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

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

[#]Correspondence should be addressed to alfredo.iacoangeli@kcl.ac.uk

1. Department of Basic and Clinical Neuroscience, Maurice Wohl Clinical Neuroscience Institute, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, SE5 9NU, UK
2. Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology & Neuroscience, King's College London, SE5 8AF, London, UK
3. Department of Pharmacology and Therapeutics, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 3BX, UK
4. NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, London, UK
5. Institute of Health Informatics, University College London, London, UK
6. NIHR Biomedical Research Centre at University College London Hospitals NHS Foundation Trust, London, UK
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, UK

Abstract

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, and 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 currently available tools is required. We evaluated the performance of a set of such tools using a variety of experimental designs and types of NGS datasets. These included 50 human short read whole-genome sequencing samples, matching long and short read NGS data, and simulated short-read NGS data. Our results highlight the performance variability of the tools across the datasets and suggest that different tools might be suitable for different study designs. Using multiple tools and a consensus approach is advisable if computationally feasible and wet-lab validation via PCR is advisable where biological samples are available.

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 (1). Human endogenous retroviruses (HERVs) comprise ~8% of the genome, whereas protein coding genes comprise only 1-2% (2). 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 (3). The HML-2 HERV-K subgroup includes some of the most recent HERV integrations, which are found as full length (or near full length) insertions in over 80 different loci (4). Though there are several full-length copies of HERV-K in the genome, none are likely to produce an infectious virus (5). HERV-K sequences can be full length proviruses, solo long terminal repeats or 2-LTR sequences and are polymorphic in the human population (6). A full length HERV-K 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 in the human genome, only the LTRs are present which contain the promoter and enhancer regions (7) (figure 1).

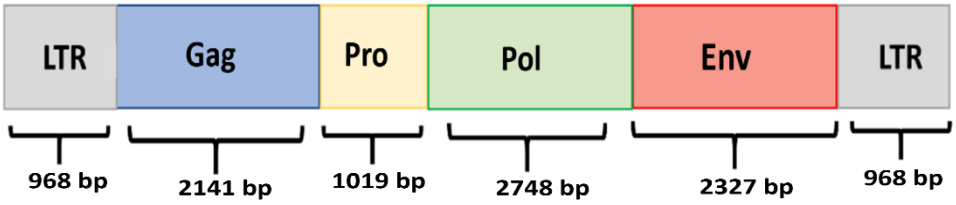

Figure 1) This schematic shows the general structure of a full length HERV-K with approximate base pair lengths for each section. The LTR length is the LTR5_Hs length given in the DFAM database (accession = DF0000558), the internal gene lengths comes from HERVK11 internal regions, again from DFAM (accession = DF0000189) (8). 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.

HERVs, and related transposable elements, have been linked to a wide range of diseases including cancer and neurodegeneration via multiple mechanisms. For example, their insertion into the human genome may alter gene expression or disrupt reading frames (9); they were reported to be upregulated in biological samples from people affected by neurodegenerative diseases and cancer (10,11,12) . Furthermore, their expression may be toxic for certain cell types such as motor neurons (13). However, characterising the HERV genomic landscape is challenging. This is due to several factors including some characteristics of their sequence which make them hard to map correctly in the human genome. For example, HERVs are thousands of bases long, they are polymorphic in humans and present a high degree of sequence similarity. As a consequence, despite the substantial body of evidence supporting their potential key role in many diseases, they remain

substantially under-investigated and the nature of their involvement with human disease is not clear.

Recent advances in next-generation sequencing (NGS) have made sequencing DNA molecules a common practice in genetic research, allowing for the investigation of a wide range of variants that spans from single nucleotide variants to large structural variants (14). This technology has also been used to study HERVs in humans (15). Currently, a number of bioinformatic tools for the identification of HERV loci in NGS data exist, mostly based on the exploitation of split and discordant reads to reveal the presence of potential HERV insertions (Figure 2). Given the growing interest in this class of elements, an independent evaluation of the current tools for HERV detection in greatly needed for the design of optimal analyses 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 short-read NGS data (short reads = 50 - 200bp). 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 samples), a simulated SR-WGS dataset, and six SR-WGS samples for which matching long-read sequencing data was available.

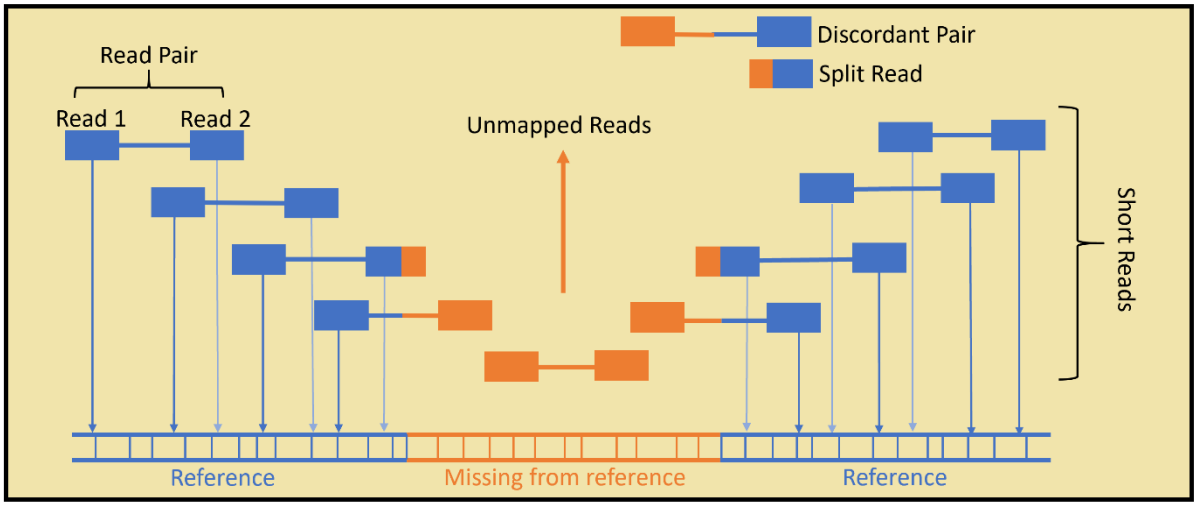

Figure 2) In standard Illumina paired end SR-WGS, the DNA sample is first sheared into small fragments. Then, both ends of the fragment are sequenced, giving paired reads. Read mapping is achieved by aligning the reads to a reference genome. HERVs that are present in the sample, but not the reference,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

Methods:

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 (16). According to its documentation, MELT was not tested for HERV detection, 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 (19,20,22).
Mobster: This tool also uses discordant reads alongside split reads and an insertional reference sequence to predict specific insertion sites. When Mobster was released, the authors reported that it was not able to identify HERV insertions. However, they tested it using 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 (17).
Retroseq: Retroseq uses discordant read pairs to identify putative insertions sites and filters for read pairs which align to a reference of interest (18).
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. These mapped fragments 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 (19).
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 (20).
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 a automatic bioinformatics pipeline, we have implemented it ourselves, following the authors' description (21).

Benchmarking experiments

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

Overlap analysis and comparison with previously reported HERVs

Each tool was applied to WGS data of 50 amyotrophic lateral sclerosis (ALS) patients from the British Project Mine dataset (23,24). This WGS was generated by sequencing DNA samples from blood, using the Illumina Hiseq 2000 platform. The resulting WGS samples had read length equal to 100bp with an average coverage depth of 40X (paired end reads). We aligned them to the hg19 reference genome using Burrows-Wheeler alignment, BWA-MEM (25). The predicted insertion sites across all genomes were compared to a list of 40 well characterised polymorphic HERV-K insertions previously described in the literature [S1] (26). They were also compared to all reference HERV loci (both HERV-Ks and wider 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 [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' [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 300 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 Figure 3.

The results of each tool were also compared to one another, giving for each tool, the proportion of its results that were also predicted by each of the other tools. This allowed us to quantify the agreement across 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 [S2].

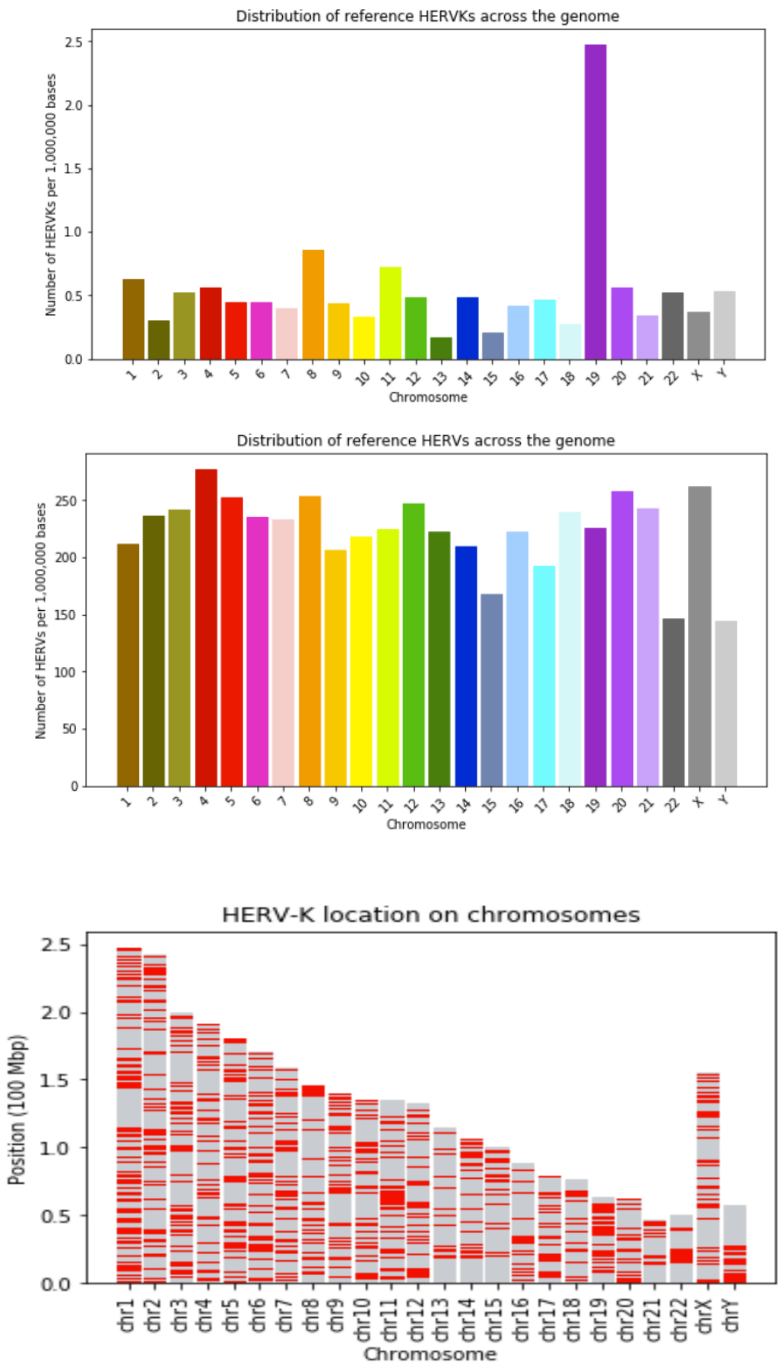

Figure 3) Upper panel - Number of HERV-Ks per million bases in each human chromosome (as given by the UCSC RMSK table). The high HERV density in chr19 has been previously reported (27) and other transposable elements are also enriched on this chromosome (28). Middle panel - Number of HERVs and LTR sequences per million bases in each human chromosome – Data is from the UCSC RMSK table. Lower panel – 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.

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 (29). Briefly, the short-read data (derived from blood) were sequenced using Illumina Hiseq 2500 giving mean 105bp paired end reads with coverage depth ranging between 15.6X and 18.8X. The long reads were sequenced using PacificBio Seqel system version 2. For these samples the read lengths were between 10KB and 18KB 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 (30,31,32). 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 (33) and RepeatMasker (34) was applied to the contigs to confirm the presence of a HERV-K LTR sequence at this site. For Melt and Steak, reference HERV-Ks were filtered out.

Simulated SR-WGS analysis

For the simulation test, short-read paired-end Illumina WGS data were simulated from the hg19 reference sequence using DWGSIM (parameters in Table 1) (35).

Hg19 is reported to have 66 HERV-K (HML-6) proviral loci (2) from which we selected 19. To simulate these 19 HERV-K sites as novel insertions, after generating the WGS data, we removed the 19 sequences from hg19 using Bedtools masking followed by deletion of the mask (36). The generated 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 19 insertions were the only non-reference elements in the simulated samples. Each tool was then applied to the simulated WGS, and the 19 sites were used as our set of true positive insertions. 15 insertions were type LTR3A and 4 were type LTR3B. The LTR3A sequence was used as target reference template. This allows 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 positive/(True positive + False negative). We defined precision as the proportion of positive results that are true: True positive/(True positive + False positive).

	Read length	No. reads	Coverage	Outer distance	-e (error)	-s (std-dev outer dist.)
Genome 1	150bp	333333333	32X	400	0.020	5
Genome 2	100bp	500000000	32X	400	0.020	5
Genome 3	150bp	105000000	10.5X	400	0.020	5
Genome 4	100bp	105000000	7X	400	0.020	5

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

Computational efficiency report

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 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.

Results:

Analysis of the 50 short-read WGS samples

Each tool was applied to 50 SR-WGS samples and the results were merged. Table 2 shows the proportion of predicted HERV-K loci that map to a known HERV loci. Total number of predictions and overlap with previously reported loci greatly varied across tools (Table 2 and Figure 4). Results varied greatly across tools. Notably, Steak gave the highest number of predictions and 84% of these results matched to previously documented HERV locations. 97.6% of the results given by Retroseq+ matched to a documented HERV loci though this tool gave the smallest number of predictions. Mobster was not able to detect any HERV-Ks in this sample.

	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%	2286
Retroseq+	0%	64%	7%	38%	97.6%	296
Melt	0.6%	26%	0%	34%	52%	638
Steak	0.8%	1.7%	56%	65%	84%	13770
Ervcaller	0%	61%	6%	33%	81%	439
Mobster	0%	0%	0%	0%	0%	0

Table 2) 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 [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 [S2-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.

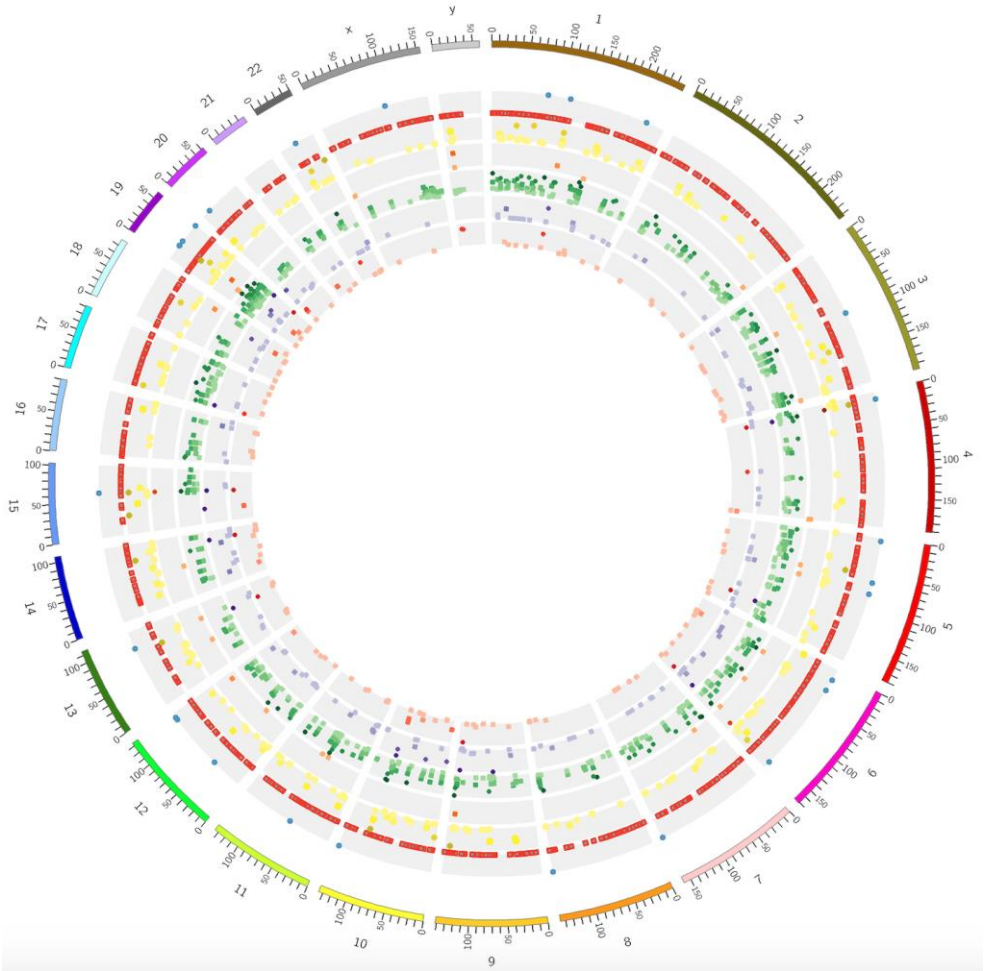

Figure 4) Overview of novel insertions predicted by the tools on the 50 genomes in a circular chromosomal plot. The order of concentric circles from the outside of the plot: circle 1 (blue dots) – known non reference insertions; circle 2 (red dots) – known reference insertions; circle 3 (yellow): Retroseq 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 colour and the height of each dot in its band is proportional to the number of subjects in whom the insertion is predicted with darker colours and higher position corresponding to a larger number.

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).

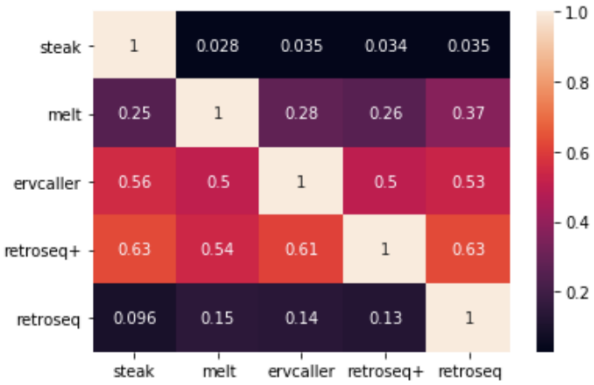

Figure 5) Heatmap reporting the proportion of insertions found by tools on the rows that were also found by the tools on the columns. E.g. the proportion in [Steak, Ervcaller] represents the proportion of Steak calls that were also called by Ervcaller. Therefore, please note that [Steak, Ervcaller] is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

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. All tools performed worse on samples with lower read depth. For example, Retroseq detected 11/15 LTR3A insertions in a 32X sample but only 7/15 LTR3A insertions in a 10.5X and 7X sample. However, the degree to which read depth affects each tool varied. ERVcaller performed equally well on 32X and 10.5X sample finding 11 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 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).

	32X										10.5X					7X				
	G1 (150bp)					G2 (100bp)					G3 (150 bp)					G4 (100bp)				
	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

Table 3) 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), the number of LTR3B insertions mistakenly classified as LTR3A insertions, and the total number of predicted insertions (Total column) for each sample.

Analysis of matching short and long-read sequencing samples

We used 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. Retroseq+ gave the smallest number of predictions; however, 78% of predicted loci were positive for an LTR5_Hs of length >850bp 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. All Retroseq+ calls 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.

	LTR5_Hs >850bp	LTR5_Hs >400bp	LTR5_Hs all	ERV >850bp	ERV >400bp	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 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 HERV-K (LTR5_Hs, the targeted HERV subgroup) or a general HERV sequence and the length of the sequence in the long-read data. The Steak results are reported both including and discarding reference HERV loci. Reference HERV loci were removed for all other tools.

CPU usage and time

Table 5 shows the CPU Time, Max virtual memory and input file type/size for each tool. Melt had a very low CPU time of 3 hours and 23 minutes; Ervcaller had the highest CPU time of 14 hours and 17 minutes. Mobster usage was not reported as it did not produce any HERV results. Though we used a SAM file with Steak, it is possible to pipe a BAM input into Steak directly. All tools produced intermediate file sizes of a negligible size (<1GB -3 GB) except for Ervcaller which produced intermediate files whose sizes summed to 87 GB. If memory limitations are a concern, tools with minimal CPU time, input file sizes and VM requirements are desirable.

	CPU Time	Max VM Size	Input File Format/Size	Max Temporary Files Size
Retroseq	03:43:01	1.15G	BAM / 77G	2GB
Retroseq+	04:01:15	1.15G	BAM / 77G	2GB
Melt	03:23:05	5.10G	BAM / 77G	3GB
Steak	04:48:35	1.17G	SAM / 287G	<1GB
Ervcaller	14:17:22	22.39G	BAM / 77G	87GB

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.

Discussion and conclusions

This study compared the performance of six computational tools for detecting HERV loci in whole genome sequencing data. Overall, our results provided evidence of their highly variable performance across SR-NGS datasets.

The first test involved applying each tool 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. 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 characterised, 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, as we would expect these to be more likely to be true insertions with respect the novel ones. 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 97.4% respectively, Table 2).

The tools were then applied 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. As expected, the tools performed better on high quality WGS data (32X and 150bp reads). 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 150bp reads).

Finally, the tools were tested on six publicly available genomes that had undergone both long and short read sequencing. Given the length of the long reads (>10Kbs), 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%). Notably, Steak showed a substantially higher precision for reference loci (>77%) than for non-reference loci (41%)

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. 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); 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, 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), 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.

Some of the variability in calling HERV-K insertion sites in our data may stem from the use of the hg19 reference genome. We expect results to differ if hg38 had been used, as it includes more alternate sequences as well as corrections to sequencing artefacts (37). However, the overarching challenge in calling HERVs remains, regardless of reference, 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 a consensus approach has shown potential, e.g. Gnomad SV (38).
Although long read sequencing can provide a better solution to the detection of large insertions and its use is on the rise, analysing short read sequencing data for large variants is

still highly relevant given the great availability of this type of data and its higher sequencing resolution that can allow for a more accurate characterization of variant sequences.

Supplementary:
All of the supplementary data and scripts used for this paper can be found at the following GitHub repository: tools_assessment_hervk_SR-WGS.

References

1. Gifford R, Tristem M. The Evolution, Distribution and Diversity of Endogenous Retroviruses. Virus Genes. 2003;26(3).
2. Pisano MP, Grandi N, Cadeddu M, Blomberg J, Tramontano E. Comprehensive Characterization of the Human Endogenous Retrovirus HERV-K(HML-6) Group: Overview of Structure, Phylogeny, and Contribution to the Human Genome. Journal of Virology. 2019 Jun 5;93(16).
3. Belshaw R, Dawson ALA, Woolven-Allen J, Redding J, Burt A, Tristem M. Genome wide Screening Reveals High Levels of Insertional Polymorphism in the Human Endogenous Retrovirus Family HERV-K(HML2): Implications for Present-Day Activity. Journal of Virology. 2005 Oct 1;79(19).
4. Subramanian RP, Wildschutte JH, Russo C, Coffin JM. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology. 2011;8(1).
5. Boller K, Schönfeld K, Lischer S, Fischer N, Hoffmann A, Kurth R, et al. Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles. Journal of General Virology. 2008 Feb 1;89(2).
6. Garcia-Montojo, M., Doucet-O'Hare, T., Henderson, L. and Nath, A., 2018. Human endogenous retrovirus-K (HML-2): a comprehensive review. Critical reviews in microbiology, 44(6), pp.715-738.
7. Klaver, B. and Berkhout, B., 1994. Comparison of 5'and 3'long terminal repeat promoter function in human immunodeficiency virus. Journal of virology, 68(6), pp.3830-3840.
8. 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. Mobile DNA, 12(1), pp.1-14.
9. Buzdin AA, Lebedev YuB, Sverdlov ED. 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. Russian Journal of Bioorganic Chemistry. 2003;29(1).
10. Jones, A.R., Iacoangeli, A., Adey, B.N., Bowles, H., Shatunov, A., Troakes, C., Garson, J.A., McCormick, A.L. and Al-Chalabi, A., 2021. A HML6 endogenous retrovirus on chromosome 3 is upregulated in amyotrophic lateral sclerosis motor cortex. Scientific Reports, 11(1), pp.1-10.
11. Dervan, E., Bhattacharyya, D.D., McAuliffe, J.D., Khan, F.H. and Glynn, S.A., 2021. Ancient Adversary–HERV-K (HML-2) in Cancer. Frontiers in Oncology, 11, p.1707.
12. Garcia-Montojo, M., Rodriguez-Martin, E., Ramos-Mozo, P., Ortega-Madueño, I., Dominguez-Mozo, M.I., Arias-Leal, A., García-Martínez, M.Á., Casanova, I., Galan, V., Arroyo, R. and Álvarez-Lafuente, R., 2020. Syncytin-1/HERV-W envelope is an early activation marker of leukocytes and is upregulated in multiple sclerosis patients. European journal of immunology, 50(5), pp.685-694.

13. Li W, Lee M-H, Henderson L, Tyagi R, Bachani M, Steiner J, et al. Human endogenous retrovirus-K contributes to motor neuron disease. Science Translational Medicine [Internet]. 2015 Sep 30 [cited 2019 Mar 19];7(307):307ra153-307ra153. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26424568
14. Iacoangeli, A., Al Khleifat, A., Sproviero, W., Shatunov, A., Jones, A.R., Morgan, S.L., Pittman, A., Dobson, R.J., Newhouse, S.J. and Al-Chalabi, A., 2019. DNAscan: personal computer compatible NGS analysis, annotation and visualisation. BMC bioinformatics, 20(1), pp.1-10.
15. Xue, B., Sechi, L.A. and Kelvin, D.J., 2020. Human endogenous retrovirus K (HML-2) in health and disease. Frontiers in Microbiology, 11, p.1690.
16. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Research. 2017 Nov;27(11).
17. Thung DT, de Ligt J, Vissers LE, Steehouwer M, Kroon M, de Vries P, et al. Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biology. 2014 Oct 28;15(10).
18. Keane TM, Wong K, Adams DJ. RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics. 2013 Feb 1;29(3).
19. Santander CG, Gambron P, Marchi E, Karamitros T, Katzourakis A, Magiorkinis G. STEAK: A specific tool for transposable elements and retrovirus detection in high-throughput sequencing data. Virus Evolution. 2017 Jul;3(2).
20. Chen X, Li D. ERVcaller: identifying polymorphic endogenous retrovirus and other transposable element insertions using whole-genome sequencing data. Bioinformatics. 2019 Oct 15;35(20).
21. Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proceedings of the National Academy of Sciences. 2016 Apr 19;113(16).
22. Feusier, J., Watkins, W.S., Thomas, J., Farrell, A., Witherspoon, D.J., Baird, L., Ha, H., Xing, J. and Jorde, L.B., 2019. Pedigree-based estimation of human mobile element retrotransposition rates. Genome research, 29(10), pp.1567-1577.
23. 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. European Journal of Human Genetics, 26(10), p.1537.
24. Iacoangeli, A., Al Khleifat, A., Jones, A.R., Sproviero, W., Shatunov, A., Opie-Martin, S., Morrison, K.E., Shaw, P.J., Shaw, C.E., Fogh, I. and Dobson, R.J., 2019. C9orf72 intermediate expansions of 24–30 repeats are associated with ALS. Acta neuropathologica communications, 7(1), pp.1-7.
25. Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997. 2013;
26. 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. 2017 Dec 27;18(1).
27. Katzourakis, A., Pereira, V. and Tristem, M., 2007. Effects of recombination rate on human endogenous retrovirus fixation and persistence. Journal of virology, 81(19), pp.10712-10717.
28. 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. International journal of molecular sciences, 20(23), p.5977.
29. Wang Y-C, Olson ND, Deikus G, Shah H, Wenger AM, Trow J, et al. High-coverage, long-read sequencing of Han Chinese trio reference samples. Scientific Data. 2019 Dec 14;6(1).
30. Chu, Chong, et al. "Comprehensive identification of transposable element insertions using

multiple sequencing technologies." Nature Communications 12.1 (2021): 1-12.
31. Troskie, Robin-Lee, et al. "Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome." Genome biology 22.1 (2021): 1-15.
32. Chu, Chong, et al. "Comprehensive identification of transposable element insertions using multiple sequencing technologies." Nature Communications 12.1 (2021): 1-12.
33. Ruan, J. and Li, H., 2020. Fast and accurate long-read assembly with wtdbg2. Nature methods, 17(2), pp.155-158.
34. Tempel S. Using and Understanding RepeatMasker. In Humana Press; 2012 [cited 2019 Mar 20]. p. 29–51. Available from: http://link.springer.com/10.1007/978-1-61779-603-6_2
35. Homer, N., 2010. Dwgsim: whole genome simulator for next-generation sequencing. GitHub repository.
36. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6).
37. Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D. and Fulton, R.S., 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome research, 27(5), pp.849-864.
38. Koch, L., 2020. Exploring human genomic diversity with gnomAD. Nature Reviews Genetics, 21(8), pp.448-448.