ERVcaller: Identifying and genotyping non-reference unfixed endogenous retroviruses (ERVs) and other transposable elements (TEs) using next-generation sequencing data

Xun Chen[1], and Dawei Li[1,2,3,4*]

[1]*Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405, USA*

[2]*Department of Computer Science, University of Vermont, Burlington, Vermont 05405, USA*

[3]*Neuroscience, Behavior, and Health Initiative, University of Vermont, Burlington, Vermont 05405, USA*

[4]*University of Vermont Cancer Center, University of Vermont, Burlington, Vermont 05405, USA*

*To whom correspondence should be addressed:

Dawei Li, Ph.D., Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405, USA. E-mail: dawei.li@uvm.edu

Number of words in the abstract: 196

Number of words in the main text (excluding online materials, acknowledgments, financial disclosures, legends, and references): 4,960

Number of Table: 1

Number of Figures: 4

Number of supplementary materials: 6 Supplementary Tables and 5 supplementary Figures and Legends

ABSTRACT

More than 8% of the human genome is derived from endogenous retroviruses (ERVs). In recent years more and more polymorphic (unfixed) ERVs were found to be associated with human diseases. However, it is still challenging to accurately detect polymorphic ERVs using next-generation sequencing (NGS) data. As such, development of new tools is needed to rectify this shortcoming. We designed a new tool, ERVcaller, to accurately detect and genotype non-reference unfixed transposable elements (TEs), including ERVs, in the human genome. We evaluated the tool using both simulated and real benchmark whole-genome sequencing datasets. The detection power and precision were ~98% and ~99%, respectively, and the genotyping accuracy was ~97% under 30X sequencing depth. We compared ERVcaller with four existing tools. ERVcaller consistently showed the highest sensitivity and precision for detecting and genotyping non-reference unfixed TEs, particularly ERVs, as well as produce the most accurate breakpoints at single nucleotide resolution, especially under low sequencing depths. By applying ERVcaller to the 1000 Genomes Project samples, we detected 100% of the ERVs and 95% of the TEs that were experimentally validated. In conclusion, ERVcaller is capable of accurately identifying and genotyping non-reference unfixed ERVs and other TEs using NGS data.

Keywords: Endogenous retrovirus; Non-reference unfixed ERV; Polymorphic ERV; Transposable element (TE); Next-generation sequencing (NGS); Low sequencing depth

INTRODUCTION

Origins of ERVs

Transposable elements (TEs), which were first found in maize, are groups of mobile DNA sequences that collectively comprise a large percentage of the genome in most species (e.g., ~45% in the human genome). TEs can be classified into two classes based on transposition mechanisms, Class I: "copy and paste" manners, and Class II: "cut and paste" manners. Endogenous retroviruses (ERVs) are a unique group of Class I TEs resulting from the fixation of ancient retroviral infections and integrations into the host genome. In the human genome, 8% of the sequences are of retroviral origin, containing > 100,000 ERV loci from approximately 50 families (1,2). The human ERV-K family (HERV-K or HML-2), a sub-group of ERVs, have most recently inserted into the human genome (i.e., ~35 million years ago), and remain replication active at some loci (3,4). Most of the other human ERVs are replication defective or inactivated because of accumulated mutations, insertions, and deletions, or solo-long terminal repeats (solo-LTRs) (2,3). However, replication or reactivation can still happen in some scenarios, including intracellular retro-transposition or trans-activation by viral infections, such as human immunodeficiency virus and Epstein-Barr virus (5,6).

ERV in human diseases

ERVs may have concurrent effects on both genome structure and biological function, having been correlated with many human diseases. ERVs may also lead to genomic rearrangements through non-allele homologous recombination with other ERV copies. Additionally, ERVs may function as a source of promoters, enhancers or transcriptional factor binding sites for regulating human gene expression levels. The viral gene expression of ERVs may interact with the human transcriptome or modulate the human immune system (2). The majority of ERVs are ancient retroviral integration events, and are defective and nonfunctional; however, under certain diseased conditions, these silenced ERVs can be re-activated and expressed. They can systematically transcribe stage-specific RNAs in early embryo development (7,8). These functional ERVs are most likely to be unfixed ERVs, i.e., the part of the ERVs that are not shared by all the human genomes and thus are polymorphic in the human populations. Although the pathogenicity of the unfixed ERVs is still poorly understood due to the complexity of ERV sequences in the human genome, unfixed ERVs have been associated with many diseases, such

as autoimmune diseases (9,10), mental disorders (6,11,12), and cancers. In various cancers, significantly upregulated expression of ERV RNAs and higher levels of proteins have been observed in tumors versus adjacent normal tissues (13,14).

Existing tools for detecting unfixed ERVs

Compared to fixed ERVs, which are shared by all the human genomes and more likely to be degraded, the newly-generated, unfixed ERVs are more likely to be intact and functional, and thus associated with human diseases (3). This study focuses on the unfixed ERVs. Thus, to determine the pathogenic effects of ERVs, it is necessary to first identify all the unfixed ERV events from the human genome. Similar to single nucleotide polymorphisms, the same unfixed ERVs can be found in many individuals, leading to common unfixed ERV loci. Indeed, many unfixed ERV loci have been recently discovered (3,4,15,16). For example, Lee *et al*. discovered a total of 15 unfixed ERV loci by screening 44 samples (17). Emanuele *et al*. found 17 unfixed ERV loci, including two additional loci, by screening more samples (3), with an average of six unfixed ERV events per human genome. By analyzing a different set of samples, Wildschutte *et al*. found 19 additional unfixed ERV loci (4). These studies indicate that more unfixed ERVs loci likely have not yet be discovered in the human populations.

With wide-spread applications of NGS technologies, bioinformatics tools were developed to discover and characterize various TEs. However, because of the long fragment insertions (e.g., ~20 kilo base pairs for ERVs) and highly abundant and repeated sequences, the detection and genotyping of ERV events are more difficult than detection of single nucleotide mutations, InDels or other structural variations. Structural variation detection tools, including VariationHunter (18,19) and Hydra (20), were first adapted for detecting unfixed TEs in whole-genome sequencing (WGS) data. Then, specific tools were developed, including RetroSeq, TIF, Mobster, Tangram, TEMP, ITIS, and STEAK (21,22,23,24,25,26,27). However, many of these tools were insufficient for accurately detecting all ERV and other TE events, were not able to genotype these events, or had other functional limitations.

A new software

In this study, we developed ERVcaller, a novel TE calling software, to accurately call and genotype "non-reference unfixed" (referred to as "unfixed" in this study) TE events, particularly ERVs, with paired-end or single-end NGS data. It considers three types of supporting reads with stringent quality control procedure for the detection and genotyping of unfixed TEs. In this study, ERVcaller was evaluated primarily using datasets with known ERV insertions. Our results showed that ERVcaller had the highest detection power and precision for detecting unfixed TE events, especially in data with low sequencing depths. It can accurately detect TE breakpoints at single nucleotide resolution. It is also capable of detecting somatic TE insertions in tumor tissues. The genotyping accuracy was above 97% when the sequencing depths were higher than 5X. Additionally, it is the only tool that can obtain high detection power and precision consistently with various TE references of different levels of sequence complexity. The performance of ERVcaller was also demonstrated in the benchmark 1000 Genome Project samples.

MATERIAL AND METHODS

Reference sequences for ERV detection

To evaluate how the use of different TE references influenced the accuracies of ERV detection, we included four different examples of TE reference sequences and compared the performance among them: (1) The HERV-K reference (GB: KU054272); (2) An ERV library containing a total of 743 diverse human or non-human ERV sequences extracted from our in-house viral database (unpublished); (3) A human TE library containing a total of 23 consensus human TE sequences, including 17 long interspersed nuclear element 1 (LINE-1 or L1) sequences, four Alu sequences, one HERV-K sequence, and one short interspersed nuclear element (SINE)/variable-number tandem-repeat (VNTR)/Alu (SVA) sequence, which were obtained from Tangram (24); (4) A eukaryotic TE library obtained from the RepBase database (28). The human reference genome GRCh37 (hg19) build was downloaded from the UCSC Genome Browser for the simulated datasets; and the GRCh38 (hg38) build was used to match the analysis of the 1000 Genomes Project samples.

Simulation

To evaluate ERV detection accuracies, we simulated unfixed ERV events by randomly inserting 500 whole and fragmented (partial) ERV genomes into the human genome using a in-house perl script. Paired-end reads with 100 base pairs (bp) length and 500 bp insert size were simulated using pirs (29) with 0.37% mutation rates. To evaluate the performance for detecting unfixed ERV events, two series of data were simulated: one with 500 ERV events of whole and partial ERV insertions derived from HERV-K library, and another with 500 ERV events of randomly inserted whole and partial genomes derived from the ERV library. The datasets were simulated using the human chromosome 1 (hg19) under sequencing depths of 1X, 2X, 3X, 4X, 5X, 10X, 15X, 20X, 30X, 40X, and 50X.

To evaluate genotyping accuracies, we simulated a series of datasets carrying unfixed ERV events with both homozygous and heterozygous insertions under sequencing depths of 5X, 10X, 30X, and 50X. For each depth, we simulated one copy of paired-end reads with 500 unfixed ERV events and another copy of paired-end reads with only 250 of these 500 unfixed ERV events. By mixing the two copies into one dataset, we obtained 250 homozygous and 250 heterozygous unfixed ERV events.

ERVcaller pipeline

ERVcaller first aligns raw FASTQ reads to the human reference genome using BWA-mem with the default parameters (30) for WGS data or using Bowtie for RNA-seq data, or directly uses the aligned read files in BAM format as the input. Then, it extracts all chimeric and improper reads (≥ 20 bp) in BAM format, and converts the extracted reads to FASTQ format using Samtools (31). Meanwhile, the paired-end reads containing split reads are extracted using Samtools (31). ERVcaller then converts all split reads (≥ 20 bp) into FASTQ reads using SE-MEI (https://github.com/dpryan79/SE-MEI). Chimeric reads are defined as paired-end reads in which one end aligns to the human reference genome and the other end aligns to a TE sequence; Split reads are reads that partially align to the human reference genome and partially align to a TE sequence (split reads must cross the insertion breakpoint); Improper reads are defined as one paired-end reads that are not aligned in proper pair (the two ends are not aligned to the same chromosome or a nearby region within the regular insert size) (**Supplementary Figure 1**). All the chimeric and split reads are aligned against each of the TE reference sequences (libraries)

using BWA mem with the customized parameters (**Supplementary Table 1**). Since one end of the improper read can be aligned to multiple locations on the human genome, it is likely that this end is derived from TEs. Hence, we also align one end of the improper reads to the TE reference library using the BWA mem with the same parameters.

The supporting reads within the two-insert size window are grouped to identify the candidate unfixed TE regions. ERVcaller further determines the unfixed TE events using the number of reads, and average alignment scores of all three types of supporting reads. The chimeric, improper, and split reads are further aligned to the candidate genomic regions only, and the TE events with no confident supporting reads are removed (i.e., supporting reads that can be fully aligned to the human reference genome are removed). To report an unfixed TE event by ERVcaller, a minimum of two supporting reads, including one uniquely mapped chimeric or improper paired-reads, are required. We use chimeric, improper, and split reads to fine-map chromosomal location of each breakpoint. If there are split reads spanning the breakpoint, the breakpoint locations can be detected at single nucleotide resolution. If there are no split reads, the breakpoint locations will be estimated according to two different cases: 1) the first nucleotide of the nearest read is considered to be the breakpoint if only one breakpoint is identified; and 2) the median of the first nucleotide of the nearest read from both the upstream and downstream breakpoints is considered to be the breakpoint if both breakpoints are identified.

The total number of chimeric, improper, and split reads supporting the hypothesis of an unfixed ERV event versus the total number of reads supporting the hypothesis of no unfixed ERV is used to determine the genotype of the ERV event. The reads that support the hypothesis of no unfixed ERV can be fully aligned to the human reference genome across the breakpoint. We extract these reads from the aligned read files in BAM format using Samtools (31). For example, if half of the reads support the presence of an unfixed ERV and the other half of the reads support no ERV at this locus, the genotype of this ERV is heterozygous (**Supplementary Figure 2**). However, genotyping may be complicated due to noise derived from the sequencing process, alignment, and subsequent analyses. Therefore, we calculated the genotype likelihood to determine the homozygous versus heterozygous state of each identified unfixed ERV event using a combined insert size and read depth approach (32). Specifically, we first calculated the count

of reads supporting no ERV insertion at the ERV breakpoint (**Supplementary Figure 2**). Second, we calculated the read counts of randomly-selected genomic locations interspersed every 50 bp within a window (i.e., 50 kilo bp) centered at the ERV breakpoint to obtain the distribution of read counts (reads within one insert size of the breakpoint were excluded). Third, we compared the read count supporting no ERV insertion (from step 1) with the distribution of the randomly sampled read counts (divided by two) from the window (from step 2) to calculate its quantile. Last, we determined the genotype based on the quantile (threshold = 0.05), i.e., quantile < 0.05 and quantile > 0.05 represent ERV homozygote and heterozygote, respectively. During the process, for better calculation of the read counts supporting no ERV insertion, the reads within 10 bp from each genomic location (applied to both of the ERV breakpoint and the genomic locations from every 50 bp window) as well as the reads outside of the insert size Poisson distribution around the ERV breakpoint were removed from the calculations. Finally, ERVcaller outputs the results into a table, which can be easily converted to Variant Call Format (33) or Plink (34) format for subsequent genetic association analyses.

Software comparison

To compare ERVcaller with other similar tools, we selected four of the most recently published tools on TE detection, including ITIS (26), TEMP (25), RetroSeq (21), and STEAK (27). After successfully installing and compiling each software, we ran each tool with the default parameters using the same datasets. BWA-MEM (30) was used to align raw reads to the human reference genome. The generated BAM files were used as the inputs for RetroSeq, TEMP, STEAK, and ERVcaller except ITIS, which only accepted the FASTQ files as the inputs. Only unfixed ERV events with two or more supporting reads were used for the comparisons.

The 1000 Genomes Project samples and data preprocessing

To further evaluate the accuracy of ERVcaller, we applied it to four (~4X sequencing depth) of the 1000 Genomes Project samples that have known ERV events validated by Sanger sequencing (4) for the detection of these unfixed ERVs, and to three other samples (~15X sequencing depth; NA12878, NA12891, and NA12892) that have known TEs validated by polymerase chain reaction (PCR) (35) for the detection of these TE events. The aligned read files in CRAM format (hg38) and raw FASTQ files were downloaded from the International Genome Sample Resource

database. The FASTQ files were aligned to the human reference genome (hg38) using BWA mem with the default parameters (30), then the output files were converted to BAM files, and sorted and indexed using Samtools (31). Samtools was also used to convert the CRAM alignment files to indexed BAM files. The latter was then submitted to ERVcaller and other tools for TE detection.

To measure the sensitivity, the PCR validated Alu, L1, and SVA insertions reported by Stewart *et al*. (35) were used as the standard. The hgLiftOver tools (https://genome.ucsc.edu/cgi-bin/hgLiftOver) were used to convert the coordinates from build hg18 to hg38. To measure the genotyping accuracy, the PCR validated genotypes of the three samples reported by Stewart *et al*. (35) were used for comparison.

RESULTS

Software development

For accurate detection and genotyping of non-reference unfixed ERVs and other TEs, we improved the calling algorithms and designed a new bioinformatics tool, ERVcaller. It is composed of three modules: **a**) extracting unmapped reads, **b**) obtaining supporting reads, and **c**) detecting and genotyping unfixed ERV and other TE events (**Figure 1**). The inputs of ERVcaller include either raw FASTQ or BAM file, the human reference genome, and a library of TE reference sequences in FASTA format. With ERVcaller, we first extract all reads that cannot be fully mapped to the human reference genome. Among the remaining unmapped reads, we obtain all the chimeric and split reads. In addition, we obtain the improper reads by extracting all reads in which the two ends are not aligned to the same chromosomes or nearby regions within regular insert size (**Supplementary Figure 1**). All the chimeric and split reads are then aligned to the TE reference library; For improper reads, only the ends that have multiple hits in the human genome are aligned to the reference library. We then use all three types of supporting reads to determine the TE chromosomal location, including breakpoints, at single nucleotide resolution. We filter low confidence events by removing each event that meets all the following criteria: 1) it has less than two supporting reads, 2) it has no reads uniquely mapped to the human genome, 3) the average alignment score of the supporting reads is lower than 30, and 4) the reads mapped to the human genome are no more than 50 bp in length. After stringent filtering, high confidence

unfixed TE events are then genotyped based on the reads across the breakpoints, as described in the Method section (**Supplementary Figure 2**).

Detection and genotyping accuracies of unfixed ERVs

We applied ERVcaller to a series of simulated datasets to measure the detection accuracies of unfixed ERVs. The simulated datasets included those containing whole HERV-K insertions and those containing partial (randomly segmented) HERV-K insertions (> 500 bp). ERVcaller was then applied to each of them using the HERV-K sequence as the reference. The detection power and precision were calculated under each sequencing depth, from 1X up to 50X (**Figure 2a**). The results from both series of datasets consistently showed that ERVcaller was able to accurately detect HERV-K whole or partial sequence insertions. For example, ERVcaller was able to detect ~97% of the simulated ERV events with ~99% precision under 5X sequencing depth (**Figure 2a**). When the sequencing depth was increased to 30X, ~98% of the ERV insertions were detected with ~99% precision.

To evaluate if repetitive and redundant sequences in the TE reference sequences influenced the detection accuracy, we applied ERVcaller to the simulated datasets, including whole and partial HERV-K insertions, using each of the four different TE sequence reference or reference libraries, including the HERV-K reference, an ERV library, a human TE library (24), and a eukaryotic TE library (28) (**Figures 2b** and **2c**). We observed slightly higher detection power and precision in the data with whole HERV-K insertions, compared to partial HERV-K insertions (**Supplementary Figure 3**). However, the detection power was 97% or higher and the precision was 99% or higher under 30X for all TE reference or reference libraries. Hence, we conclude that ERVcaller has high detection power and precision, with little influence from selection of different TE sequence references.

To estimate the accuracy of ERVcaller for genotyping detected unfixed ERV events, we generated a series of simulated datasets containing homozygous and heterozygous unfixed ERV events. We determined the genotype of each ERV event by calculating the number of reads supporting the presence of the event versus the number of reads supporting the absence of the event across the breakpoint. If there are one or more split reads, the breakpoint will be detected

directly. Without any split reads, the breakpoint will be estimated using the chimeric and improper reads, and the genotype will be calculated based on the estimated breakpoint. Our results showed that the genotyping accuracies were 98% for datasets with sequencing depths under 5X; and above 97% for datasets with depths 30X or higher. Within each dataset, there were no significant differences in accuracy between homozygous and heterozygous ERV events, no matter the ERV insertions were whole or partial HERV-K sequences (**Table 1**).

Software comparison

We performed a comprehensive comparison of the performance of ERVcaller with four recently published tools for TE detection, including RetroSeq, STEAK, ITIS, and TEMP (21,25,26,27). We first used the HERV-K sequence as the only reference to detect the simulated candidate HERV-K events, then used the large ERV library that contained diverse human and non-human ERV sequences as the references to detect the same HERV-K events, and last used this ERV reference library to detect all simulated events of different HERV families. The simulated whole and partial HERV-K insertions were evaluated separately. Datasets of various sequencing depths were used for the evaluation of detection power and precision. Additionally, we compared the accuracies for detecting ERV breakpoints.

We first measured the accuracy of each tool for detecting the candidate HERV-K events using the HERV-K sequence as the reference. The results showed that STEAK and ITIS had extremely low (< 10%) power for detecting the partial insertions (**Figure 3b**). STEAK and TEMP had low precision for both whole and partial insertions, and the precision decreased with increased depths (**Figure 3e**). Only ERVcaller and RetroSeq were able to accurately detect the candidate HERV-K events with ERVcaller showing relatively higher detection power (**Figure 3a** and **Figure 3b**). We then measured the accuracy for detecting the candidate HERV-K events using the large ERV library as the references. Compared to the analyses described above, the precision of TEMP and ITIS decreased (**Supplementary Figure 4b**). STEAK was not applicable to the use of multiple reference sequences. Only ERVcaller and RetroSeq showed consistently high detection power and precision. We further measured the accuracy for detecting all ERV insertions using the large ERV library. Compared to the analyses of detecting only HERV-K insertions, the detection power decreased slightly in TEMP, RetroSeq, and ERVcaller (**Figure 3c**

and **3d**); however, regarding the precision, only ERVcaller was consistently around 99%, while the precision decreased significantly (< 90% under 30X) in RetroSeq, TEMP, and ITIS (**Figure 3g** and **3h**). STEAK was not applicable for multiple references and thus not analyzed. Thus, only ERVcaller was capable of detecting a wide range of ERV events with both high precision and detection power.

As precise breakpoints of detected ERV events are also important, we compared the five tools for their accuracies of breakpoint detection. Both HERV-K and ERV insertions as well as the HERV-K reference and the large ERV library were evaluated. The results showed that only ERVcaller, RetroSeq, and TEMP were able to report breakpoints at single nucleotide resolution. When stringent criteria, such as a two-base pair window, were used to measure whether the breakpoints were correctly detected, ERVcaller showed the highest accuracy (**Supplementary Table 2**). Thus, ERVcaller was able to detect ERV breakpoints with the highest accuracy, regardless of length of ERV insertion and selection of reference sequences. Moreover, we analyzed the distribution of the distances from the detected breakpoints (i.e., > 2 bp in length) to the simulated breakpoints (**Supplementary Figure 5**). Shorter distances represent higher accuracy for breakpoint detection. As expected, ERVcaller showed a distribution of the shortest distances compared to other tools. For example, TEMP showed a hotspot at around 250 bp, suggesting a significant deviation of detected breakpoints. In conclusion, ERVcaller was capable of accurately detecting ERV breakpoints at single nucleotide resolution.

Finally, we compared the speed of ERVcaller with the other four tools using the HERV-K sequence and the ERV library separately. We calculated the runtime of each tool using the Vermont Advanced Computing Core cluster with computing nodes having 12 cores and 48 GB memory. Aside from ITIS, which required FASTQ files as the input, all the other tools used BAM files as input. Our results showed that when the HERV-K reference was used, ERVcaller was slightly slower than TEMP, but faster than the other three tools analyzed in this study, including RetroSeq, STEAK, and ITIS. When the ERV library was used, ERVcaller and TEMP were the fastest, followed by RetroSeq and ITIS (STEAK was not analyzed because it was not applicable to multiple references). The runtimes of ITIS and RetroSeq increased significantly when the ERV library was used compared to the HERV-K reference (**Figure 4**).

Detection and genotyping of unfixed ERVs and other TEs in the 1000 Genomes Project samples

To evaluate the sensitivity of ERVcaller for detecting non-reference unfixed ERVs and other TEs in real WGS data, we first applied it to detect the six unfixed ERV events that have been identified and validated by PCR in four benchmark 1000 Genomes Project samples (4). Our results showed that ERVcaller detected all of these events using the default parameters (**Supplementary Table 3**). To measure the sensitivity of ERVcaller for detecting other TE events, we then applied it to detect the 2,670 unfixed TE events that were identified and validated by PCR in a trio of the 1000 Genomes Project samples (35). Consistently, ERVcaller detected almost all of them, i.e., 96%, 88%, and 100% for Alu, L1, and SVA events, respectively (**Supplementary Table 4**). By comparison, RetroSeq obtained 94% and 78% sensitivity for Alu and L1, respectively; TEA obtained 93% and 82%, respectively; and Tangram showed 98.8% and 86.6%, respectively, based on the sensitivity described in a previous study (24). We further measured the genotyping accuracy of ERVcaller using the unfixed TEs that have been validated by PCR in a previous study (35). Among a total of 999 TE genotypes, ERVcaller correctly genotyped ~98% of them (**Supplementary Table 4**), which was significantly higher than the accuracies from other tools based on the analysis by Wu *et al*., 2015 (24).

DISCUSSION

Non-reference unfixed (polymorphic) ERVs and other TEs are an important part of structural variations and potentially critical in human evolution and the development of human diseases (2,36). They are typically repetitive sequences, and thus difficult to detect using short NGS reads. In this study, we improved the TE calling algorithm and developed ERVcaller to accurately detect and genotype non-reference unfixed TE events using NGS data. As we considered three types of supporting reads: chimeric, improper, and split reads, as well as implemented the stringent quality control procedure, ERVcaller achieved reasonably high detection power and precision. When we applied it to series of simulated datasets, we observed ~98% detection power and ~99% precision under 30X sequencing depth (**Figure 2**). Under low sequencing depths (< 5X), ERVcaller still showed the highest detection power and precision compared to other tools analyzed in this study (**Figure 3** and **Supplementary Figure 4**). In

addition, ERVcaller is capable of fine-mapping the breakpoints at single nucleotide resolution, with an average of > 90% accuracy under 30X sequencing depth, which is significantly higher than other tools (**Supplementary Table 3**). More importantly, ERVcaller achieved high genotyping accuracies, e.g., 97% accuracy under 30X sequencing depth. ERVcaller can be applied to both paired-end and single-end data from WGS, RNA-sequencing or targeted DNA sequencing.

ERVcaller accurately detected and genotyped unfixed TE events, particularly ERVs. ERVcaller achieved high sensitivity due to several reasons, such as implementation of the high mapping rate aligner BWA-mem; use of all three types of reads to detect unfixed ERV events; adoption of a multiple-mapping approach to simultaneously align each read to multiple references or locations; and ranking to determine the top candidate location based on the average alignment scores of all supporting reads. Meanwhile, we implemented stringent filtering steps, which significantly increased the precision of TE detection and genotyping. By comparison, each of the tools analyzed here showed one or multiple major limitations. For example, TEMP had no function for genotyping (25); ITIS was not practical for large datasets because ITIS screened each TE reference separately and the computational time increased significantly with increased number of TEs (26); RetroSeq was able to genotype unfixed ERV events (21,24); however, the genotyping function has been removed from the latest version of RetroSeq. Neither ITIS nor STEAK was able to detect partial ERV insertions (**Figure 3**) or to fine-map breakpoints. By comparing with these tools (**Figure 3** and **Supplementary Figure 4**), ERVcaller detected TE events, particularly ERVs, with reasonably high detection power and precision, and with little influence from number of TEs or selection of TE references.

ERVcaller shares some of the limitations common to other NGS-based approaches. As it is alignment-based, ERVcaller requires a TE reference or reference library. As a trade-off of precision, ERVcaller potentially misses a certain number of ERVs located in repetitive sequence regions, which is common to all similar tools. The structure of non-reference ERV insertions can be complex because of target site duplications (15), making these events even harder to detect. These issues may be partially addressed by increasing insert size, read length, or both. In this study, only unfixed ERVs with two or more supporting reads were used for the comparisons,

which might be slightly different if more or less stringent criteria were used; however, we did not expect significant changes in the comparisons. In addition, not all the published tools were compared in this study, such as TIF (22) and MELT (37). The former only uses split reads for detection. **Supplementary Table 5** shows a comparison of the characteristics of seven tools, including those not analyzed in this study.

Many high frequency unfixed ERV loci, i.e., loci observed in many individuals in a population, have been discovered previously (3,4). However, evidence suggests that there are still a large number of undiscovered unfixed ERV loci in the human genome, including population- or disease-specific loci (4). For example, by screening more than 2,000 samples, Wildschutte *et al*. detected many rare, less-common, and population-specific unfixed ERV events (allele frequency < 5%). Thus, more unfixed ERV events may be discovered as more samples are analyzed appropriately. For instance, among the 19 analyzed samples, ERVcaller identified a total of 311 ERV events with an average of 16.4 non-reference unfixed ERV events (ranging from 9 to 31) per individual.

More importantly, there is a considerable amount of variation of ERV properties in the general population, such as the number of unfixed ERVs per human genome, and the frequencies of these ERVs (up to 0.75) in the general populations (4). This indicates that polymorphic ERVs may also play important roles in human disease and health. Thus, for subsequent genetic association analyses using the identified ERV information, we calculated a variety of measurements of ERVs as an output file of the ERVcaller software. For example, for each analyzed sample, ERVcaller provides the total number (or copy number) of unfixed ERV and other TE events. For each ERV event, ERVcaller provides several items, including ERV family, ERV location on the human genome, genotype of the ERV event, and the length of inserted sequence at the breakpoint (i.e., target site duplication). Other information, such as population frequency, can be obtained using other sources (4). **Supplementary Table 6** shows the variables measured by ERVcaller for future genetic association studies.

To conclude, we developed a new tool for identifying and genotyping non-reference unfixed TEs, particularly ERVs, using NGS data. As it uses multiple types of supporting reads

and implements stringent quality control procedure, ERVcaller is accurate, especially for data with low sequencing depths, and thus, can be used to discover and genotype unfixed ERV and other TE events in the human genomes, and potentially the genomes of other species.

Software availability

ERVcaller is an open-source software. ERVcaller v0.1 source codes, documentation, and example data are available at www.uvm.edu/genomics/software/ERVcaller.

Electronic database information

Accession numbers and URLs for data presented herein are as follows:
Repbase database: http://www.girinst.org/;
UCSC Genome Browser database: http://hgdownload.soe.ucsc.edu/downloads.html#human;
International Genome Sample Resource database: http://www.internationalgenome.org/home.

ACKNOWLEDGEMENT

The authors thank Drs. Zhiping Weng and Akio Miyao for their suggestions and discussions on the analysis. We also thank Jason Kost, M.S., David Dewhurst, M.S., and John Baronas for their careful reviews of the manuscript.

FUNDING

This work was supported by the Start-up Fund of The University of Vermont, the Institutional Research Grant 126773-IRG 14-196-01 awarded to the University of Vermont Cancer Center from the American Cancer Society.

Competing financial interests

The authors declare no competing financial interests.

REFERENCES

1. International Human Genome Sequencing, C. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860.

2. Jern, P. and Coffin, J.M. (2008) Effects of retroviruses on host genome function. *Annu Rev Genet*, **42**, 709-732.
3. Marchi, E., Kanapin, A., Magiorkinis, G. and Belshaw, R. (2014) Unfixed Endogenous Retroviral Insertions in the Human Population. *Journal of Virology*, **88**, 9529-9537.
4. Wildschutte, J.H., Williams, Z.H., Montesion, M., Subramanian, R.P., Kidd, J.M. and Coffin, J.M. (2016) Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proceedings of the National Academy of Sciences of the United States of America*.
5. Garrison, K.E., Jones, R.B., Meiklejohn, D.A., Anwar, N., Ndhlovu, L.C., Chapman, J.M., Erickson, A.L., Agrawal, A., Spotts, G., Hecht, F.M. *et al.* (2007) T cell responses to human endogenous retroviruses in HIV-1 infection. *PLoS Pathog*, **3**, e165.
6. Slokar, G. and Hasler, G. (2015) Human Endogenous Retroviruses as Pathogenic Factors in the Development of Schizophrenia. *Front Psychiatry*, **6**, 183.
7. Goke, J., Lu, X., Chan, Y.S., Ng, H.H., Ly, L.H., Sachs, F. and Szczerbinska, I. (2015) Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell*, **16**, 135-141.
8. Robbez-Masson, L. and Rowe, H.M. (2015) Retrotransposons shape species-specific embryonic stem cell gene expression. *Retrovirology*, **12**, 45.
9. Gröger, V. and Cynis, H. (2018) Human Endogenous Retroviruses and Their Putative Role in the Development of Autoimmune Disorders Such as Multiple Sclerosis. *Frontiers in Microbiology*, **9**, 265.
10. Brodziak, A., Ziolko, E., Muc-Wierzgon, M., Nowakowska-Zajdel, E., Kokot, T. and Klakla, K. (2012) The role of human endogenous retroviruses in the pathogenesis of autoimmune diseases. *Med Sci Monit*, **18**, Ra80-88.
11. Douville, R.N. and Nath, A. (2014) Human endogenous retroviruses and the nervous system. *Handbook of clinical neurology*, **123**, 465-485.
12. Li, W., Lee, M.-H., Henderson, L., Tyagi, R., Bachani, M., Steiner, J., Campanac, E., Hoffman, D.A., von Geldern, G., Johnson, K. *et al.* (2015) Human endogenous retrovirus-K contributes to motor neuron disease. *Science Translational Medicine*, **7**, 307ra153-307ra153.

13. Kassiotis, G. (2014) Endogenous retroviruses and the development of cancer. *Journal of immunology*, **192**, 1343-1349.
14. Gonzalez-Cao, M., Iduma, P., Karachaliou, N., Santarpia, M., Blanco, J. and Rosell, R. (2016) Human endogenous retroviruses and cancer. *Cancer Biology & Medicine*, **13**, 483-488.
15. Kahyo, T., Yamada, H., Tao, H., Kurabe, N. and Sugimura, H. (2017) Insertionally polymorphic sites of human endogenous retrovirus-K (HML-2) with long target site duplications. *BMC Genomics*, **18**, 487.
16. Macfarlane, C. and Simmonds, P. (2004) Allelic variation of HERV-K(HML-2) endogenous retroviral elements in human populations. *J Mol Evol*, **59**, 642-656.
17. Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L.J., Lohr, J.G., Harris, C.C., Ding, L., Wilson, R.K. *et al.* (2012) Landscape of Somatic Retrotransposition in Human Cancers. *Science*, **337**, 967-971.
18. Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., Eichler, E.E. and Sahinalp, S.C. (2010) Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, **26**, i350-357.
19. Hormozdiari, F., Alkan, C., Eichler, E.E. and Sahinalp, S.C. (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res*, **19**, 1270-1278.
20. Quinlan, A.R., Clark, R.A., Sokolova, S., Leibowitz, M.L., Zhang, Y., Hurles, M.E., Mell, J.C. and Hall, I.M. (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Research*, **20**, 623-635.
21. Keane, T.M., Wong, K. and Adams, D.J. (2013) RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics*, **29**, 389-390.
22. Nakagome, M., Solovieva, E., Takahashi, A., Yasue, H., Hirochika, H. and Miyao, A. (2014) Transposon Insertion Finder (TIF): a novel program for detection of de novo transpositions of transposable elements. *BMC bioinformatics*, **15**, 1-9.
23. Thung, D.T., de Ligt, J., Vissers, L.E., Steehouwer, M., Kroon, M., de Vries, P., Slagboom, E.P., Ye, K., Veltman, J.A. and Hehir-Kwa, J.Y. (2014) Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome Biology*, **15**, 1-11.

24. Wu, J., Lee, W.-P., Ward, A., Walker, J.A., Konkel, M.K., Batzer, M.A. and Marth, G.T. (2014) Tangram: a comprehensive toolbox for mobile element insertion detection. *BMC Genomics*, **15**, 1-15.

25. Zhuang, J., Wang, J., Theurkauf, W. and Weng, Z. (2014) TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Research*, **42**, 6826-6838.

26. Jiang, C., Chen, C., Huang, Z., Liu, R. and Verdier, J. (2015) ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. *BMC bioinformatics*, **16**, 1-8.

27. Santander, C.G., Gambron, P., Marchi, E., Karamitros, T., Katzourakis, A. and Magiorkinis, G. (2017) STEAK: A specific tool for transposable elements and retrovirus detection in high-throughput sequencing data. *Virus Evolution*, **3**, vex023-vex023.

28. Bao, W., Kojima, K.K. and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 11.

29. Hu, X., Yuan, J., Shi, Y., Lu, J., Liu, B., Li, Z., Chen, Y., Mu, D., Zhang, H., Li, N. *et al.* (2012) pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics*, **28**, 1533-1535.

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

31. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

32. Handsaker, R.E., Korn, J.M., Nemesh, J. and McCarroll, S.A. (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature genetics*, **43**, 269.

33. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156-2158.

34. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. *et al.* (2007) PLINK: A Tool Set for

Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, **81**, 559-575.

35. Stewart, C., Kural, D., Strömberg, M.P., Walker, J.A., Konkel, M.K., Stütz, A.M., Urban, A.E., Grubert, F., Lam, H.Y.K., Lee, W.-P. *et al.* (2011) A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. *PLOS Genetics*, **7**, e1002236.
36. Burns, K.H. (2017) Transposable elements in cancer. *Nat Rev Cancer*, **17**, 415-424.
37. Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E.C., Pittard, W.S., Mills, R.E., Consortium, G.P. and Devine, S.E. (2017) The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome Research*.

Table

Table 1 Detection and genotyping accuracies of HERV-K events

Sequencing depth	Whole HERV-K insertions			Partial HERV-K insertions		
	Detection power	Precision	Genotyping accuracy	Detection power	Precision	Genotyping accuracy
5X	98.2%	100%	98.0%	97.6%	100%	96.5%
10X	98.6%	100%	98.8%	98.4%	100%	98.0%
30X	99.2%	100%	98.6%	99.0%	99.8%	97.1%
50X	99.2%	99.8%	98.2%	99.0%	100%	97.3%

We simulated two series of datasets with whole and partial HERV-K insertions under sequencing depths of 1X, 5X, 10X, 30X, and 50X. A total of 500 unfixed HERV-K events included 250 homozygous and 250 heterozygous ERV events. The genotyping accuracies were calculated based on the detected unfixed HERV-K events compared to the simulated data.

Figure Legends

Figure 1 ERVcaller workflow. The three components include **a**) extracting unmapped reads; **b**) obtaining supporting reads, and **c**) detecting and genotyping unfix ERV and TE events.

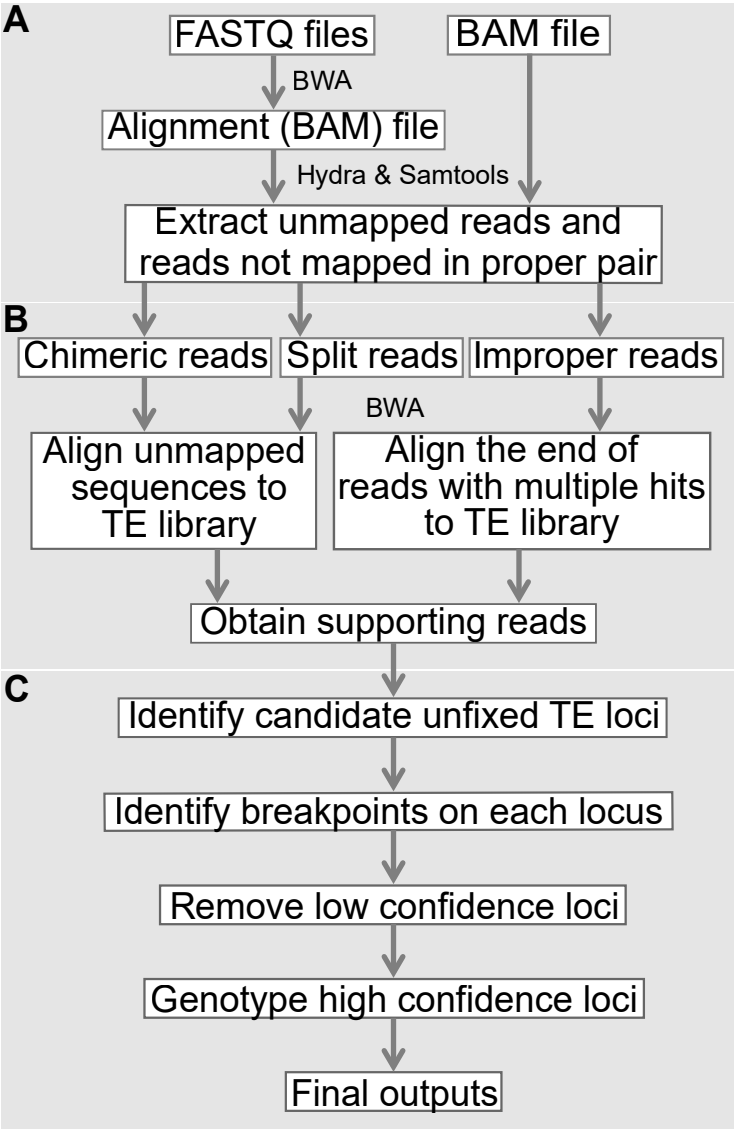

A
FASTQ files
BAM file
BWA
Alignment (BAM) file
Hydra & Samtools
Extract unmapped reads and reads not mapped in proper pair
B
Chimeric reads
Split reads
Improper reads
BWA
Align unmapped sequences to TE library
Align the end of reads with multiple hits to TE library
Obtain supporting reads
C
Identify candidate unfixed TE loci
Identify breakpoints on each locus
Remove low confidence loci
Genotype high confidence loci
Final outputs

Figure 2 Detection power and precision under different sequencing depths. **a**) The detection power and precision of ERVcaller applied to the simulated datasets with whole and partial HERV-K insertions using the HERV-K sequence as the only reference. **b**) The detection power of ERVcaller applied to the datasets with HERV-K insertions using four different types of references, including the HERV-K sequence as the only reference; a larger ERV library containing 743 human and non-human ERV sequences; a human TE library containing only 23 consensus human TEs; and an eukaryotic TE library from the RepBase database. **c**) The precision of ERVcaller applied to the datasets.

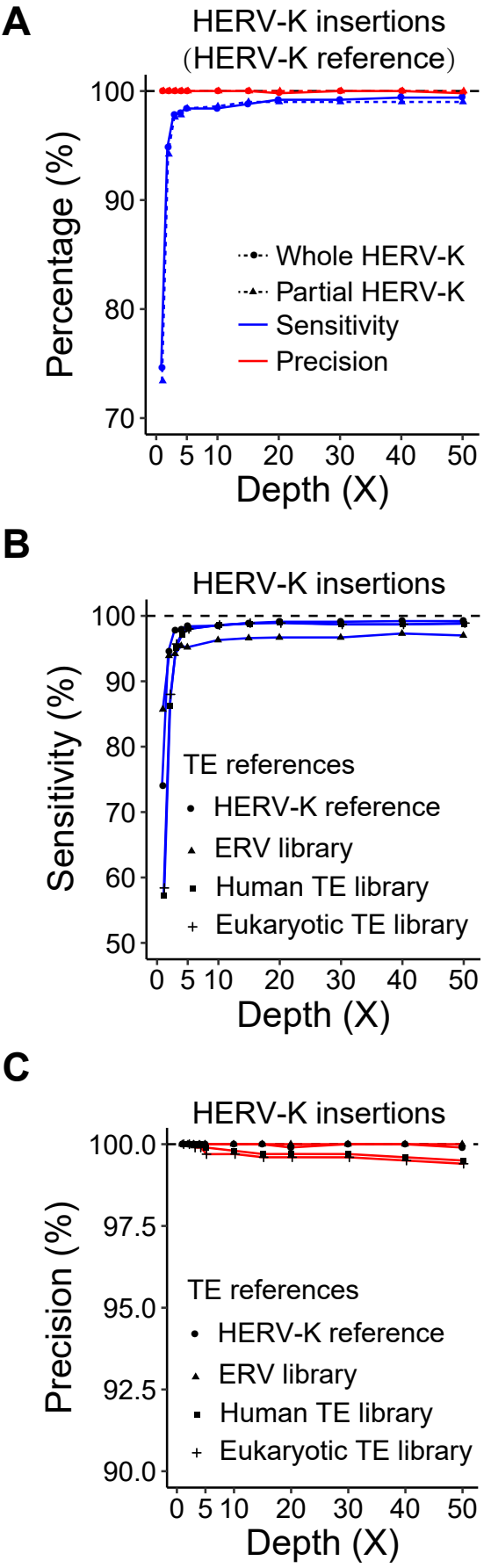

A
HERV-K insertions
(HERV-K reference)
Percentage (%)
100
90
80
70
Whole HERV-K
Partial HERV-K
Sensitivity
Precision
0 5 10 20 30 40 50
Depth (X)
B
HERV-K insertions
Sensitivity (%)
100
90
80
70
60
50
TE references
HERV-K reference
ERV library
Human TE library
Eukaryotic TE library
0 5 10 20 30 40 50
Depth (X)
C
HERV-K insertions
Precision (%)
100.0
97.5
95.0
92.5
90.0
TE references
HERV-K reference
ERV library
Human TE library
Eukaryotic TE library
0 5 10 20 30 40 50
Depth (X)

Figure 3 Comparison of the accuracies of ERVcaller with four other tools. **a**,**e**) The detection power and precision of the tools applied to the datasets with whole HERV-K insertions using the HERV-K sequence as the reference; **b**,**f**) The detection power and precision of the tools applied to the datasets with partial HERV-K insertions using the HERV-K sequence as the reference; **c**,**g**) The detection power and precision of the tools applied to the datasets with whole ERV (from the ERV library) insertions using the ERV library as the references; **d**,**h**) The detection power and precision of the tools applied to the datasets with partial ERV genome (from the ERV library) insertions using the ERV library as the references.

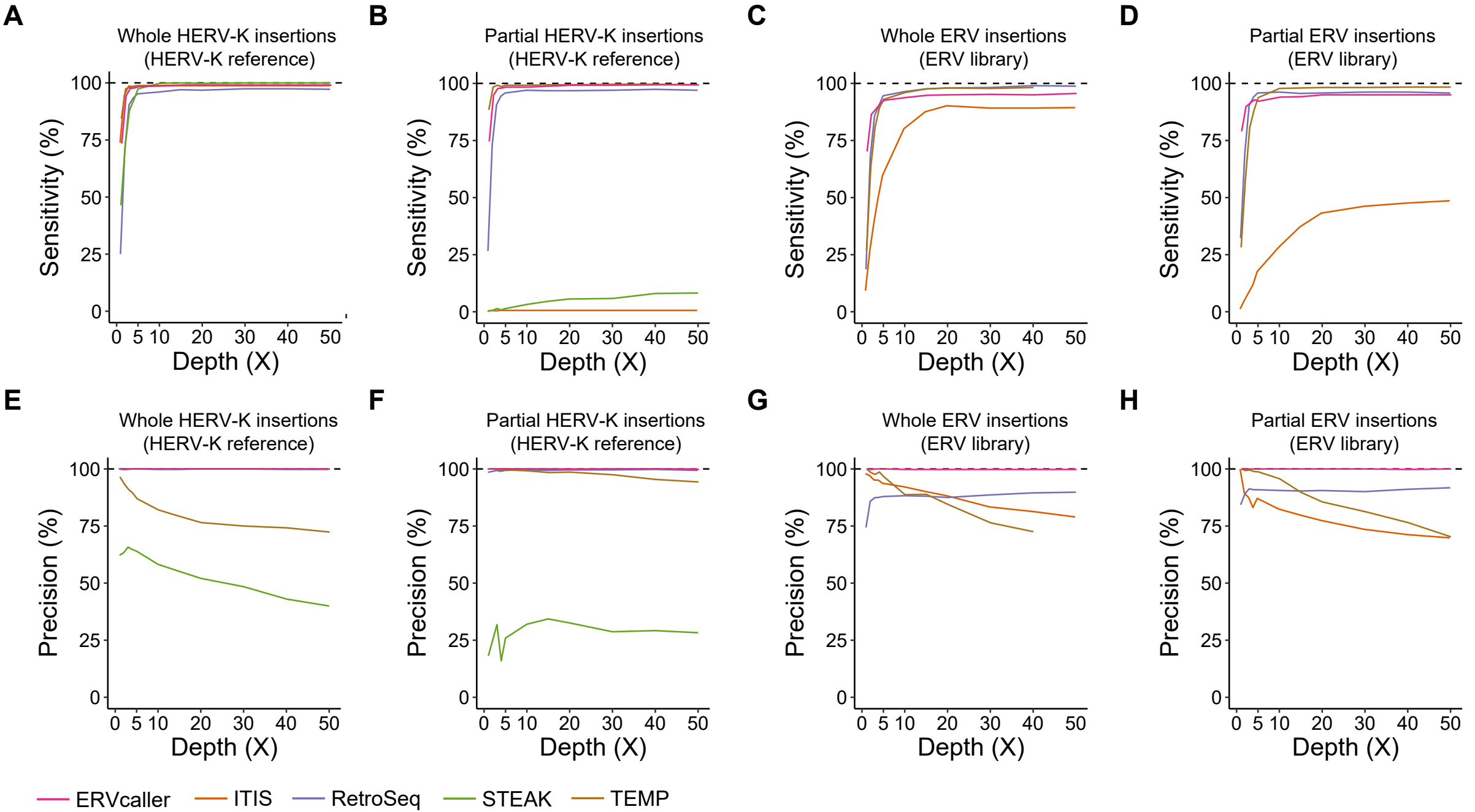
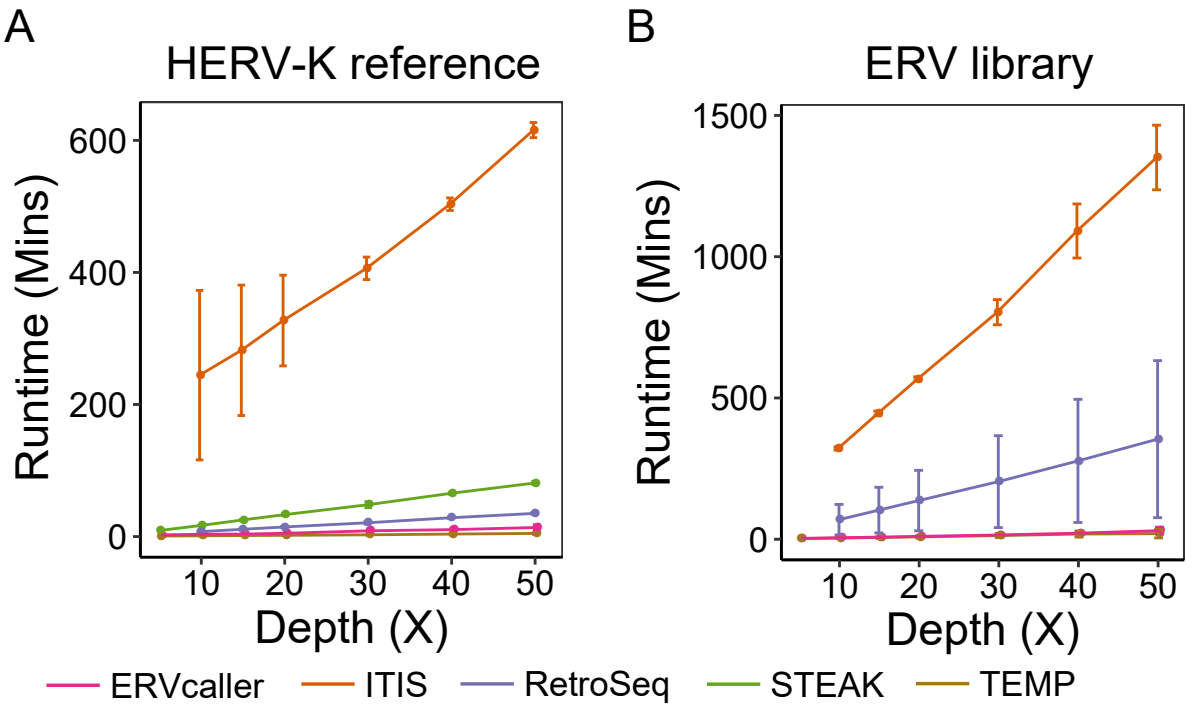

A
Whole HERV-K insertions
(HERV-K reference)
B
Partial HERV-K insertions
(HERV-K reference)
C
Whole ERV insertions
(ERV library)
D
Partial ERV insertions
(ERV library)
E
Whole HERV-K insertions
(HERV-K reference)
F
Partial HERV-K insertions
(HERV-K reference)
G
Whole ERV insertions
(ERV library)
H
Partial ERV insertions
(ERV library)
Sensitivity (%)
Precision (%)
Depth (X)
0
25
50
75
100
0 5 10 20 30 40 50
ERVcaller
ITIS
RetroSeq
STEAK
TEMP

Figure 4 Comparison of runtime of ERVcaller with four other tools using **a**) the HERV-K reference and **b**) the ERV library, respectively.


A
HERV-K reference
Runtime (Mins)
600
400
200
0
10
20
30
40
50
Depth (X)
B
ERV library
Runtime (Mins)
1500
1000
500
0
10
20
30
40
50
Depth (X)
ERVcaller
ITIS
RetroSeq
STEAK
TEMP

Supplementary Tables and Figures

Supplementary Tables

Supplementary Table 1 BWA-mem parameters used to align the supporting reads against the TE references by ERVcaller

Parameter	Value	Description of the parameter (BWA)
-c	100,000	*Skip seeds with more than INT occurrences (default: 500)*
-T	20	*Minimum score to output (default: 30)*
-h	10,000	*If there are < INT hits with score > 80% of the max score, output all in XA (default: 5, 200)*
-a	-	*Output all alignments for single-end or unpaired paired-end reads*
-Y	-	*Use soft clipping for supplementary alignments*
-M	-	*Mark shorter split hits as secondary*

Adopted from BWA. For the other parameters that are not listed here, the default values were used.

Supplementary Table 2 Percentage of accurately detected breakpoints by ERVcaller, TEMP, and RetroSeq at single nucleotide resolution using a two base pair window *(Excel Table)*

Supplementary Table 3 Detection of six experimentally validated unfixed ERV loci in the 1000 Genomes Project samples *(Excel Table)*

Supplementary Table 4 Detection of other experimentally validated unfixed TEs, including Alu, L1, and SVA insertions detected by ERVcaller *(Excel Table)*

Supplementary Table 5 Comparison of the characteristics of ERVcaller with seven other published tools for detecting TEs *(Excel Table)*

Supplementary Table 6 Summary of variables measured by ERVcaller for future genetic association study

	Fixed ERVs	Unfixed ERVs	Variation in population
	Variables measured in each ERV event, in each person		

Family/strain[1]	-	+	HERV-K
Location[2]	-	+	Genome-wide
ERV genotype	-	+	Homozygous ERV insertion, heterozygous ERV insertion, and no ERV insertion
Target site duplication[3]	-	+	3 - 250 bp
SNPs/Indels/CNVs	+	+	Same as the rest of the human genome
	Variable measured in each person		
Number of ERVs[4]	-	+	1 - 13 (mean = 6.4) ERVs per individual based on 27 ERV loci
	Variable measured in each ERV event, in each population		
Population frequency[4]	-	+	0.0002 - 0.75 (mean = 0.18)

[1], the information must be provided in the TE reference library.

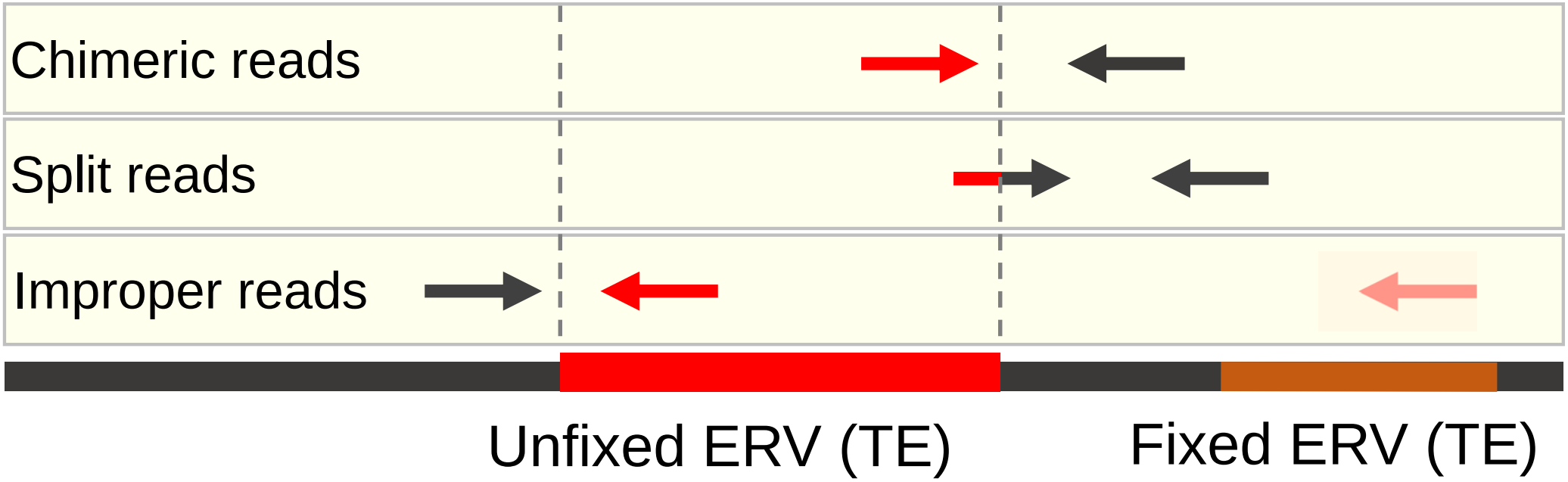
[2], predicted functionality may be obtained through leverage of publicly available database.

[3], target site duplications are often found at the ERV breakpoint.

[4], the study by Wildschutte *et al*. (4) was used to obtain the ERV frequencies in the general populations.

Supplementary Figure Legends

Supplementary Figure 1 Chimeric, split, and improper reads used to identify unfixed ERVs and other TEs by ERVcaller.



Reads uniquely mapped to the human genome

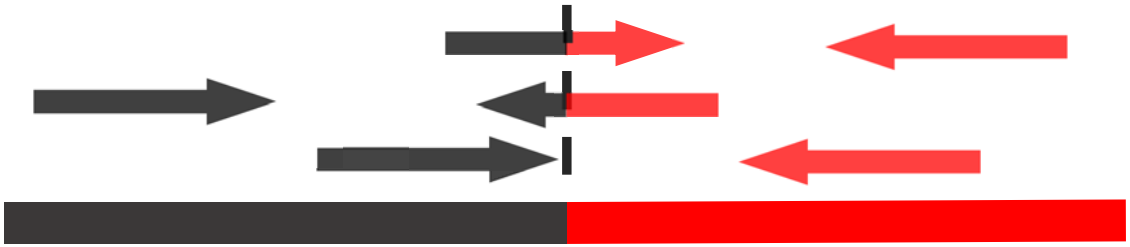
Reads mapped to the ERV (TE) sequence

Reads mapped to the human genome with multiple hits

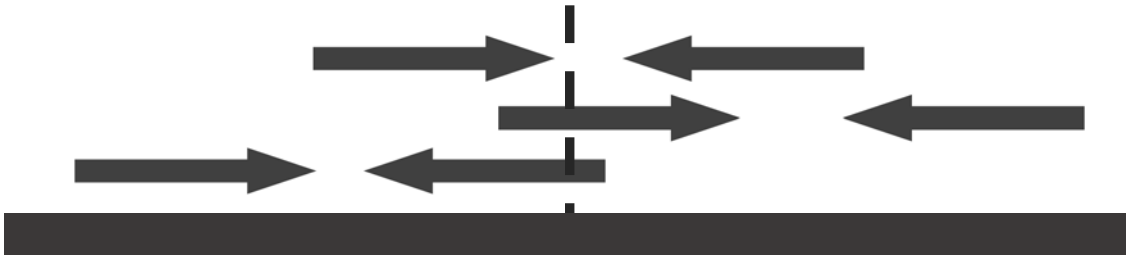
Supplementary Figure 2 Reads supporting the presence of an unfixed TE event and reads supporting no such TE event. These reads are used to genotype each unfixed TE event.

ERV (TE) genotyping

a) ERV (TE) insertion



b) No insertion



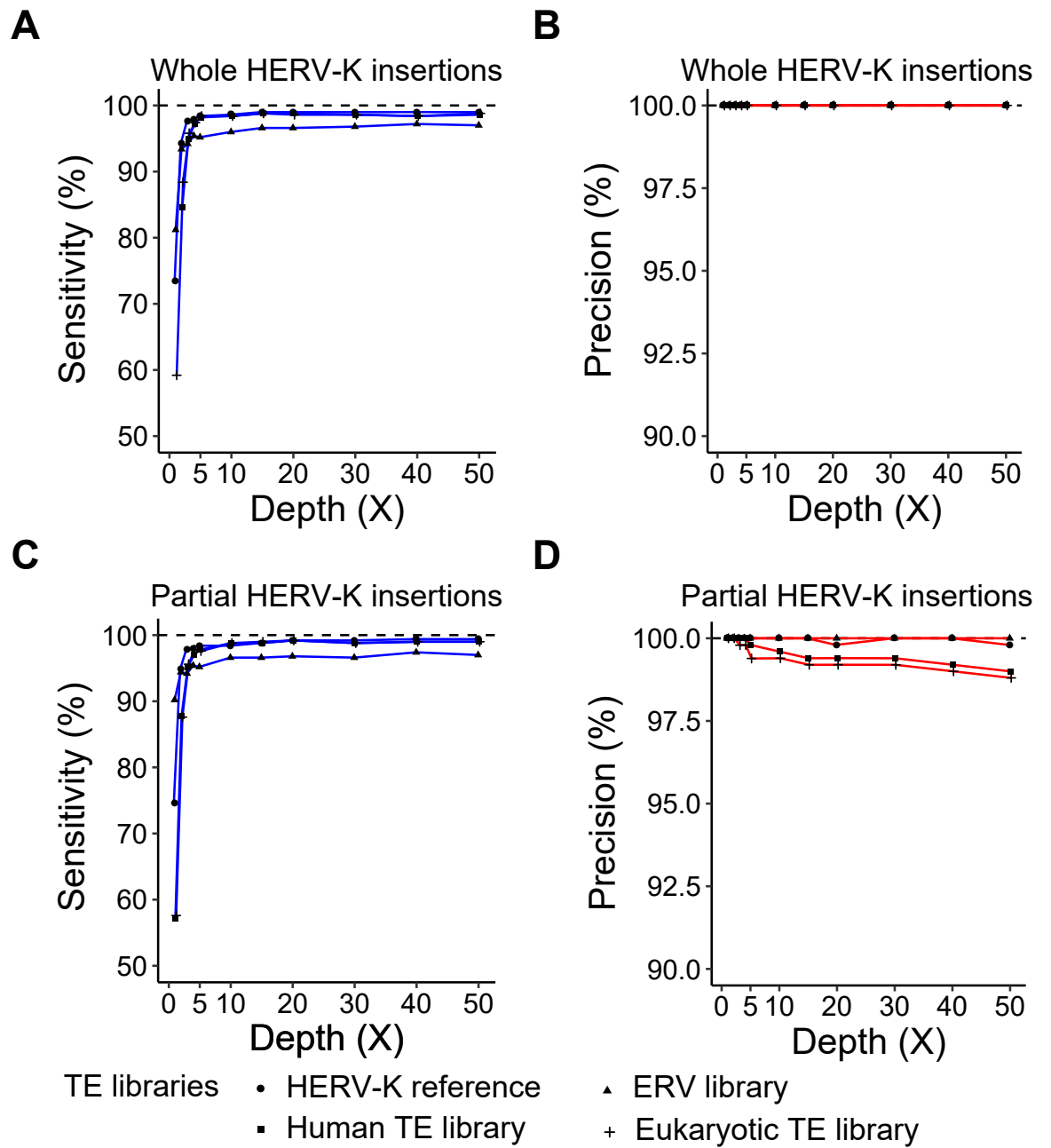
Unfixed ERV (TE) sequence

Human genome

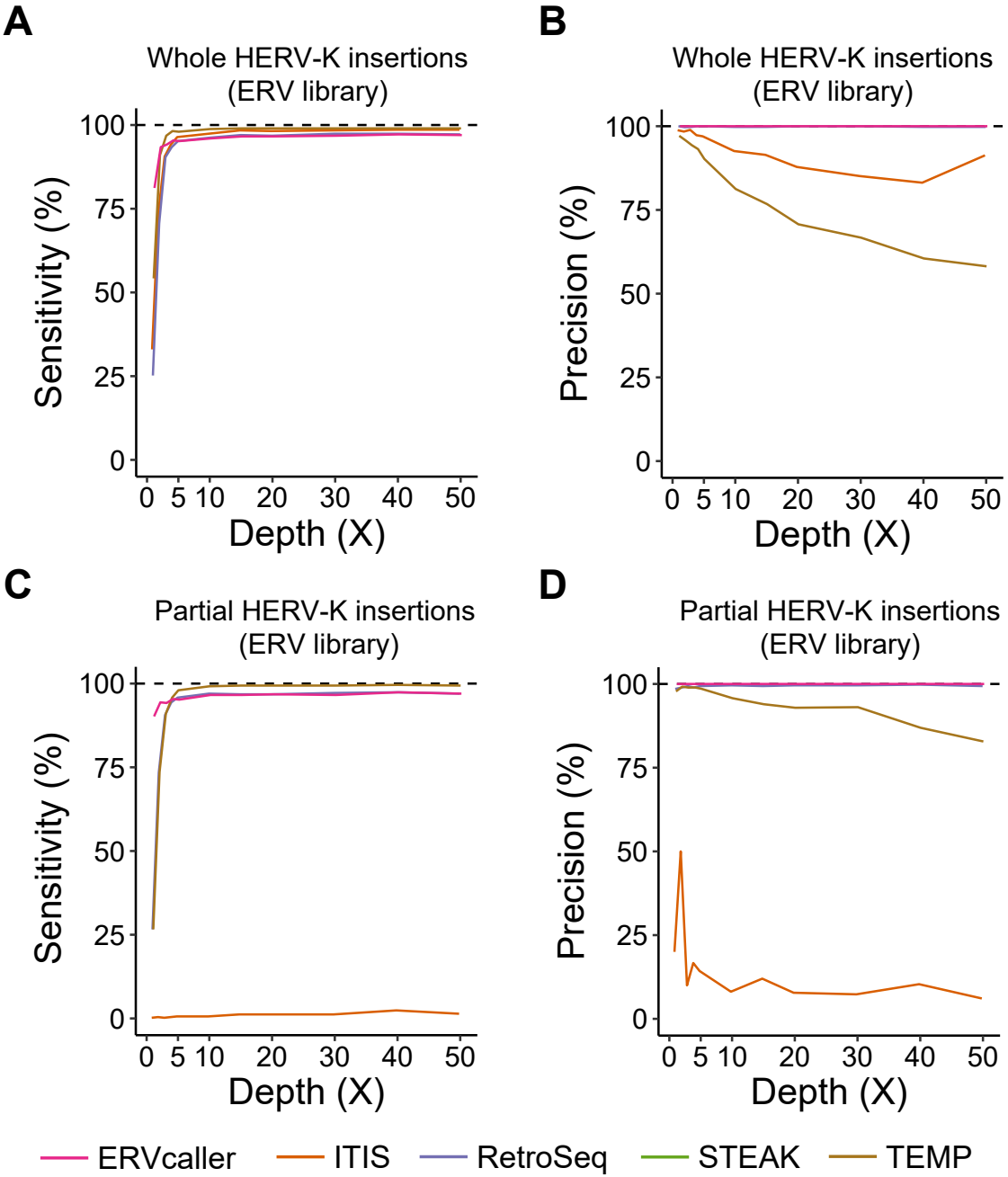
Reads mapped to the ERV (TE) sequence

Reads mapped to the human genome

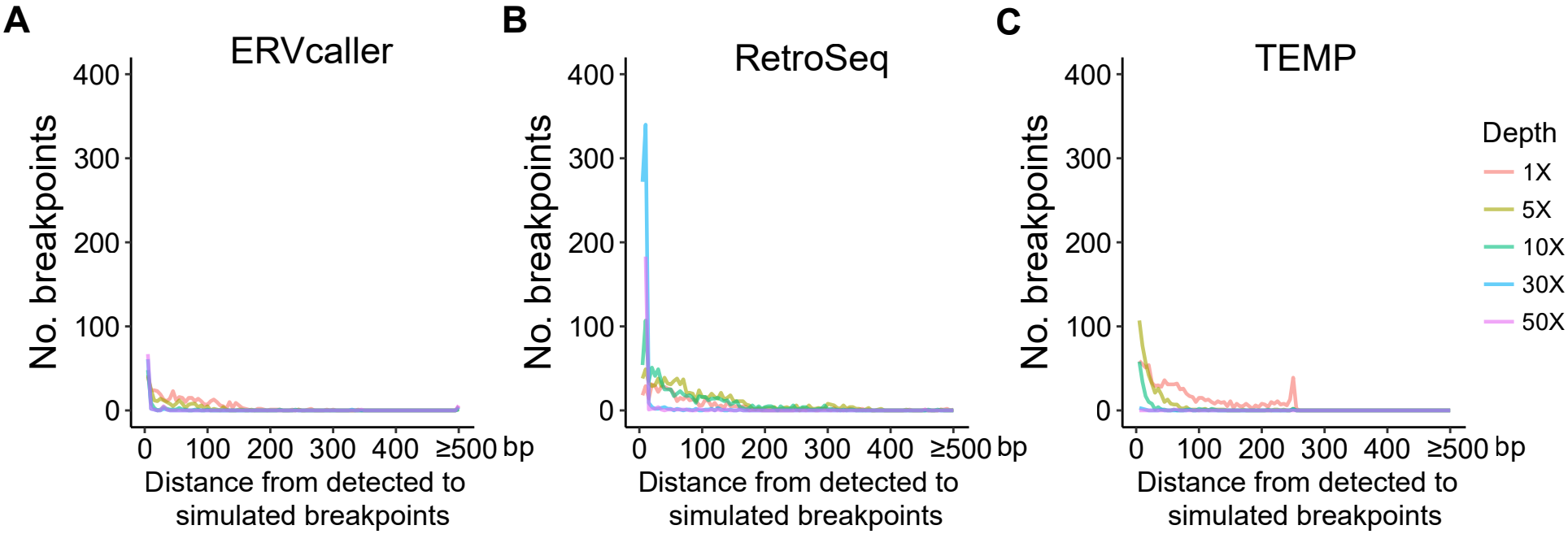
Supplementary Figure 3 Detection power and precision of ERVcaller applied to the simulated datasets. **a,b**) Whole HERV-K insertions and **b,d**) partial HERV-K insertions using different references.


A
Whole HERV-K insertions
Sensitivity (%)
100
90
80
70
60
50
0 5 10 20 30 40 50
Depth (X)
B
Whole HERV-K insertions
Precision (%)
100.0
97.5
95.0
92.5
90.0
0 5 10 20 30 40 50
Depth (X)
C
Partial HERV-K insertions
Sensitivity (%)
100
90
80
70
60
50
0 5 10 20 30 40 50
Depth (X)
D
Partial HERV-K insertions
Precision (%)
100.0
97.5
95.0
92.5
90.0
0 5 10 20 30 40 50
Depth (X)
TE libraries
HERV-K reference
Human TE library
ERV library
Eukaryotic TE library

Supplementary Figure 4 Comparison of the detection power and precision of ERVcaller with RetroSeq, STEAK, and TEMP. **a**) Detection power and **b**) Precision of four tools applied to the whole HERV-K insertions using the ERV library as the references; **c**) Detection power and **d**) Precision of the four tools applied to partial HERV-K insertions using the ERV library as the references.


A
Whole HERV-K insertions
(ERV library)
Sensitivity (%)
100
75
50
25
0
0 5 10 20 30 40 50
Depth (X)
B
Whole HERV-K insertions
(ERV library)
Precision (%)
100
75
50
25
0
0 5 10 20 30 40 50
Depth (X)
C
Partial HERV-K insertions
(ERV library)
Sensitivity (%)
100
75
50
25
0
0 5 10 20 30 40 50
Depth (X)
D
Partial HERV-K insertions
(ERV library)
Precision (%)
100
75
50
25
0
0 5 10 20 30 40 50
Depth (X)
ERVcaller
ITIS
RetroSeq
STEAK
TEMP

Supplementary Figure 5 Distribution of the distances from the detected breakpoints to the simulated breakpoints for **a**) ERVcaller; **b**) RetroSeq; and **c**) TEMP.


A
ERVcaller
B
RetroSeq
C
TEMP
No. breakpoints
400
300
200
100
0
0
100
200
300
400
≥500
bp
Distance from detected to simulated breakpoints
Depth
1X
5X
10X
30X
50X