ERVcaller: Identifying and genotyping unfixed endogenous retrovirus (ERV) and other transposable element (TE) insertions 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: 220

Number of words in the main text (excluding online materials, acknowledgments, financial disclosures, legends, and references): 5,448

Number of Table: 0

Number of Figures: 5

Number of supplementary materials: 7 Supplementary Tables and 7 supplementary Figures and Legends

Abstract

Motivation: More than 8% of the human genome is derived from endogenous retroviruses (ERVs). In recent years, an increasing number of human diseases have been found to be associated with ERVs. However, it is still challenging to accurately detect the full spectrum of polymorphic (unfixed) ERVs using next-generation sequencing (NGS) data.

Results: We designed a new tool, ERVcaller, to detect and genotype unfixed transposable element (TE) insertions, including ERVs, in the human genome. We evaluated the tool using both simulated and real benchmark whole-genome sequencing datasets. ERVcaller achieved > 97% sensitivity and > 99% precision for detecting, and > 96% accuracy for genotyping the simulated HERV-K insertions (sequencing depth > 5X). We compared ERVcaller with four existing tools, and ERVcaller consistently showed the highest sensitivity and precision for detecting unfixed ERV insertions, especially under low sequencing depths. ERVcaller also achieved the most precise determination of ERV breakpoints at single-nucleotide resolution. By applying ERVcaller to a subset of the 1000 Genomes Project samples, we detected 100% of the known unfixed ERV insertions and 95% of other unfixed TE insertions. We also detected almost all the known genotypes (100% for ERVs and 98% for other TEs). In conclusion, ERVcaller is capable of identifying and genotyping TE insertions using NGS data with high sensitivity and precision. This tool can be applied broadly to other species.

Availability: www.uvm.edu/genomics/software/ERVcaller.html

Contact: dawei.li@uvm.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Keywords: Transposable element (TE); Endogenous retrovirus (ERV); Polymorphic ERV; Unfixed TE insertion; Next-generation sequencing (NGS)

Introduction

Origin of endogenous retroviruses

Transposable elements (TEs), which were first found in maize, are groups of mobile DNA sequences that collectively comprise a large percentage of most eukaryotic genomes (e.g., ~45% of the human genome)[1]. TEs can be categorized into two classes based on transposition mechanisms, Class I: "copy and paste" TEs, and Class II: "cut and paste" TEs[2]. 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[3]. In the human genome, 8% of the sequences are of retroviral origin, containing > 100,000 ERV loci from approximately 50 groups[1, 4]. Almost all human ERVs (HERVs) are replication defective, potentially due to accumulated mutations, insertions, deletions, or becoming solo-long terminal repeats (solo-LTRs)[4, 5]. To date, no replication-competent HERVs were reported which may due to the relatively small number of individuals examined or because they are rare events within the human population[6-8]. However, the HERV-K (HML-2) group, which was relatively recently inserted into the human genome (i.e., ~35 million years ago), has been found to be polymorphic[6, 8-10].

Endogenous retrovirus and human disease

Polymorphic ERVs have been associated with many human diseases because ERVs may have concurrent effects on both structure and biological function of the human genome. ERVs may lead to genomic rearrangements through non-allele homologous recombination with other ERV copies; they may also behave as a source of promoters, enhancers or transcriptional factor binding sites for regulating human gene expression levels[4]. They can either systematically transcribe stage-specific RNAs in early embryo development[11, 12] or be re-activated and expressed under certain disease conditions[13] or with infections of viruses, such as human immunodeficiency virus and Epstein-Barr virus[14, 15]. The expression of viral genes (e.g., *gap* and *pol*) may interact with the human transcriptome or modulate the human immune system[4]. Thus, ERVs have been associated with autoimmune diseases[16-18], mental disorders[19-21], cancers, and many other diseases. For example, significantly upregulated expression of ERV RNAs and higher levels of proteins have been observed in tumors versus adjacent normal tissues[22, 23].

The fixed ERVs, which are usually ancient retroviral integration events, are shared by all the human genomes and are more likely to be degraded. The unfixed ERVs, which are typically more recently integrated, are not shared by all the human genomes and thus are polymorphic in the human population[8], and are more likely to be intact and functional. The speculative pathogenic mechanisms of polymorphic ERVs include disrupting functions of the human genes at or near their integration sites, expressing accessory proviral proteins, or altering the adaptive and innate immune response[24]. Although the pathogenicity of the unfixed ERVs is still poorly understood due to the complexity of ERV sequences, the unfixed ERVs are more likely to be associated with human diseases[5, 8]. This study focuses on the unfixed ERVs.

Existing tools for detecting unfixed ERVs

To determine the pathogenic effects of ERVs, it is necessary to first identify all the unfixed ERV insertions in the human genome. Similar to single nucleotide polymorphisms (SNPs), the same unfixed ERV insertions can be found in one or more individuals, leading to unfixed ERV loci. Indeed, many unfixed ERV loci have been recently discovered[5, 9, 10, 25]. For example, Lee *et al*. discovered a total of 15 unfixed ERV loci by screening 44 samples[26]. By screening more samples, Emanuele *et al*. found 17 unfixed ERV loci, including two novel loci[5], with an average of six unfixed ERV insertions per human genome. By analyzing a different set of samples, Wildschutte *et al*. found 19 new unfixed ERV loci[10]. These studies indicate that more unfixed ERV loci likely exist which have not yet been discovered in the human population.

With wide-spread applications of next-generation sequencing (NGS) technologies, bioinformatics tools have been developed to discover unfixed TEs. For example, structural variation detection tools, including VariationHunter[27, 28] and Hydra[29], were first adapted for detecting unfixed TE insertions using whole-genome sequencing (WGS) data. However, because of the long fragment insertions and highly abundant repeated sequences, the detection of unfixed TEs, including ERVs, are more difficult than the detection of SNPs, small insertions and deletions (InDels), or other structural variations. Specific tools were then developed, including TEA, RetroSeq, TIF, Mobster, Tangram, TEMP, ITIS, and STEAK[26, 30-36] (**Supplementary Table 1**). Although each software had its own merits, many of these tools were insufficient for

accurately detecting the full spectrum of the unfixed TE insertions or genotyping them, or had other limitations.

A new software

In this study, we developed a novel software, ERVcaller, to detect and genotype unfixed TE insertions, particularly ERVs, with paired-end or single-end NGS data. It considers three types of supporting reads with stringent quality control procedure, leading to more efficient and accurate detection of unfixed TE insertions. Compared to the other analyzed tools, ERVcaller had the highest detection sensitivity and precision, especially with low sequencing depths. It also detected the most precise breakpoints. The performance of ERVcaller was also demonstrated in the benchmark 1000 Genomes Project samples containing known unfixed ERV or other TE insertions. It is the only tool that has achieved both high sensitivity and precision consistently with various TE references of different levels of sequence complexity.

Methods

Reference sequences for detecting unfixed ERV and other TE insertions

To evaluate how the use of different TE references influence the detection of unfixed ERV and other TE insertions, we compared the performance of ERVcaller using four different collections of TE sequences as the reference. These included: (1) The HERV-K reference (KU054272.1); (2) An ERV library containing a total of 743 diverse human and 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 obtained from Tangram[33], including 17 long interspersed nuclear element 1 (L1) sequences, four Alu sequences, one HERV-K sequence, and one short interspersed nuclear element/variable number tandem repeat/Alu (SVA) sequence; and (4) A eukaryotic TE library containing all the repetitive DNA elements from the RepBase database[37]. The human reference genome GRCh37 (hg19) build was downloaded from the UCSC Genome Browser for the analysis of simulated datasets, and the GRCh38 (hg38) build was downloaded and used for the analysis of the 1000 Genomes Project samples.

Simulation

To evaluate the sensitivity and precision for detecting unfixed TE insertions, we simulated ERV insertions by randomly inserting 500 whole and partial (randomly fragmented) ERV genomes into human chromosome 1 (hg19) using an in-house Perl script. Paired-end reads of 100 base pairs (bp) length with 500 bp insert size were simulated using pirs[38] with default parameters. Two series of data were simulated, including one with 500 whole and partial ERV insertions derived from the HERV-K reference, and the other with 500 whole and partial ERV insertions derived from the ERV library. Each of these series was simulated with sequencing depths of 1X, 2X, 3X, 4X, 5X, 10X, 15X, 20X, 30X, 40X, and 50X.

To evaluate genotyping accuracies, we simulated another series of datasets carrying both homozygous and heterozygous HERV-K insertions with sequencing depths of 5X, 10X, 30X, and 50X. For each depth, we simulated one set of paired-end reads with 500 HERV-K insertions, and another set of paired-end reads but with only half of these insertions. By combining the two sets of reads into one dataset, we obtained 250 homozygous and 250 heterozygous HERV-K insertions.

ERVcaller pipeline

ERVcaller first aligns raw FASTQ reads to the human reference genome using BWA-MEM (arXiv:1303.3997) with the default parameters for WGS data, using Tophat2[39] for RNA-sequencing (RNA-Seq) data, or directly uses a pre-aligned BAM file as input. Chimeric reads (also known as discordant reads) are defined as a read pair having one end aligned to the human reference genome while the other end does not align to the human reference genome but aligns to a TE reference sequence(s)); Split reads (also known as soft-clipped reads) are reads with one part aligned to the human reference genome and the other part aligned to a TE reference sequence(s); Improper reads are defined as a read pair without both ends aligned to the same chromosome or within the insert size, and one end is aligned to a TE sequence (**Supplementary Figure 1**). Chimeric reads are obtained using Samtools[40] by extracting reads flagged as unmapped (SAM flag=4) and then removing reads flagged as mate unmapped and non-primary alignment (SAM flag=264); meanwhile, reads flagged as mate unmapped (SAM flag=8) are extracted and the reads flagged as either unmapped, or non-primary alignment if it is mapped (SAM flag=260) are removed. Improper reads are obtained using Samtools[40] by removing all

reads flagged as either mapped in a proper pair, or read and mate unmapped (SAM flag=14). Split reads are obtained by first extracting reads flagged as mapped in a proper pair (SAM flag=2), then SE-MEI (https://github.com/dpryan79/SE-MEI) is used to extract the sequences of split reads (≥ 20 bp) in FASTQ format. Meanwhile, once a putative TE insertion is identified, regardless of whether the breakpoint is mapped at nucleotide resolution, the split reads (< 20 bp) are kept to potentially improve breakpoint detection. All chimeric and split reads are aligned concurrently against all TE sequences included in the TE reference (library) using BWA-MEM with the customized parameters (**Supplementary Table 2**). If one end of an improper read can be aligned to multiple locations on the human reference genome, it is likely that this end is derived from a TE, thus we also aligned this end to the TE references using BWA-MEM with the same customized parameters.

After aligning supporting reads against the human reference genome, reads are grouped by two insert-sized windows to identify candidate genomic regions for the unfixed TE insertions. ERVcaller further determines the TE insertions using the number of reads and the average alignment score of all supporting reads. The supporting reads are further aligned back to each candidate genomic region, and the TE insertions with no confident supporting reads are removed (i.e., supporting reads that can be fully aligned to each candidate genomic region are removed). To report an unfixed TE insertion, a minimum of two supporting reads, including at least one chimeric or improper read, are required. We then use chimeric, improper, and split reads to determine the chromosomal location of each breakpoint. If split reads spanning the breakpoints are detected, the breakpoint locations can be precisely determined. Without split reads, the breakpoint locations can be estimated according to two different cases: 1) if either upstream (5') or downstream (3') breakpoint is identified, the first nucleotide of the nearest read to the breakpoint is considered to be the estimated breakpoint; 2) if both the upstream and downstream breakpoints are identified, the median of the first nucleotide of the nearest read to the breakpoint is considered to be the estimated breakpoint.

The total number of chimeric, improper, and split reads supporting the hypothesis of the unfixed TE insertion versus the total number of reads supporting the hypothesis of no insertion is used to determine the genotype of an unfixed TE insertion. The reads supporting the hypothesis

of no insertion are the reads that can be fully aligned to the human reference genome across the breakpoint. We extract these reads from the aligned BAM file using Samtools[40] and Perl scripts. For example, if the number of reads supporting the presence of an unfixed TE insertion and the number of reads support no insertion at a locus are the same, the genotype of this TE insertion is heterozygous (**Supplementary Figure 2**). However, genotyping may be complicated due to the noise derived from the sequencing process, alignment, and subsequent analyses. Therefore, we calculate the genotype likelihood to determine the homozygous versus heterozygous state of each identified unfixed TE insertion using a combined insert-size and read-depth approach[14]. Specifically, we first calculate the count of reads supporting no insertion at the TE breakpoint. Second, we calculate the read counts of randomly-selected genomic locations interspersed every 50 bp within a window (e.g., 10 kilo bp) centered at the TE breakpoint, and then obtained the distribution of read counts (genomic locations within one insert size of the breakpoint are excluded). Third, we compare the read counts supporting no 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 determine the genotype based on the quantile (e.g., threshold = 0.1), i.e., quantile > 0.1 and quantile < 0.1 represent homozygous and heterozygous TE insertions, respectively. During the process, for better calculation of the read counts supporting no insertion, the reads either outside of the insert size Poisson distribution or mapped within 10 bp of the feature location (applied to both the TE breakpoint and the randomly-selected locations) are removed from the calculations. Finally, ERVcaller outputs the results in a table, which can be easily converted to Variant Call Format[41] or Plink[42] format for subsequent genetic association analyses.

Software comparison

To compare ERVcaller with other existing tools for detecting unfixed TE insertions, we selected four of the most recently published tools, including ITIS[35], TEMP[34], RetroSeq[30], and STEAK[36]. After successfully installing and compiling these tools, we ran each tool with the default parameters using the simulated datasets. BWA-MEM was used to align raw reads to the human reference genome. Aside from ITIS, which required FASTQ files as input, ERVcaller and the other three tools used BAM files as input. Only the unfixed TE insertions with two or more supporting reads were kept for the comparative analysis. For comparing the runtimes of these

tools, we used the Vermont Advanced Computing Core cluster computing nodes having 12 cores and 48 GB memory.

1000 Genomes Project samples and data preprocessing

To further evaluate the performance of ERVcaller, we applied it to two subsets of the 1000 Genomes Project samples[43]. The first set included 15 samples (~4X sequencing depth), and in four of them unfixed ERV insertions have been previously identified and validated by Sanger sequencing[10]. The second set included three samples (~15X sequencing depth), and in all of them unfixed TE (i.e., Alu, L1, and SVA) insertions have been previously genotyped and validated by polymerase chain reaction (PCR)[44]. Those experimentally validated insertions were used to measure the detection and genotyping accuracy of ERVcaller.

The aligned CRAM files (hg38) for the first set of samples, and the raw FASTQ files for the second set of samples 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, then the output SAM files were converted (to BAM files), sorted, and indexed using Samtools[40]. Samtools was also used to convert the aligned CRAM files to indexed BAM files. The hgLiftOver tools (https://genome.ucsc.edu/cgi-bin/hgLiftOver) was used to convert the coordinates from reference build hg18, which was used in the original paper[44], to hg38.

PCR and Sanger sequencing validation

Two of the analyzed 1000 Genomes project samples[43] that contained newly detected unfixed ERV insertions were selected for PCR and Sanger sequencing validation. The primers for PCR were described previously[10]. The PCR products were purified and cloned using the StrataClone Blunt PCR Cloning Vector pSC-B-amp/kan plasmid (Agilent Technologies, La Jolla, CA). Both alleles (i.e., with and without insertion) were confirmed by Sanger sequencing using the universal T3 primers. The ERV genotypes were verified by PCR in seven of the analyzed samples using the same primers.

Results

Software development

For accurate detection and genotyping of unfixed ERV and other TE insertions, we improved upon existing calling algorithms and further 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 TE insertions (**Figure 1**). The inputs of ERVcaller include either raw FASTQ or BAM file(s), the human reference genome, and TE reference sequence(s). With ERVcaller, we first extracted all reads that could not be fully mapped to the human reference genome. Among these unmapped reads, we obtained all chimeric and split reads, and then aligned them to the TE reference library. We also obtained the improper reads by extracting all reads in which the two ends were not aligned in proper pairs (**Supplementary Figure 1**), and aligned those with ends having multiple hits on the human genome to the TE reference library. We then used the three types of supporting reads to determine the chromosomal location of each TE insertion, including both upstream and downstream breakpoints. We identify confident TE insertions by only including those meeting all the following criteria: 1) it had at least two supporting reads; 2) it had at least one chimeric or improper read; 3) the average alignment score of each of the supporting reads was greater than 30; and 4) each of the reads mapped to the human genome was at least 50 bp in length. If no reads could be uniquely mapped to the human genome, the detected TE insertions were likely to be false-positives; however, as some unfixed ERV insertions had been found within repeat regions (i.e., segmental duplications and Alu elements)[10], we annotated and kept those results as a separate group for further confirmation. After stringent filtering, high confidence unfixed TE insertions were then genotyped based on the reads crossing the breakpoints, as described in the Methods section (**Supplementary Figure 2**).

Detection and genotyping of simulated ERV insertions

To measure the sensitivity and precision of ERVcaller for detecting unfixed ERV insertions, we applied it to a series of simulated datasets. The datasets consisted of those containing whole HERV-K insertions and those containing partial (> 500 bp) HERV-K insertions. ERVcaller was applied to each of them using the HERV-K sequence as the reference. The detection of the whole and partial ERV insertions were evaluated separately. The sensitivity and precision were calculated at each sequencing depth, from 1X to 50X. Sensitivity (detection power) was defined to be the ratio of correctly identified versus total simulated TE insertions. Precision was defined

to be the ratio of correctly identified versus total identified TE insertions. The results showed that ERVcaller was able to consistently detect > 97% of the whole and partial HERV-K insertions when the depths were 5X or higher with almost 100% precision (**Figure 2** and **Supplementary Table 3**).

To further evaluate whether repetitive and redundant sequences in a TE reference library influenced the sensitivity and precision, we applied ERVcaller to the same datasets using four TE reference libraries as the references, including the HERV-K reference, an ERV library, a human TE library[33], and a eukaryotic TE library[37]. All of them showed consistently > 95% sensitivity and > 99% precision when depth > 5X (**Figure 3A**, **3B** and **Supplementary Figure 3**). Hence, we conclude that ERVcaller achieves high sensitivity and precision, with little influence from the selection of different TE references, suggesting that larger TE reference libraries, presumably including all potential TEs, may be used as the references for ERVcaller to detect highly divergent or novel TE insertions.

To measure the genotyping accuracy of ERVcaller, we applied it to a series of simulated datasets containing homozygous and heterozygous HERV-K insertions. Our results showed that the genotyping accuracies were consistently > 96% for all simulated sequencing depths (> 98% for genotyping the whole HERV-K insertions) (**Supplementary Table 3**). No significant differences were observed for genotyping accuracy between homozygous and heterozygous ERV insertions.

Software comparison for detection of unfixed ERV insertions

We performed a comprehensive comparison of ERVcaller with four existing tools for detecting unfixed ERV insertions, including RetroSeq, STEAK, ITIS, and TEMP[30, 34-36]. We first used the HERV-K sequence as the reference to detect the simulated HERV-K insertions, then used the ERV library containing diverse human and non-human ERV sequences as the references to detect the HERV-K insertions, and last used the same ERV library to detect insertions derived from the ERV library. Additionally, we compared the accuracies for detecting ERV breakpoints and runtimes among these tools.

We first measured the sensitivity and precision of each tool for detecting the HERV-K insertions using the HERV-K sequence as the reference. For detecting the whole HERV-K insertions, all tools had high sensitivity (> 95%) as long as the sequencing depths were 5X or higher (**Figure 4A**). Other than TEMP and STEAK, which had 88% and 41% precision, respectively, ERVcaller, RetroSeq, and ITIS achieved almost 100% precision (**Figure 4B**). For detecting the partial HERV-K insertions, other than ITIS and STEAK (< 10% sensitivity), ERVcaller, RetroSeq, and TEMP consistently had high sensitivity (96%) (**Figure 4C**). Except for STEAK (< 34%), all other tools showed high precision (> 94%) (**Figure 4D**). *Thus, among the five tools analyzed, only ERVcaller and RetroSeq were able to detect whole and partial HERV-K insertions with both high sensitivity and precision. Among these two, ERVcaller consistently achieved slightly higher sensitivity than RetroSeq.*

We then measured the sensitivity and precision for detecting HERV-K insertions using the ERV library as the reference. Only four tools were analyzed as STEAK was not applicable when using multiple references. For detecting the whole HERV-K insertions, all four tools had high sensitivity (> 95% when depth > 5X) (**Figure 4E**). ERVcaller and RetroSeq achieved 100% precision, while ITIS and TEMP had significantly decreased precision (90% and 72%, respectively) (**Figure 4F**). For detecting the partial HERV-K insertions, except ITIS (< 3%), the other three tools showed > 95% sensitivity when depth > 5X (**Figure 4G**). ERVcaller and RetroSeq had almost 100% precision, while TEMP (92% on average) and ITIS (< 15%) had significantly lower precision (**Figure 4H**). *Thus, among the four tools analyzed, only ERVcaller and RetroSeq were not affected by increasing the number of sequences in the TE reference library.*

We further measured the sensitivity and precision for detecting ERV insertions (derived from the ERV library) using the ERV library as the reference. Similarly, only four tools were compared as STEAK was not applicable with multiple references. For detecting the whole ERV insertions, except ITIS (84% on average), the other three tools had > 93% sensitivity when depth > 5X (**Figure 4I**). Only ERVcaller achieved almost 100% precision (**Figure 4J**). For detecting the partial ERV insertions, except ITIS (< 50%), the other three had > 92% sensitivity when depth > 5X (**Figure 4K**). Consistently, only ERVcaller achieved nearly 100% precision

(**Figure 4L**). *Thus, among these tools, only ERVcaller was capable of detecting a wide spectrum of unfixed ERV insertions with both high sensitivity and precision.*

We further compared the accuracies for detecting ERV breakpoints. Only three tools were compared as ITIS and STEAK were not designed for detecting breakpoints at single-nucleotide resolution. If the detected breakpoint coordinate was within a two-base pair window of the simulated coordinate, it was considered correct. When the HERV-K sequence was used as the reference to detect the HERV-K breakpoints, TEMP and ERVcaller showed comparable accuracies (e.g., > 90% when depth > 15X) (**Supplementary Figure 4A** and **B**). When the ERV library was used as the references (to detect the HERV-K breakpoints), only ERVcaller achieved high accuracy for detecting breakpoints of the whole and partial HERV-K insertions (e.g., > 90% when depth > 15X) (**Supplementary Figure 4C** and **D**). When the ERV library was used as the reference to detect the wide spectrum of ERV insertions, ERVcaller achieved significantly higher accuracy than the other tools (e.g., > 80% when depth > 15X) (**Supplementary Figure 4E** and **F**). Additionally, compared to ITIS and STEAK, ERVcaller was capable of consistently detecting the breakpoints when the sequencing depth <10X (**Supplementary Figure 4**). Moreover, ERVcaller consistently detected the breakpoints, regardless of the inserted ERV sequence, the length of it or the TE reference (e.g., > 80% when depth > 15X) (**Supplementary Figure 5**). We also analyzed the distribution of the distance from the detected breakpoint (i.e., > 2 bp in distance) to the simulated breakpoint. Shorter distances represent higher accuracy. Consistently, ERVcaller revealed a distribution of the shortest distances compared to the other two tools (**Supplementary Figure 6**). For example, TEMP showed a hotspot around the 250 bp distance, suggesting deviation of the detected breakpoints. *Thus, ERVcaller was able to correctly detect ERV breakpoints, regardless of the lengths of ERV insertions or the number of TE sequences in the reference library.*

We also compared the runtime among the five tools. The whole and partial HERV-K and ERV insertions were combined in this analysis. When the HERV-K reference was used for detecting the HERV-K insertions, TEMP had the shortest runtime (2.7 mins) at 30X depth, followed by ERVcaller (8.0 mins), RetroSeq (20.0 mins), STEAK (46.3 mins), and ITIS (410.0 mins). When the ERV library was used for detecting the wide spectrum ERV insertions (STEAK

was not analyzed as it was not applicable), ERVcaller was the fastest (18.0 mins) at 30X depth, followed by TEMP (21.1 mins), RetroSeq (394.7 mins), and ITIS (813.8 mins) (**Figure 5**). *Thus, the high speed of ERVcaller allows for identifying the full spectrum of ERV and other TE insertions with a large sample size.*

Detection and genotyping of unfixed ERV and other TE insertions in the 1000 Genomes Project samples

To measure the sensitivity of ERVcaller for detecting and genotyping unfixed ERV and other TE insertions in real WGS data, we first applied it to detect the six unfixed ERV insertions in four benchmark 1000 Genomes Project samples that had been previously confirmed by Sanger sequencing[10]. Our results showed that ERVcaller detected all of the six insertions using the default parameters (**Supplementary Table 4**). We then analyzed an additional 11 samples and verified two new insertions using PCR and Sanger sequencing. Both of them, including the homogeneous and homozygous insertion alleles, were confirmed (**Supplementary Figure 57**). We then randomly selected seven out of the 11 samples and genotyped each sample for the two ERV loci using PCR. All 14 genotypes were confirmed (**Supplementary Table 5**). Among the 15 analyzed samples, ERVcaller identified a total of 233 unfixed ERV insertions with an average of 15.5 insertions per individual (ranging from 9 to 31 insertions).

To measure the sensitivity of ERVcaller for detecting other unfixed TE insertions, we then applied it to detect the 1,136 unfixed TE insertions in a trio of the 1000 Genomes Project samples that have been previously confirmed by PCR[44]. ERVcaller consistently detected nearly all of them, i.e., 96%, 88%, and 100% for Alu, L1, and SVA insertions, respectively (**Supplementary Table 6**). Based on a previous study[33], RetroSeq, TEA, and Tangram obtained 94%, 93%, and 98.8% sensitivity for Alu, and 78%, 82%, and 86.6% sensitivity for L1, respectively. We further measured the genotyping accuracy of ERVcaller using the 999 unfixed TE insertions that have been previously confirmed by PCR[44]. ERVcaller correctly genotyped 98% of them (**Supplementary Table 6**). Based on the previous study[33], Tangram and RetroSeq obtained 92.9% and 71.8% for Alu, and 91.1% and 66.7% for L1, respectively (TEA was not included because it was not applicable for genotyping).

Discussion

Polymorphic (unfixed) ERVs and other TEs are an important category of structural variation, potentially critical in human evolution and the development of human disease[4, 45]. They are typically repetitive sequences and are thus difficult to detect using short NGS reads. In this study, we improved upon existing TE calling algorithms and developed ERVcaller to accurately detect and genotype unfixed TE insertions using NGS data. ERVcaller can be applied to both paired-end and single-end WGS, RNA-Seq or targeted DNA sequencing data. By applying to series of simulated datasets, we observed > 97% sensitivity and > 99% precision when the sequencing depths were > 5X (**Figure 2**). Under low sequencing depths (< 5X), ERVcaller still showed the highest sensitivity and precision when the ERV library was used, compared to other tools analyzed in this study (**Figure 4**). In addition, ERVcaller is capable of correctly detecting the insertion breakpoints at single-nucleotide resolution (e.g., > 90% accuracy when depth > 15X), especially with the use of the ERV library as the reference (**Supplementary Figure 4**). More importantly, ERVcaller achieved high genotyping accuracies, i.e., > 96% accuracy as long as the sequencing depths were > 5X (**Supplementary Table 3**). By applying ERVcaller to a subset of the 1000 Genomes project samples, we detected most of the known unfixed TE insertions (100% for ERVs and 96% for other TEs) and their genotypes (100% for ERVs and 98% for other TEs) (**Supplementary Table 6**).

ERVcaller achieves high sensitivity for several reasons, such as implementation of the high mapping-rate aligner BWA-MEM; use of three types of supporting reads; adoption of customized parameters to simultaneously align each read to multiple references or locations and determine the top candidate location based on the average alignment scores of all supporting reads. For example, due to the customized BWA-MEM parameters, ERVcaller significantly increases the sensitivity for detecting highly divergent and novel unfixed TE insertions (e.g., when we used the LTR5_Hs sequence as the only reference, ERVcaller was still capable of discovering the insertions derived from related sequences, such as LTR5A/B). Because it allows for the use of redundant repetitive sequences as the TE reference (e.g., the entire RepBase) with little influence, ERVcaller further increases its detection sensitivity. As we considered three types of supporting reads: chimeric, improper, and split reads, as well as implemented stringent quality control procedures, ERVcaller achieved both high sensitivity and precision. By

comparison, each of the other tools analyzed here showed one or more major limitations. For example, TEMP, ITIS, and STEAK had no functions for genotyping[34-36]. RetroSeq was able to genotype unfixed ERV insertions[30, 33]; however, the genotyping function has been removed from its latest version. ITIS was not practical for large datasets because it screened each TE reference separately, and thus the computing time increased significantly with the increasing number of sequences in the TE reference library[35]. Neither ITIS nor STEAK was able to efficiently detect partial ERV insertions (**Figure 4**) or to detect breakpoints at single-nucleotide resolution.

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 with precision, ERVcaller potentially misses a certain number of unfixed TE insertions in repetitive sequence regions, which is common to all tools of this kind. The unfixed ERV insertions can be complex because of target site duplications at the breakpoints[9], making these insertions even harder to detect. These issues may be partially addressed by increasing insert size, read length, and sequencing depth. Similar to other existing tools, ERVcaller is designed to detect unfixed ERVs that are not present in the human reference genome. To detect the unfixed ERVs existing in the human reference genome, an alternate reference assembly may be created with all unfixed ERV insertions of interest removed. Alternatively, structural variation detection tools, such as Breakdancer[46], can be used to detect those reference unfixed TEs as deletions. In this study, only unfixed ERV insertions with two or more supporting reads were used for the software comparisons with default parameters; however, we did not expect significant changes if different parameters were used. In addition, not all the published tools were compared in this study, such as TIF[31] and MELT[47]. The former only uses split reads for TE detection. **Supplementary Table 1** shows a comparison of the characteristics of existing tools, including those not analyzed in this study.

Many high frequency unfixed ERV loci (i.e., the unfixed ERVs observed in many individuals within a population) have been discovered previously[5, 10]. However, evidence suggests that there are still a large number of undiscovered unfixed ERV loci in the human genome[24], including population- or disease-specific loci[10]. For example, by screening more than 2,000 samples with low sequencing depths, Wildschutte *et al*. detected many rare or population

specific unfixed ERV insertions. Thus, more unfixed ERV insertions may be discovered as more samples are analyzed appropriately. For instance, in this study, we identified a total of 233 unfixed ERV insertions among 15 samples.

More importantly, there is a considerable amount of variation of ERV properties in human populations, such as the number of unfixed ERVs per human genome and frequencies of these ERVs[10]. This indicates that polymorphic ERVs may also be analyzed for their roles in human diseases and health. Thus, for subsequent genetic association analyses using the identified ERV information, we calculated a variety of measurements of ERVs as output from the ERVcaller software. For example, for each analyzed sample, ERVcaller provides the total number of unfixed ERV and other TE insertions and copy numbers for each family (or group). For each insertion, ERVcaller provides several items: including TE group or sub-group, location on the human genome, the genotype, and the length of inserted sequence at the breakpoint (i.e., target site duplication). Other information, such as population frequency, can be obtained using existing sources[10]. **Supplementary Table 7** shows the variables measured by ERVcaller. Thus, ERVcaller can also serve as a "genotyper" of ERV and other TE insertions for genetic association studies.

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.html.

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 Thomas Buttolph and Sheryl White, Ph.D. in the COBRE Neuroscience Cellular and Molecular Core at the University of Vermont for their technical expertise in

performing PCR and cloning. The authors thank the Vermont Integrative Genomics Resource for their technical expertise in performing Sanger sequencing. 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., Cong Gao, 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. Initial sequencing and analysis of the human genome. *Nature* **409**, 860 (2001).
2. Wessler, S.R. Transposable elements and the evolution of eukaryotic genomes. *Proceedings of the National Academy of Sciences* **103**, 17600-17601 (2006).
3. Katzourakis, A., Pereira, V. & Tristem, M. Effects of Recombination Rate on Human Endogenous Retrovirus Fixation and Persistence. *Journal of Virology* **81**, 10712-10717 (2007).
4. Jern, P. & Coffin, J.M. Effects of retroviruses on host genome function. *Annu Rev Genet* **42**, 709-732 (2008).
5. Marchi, E., Kanapin, A., Magiorkinis, G. & Belshaw, R. Unfixed endogenous retroviral insertions in the human population. *J Virol* **88**, 9529-9537 (2014).
6. Macfarlane, C.M. & Badge, R.M. Genome-wide amplification of proviral sequences reveals new polymorphic HERV-K(HML-2) proviruses in humans and chimpanzees that are absent from genome assemblies. *Retrovirology* **12**, 35 (2015).

7. Subramanian, R.P., Wildschutte, J.H., Russo, C. & Coffin, J.M. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology* **8**, 90 (2011).
8. Belshaw, R. et al. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. *J Virol* **79**, 12507-12514 (2005).
9. Kahyo, T., Yamada, H., Tao, H., Kurabe, N. & Sugimura, H. Insertionally polymorphic sites of human endogenous retrovirus-K (HML-2) with long target site duplications. *BMC Genomics* **18**, 487 (2017).
10. Wildschutte, J.H. et al. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E2326-2334 (2016).
11. Goke, J. et al. Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* **16**, 135-141 (2015).
12. Robbez-Masson, L. & Rowe, H.M. Retrotransposons shape species-specific embryonic stem cell gene expression. *Retrovirology* **12**, 45 (2015).
13. Rooney, M.S., Shukla, S.A., Wu, C.J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48-61 (2015).
14. Handsaker, R.E., Korn, J.M., Nemesh, J. & McCarroll, S.A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**, 269-276 (2011).
15. Garrison, K.E. et al. T cell responses to human endogenous retroviruses in HIV-1 infection. *PLoS Pathog* **3**, e165 (2007).
16. Groger, V. & Cynis, H. Human Endogenous Retroviruses and Their Putative Role in the Development of Autoimmune Disorders Such as Multiple Sclerosis. *Front Microbiol* **9**, 265 (2018).
17. Brodziak, A. et al. The role of human endogenous retroviruses in the pathogenesis of autoimmune diseases. *Med Sci Monit* **18**, RA80-88 (2012).

18. Marguerat, S., Wang, W.Y., Todd, J.A. & Conrad, B. Association of human endogenous retrovirus K-18 polymorphisms with type 1 diabetes. *Diabetes* **53**, 852-854 (2004).
19. Douville, R.N. & Nath, A. Human endogenous retroviruses and the nervous system. *Handb Clin Neurol* **123**, 465-485 (2014).
20. Li, W. et al. Human endogenous retrovirus-K contributes to motor neuron disease. *Sci Transl Med* **7**, 307ra153 (2015).
21. Slokar, G. & Hasler, G. Human Endogenous Retroviruses as Pathogenic Factors in the Development of Schizophrenia. *Front Psychiatry* **6**, 183 (2015).
22. Kassiotis, G. Endogenous retroviruses and the development of cancer. *Journal of immunology* **192**, 1343-1349 (2014).
23. Gonzalez-Cao, M. et al. Human endogenous retroviruses and cancer. *Cancer Biol Med* **13**, 483-488 (2016).
24. Moyes, D., Griffiths, D.J. & Venables, P.J. Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. *Trends Genet* **23**, 326-333 (2007).
25. Macfarlane, C. & Simmonds, P. Allelic variation of HERV-K(HML-2) endogenous retroviral elements in human populations. *J Mol Evol* **59**, 642-656 (2004).
26. Lee, E. et al. Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967-971 (2012).
27. Hormozdiari, F. et al. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* **26**, i350-357 (2010).
28. Hormozdiari, F., Alkan, C., Eichler, E.E. & Sahinalp, S.C. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* **19**, 1270-1278 (2009).
29. Quinlan, A.R. et al. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* **20**, 623-635 (2010).
30. Keane, T.M., Wong, K. & Adams, D.J. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* **29**, 389-390 (2013).
31. Nakagome, M. et al. Transposon Insertion Finder (TIF): a novel program for detection of de novo transpositions of transposable elements. *BMC bioinformatics* **15**, 71 (2014).
32. Thung, D.T. et al. Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome biology* **15**, 488 (2014).

33. Wu, J. et al. Tangram: a comprehensive toolbox for mobile element insertion detection. *BMC Genomics* **15**, 795 (2014).
34. Zhuang, J., Wang, J., Theurkauf, W. & Weng, Z. TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Res* **42**, 6826-6838 (2014).
35. Jiang, C., Chen, C., Huang, Z., Liu, R. & Verdier, J. ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. *BMC bioinformatics* **16**, 72 (2015).
36. Santander, C.G. et al. STEAK: A specific tool for transposable elements and retrovirus detection in high-throughput sequencing data. *Virus Evol* **3**, vex023 (2017).
37. Bao, W., Kojima, K.K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11 (2015).
38. Hu, X. et al. pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics* **28**, 1533-1535 (2012).
39. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**, R36 (2013).
40. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
41. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
42. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-575 (2007).
43. Genomes Project, C. et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
44. Stewart, C. et al. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* **7**, e1002236 (2011).
45. Burns, K.H. Transposable elements in cancer. *Nat Rev Cancer* **17**, 415-424 (2017).
46. Chen, K. et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**, 677-681 (2009).
47. Gardner, E.J. et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res* **27**, 1916-1929 (2017).

Figures and Figure Legends

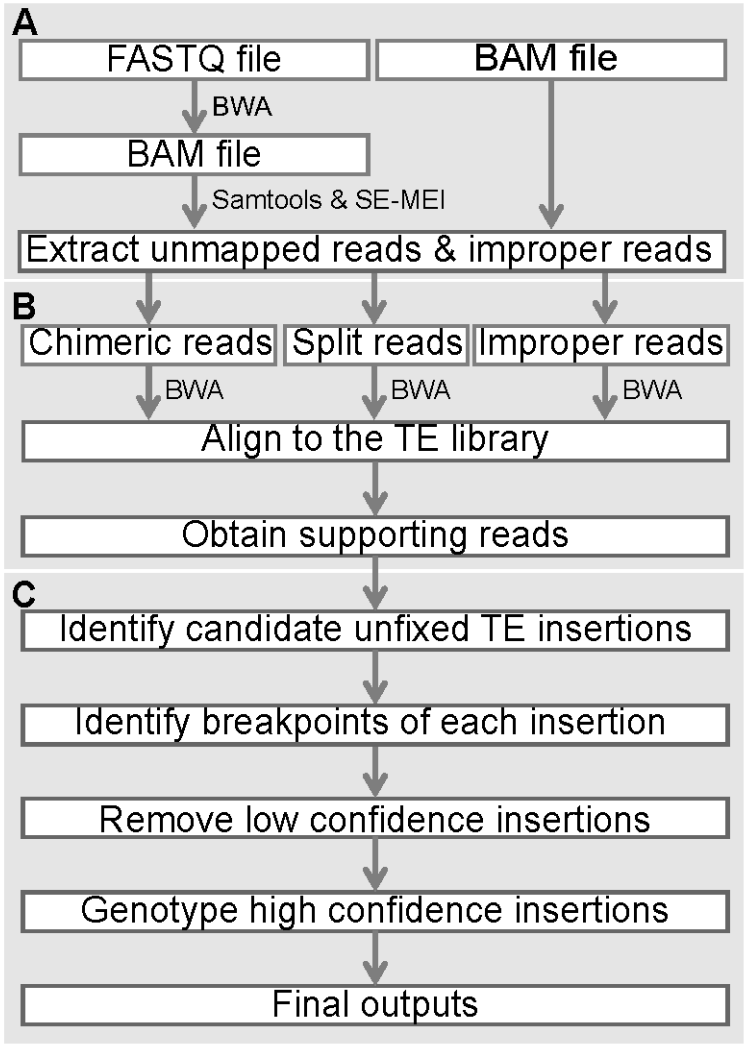

Figure 1 ERVcaller workflow. The three components include **A**) extracting unmapped reads; **B**) obtaining supporting reads; and **C**) detecting and genotyping unfix TE insertions.

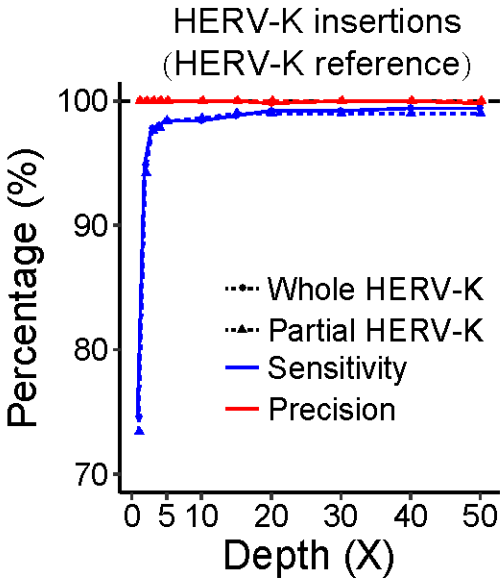

Figure 2 Sensitivity and precision under different sequencing depths. The sensitivity and precision of ERVcaller applied to the simulated datasets with whole and partial HERV-K insertions using the HERV-K sequence as the only reference.

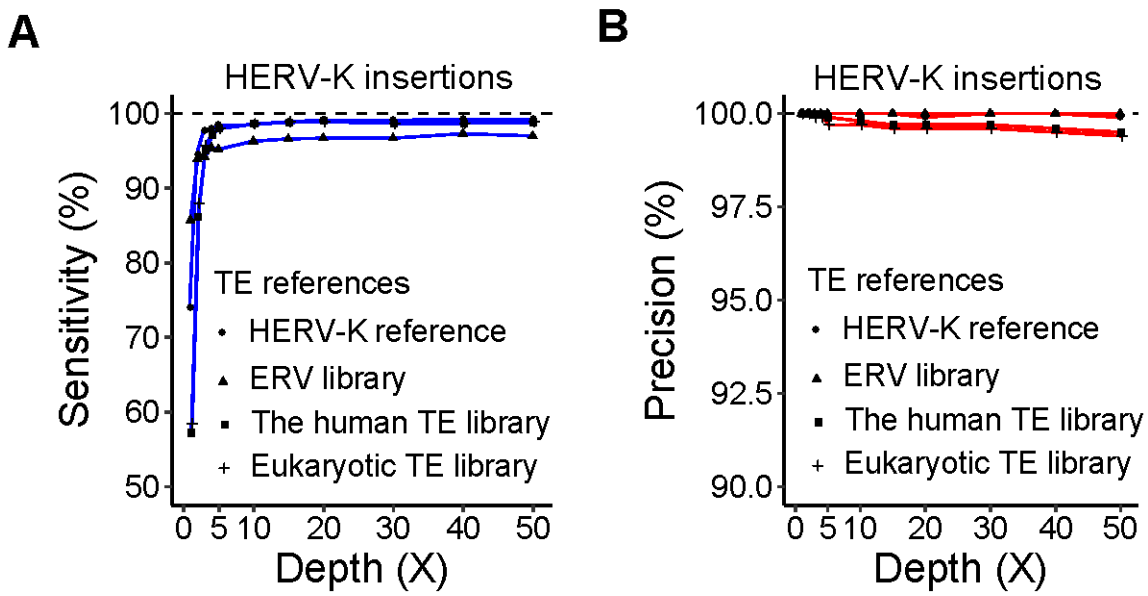

Figure 3 Sensitivity and precision with four different types of TE references or libraries. **A**) Sensitivity of ERVcaller applied to the datasets with HERV-K insertions using four different types of TE references or libraries, including the HERV-K sequence as the only reference; an ERV library containing 743 human and non-human ERV sequences; a human TE library containing only 23 consensus human TEs; and a eukaryotic TE library from the RepBase database; **B**) Precision of ERVcaller applied to the datasets.

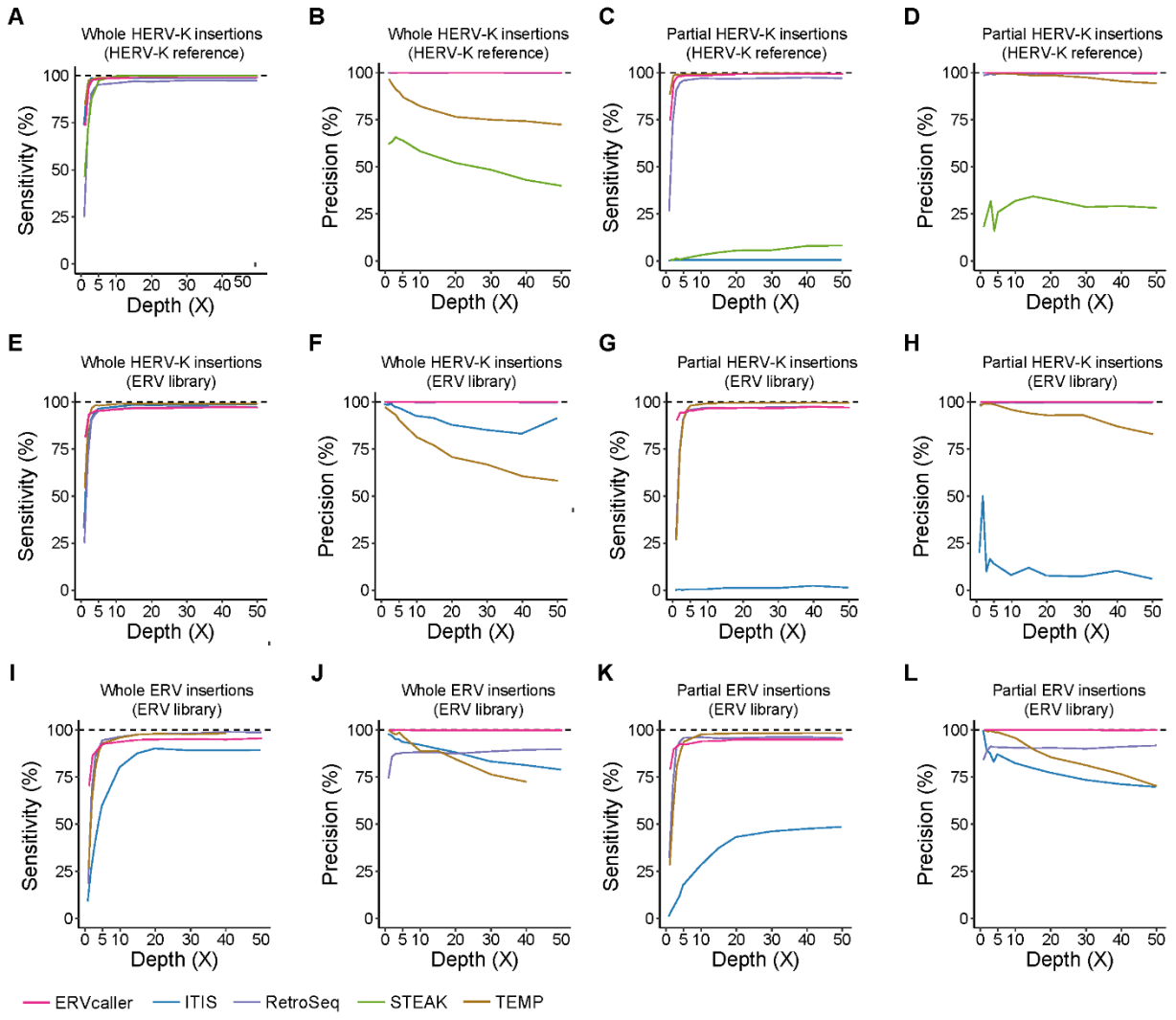

Figure 4 Comparison of sensitivity and precision of ERVcaller with the other four tools. **A**,**B**) Sensitivity and precision of the tools applied to the datasets with whole HERV-K insertions using the HERV-K sequence as the reference; **C**,**D**) Sensitivity and precision of the tools applied to the datasets with partial HERV-K insertions using the HERV-K sequence as the reference; **E**,**F**) Sensitivity and precision of the tools applied to the datasets with whole HERV-K insertions using the ERV library as the references; **G**,**H**) Sensitivity and precision of the tools applied to the datasets with partial HERV-K insertions using the ERV library as the references; **I**,**J**) Sensitivity and precision of the tools applied to the datasets with whole ERV (from the ERV library) insertions using the ERV library as the references; **K**,**L**) Sensitivity 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.

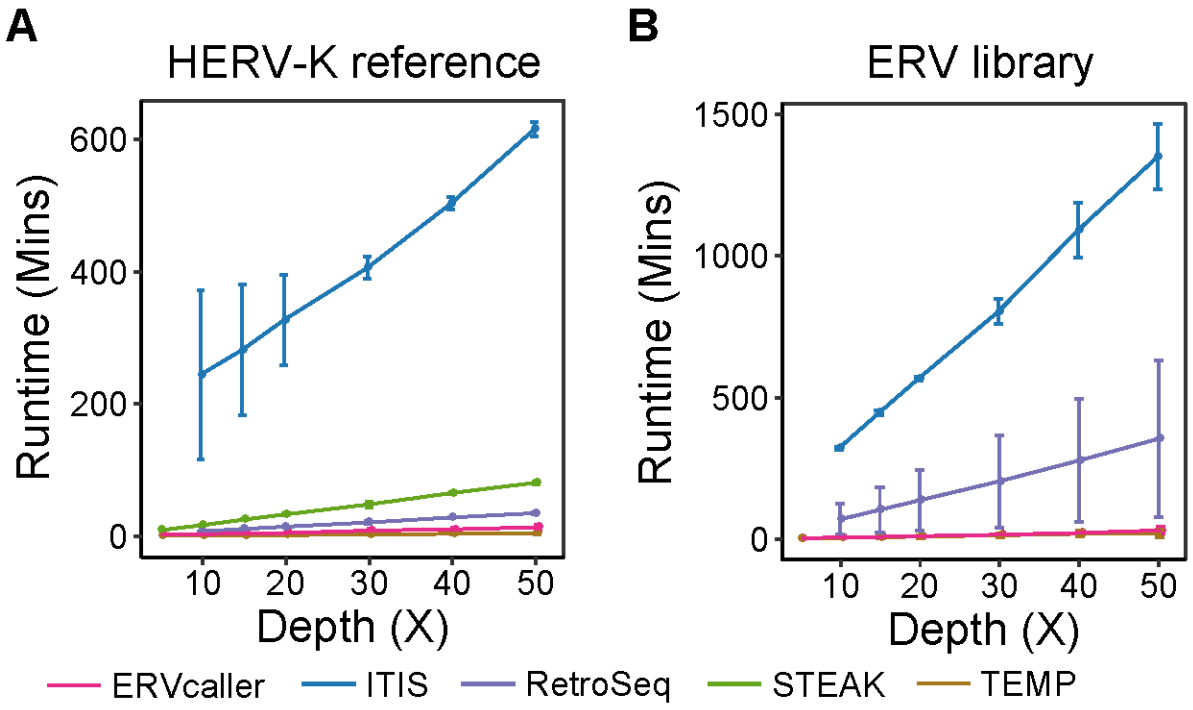

Figure 5 Comparison of runtime of ERVcaller with four published tools using **A**) the HERV-K reference, and **B**) the ERV library, as the reference(s), respectively.