3 (1990), 403–410.

[2] W Bao, KK Kojima, and O Kohany. [n. d.]. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015; 6: 11.

[3] Victoria P Belancio, Dale J Hedges, and Prescott Deininger. 2008. Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. *Genome research* 18, 3 (2008), 343–358.

[4] Jinfeng Chen, Travis R Wrightsman, Susan R Wessler, and Jason E Stajich. 2017. RelocaTE2: a high resolution transposable element insertion site mapping tool for population resequencing. *PeerJ* 5 (2017), e2942.

[5] Jian-Min Chen, Peter D Stenson, David N Cooper, and Claude Férec. 2005. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Human genetics* 117, 5 (2005), 411–427.

[6] International Human Genome Sequencing Consortium et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 6822 (2001), 860.

[7] Mouse Genome Sequencing Consortium et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 6915 (2002), 520.

[8] Paolo Ferragina and Giovanni Manzini. 2000. Opportunistic data structures with applications. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on.* IEEE, 390–398.

[9] Eugene J Gardner, Vincent K Lam, Daniel N Harris, Nelson T Chuang, Emma C Scott, William S Pittard, Ryan E Mills, Scott E Devine, 1000 Genomes Project Consortium, et al. 2017. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome research* (2017), gr–218032.

[10] Seth Greenstein, James Holt, and Leonard McMillan. 2015. Short read error correction using an FM-index. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on.* IEEE, 101–104.

[11] James Holt and Leonard McMillan. 2014. Constructing Burrows-Wheeler transforms of large string collections via merging. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics.* ACM, 464–471.

[12] Chuan Jiang, Chao Chen, Ziyue Huang, Renyi Liu, and Jerome Verdier. 2015. ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. *BMC bioinformatics* 16, 1 (2015), 72.

[13] Thomas M Keane, Kim Wong, and David J Adams. 2012. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* 29, 3 (2012), 389–390.

[14] W James Kent. 2002. BLATâĂŤthe BLAST-like alignment tool. *Genome research* 12, 4 (2002), 656–664.

[15] Ben Langmead and Steven L Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* 9, 4 (2012), 357.

[16] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 10, 3 (2009), R25.

[17] H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, and R Durbin. [n. d.]. 692 (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 16 ([n. d.]), 2078–693.

[18] Jingtao Lilue, Anthony G Doran, Ian T Fiddes, Monica Abrudan, Joel Armstrong, Ruth Bennett, William Chow, Joanna Collins, Stephan Collins, Anne Czechanski, et al. 2018. Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nature genetics* (2018), 1.

[19] Paolo Mita and Jef D Boeke. 2016. How retrotransposons shape genome regulation. *Current opinion in genetics & development* 37 (2016), 90–100.

[20] Koji Muratani, Toshikazu Hada, Yoshihiro Yamamoto, Tadashi Kaneko, Yoshihisa Shigeto, Toru Ohue, Junichi Furuyama, and Kazuya Higashino. 1991. Inactivation of the cholinesterase gene by Alu insertion: possible mechanism for human gene transposition. *Proceedings of the National Academy of Sciences* 88, 24 (1991), 11315–11319.

[21] Matthew T Reilly, Geoffrey J Faulkner, Joshua Dubnau, Igor Ponomarev, and Fred H Gage. 2013. The role of transposable elements in health and diseases of the central nervous system. *Journal of Neuroscience* 33, 45 (2013), 17577–17586.

[22] Lavanya Rishishwar, Leonardo Mariño-Ramírez, and I King Jordan. 2016. Benchmarking computational tools for polymorphic transposable element detection. *Briefings in Bioinformatics* 18, 6 (2016), 908–918.

[23] Kamil Salikhov, Gustavo Sacomoto, and Gregory Kucherov. 2013. Using cascading Bloom filters to improve the memory usage for de Brujin graphs. In *International Workshop on Algorithms in Bioinformatics.* Springer, 364–376.

[24] Phillip SanMiguel, Alexander Tikhonov, Young-Kwan Jin, Natasha Motchoulskaia, Dmitrii Zakharov, Admasu Melake-Berhan, Patricia S Springer, Keith J Edwards, Michael Lee, Zoya Avramova, et al. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274, 5288 (1996), 765–768.

[25] Jared T Simpson and Richard Durbin. 2010. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* 26, 12 (2010), i367–i373.

[26] José AJM van den Hurk, Dorien JR van de Pol, Bernd Wissinger, Marc A van Driel, Lies H Hoefsloot, Ilse J de Wijs, L Ingeborgh van den Born, John R Heckenlively, Han G Brunner, Eberhart Zrenner, et al. 2003. Novel types of mutation in the choroideremia (CHM) gene: a full-length L1 insertion and an intronic mutation activating a cryptic exon. *Human genetics* 113, 3 (2003), 268–275.

[27] Margaret R Wallace, Lone B Andersen, Ann M Saulino, Paula E Gregory, Thomas W Glover, and Francis S Collins. 1991. A de novo Alu insertion results in neurofibromatosis type 1. *Nature* 353, 6347 (1991), 864.

[28] Jeremy R Wang, James Holt, Leonard McMillan, and Corbin D Jones. 2018. FMLRC: Hybrid long read error correction using an FM-index. *BMC bioinformatics* 19, 1 (2018), 50.

[29] Travis J Wheeler, Jody Clements, Sean R Eddy, Robert Hubley, Thomas A Jones, Jerzy Jurka, Arian FA Smit, and Robert D Finn. 2012. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic acids research* 41, D1 (2012), D70–D82.

[30] Jiali Zhuang, Jie Wang, William Theurkauf, and Zhiping Weng. 2014. TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic acids research* 42, 11 (2014), 6826–6838.