frontiers | Frontiers in Bioinformatics

TYPE Original Research
PUBLISHED 08 February 2023
DOI 10.3389/fbinf.2022.1062328

OPEN ACCESS

EDITED BY
Richard Allen White III,
University of North Carolina at Charlotte, United States

REVIEWED BY
Emma Fay Harding,
University of New South Wales, Australia
Robert James Gifford,
University of Glasgow, United Kingdom
Stephen Francis,
University of California, San Francisco, United States

*CORRESPONDENCE
Alfredo Iacoangeli,
✉ alfredo.iacoangeli@kcl.ac.uk

SPECIALTY SECTION
This article was submitted to Genomic Analysis, a section of the journal Frontiers in Bioinformatics

RECEIVED 05 October 2022
ACCEPTED 12 December 2022
PUBLISHED 08 February 2023

CITATION
Bowles H, Kabiljo R, Al Khleifat A, Jones A, Quinn JP, Dobson RJB, Swanson CM, Al-Chalabi A and Iacoangeli A (2023), An assessment of bioinformatics tools for the detection of human endogenous retroviral insertions in short-read genome sequencing data. *Front. Bioinform.* 2:1062328.
doi: 10.3389/fbinf.2022.1062328

An assessment of bioinformatics tools for the detection of human endogenous retroviral insertions in short-read genome sequencing data

Harry Bowles[1], Renata Kabiljo[1,2], Ahmad Al Khleifat[1], Ashley Jones[1], John P. Quinn[3], Richard J. B. Dobson[2,4,5,6], Chad M. Swanson[7], Ammar Al-Chalabi[1,8] and Alfredo Iacoangeli[1,2,4]*

[1]Department of Basic and Clinical Neuroscience, King's College London, Maurice Wohl Clinical Neuroscience Institute, Institute of Psychiatry, Psychology and Neuroscience, London, United Kingdom, [2]Department of Biostatistics and Health Informatics, King's College London, Institute of Psychiatry, Psychology and Neuroscience, London, United Kingdom, [3]Department of Pharmacology and Therapeutics, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, United Kingdom, [4]NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, London, United Kingdom, [5]Institute of Health Informatics, University College London, London, United Kingdom, [6]NIHR Biomedical Research Centre, University College London Hospitals NHS Foundation Trust, London, United Kingdom, [7]Department of Infectious Diseases, School of Immunology and Microbial Sciences, King's College London, London, United Kingdom, [8]Department of Neurology, King's College Hospital, London, United Kingdom

There is a growing interest in the study of human endogenous retroviruses (HERVs) given the substantial body of evidence that implicates them in many human diseases. Although their genomic characterization presents numerous technical challenges, next-generation sequencing (NGS) has shown potential to detect HERV insertions and their polymorphisms in humans. Currently, a number of computational tools to detect them in short-read NGS data exist. In order to design optimal analysis pipelines, an independent evaluation of the available tools is required. We evaluated the performance of a set of such tools using a variety of experimental designs and datasets. These included 50 human short-read whole-genome sequencing samples, matching long and short-read sequencing data, and simulated short-read NGS data. Our results highlight a great performance variability of the tools across the datasets and suggest that different tools might be suitable for different study designs. However, specialized tools designed to detect exclusively human endogenous retroviruses consistently outperformed generalist tools that detect a wider range of transposable elements. We suggest that, if sufficient computing resources are available, using multiple HERV detection tools to obtain a consensus set of insertion loci may be ideal. Furthermore, given that the false positive discovery rate of the tools varied between 8% and 55% across tools and datasets, we recommend the wet lab validation of predicted insertions if DNA samples are available.

KEYWORDS

herv-k, benchmarking, whole-genome sequencing, retrovirus, bioinformatics

1 Introduction

Endogenous retroviruses (ERVs) integrated into the genome of vertebrates as a result of ancient exogenous infections. They invaded the germ cell lines of all vertebrates including humans, becoming an integral part of the germline transmission and therefore replicate in a Mendelian fashion (Gifford and Tristem, 2003). Human endogenous retroviruses (HERVs) comprise ~8% of the genome, whereas protein coding genes comprise only 1%–2% (Pisano et al., 2019). Although they make up a striking portion of the human genome, most of them are inactive as a consequence of the accumulation of mutations and DNA methylation (Belshaw et al., 2005). The HML-2 HERV-K subgroup includes some of the most recent HERV integrations, which are found as full length (or near full length) sequences in over 80 different loci (Subramanian et al., 2011). Though there are several full-length copies of HERV-K in the genome, none are likely to produce an infectious virus (Boller et al., 2008). HERV-K sequences can be full length proviruses, solo long terminal repeats or 2-LTR sequences and are polymorphic in the human population (Garcia-Montojo et al., 2018). A full length ERV provirus consists of long terminal repeats (LTRs) flanking the viral genes (gag, pro, pol and env). In the majority of elements defined as HERV loci, only the LTRs are present and these contain the promoter and enhancer regions (Klaver and Berkhout, 1994) (Figure 1).

HERVs are classified as transposable elements (TEs) but there are notable differences between HERVs and other TEs. HERVs are less common than Alus and LINEs (long interspersed nuclear elements) and a full length HERV-K provirus is approximately 10 kb in length, compared to full length LINEs which are 6 kb in length, while Alu insertions are approximately 300 bp (Payer and Burns, 2019). LINEs are highly capable of transposition and Alus can hijack LINE mechanisms to the same end, in fact Alus are the most abundant transposable element in the human genome. In contrast to this, HERVs generally lack transposon function, though they can transcribe RNA (Dolei et al., 2019). A further difference is that HERVs are flanked by LTR sequences, containing promoter and enhancer regions, which are not found in LINEs or Alus (Larsen et al., 2018). The SVA (Sine VNTR Alu) is another TE reliant on LINE1 for transposon function and is a composite TE containing ALU and HERV LTR sequence. In contrast to HERVs, most SVA elements are full length while the majority of HERVs contain only the flanking LTRs (Hancks and Kazazian, 2010).

Characterizing the HERV genomic landscape is challenging. HERVs are thousands of bases long and highly repetitive. This means that short-read genome sequencing cannot characterize the sequence of HERVs that are not present in the reference genome, beyond a limited number of bases (Ewing, 2015). Furthermore, reads from repetitive regions can introduce ambiguities in the mapping step of genome alignment where there are multiple putative matches (Teissandier et al., 2019). This issue extends to biological tests of transposable elements, as short oligos designed to target a specific TE locus may sit down at multiple locations on the genome (Bourque et al., 2018).

HERV-Ks, and related transposable elements, have been linked to a wide range of diseases including cancer and neurodegenerative diseases, such as amyotrophic lateral sclerosis (ALS), *via* multiple mechanisms. For example, their insertion into the human genome may alter gene expression or disrupt reading frames (Buzdin et al., 2003); they were reported to be upregulated in biological samples from people affected by neurodegenerative diseases and cancer (Garcia-Montojo et al., 2020; Dervan et al., 2021; Jones et al., 2021); furthermore, their expression may be toxic for certain cell types such as motor neurons (Li et al., 2015). Given their proposed broad role in human diseases, we focused our work on HERV-Ks.

Recent advances in next-generation sequencing (NGS) have made sequencing large DNA molecules a common practice in genetic research, allowing for the investigation of a wide range of variants from single nucleotide variants to large structural variants (Iacoangeli et al., 2019a). This technology has also been used to study HERVs (Xue et al., 2020a). It is established that HERVs can express RNA which can be captured in NGS experiments. For example, RNA sequencing experiments have quantified HERV RNA in healthy and tumor cell lines (Rezaei et al., 2021) and have highlighted HERV RNA as a biomarker for cell pluripotency (Santoni et al., 2012). Chip-seq experiments have highlighted a role for HERV-H loci in

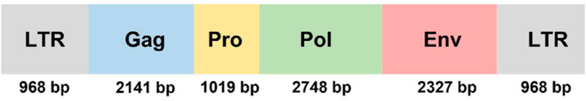

FIGURE 1
This schematic representation shows the general structure of a full length HERV-K with approximate base pair lengths for each section. The LTR length is the LTR5_Hs length given in the DFAM database (accession = DF0000558), the internal gene lengths come from HERVK11 internal regions (DFAM accession = DF0000189) (Storer et al., 2021). The gag gene encodes a polyprotein that is cleaved to produce the structural proteins that make up a viral particle. The pro gene encodes the viral protease while pol codes for the viral enzymes required for reverse transcription and integrase. The env gene codes for the viral envelope protein that mediates viral entry into a target cell. The gene products require post translational cleavage, for example, pro cleaves the gag polyprotein into "matrix", "capsid" and "nucleocapsid" subunits. The majority of HERV-K loci only contain the LTR regions which contain the enhancer and promoter sequences that regulate transcription.

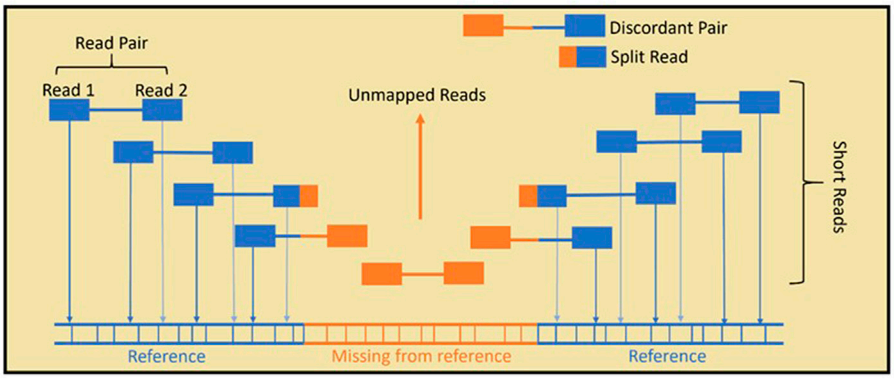

FIGURE 2
In standard Illumina paired-end SR-WGS, the DNA sample is first sheared into small fragments. Then, both ends of the fragment are sequenced, giving paired reads. Read mapping is achieved by aligning the reads to a reference genome. HERVs that are present in the sample but not the reference, cannot be aligned fully and remain largely unmapped, which means they are often overlooked in genome analysis. Specialised bioinformatics tools use unmapped, split and discordant reads to predict non-reference HERVs.

chromatin restructuring during cell differentiation (Zhang et al., 2019). Genome wide methylation studies may prove to be very useful in understanding HERV regulation as hypomethylation of HERVs correlates with increased expression (Chiappinelli et al., 2015).

Currently, a number of bioinformatic tools for the identification of HERV insertion loci in short-read whole-genome sequencing (SR-WGS) data exist, mostly based on the exploitation of split and discordant reads to reveal the presence of potential HERV insertions (Figure 2). Some of these tools have been developed to detect a range of TEs, including HERVs, Alus and LINEs, while others are specifically targeted to HERV detection. Given the lack of a comprehensive and independent assessment of their performance, an evaluation of the current tools for HERV detection is greatly needed for the design of optimal analysis pipelines and to promote the discussion necessary for the scientific community to establish best practice protocols. On this basis, we designed a set of experiments to benchmark widely used computational tools and protocols for the detection of HERVs in SR-WGS data (short reads = 50–200 bp). We hypothesize that specialist HERV tools might perform better than general TE detectors and we aim to quantify the benefits and limitations of using generalist or specialized tools to study HERVs in NGS. Considering their proposed role in human disease, we focused our experiment on the identification of HERV-K insertions that are not present in the human reference genome (non-reference HERV-Ks). We tested six widely used tools on three short-read sequencing datasets: a large short-read whole-genome sequencing (SR-WGS) dataset (50 human samples), a simulated SR-WGS dataset, and six SR-WGS samples for which matching long-read sequencing data was available.

2 Methods

2.1 Overview of the tested tools

MELT: MELT scans WGS data for clusters of discordant read pairs and split reads. Split and discordant reads can be mapped to an insertional element reference sequence provided by the user, to allow the detection of specific insertion types. It can also genotype reference mobile elements (Gardner et al., 2017). MELT has previously been used to integrate TE insertion and TE expression data in cancer lines (Clayton et al., 2016) and to elucidate the evolutionary mechanisms underlying TE diversity (Rishishwar et al., 2018). According to its documentation, MELT was not tested for HERV detection in its original publication, and the authors predicted that it may perform poorly on LTR elements compared to non-LTR transposable elements (such as Alus). Nevertheless, MELT has been used to detect HERVs (Santander et al., 2017; Chen and Li., 2019; Feusier et al., 2019).

Mobster: This tool also uses discordant reads alongside split reads and an insertional reference sequence to predict specific insertion sites. Mobster has also been used to highlight the role of TEs in cancer (Clayton et al., 2016) and has been used to show an association between TEs and autism (Borges-Monroy et al., 2021). When Mobster was released, the authors reported that it was not able to identify HERV insertions. However, they tested it using just two WGS paired-end samples and since its first publication in 2014, Mobster has been extensively upgraded and gained a considerable popularity. We included it in our experiment, tested its updated version on a larger sample and used a different HERV-K template sequence to that used in the authors' benchmarking work (Thung et al., 2014).

Retroseq: Retroseq uses discordant read pairs to identify putative insertion sites and filters for read pairs which align to a reference of interest (Keane et al., 2013). Retroseq has been used for mapping transposable elements in evolutionary studies (Dennenmoser et al., 2019) and extensively adopted as the starting point of more advanced pipelines (Chen and Li, 2019).

Steak: Steak annotates both reference and non-reference mobile elements. Unlike the other tools, it first identifies reads that partially map to the target HERV reference sequence. These reads are assumed to map to the edge of the insertion. The mapped fragments of the reads are removed, and a library of host reference flank reads and their mates is created. These reads are mapped onto the human reference genome to identify the presence of both reference and non-reference HERV loci (Santander et al., 2017). Steak has been less widely used than MELT, Mobster and Retroseq, though it has been used in combination with PCR amplification to map HERV-K loci in the genome (Xue et al., 2020b) and is regularly used as a benchmarking comparison for new tool publications.

ERVcaller: The tool extracts incorrectly mapped read pairs and split reads to identify likely insertions sites, that are then aligned to a reference sequence to allow for the detection of specific insertion types (Chen and Li, 2019). ERVcaller has been used in combination with chip-seq and RNAseq analysis to quantify the contribution of TEs to epigenetic regulation (Groza et al., 2022).

Retroseq+: This pipeline is our in-house implementation of a protocol described by Wildschutte et al. It uses Retroseq as a base for predicting HERV-K insertion loci. It then refines the results through insertion junction reconstruction and secondary scanning of the junction for HERV-K sequence by RepeatMasker. Because this protocol is not available as an automatic bioinformatics pipeline, we have implemented it ourselves, following the authors' description (Wildschutte et al., 2016; Kabiljo et al., 2022).

These chosen tools include widely used, established TE detectors (Mobster, MELT and Retroseq) as well as newer tools specifically developed for HERV detection (ERVcaller, Steak, Retroseq+). We did not include tools designed for the analysis of tumor cell lines, or tools designed to analyze data other than short-read NGS, or evolutionary aspects of HERVs as they fell outside our scope to test tools for the detection of germline non-reference HERV-K insertions in short-read NGS data. Scripts for these tools are available as supplementary materials as are flow diagrams explaining each tool in more detail.

2.2 Benchmarking experiments overview

In order to assess the performance of these tools, we set up four experiments: i) we estimated the performance of the tools using simulated NGS data; ii) using the HERV-K calls from the 50 SR-WGS samples, we attempted to validate the tools by quantifying the proportion of predicted HERV-K insertions that were previously reported in literature; iii) using the tools on 50 SR-WGS samples, we measured the agreement between tools; iv) finally, we assessed the specificity of each tool by using long-read data for the validation of HERV-K calls from matching short-read data.

2.3 Simulated short-read WGS analysis

The purpose of this test was to assess the sensitivity and specificity of each tool using simulated data with a known set of insertions. For the simulation test, short-read paired-end Illumina WGS data were simulated from the hg19 reference sequence using DWGSIM (parameters in Table 1) (Homer, 2010). Hg19 is reported to have 66 HERV-K (HML-6) full length, proviral loci (2) from which we randomly selected 15 of type LTR3A to use as target LTRs (Supplementary File S4). Furthermore, in order to test whether the tools are able to distinguish between LTR types, we also randomly selected four LTR3B type HML-6 proviral loci (Supplementary File S5). Both LTR3A and LTR3B are HML6 insertions but have been shown to cluster separately in phylogenetic analyses (Pisano et al., 2019). We expect that including both subtypes in our test may provide a higher degree of resolution into the evaluation of the tools' accuracy.

To simulate these HERV-K sites as novel insertions, after generating the WGS data, we removed these proviruses from the hg19 reference using Bedtools masking followed by deletion of the mask (Quinlan and Hall, 2010). The simulated FASTQ files were then aligned to the edited hg19 using BWA-MEM. Thus, the simulated data contained 19 known HERV-K insertions that were not present in our edited reference genome, and these insertions were the only non-reference HERV elements in the simulated samples. Each tool was then applied to the simulated WGS. Only the LTR3A sequence was used as target reference sequence template, meaning each tool should have only detected the 15 LTR3A loci, not the 4 LTR3B loci. This allowed us to see how well each tool can distinguish specific insertion types as well as assess general sensitivity. We defined sensitivity as the proportion of the known, non-reference insertions which were successfully detected: True positives/(True positives + False negatives). We defined precision as the proportion of positive results that are true: True positives/(True positives + False positives).

2.4 Overlap analysis and comparison with previously reported HERVs

Each tool was applied to WGS data of 50 ALS patients from the British Project Mine dataset (Project MinE ALS Sequencing Consortium, 2018; Iacoangeli et al., 2019b). This WGS was generated from blood samples, using the Illumina Hiseq

TABLE 1 Custom parameters used to generate the simulated data. All parameters not included in this table were kept as default.

	Read length	No. reads	Coverage	Outer distance	-e (error)	-s (std-dev outer dist)
Genome 1	150 bp	333,333,333	32X	400	0.020	5
Genome 2	100 bp	500,000,000	32X	400	0.020	5
Genome 3	150 bp	105,000,000	10.5X	400	0.020	5
Genome 4	100 bp	105,000,000	7X	400	0.020	5

2000 platform. The resulting WGS samples had read length equal to 100 bp with an average coverage depth of 40X (paired-end reads). We aligned them to the hg19 reference genome using Burrows-Wheeler alignment, BWA-MEM (Li, 2013). The predicted insertion sites across all genomes were compared to a list of 40 well characterized polymorphic HERV-K insertions previously described in the literature (Supplementary File S1) (Kahyo et al., 2017). They were also compared to all reference HERV loci (both HERV-Ks and other HERV/LTR subgroups) obtained through the UCSC table browser using the RepeatMasker (RMSK) track for the hg19 genome build. The following identifiers were used to retrieve HERV-K reference loci: LTR5_Hs, LTR5A, LTR5B, HERV-K and HERV-K-int (Supplementary File S2); while the set of all reference HERVs was obtained by taking the entire UCSC RMSK hg19 track and extracting those which had the identifiers “ERV” or “LTR” (Supplementary File S3). HML-2 type HERV-Ks, which are targeted in this analysis can be subclassified based on their LTR sequence. LTR5B is the phylogenetically oldest HML-2 LTR, LTR5_Hs is younger and human specific.

An overlap was defined as a predicted insertion being within 500 base pairs of a known ERV locus. The number of reference HERV-Ks and HERVs per million bases across the human chromosomes is shown in Figures 3A–C.

This test allowed us to quantify the proportion of predicted insertions of each tool that matched known and validated HERV-Ks under the assumption that such calls are more likely to be true positives and therefore tools which show a higher proportion are more reliable.

The results of each tool were also compared to one another, giving for each one of them, the proportion of its results that were also predicted by each of the other tools. For Steak and Melt reference results were filtered out from the total results using the UCSC RMSK table of hg19 reference HERV-K loci (Supplementary Table S2). This allowed us to quantify the agreement across tools.

2.5 Long-read sequencing data

We used a set of six samples from Wang et al., for which both short and long-read genome sequencing data were available (Wang et al., 2019) (GIAB data, IDs: HG002, HG003, HG004, HG005, HG006, HG007). Briefly, the short-read data (derived from blood) were sequenced using Illumina Hiseq 2500 giving mean 105 bp paired end reads with coverage depth ranging between 15.6X and 18.8X. The long reads were sequenced using PacificBio Sequel system version 2. For these samples the read lengths were between 10 KB and 18 KB and the mean coverage depth of samples ranged between 28.5X and 69X.

Long-read sequencing (read length >10,000 base-pairs) can capture a large overhang, if not the whole, of the HERV-K allowing for their accurate identification (Chu et al., 2021; Chu et al., 2021; Troskie et al., 2021). We applied each tool to the short-read WGS data to predict LTR5_Hs HERV-K insertions and used the long-read data for validation as follows. For each predicted insertion, we extracted long reads mapped at the corresponding locus from the matching long-read WGS sample. These long reads were then assembled into contigs using wtdbg2 (Ruan and Li, 2020). RepeatMasker (Tempel, 2012) can detect and classify repetitive elements in genomic sequences. It was applied to the long-read assembled contigs to confirm the presence of the HERV-K LTR5_Hs sequence at each predicted locus.

If the contig, at a given locus where a short-read based prediction was made, tested positive for HERV-K when analysed with RepeatMasker, the predicted HERV-K insertion was considered true. The proportion of each tool’s predictions which are successfully validated is an indicator of the tool accuracy.

2.6 Computational efficiency report

The computational efficiency of each tool is an important factor, especially if the users have limited resources and large datasets. The purpose of this test was to quantify the computational resources required by each tool.

We tested the memory usage and time taken for each tool to run on a single WGS sample from the Project MinE dataset. Slurm was the scheduling system on the Linux HPC platform used for this project. Slurm has its own command (sacct) for timing scripts and assessing memory usage. Each tool was applied to a single WGS sample from Project MinE and sacct was used to report the tools memory and cpu usage. To determine the sizes of intermediate files produced by each

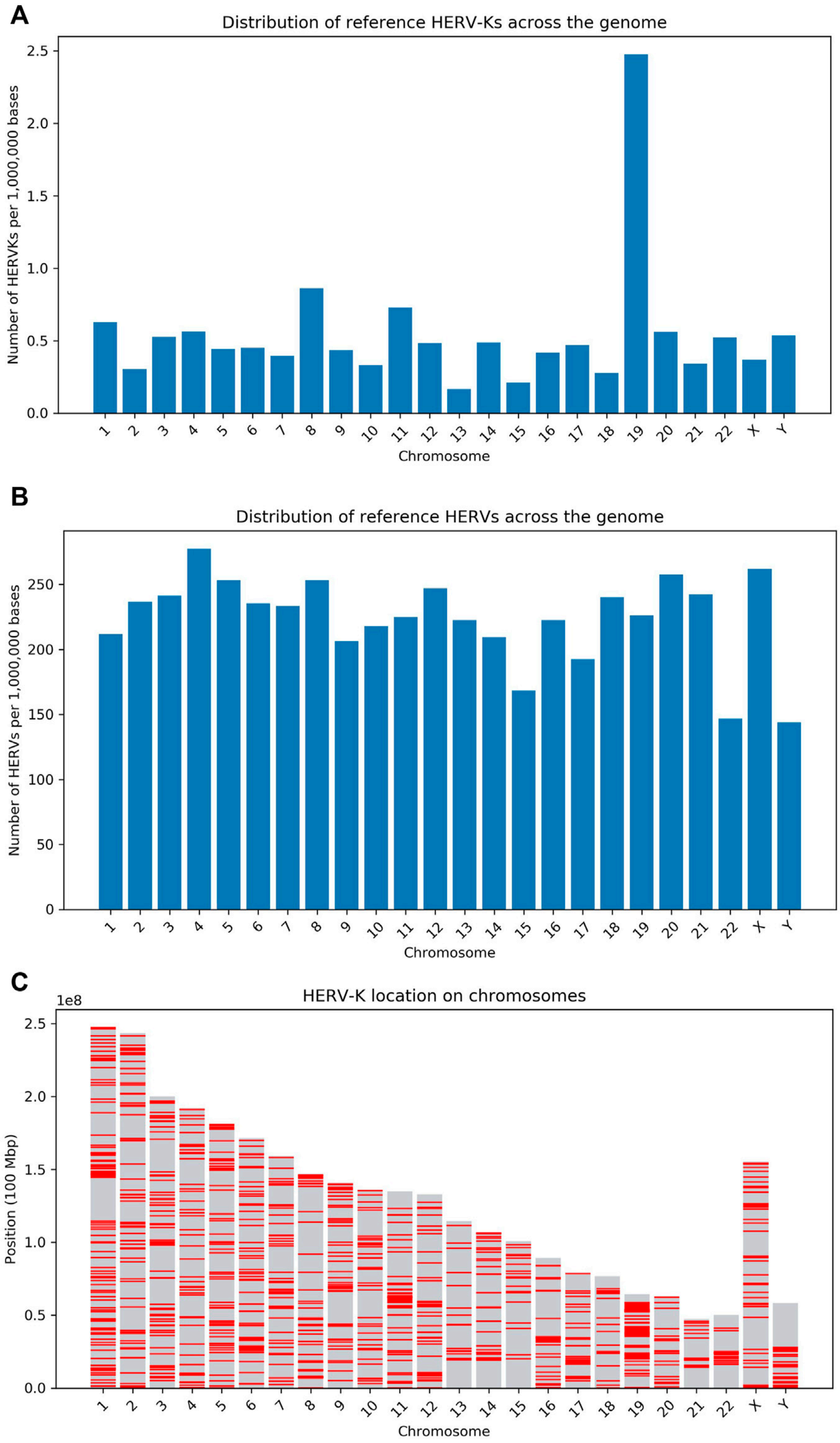

FIGURE 3
Overview of the density of HERV loci in the human genome. **(A)** Number of HERV-Ks per million bases in each human chromosome (as given by the UCSC RMSK table). The high HERV density in chr19 has been previously reported (Katzourakis et al., 2007) and other transposable elements are also enriched on this chromosome (Gianfrancesco et al., 2019). **(B)** Number of HERVs and LTR sequences per million bases in each human chromosome–Data is from the UCSC RMSK table. **(C)** This panel shows the distribution of HERV-K on each chromosome, with a red line indicating the presence of the HERV-K LTR. These results are obtained from the LTR5/HERV-K UCSC RMSK table.

TABLE 2 Results of analysis of simulated WGS data. Each row corresponds to a different tool. The table reports the number of correctly identified LTR3A insertions (the target insertions, 15 loci in the simulated genome), the number of LTR3B insertions (4 loci in the simulated genome) mistakenly classified as LTR3A insertions, and the total number of predicted insertions, including the ones that did not correspond to any of the 19 simulated insertion loci, for each sample.

	Genome 1 (150 bp, 32X)					Genome 2 (100 bp, 32X)					Genome 3 (150 bp, 10.5X)					Genome 4 (100 bp, 7X)				
	LTR3A	LTR3B	Total	Precision	Sensitivity	LTR3A	LTR3B	Total	Precision	Sensitivity	LTR3A	LTR3B	Total	Precision	Sensitivity	LTR3A	LTR3B	Total	Precision	Sensitivity
Retroseq	11	2	16	0.68	0.73	11	2	18	0.61	0.73	7	0	8	0.88	0.46	7	0	8	0.88	0.46
Retroseq+	10	0	18	0.56	0.67	9	0	20	0.45	0.60	8	0	12	0.67	0.53	5	0	6	0.83	0.33
Melt	7	0	9	0.78	0.46	7	0	9	0.78	0.46	8	0	9	0.89	0.53	7	0	9	0.78	0.46
Steak	3	0	5	0.6	0.20	2	0	2	1.00	0.13	2	0	2	1.00	0.13	1	0	1	1.00	0.07
ERVcaller	12	0	13	0.92	0.80	12	0	14	0.86	0.80	12	0	13	0.92	0.80	10	0	11	0.91	0.67
Mobster	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

tool, the "du" command was run on a loop, executing every second, over the directory in which each tool was running. The difference between the starting directory size and largest directory size reported by "du" is reported.

3 Results

3.1 Simulated data results

Each tool was applied to a set of four simulated WGS samples, with known HERV insertions of different types (LTR3A and LTR3B). Each tool was given LTR3A as target element. This allowed us to estimate the sensitivity and precision of each tool. Precision and sensitivity highly varied across tools, ranging between 0.56–0.92, and 0.20–0.80 respectively on the higher quality samples (32X and 150 bp reads). Although the best performing tool was ERVcaller, HERV specialist tools did not perform consistently better than generalist tools, and Steak showed the lowest precision and sensitivity across all simulated genomes. All tools performed worse on samples with lower read depth (Table 2). For example, Retroseq detected 11/15 LTR3A insertions in a 32X sample but only 7/15 LTR3A insertions in the 10.5X and 7X samples. However, the degree to which read depth affects each tool varied. ERVcaller performed equally well on 32X and 10.5X samples finding 12 of the 15 LTR3A insertions and, in the 7X sample, it still identified 10 insertions. The only tool to mistakenly detect an LTR3B insertion was Retroseq. Sensitivity and precision greatly varied across tools. E.g. Steak and ERVcaller had the highest average precision (0.9) while Retroseq+ had the lowest average precision (0.63). However, Steak had the lowest average sensitivity (0.13) while ERVcaller had the highest (0.77). Mobster did not detect any insertions in this experiment.

3.2 Analysis of the 50 short-read WGS samples

Each tool was applied to 50 SR-WGS samples and the results were merged. Table 3 shows the proportion of predicted HERV-K insertions that map to a known HERV locus. Tools with a higher rate of predicted insertions matching to documented previously reported loci are expected to have a higher accuracy as such insertions are more likely to be true positives. It is also important to consider the number of loci given by each tool as they may sacrifice sensitivity to increase the true positive rate. Total number of predictions and overlap with previously reported loci greatly varied across tools (Table 3 and Figure 4) but two of the HERV specific tools (Retroseq+ and ERVcaller) appear to have the highest proportion of predicted insertions that overlapped with previously reported ones. Notably, Steak gave the highest number of predictions and

TABLE 3 The “Known Polymorphic” column shows the proportion of predicted HERV-K insertion loci that matched to HERV-Ks from the literature reported to be polymorphic (Supplementary Table S1). “UCSC HERV-K” and “UCSC HERV” columns show the proportion which matched to hg19 reference HERV-Ks and HERVs given by the UCSC table browser (Supplementary File S2, Supplementary File S3). Total percent previously reported shows the proportion of predictions that are present in the polymorphic set or in the UCSC sets. “Total No. predictions” is the total number of predictions given across all 50 genomes. *There is an overlap between the Non-reference polymorphic HERV-Ks and the “UCSC HERV/LTR” set, this explains why the “Total % previously reported” column is less than a sum of these two columns for most tools. Top performing and lowest performing tools are highlighted in blue and red respectively.

	Known polymorphic HERV-Ks (S1)		UCSC HERV-K (S2)	UCSC HERV/LTR (S3)	Total % previously reported*	Total no. predictions
	Reference (%)	Non-reference (%)	Reference (%)	Reference (%)		
Retroseq	0	11	2	31	39%	2,286
Retroseq+	0	64	7	38	97.6%	296
Melt	0.6	26	0	34	52%	638
Steak	0.8	1.7	56	65	84%	13,770
ERVcaller	0	61	6	33	81%	439
Mobster	0	0	0	0	0%	0

84% of these results matched to previously documented HERV locations. 39% of Retroseq’s predictions and 52% of Melt’s predictions were previously reported. ERVcaller and Retroseq+ generated sets of predicted loci that greatly matched to previously reported ones (81% and 97.6% respectively) (Kahyo et al., 2017). Mobster was not able to detect any HERV-Ks in this sample. The proportion of HERV-Ks being present in introns, exons and intergenic regions was broadly consistent across tools and in line with results from previous studies (Supplementary Table S1). We also report the frequencies of HERV-K integrations for each tool (Supplementary Table S2).

The agreement between tools (Figure 5) greatly varied, ranging between 2.8% (proportion of Steak calls that were also called by Melt) and 63% (proportion of Retroseq+ calls that were also called by Steak). The number of insertions predicted ranged between 296 (Retroseq+) and 13,770 (Steak).

3.3 Analysis of matching short and long-read sequencing samples

We ran each tool on six SR-WGS samples and used the matching long-read sequencing data for validation (Table 4). Consistently with the other tests, the tools’ performance varied. Generally, the HERV specific tools outperformed the generalist tools in this test, though Retroseq had slightly higher proportions of confirmed calls than ERVcaller. Retroseq+ gave the smallest number of predictions, however, 78% of predicted loci were positive for an LTR5_Hs of length >850 bp in the corresponding long-read sample. We are particularly interested in the larger insertions as they would suggest a complete LTR (968 bases) that will most likely contain regions of biological importance such as the LTR promoters and enhancers. The great majority of the loci predicted by the tools were confirmed to contain ERV sequences in the long-read data. However, considering only the predicted insertions that correctly contained LTR5_Hs (the target HERV-K element), the performance of the tools varied greatly. For example, 78% of insertions called by Retroseq+ were LTR5_Hs, while only 13% of the Melt calls were LTR5_Hs. Most Retroseq+ calls (94%) were >850 bases while a substantial proportion of the loci identified by the other tools were smaller. Moreover, the number of predicted insertions also varied, ranging between 18 (Retroseq+) and 481 (ERVcaller). Notably, Steak identified a large number of long LTR5_Hs insertions but over two-thirds were reference loci and Steak showed a substantially higher precision for reference loci (>77%) than for non-reference loci (41%). Supplementary table 4 shows the proportion of predicted loci for which RepeatMasker reports either HERV-K internal gene sequence or an SVA of at least 50 bps in length. ERVcaller has the highest number of SVA positive loci (61%) while Retroseq+ and Steak have the highest proportion of HERV-K internal gene positive loci (11% and 12% respectively).

3.4 CPU usage and time

Finally, the tools were run on a single short-read whole-genome sequencing sample from Project MinE to quantify their memory, CPU and storage usage efficiency. Time, memory and space used were recorded (Table 5). This was achieved using the in-built Slurm HPC scheduling system. All of the tools had a relatively similar CPU time (mean = 3:59 CPU hours) and hard disk usage (mean = 2 GB) with the exception of ERVcaller which had a much higher CPU time (14:17 CPU hours) and used a lot

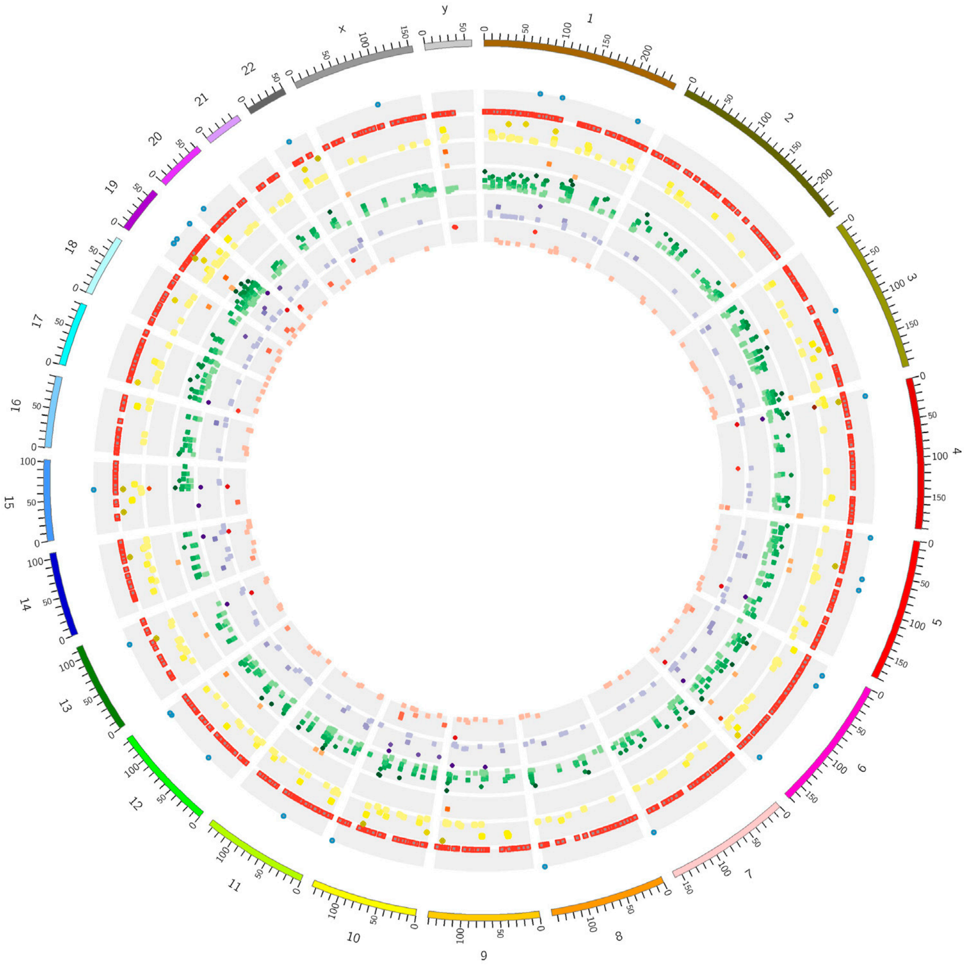

FIGURE 4
Overview of novel insertions predicted by the tools on the 50 genomes in a circular chromosomal plot. The order of concentric circles from the outside of the plot: circle 1 (blue dots)—known non-reference insertions; circle 2 (red dots)—known reference insertions; circle 3 (yellow): Retroseq predictions; circle 4 (orange): Retroseq plus predictions; circle 5 (green): Steak predictions; circle 5 (purple) ERVcaller predictions, circle 6 (red): Melt predictions. The intensity of the color and the height of each dot in its band is proportional to the number of subjects in whom the insertion is predicted with darker colors and higher position corresponding to a larger number.

more storage space (87 GB). This contrasts with the results of the original ERVcaller paper which showed that ERVcaller was faster than Retroseq and Melt. A key difference is that, in our test, ERVcaller was run on two CPUs but in the original paper it was run on 12. If the user has a large number of samples, or limited computational resources, ERVcaller may not be appropriate.

4 Discussion

This study compared the performance of six computational tools for detecting HERV loci in whole-genome sequencing data. Three of the tools we tested, ERVcaller, Steak and Retroseq+, were developed to identify exclusively HERVs, while the other three, Retroseq, Mobster and MELT, were designed to identify a broader range of TEs. Our results provided evidence of their highly variable performance across SR-NGS datasets, however, in all experiments HERV specialist tools generally performed better than generalist TE callers in calling HERVs.

The first test involved applying each tool to simulated WGS. In order to simulate potentially realistic (proviral) insertions, we first generated WGS samples using hg19 varying read length and coverage depth. Then we used a copy of hg19 in which we removed a set of known reference HERV loci, for read mapping of the simulated samples and HERV detection. Therefore, this experiment allowed us to assess how read length and coverage depth affect the tools' performance and to quantify the tools' precision and sensitivity (Table 3). As expected, the tools performed better on high quality WGS data (32X and 150 bp reads). Steak's sensitivity was lower than the other tools for detecting insertions in all simulated genomes

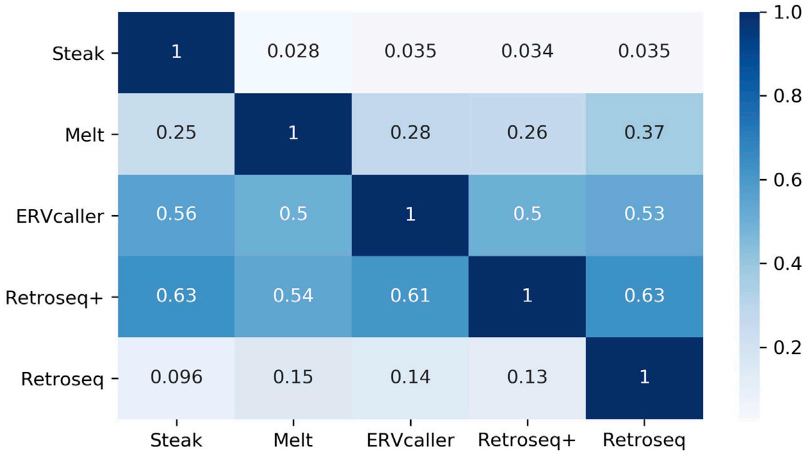

FIGURE 5
Heatmap reporting the proportion of insertions found by tools on the rows that were also found by the tools on the columns. E.g. the proportion in (Steak, ERVcaller) represents the proportion of Steak calls that were also called by ERVcaller. Therefore, please note that (Steak, ERVcaller) is not equal to (ERVcaller, Steak). Reference HERV-Ks have been filtered out.

TABLE 4 This table shows the proportion of results predicted by each tool in the short-read samples, that are positive for a HERV sequence in the associated long-read data. Each column shows the proportion of results that were positive for HERV-K (LTR5_Hs, the targeted HERV subgroup) or a general HERV sequence in the long-read contig data. The results are stratified by the length of the HERV sequence found in the long-read contig data. The Steak results are reported both including and discarding reference HERV loci. Reference HERV loci were removed for all other tools. Top performing and lowest performing tools are highlighted in blue and red respectively.

	LTR5_Hs > 850 bp	LTR5_Hs > 400 bp	LTR5_Hs all	ERV >850 bp	ERV >400 bp	ERV all	Total no. predictions
Retroseq	17%	21%	23%	52%	86%	96%	247
Retroseq+	78%	78%	78%	94%	100%	100%	18
Melt	10%	11%	13%	42%	79%	95%	412
Steak Non-reference	33%	37%	41%	57%	73%	92%	51
Steak ref + non-ref	74%	76%	77%	86%	92%	98%	172
ERVCaller	14%	15%	16%	34%	75%	92%	481
Mobster	NA	NA	NA	NA	NA	NA	NA

TABLE 5 This table shows how memory is used by each tool. The CPU time is equal to the number of CPUs * time. MAX VM size is the maximum virtual memory used at any one time by any part of the job. The input file size column reports the size of input sequencing data in the format required by each tool. The Max Temporary Files Size shows the maximum temporary storage required by each tool while running. For tools where there is an option to remove (clean up) temporary files, this option was not used. Lowest performing tool is highlighted in red.

	CPU time	Max vm size (GB)	Input file format/Size (GB)	Max temporary files size (GB)
Retroseq	03:43:01	1.15	BAM/77	2
Retroseq+	04:01:15	1.15	BAM/77	2
Melt	03:23:05	5.10	BAM/77	3
Steak	04:48:35	1.17	SAM/287	<1
ERVcaller	14:17:22	22.39	BAM/77	87

(≤20%). Although apparently surprising, this result is consistent with another independent evaluation of Steak (Chen and Li, 2019).

Following this, each tool was applied to 50 WGS samples of real individuals from an ALS cohort. This allowed us to quantify the agreement between tools and the proportion of results that matched to known HERV-K loci (Table 2). In this experiment Mobster could not identify any HERV insertions confirming its inability to detect this type of elements as was suggested by the authors in their original benchmarking analysis. The agreement between tools ranged between 3% and 63%, and the number of insertions predicted ranged between 296 (Retroseq+) and 13,770 (Steak). A part of this variability can be explained by the fact that Steak was designed to detect the presence of both reference and non-reference HERV insertions, however, the tools' accuracy might also contribute substantially. Indeed, although 65% of Steak's predictions matched reference HERV loci, only 1.7% overlapped with the highly characterized, non-reference, polymorphic loci. Looking at the proportion of insertions that matched the non-reference HERV-Ks previously reported in the literature can inform us about the quality of the predictions made by the tools. We have greater confidence that these known HERV-Ks are true compared to novel HERV-Ks which have not been previously reported or validated.

The tools were also tested on six publicly available genomes that had undergone both long and short-read sequencing (Table 4). Given the length of the long reads (>10 kbs), this dataset allowed us to confirm the insertions called in the short-read data using the long-reads. In this experiment the great majority (>92%) of all insertions predicted by the tools were confirmed HERVs in the long-read data. However, only Retroseq+ insertions were largely (78%) confirmed to be LTR5_Hs (the target HERV-K element), while the other tools showed a lower ability to distinguish between different HERV LTRs (13%–41%).

Finally, the tools were tested on a single WGS sample, and the time, memory and space used for temporary files were recorded. All of the tools had a relatively similar CPU time and hard disk usage with the exception of ERVcaller which had a much higher CPU time (14:17 CPU hours) and used a lot more storage space for temporary files (87 GB). This contrasts with the results of the original ERVcaller paper which showed that ERVcaller was faster than Retroseq and Melt (Chen and Li, 2019).

In conclusion, our analyses showed that tools and protocols developed specifically for the detection of HERV-Ks, such as ERVcaller, Retroseq+, and Steak, generally outperformed generalist tools such as Mobster and MELT. This trend is clearly visible in Supplementary Table S3 that reports an overview of key results across all benchmarking experiments. This finding is consistent with MELT documentation and supported by a recent paper from Niu et al. (Niu et al., 2021). Niu and colleagues found that HERV-K integrations detected by MELT had a 23% false discovery rate (FDR) when tested using PCR, which was a much higher FDR than the other transposable elements. HERV-K insertions in databases based on MELT, including the widely used GNOMAD-SV (Koch, 2020) and the newer HMEID database (Niu et al., 2021), are likely to be unreliable for use in HERV-K focused studies.

Moreover, the experiments highlighted important characteristics of the tools that the users should consider when designing their analysis pipeline: our implementation of the protocol developed by Wildschutte and colleagues (Retroseq+) produced the most reliable predictions but also the smallest number (296 predictions across 50 genomes, Table 2); Steak was the only tool able to comprehensively capture the presence of reference HERVs but its performance was substantially higher on reference HERVs than on non-reference HERVs; ERVcaller and Retroseq showed a good balance between number of detected insertions and their quality, however, their performance greatly varied across experiments. For example, they showed high precision and sensitivity in the simulated data (Table 3), but when applied to real data that is expected to include a large number of other types of insertions (the initial large SR-WGS dataset and the matching short and long-read data, Tables 2, 4), both of them showed high sensitivity but low specificity.

Given that all tools presented strengths and weaknesses, we recommend the users to base their choice on the requirements and objectives of the study and to consider combining multiple tools and a consensus approach if computationally feasible. For example, for rare genetic diseases in which both common polymorphisms and rare disruptive variants contribute to their genetics, such as ALS and other neurodegenerative disorders, one could combine the ability of Steak to call reference HERVs, with one of the other tools that showed a higher performance on non-reference insertions. Moreover, according to the availability of biological samples for wet-lab validation, one might choose a more conservative caller such as Retroseq+ or a more sensitive tool such as ERVcaller.

A limitation of this study is that it is focused on the detection of non-reference HERV insertions and it does not consider HERV annotation. HERV annotation could provide key pieces of information such as HERV family, subtype, location of promoter and enhancer regions, genotype, truncations and other polymorphisms, and whether they have potential for transcription. These are essential for their study and biological interpretation (Grandi et al., 2021; Jia et al., 2022). However, while this type of analysis can be performed for reference HERV loci, it is not possible for non-reference HERV detected in short-read NGS given that this technology does not allow for the characterization of the insertion sequence beyond the read-length.

In interpreting our results, it is important to note that our data may stem from the use of the hg19 reference genome. Results might be slightly different using hg38 as it includes more alternate sequences as well as corrections to sequencing artefacts (Schneider et al., 2017). However, the overarching challenge in calling HERVs remains, regardless of which reference is used, as short-read sequencing presents intrinsic limitations to capture large insertions. This challenge applies to most types of variants larger than some tens of base pairs and consensus approaches have shown potential, e.g. Gnomad SV (Koch, 2020). Long-read sequencing can provide a better solution to the detection of large insertions and its use is on the rise, analyzing short-read sequencing data for large variants is still highly relevant given the great availability of this type of data and its higher per base sequencing resolution.

Data availability statement

The original contributions presented in the study are included in the article and in the Supplementary Material, further inquiries can be directed to the corresponding author. Supplementary materials, including all supplementary tables, figures and the scripts to run the analyses, are available on GitHub: https://github.com/KHP-Informatics/tools_assessment_hervk_SR-WGS.

Author contributions

Conceptualization, AI, CS, JQ, AJ and AAC; methodology, AI, HB, RK; software, AI, HB, RK and RD; validation, AI, HB and RK; formal analysis, AI, HB and RK; investigation, AI, HB, RK and CS; resources, AI and RD; data curation, AI, AAC, AK, AJ, and RD; writing—original draft preparation, AI, HB, RK; writing—review and editing, CS, AI, JQ; visualization, HB and RK; supervision, AI, AAC, CMS, JQ; project administration, AAC, AI; funding acquisition, AI, AAC. All authors have read and agreed to the published version of the manuscript.

Funding

United Kingdom Research and Innovation; Medical Research Council; South London and Maudsley NHS Foundation Trust; MND Scotland; Motor Neurone Disease Association; Rosetrees Trust; National Institute for Health Research; China Scholarship Council; Spastic Paraplegia Foundation. Funding for open access charge: UKRI. AI is funded by the Motor Neurone Disease Association, MND Scotland, Spastic Paraplegia Foundation, Darby Rimmer MND Foundation, Rosetrees Trust and the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RK is funded by the MND Scotland. This is an EU Joint Programme-Neurodegenerative Disease Research (JPND) project. The project is supported through the following funding organizations under the aegis of JPND-https://www.neurodegenerationresearch.eu/ [United Kingdom, Medical Research Council MR/L501529/1 to AC, principal investigator (PI) and MR/R024804/1 to AC, (PI); Economic and Social Research Council ES/L008238/1 to AC (co-PI)] and through the Motor Neurone Disease Association. This study represents independent research partly funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The work leading up to this publication was funded by the European Community's Horizon 2020 Programme (H2020-PHC-2014-two-stage; grant 633413). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, King's College London, or the Department of Health and Social Care. AK is funded by ALS Association Milton Safenowitz Research Fellowship, The Motor Neurone Disease Association (MNDA) Fellowship, The Darby Rimmer Foundation and The NIHR Maudsley Biomedical Research Centre. This research was supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London.

Acknowledgments

We acknowledge use of the research computing facility at King's College London, Rosalind (https://rosalind.kcl.ac.uk), which is delivered in partnership with the National Institute for Health Research (NIHR) Biomedical Research Centres at South London & Maudsley and Guy's and St. Thomas' NHS Foundation Trusts and part-funded by capital equipment grants from the Maudsley Charity (award 980) and Guy's and St Thomas' Charity (TR130505). We also want to thank all people with MND and their families for contributing to the studies that generated the whole-genome sequencing data.

Conflict of interest

AC is the Principal Investigator of the Lighthouse 2 trial of Triumeq in ALS.

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

References

Belshaw, R., Dawson, A. L. A., Woolven-Allen, J., Redding, J., Burt, A., and Tristem, M. (2005). Genome wide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): Implications for present-day activity. *J. Virology* 79 (19), 12507–12514. doi:10.1128/jvi.79.19.12507-12514.2005

Boller, K., Schönfeld, K., Lischer, S., Fischer, N., Hoffmann, A., Kurth, R., et al. (2008). Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles. *J. General Virology* 89 (2), 567–572. doi:10.1099/vir.0.83534-0

Borges-Monroy, R., Chu, C., Dias, C., Choi, J., Lee, S., Gao, Y., et al. (2021). Whole-genome analysis reveals the contribution of non-coding de novo transposon insertions to autism spectrum disorder. *Mob. DNA* 12 (1), 28–15. doi:10.1186/s13100-021-00256-w

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., et al. (2018). Ten things you should know about transposable elements. *Genome Biol.* 19 (1), 199–212. doi:10.1186/s13059-018-1577-z

Buzdin, A. A., Lebedev, YuB., and Sverdlov, E. D. (2003). Human genome-specific HERV-K intron LTR genes have a random orientation relative to the direction of transcription, and, possibly, participated in antisense gene expression regulation. *Russ. J. Bioorg. Chem.* 29 (1), 103–106. doi:10.1023/a:1022294906202

Chen, X., and Li, D. (2019). ERVcaller: Identifying polymorphic endogenous retrovirus and other transposable element insertions using whole-genome sequencing data. *Bioinformatics* 35 (20), 3913–3922. doi:10.1093/bioinformatics/btz205

Chiappinelli, K. B., Strissel, P. L., Desrichard, A., Li, H., Henke, C., Akman, B., et al. (2015). Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. *Cell* 162 (5), 974–986. doi:10.1016/j.cell.2015.07.011

Chu, C., Lee, S., Borges-Monroy, R., Viswanadham, V. V., Li, H., Lee, E. A., et al. (2021). Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat. Commun.* 12, 3836–3912. doi:10.1038/s41467-021-24041-8

Clayton, E. A., Wang, L., Rishishwar, L., Wang, J., McDonald, J. F., and Jordan, I. K. (2016). Patterns of transposable element expression and insertion in cancer. *Front. Mol. Biosci.* 3, 76. doi:10.3389/fmolb.2016.00076

Dennenmoser, S., Sedlazeck, F. J., Schatz, M. C., Altmüller, J., Zytnicki, M., and Nolte, A. W. (2019). Genome-wide patterns of transposon proliferation in an evolutionary young hybrid fish. *Mol. Ecol.* 28 (6), 1491–1505. doi:10.1111/mec.14969

Dervan, E., Bhattacharyya, D. D., McAuliffe, J. D., Khan, F. H., and Glynn, S. A. (2021). Ancient adversary–HERV-K (HML-2) in cancer. *Front. Oncol.* 11, 658489. doi:10.3389/fonc.2021.658489

Dolei, A., Ibba, G., Piu, C., and Serra, C. (2019). Expression of HERV genes as possible biomarker and target in neurodegenerative diseases. *Int. J. Mol. Sci.* 20 (15), 3706. doi:10.3390/ijms20153706

Ewing, A. D. (2015). Transposable element detection from whole genome sequence data. *Mob. DNA* 6 (1), 24–29. doi:10.1186/s13100-015-0055-3

Feusier, J., Watkins, W. S., Thomas, J., Farrell, A., Witherspoon, D. J., Baird, L., et al. (2019). Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res.* 29 (10), 1567–1577. doi:10.1101/gr.247965.118

Garcia-Montojo, M., Doucet-O'Hare, T., Henderson, L., and Nath, A. (2018). Human endogenous retrovirus-K (HML-2): A comprehensive review. *Crit. Rev. Microbiol.* 44 (6), 715–738. doi:10.1080/1040841x.2018.1501345

Garcia-Montojo, M., Rodriguez-Martin, E., Ramos-Mozo, P., Ortega-Madueño, I., Dominguez-Mozo, M. I., Arias-Leal, A., et al. (2020). Syncytin-1/HERV-W envelope is an early activation marker of leukocytes and is upregulated in multiple sclerosis patients. *Eur. J. Immunol.* 50 (5), 685–694. doi:10.1002/eji.201948423

Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Pittard, W. S., et al. (2017). The mobile element locator tool (MELT): Population-scale mobile element discovery and biology. *Genome Res.* 27 (11), 1916–1929. doi:10.1101/gr.218032.116

Gianfrancesco, O., Geary, B., Savage, A. L., Billingsley, K. J., Bubb, V. J., and Quinn, J. P. (2019). The role of SINE-VNTR-alu (SVA) retrotransposons in shaping the human genome. *Int. J. Mol. Sci.* 20 (23), 5977. doi:10.3390/ijms20235977

Gifford, R., and Tristem, M. (2003). The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes* 26 (3), 291–315. doi:10.1023/a:1024455415443

Grandi, N., Pisano, M. P., Pessiu, E., Scognamiglio, S., and Tramontano, E. (2021). HERV-K (HML7) integrations in the human genome: Comprehensive characterization and comparative analysis in non-human primates. *Biology* 10 (5), 439. doi:10.3390/biology10050439

Groza, C., Chen, X., Pacis, A., Simon, M. M., Pramatarova, A., Aracena, K. A., et al. (2022). Genome graphs detect human polymorphisms in active epigenomic state during influenza infection. bioRxiv. 2021–2109.

Hancks, D. C., and Kazazian, H. H. (2010). SVA retrotransposons: Evolution and genetic instability. *Seminars Cancer Biol.* 20 (4), 234–245. doi:10.1016/j.semcancer.2010.04.001

Homer, N. (2010). Dwgsim: Whole genome simulator for next-generation sequencing. Version 0.1.13. Available at: https://github.com/nh13/DWGSIM.

Iacoangeli, A., Al Khleifat, A., Jones, A. R., Sproviero, W., Shatunov, A., Opie-Martin, S., et al. (2019). C9orf72 intermediate expansions of 24–30 repeats are associated with ALS. *Acta Neuropathol. Commun.* 7 (1), 115–117. doi:10.1186/s40478-019-0724-4

Iacoangeli, A., Al Khleifat, A., Sproviero, W., Shatunov, A., Jones, A. R., Morgan, S. L., et al. (2019). DNAscan: Personal computer compatible NGS analysis, annotation and visualisation. *BMC Bioinforma.* 20 (1), 213–310. doi:10.1186/s12859-019-2791-8

Jia, L., Liu, M., Yang, C., Li, H., Liu, Y., Han, J., et al. (2022). Comprehensive identification and characterization of the HERV-K (HML-9) group in the human genome. *Retrovirology* 19 (1), 11–18. doi:10.1186/s12977-022-00596-2

Jones, A. R., Iacoangeli, A., Adey, B. N., Bowles, H., Shatunov, A., Troakes, C., et al. (2021). A HML6 endogenous retrovirus on chromosome 3 is upregulated in amyotrophic lateral sclerosis motor cortex. *Sci. Rep.* 11 (1), 14283–14310. doi:10.1038/s41598-021-93742-3

Kabiljo, R., Bowles, H., Marriott, H., Jones, A. R., Bouton, C. R., Dobson, R. J., et al. (2022). RetroSnake: A modular pipeline to detect human endogenous retroviruses in genome sequencing data. *Iscience* 25 (11), 105289. doi:10.1016/j.isci.2022.105289

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 (1), 487. doi:10.1186/s12864-017-3872-6

Katzourakis, A., Pereira, V., and Tristem, M. (2007). Effects of recombination rate on human endogenous retrovirus fixation and persistence. *J. virology* 81 (19), 10712–10717. doi:10.1128/jvi.00410-07

Keane, T. M., Wong, K., and Adams, D. J. (2013). RetroSeq: Transposable element discovery from next-generation sequencing data. *Bioinformatics* 29 (3), 389–390. doi:10.1093/bioinformatics/bts697

Klaver, B., and Berkhout, B. (1994). Comparison of 5'and 3'long terminal repeat promoter function in human immunodeficiency virus. *J. Virology* 68 (6), 3830–3840. doi:10.1128/jvi.68.6.3830-3840.1994

Koch, L. (2020). Exploring human genomic diversity with gnomAD. *Nat. Rev. Genet.* 21 (8), 448. doi:10.1038/s41576-020-0255-7

Larsen, P. A., Hunnicutt, K. E., Larsen, R. J., Yoder, A. D., and Saunders, A. M. (2018). Warning SINEs: Alu elements, evolution of the human brain, and the spectrum of neurological disease. *Chromosome Res.* 26 (1), 93–111. doi:10.1007/s10577-018-9573-4

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997.

Li, W., Lee, M-H., Henderson, L., Tyagi, R., Bachani, M., Steiner, J., et al. (2015). Human endogenous retrovirus-K contributes to motor neuron disease. *Sci. Transl. Med.* 7, 307ra153. [Internet]. doi:10.1126/scitranslmed.aac8201

Niu, Y., Teng, X., Shi, Y., Li, Y., Tang, Y., and Zhang, P.1. (2021). Genome-wide analysis of mobile element insertions in human genomes. bioRxiv.

Payer, L. M., and Burns, K. H. (2019). Transposable elements in human genetic disease. *Nat. Rev. Genet.* 20 (12), 760–772. doi:10.1038/s41576-019-0165-8

Pisano, M. P., Grandi, N., Cadeddu, M., Blomberg, J., and Tramontano, E. (2019). Comprehensive characterization of the human endogenous retrovirus HERV-K(HML-6) group: Overview of structure, phylogeny, and contribution to the human genome. *J. Virology* 93 (16), e00110. doi:10.1128/jvi.00110-19

Project MinE ALS Sequencing Consortium (2018). Project MinE: Study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *Eur. J. Hum. Genet.* 26 (10), 1537–1546. doi:10.1038/s41431-018-0177-4

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26 (6), 841–842. doi:10.1093/bioinformatics/btq033

Rezaei, S. D., Hayward, J. A., Norden, S., Pedersen, J., Mills, J., Hearps, A. C., et al. (2021). HERV-K gag RNA and protein levels are elevated in malignant regions of the prostate in males with prostate cancer. *Viruses* 13 (3), 449. doi:10.3390/v13030449

Rishishwar, L., Wang, L., Wang, J., Soojin, V. Y., Lachance, J., and Jordan, I. K. (2018). Evidence for positive selection on recent human transposable element insertions. *Gene* 675, 69–79. doi:10.1016/j.gene.2018.06.077

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* 17 (2), 155–158. doi:10.1038/s41592-019-0669-3

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 Evol.* 3 (2), vex023. doi:10.1093/ve/vex023

Santoni, F. A., Guerra, J., and Luban, J. (2012). HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* 9 (1), 111–115. doi:10.1186/1742-4690-9-111

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., Kitts, P. A., et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* 27 (5), 849–864. doi:10.1101/gr.213611.116

Storer, J., Hubley, R., Rosen, J., Wheeler, T. J., and Smit, A. F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* 12 (1), 2–14. doi:10.1186/s13100-020-00230-y

Subramanian, R. P., Wildschutte, J. H., Russo, C., and Coffin, J. M. (2011). Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology* 8 (1), 90. doi:10.1186/1742-4690-8-90

Teissandier, A., Servant, N., Barillot, E., and Bourc'his, D. (2019). Tools and best practices for retrotransposon analysis using high-throughput sequencing data. *Mob. DNA* 10 (1), 52–12. doi:10.1186/s13100-019-0192-1

Tempel, S. (2012). Using and understanding RepeatMasker. *Methods Mol. Biol.* 859, 29. doi:10.1007/978-1-61779-603-6_2

Thung, D. T., de Ligt, J., Vissers, L. E., Steehouwer, M., Kroon, M., de Vries, P., et al. (2014). Mobster: Accurate detection of mobile element insertions in next generation sequencing data. *Genome Biol.* 15 (10), 488. doi:10.1186/s13059-014-0488-x

Troskie, R-L., Jafrani, Y., Mercer, T. R., Ewing, A. D., Faulkner, G. J., and Cheetham, S. W. (2021). Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome. *Genome Biol.* 221, 146–215. doi:10.1186/s13059-021-02369-0

Wang, Y-C., Olson, N. D., Deikus, G., Shah, H., Wenger, A. M., Trow, J., et al. (2019). High-coverage, long-read sequencing of Han Chinese trio reference samples. *Sci. Data* 6 (1), 91. doi:10.1038/s41597-019-0098-2

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. *Proc. Natl. Acad. Sci.* 113 (16), E2326–E2334. doi:10.1073/pnas.1602336113

Xue, B., Sechi, L. A., and Kelvin, D. J. (2020). Human endogenous retrovirus K (HML-2) in health and disease. *Front. Microbiol.* 11, 1690. doi:10.3389/fmicb.2020.01690

Xue, B., Zeng, T., Jia, L., Yang, D., Lin, S. L., Sechi, L. A., et al. (2020). Identification of the distribution of human endogenous retroviruses K (HML-2) by PCR-based target enrichment sequencing. *Retrovirology* 17 (1), 10–15. doi:10.1186/s12977-020-00519-z

Zhang, Y., Li, T., Preissl, S., Amaral, M. L., Grinstein, J. D., Farah, E. N., et al. (2019). Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat. Genet.* 51 (9), 1380–1388. doi:10.1038/s41588-019-0479-7