RetroSnake: A modular pipeline to detect human endogenous retroviruses in genome sequencing data

List of annotated HERVs with clickable links

| AnnotSV ID     | SV type | Gene name | Location | OMIM ID |
|----------------|---------|-----------|----------|---------|
| 14 780455146   | MEI     | SPTLC2    |          | 600713  |
| 10 101587719   | MEI     | ABCC2     |          | 601187  |
| 9 33110539     | MEI     | BHDALT1   |          | 137006  |
| 6 129409221     | MEI     | LAMA2     |          | 156225  |
| 5 80099344     | MEI     | MSH4      |          | 600887  |

Highlights

- RetroSnake is an end-to-end pipeline for detection of HERV-K insertions
- Modular and computationally efficient (~4 h per genome)
- Easy setup and installation with Snakemake
- Can be installed and used by users with limited computational experience

Renata Kabiljo, Harry Bowles, Heather Marriott, ..., Chad M. Swanson, Ammar Al-Chalabi, Alfredo Iacoangeli

renata.kabiljo@kcl.ac.uk (R.K.)
alfredo.iacoangeli@kcl.ac.uk (A.I.)
Human endogenous retroviruses (HERVs) integrated into the human genome as a result of ancient exogenous infections and currently comprise ~8% of our genome. The members of the most recently acquired HERV family, HERV-Ks, still retain the potential to produce viral molecules and have been linked to a wide range of diseases including cancer and neurodegeneration. Although a range of tools for HERV detection in NGS data exist, most of them lack wet lab validation and they do not cover all steps of the analysis. Here, we describe RetroSnake, an end-to-end, modular, computationally efficient, and customizable pipeline for the discovery of HERVs in short-read NGS data. RetroSnake is based on an extensively wet-lab validated protocol, it covers all steps of the analysis from raw data to the generation of annotated results presented as an interactive html file, and it is easy to use by life scientists without substantial computational training.

Availability and implementation: The Pipeline and an extensive documentation are available on GitHub.
Here, we present RetroSnake, a comprehensive Snakemake pipeline for the detection of HERVs in short-read NGS data. RetroSnake covers all commonly needed steps, from the pre-processing of sequencing data in CRAM/BAM format, to the annotation of results with information from a range of biological data-bases, and includes Retroseq and the Wildshutte protocol, in a single, flexible, and modular pipeline that requires minimal setup and informatics skills.

RESULTS

RetroSnake pipeline architecture

RetroSnake uses a modern computational workflow management system, Snakemake (Koster and Rahmann, 2018), to combine all steps employed in the detection of transposable elements into a single, fast, and easy to use pipeline. The core steps of the RetroSnake pipeline are as follows: (i) file conversion to a BAM format required for Retroseq; (ii) running Retroseq; (iii) filtering the predicted insertions; (iv) further refining of the results through insertion junction reconstruction and scanning of the reconstructed contigs for presence of the target HERV; (v) comparing the predicted insertions with known insertions; (vi) annotation of the results with information from a range of biological databases (e.g. overlapping genes and regulatory elements, predicted functional effect); (vii) generation of an interactive html report.

The Snakemake workflows consist of rules that define how to create output files from input files. The workflow is inferred by dependencies between the rules that arise from one rule requiring an output file of another as input. Upon execution, Snakemake determines the combination of rules necessary to achieve a requested output. This combination of rules is a directed acyclic graph (DAG) of jobs. A graphical presentation of RetroSnake DAG is shown in Figure 1. Users can request many different levels and formats of output: e.g., a bed file with novel verified insertions, or an html file with annotated insertions that can

Figure 1. Graphical overview of the full RetroSnake pipeline
CramToBam: conversion of cram files to bam files; retroseqDiscover and retroseqCall: two steps of retroSeq for detection of HERV-K insertions; filterCalls: quality filtering of predicted insertions based on flags produced by RetroSeq; verify: use RepeatMasker to verify the predicted HERV-K insertion in contigs assembled from reads surrounding the insertion site; markKnownFiltered, markKnownVerified: report all previously reported HERV-K predictions; markNovelFiltered, markNovelVerified: report all HERV-K predictions which have not been reported previously; annotateFiltered and annotateVerified: use AnnotSV to annotate predictions with biologically meaningful information. See STAR Methods for more details about each step.
be opened in any browser and displays hyperlink to other resources. The full list of options for RetroSnake output is in Table S1, and an illustrative figure of annotated insertions is shown in Figure 2.

RetroSnake utilizes the Conda package management system (https://conda.io) with its Bioconda channel (Gruning et al., 2018), which can be installed and updated easily with only a few shell commands. Through Conda, it is possible to define isolated software environments per rule. Upon execution of a workflow, Conda obtains and deploys the defined software packages. Many RetroSnake steps rely solely on software distributed via Conda, and as such, require no further installation. The only dependencies that require manual installations are RepeatMasker and AnnotSV. Detailed guidance for their deployment is provided on GitHub (https://github.com/KHP-Informatics/RetroSnake).

Results on 162 human short-read whole-genome sequencing samples

To illustrate the utility of RetroSnake, we applied it to the detection of HERV-Ks in a dataset of 162 human whole-genome sequencing samples from the MRC London Neurodegenerative Diseases Brain Bank based at the Institute of Psychiatry, Psychology & Neuroscience, King’s College London. These samples were sequenced as part of Project MinE. The data are described in detail in previous publications (Iacoangeli et al., 2019; Iacoangeli et al., 2021; Project MinE ALS Sequencing Consortium, 2018). Out of 36 known non-reference HERV-K insertions (Table S2), our pipeline identified 19 in at least one subject. These insertions and their frequencies at different stringency levels of RetroSnake are in Table 1. For these 19 insertions, frequencies of the predictions were compared with frequencies previously reported for the same insertions in multiple publications, where available (Figure 3). Most HERV-K insertions of highest frequency have indeed been previously reported. RetroSnake also found 45 novel insertions when its most stringent verification level was used, 18 of which were found in at least two samples (Figure S1). Table S3 includes a list of all predicted novel insertions.

Lab validation

Nested PCR was used to validate one novel high frequency HERV-K detected by RetroSnake located approximately at chr6:153429801 of Hg19 (Figure S1), falling inside an intron of the RGS17 gene. This was the novel insertion with the highest frequency in our dataset. It was detected in 24 samples, and we were able to retrieve a DNA sample for one of them. The lab validation protocol was applied to the sample predicted to contain the HERV-K insertion and two samples in which it was predicted to be absent. DNA was obtained from the MRC London Neurodegenerative Diseases Brain Bank at King’s College London. The nested PCR protocol gave a specific product at the expected size in the sample predicted to have the HERV-K insertion (Figure 4). This product was not observed in the two samples in which RetroSnake did not predict the insertion. The PCR product sequence aligns to the correct chromosome 6 location.
up to base pair ~286 in the sequencing read; from this position, it then aligns well to LTR5_Hs (Figure S2), validating the predicted HERV-K integration site.

### Computational efficiency

The whole pipeline took approximately 4 hours of CPU time per sample, with an additional 2 hours if CRAM to BAM conversion was needed. The CPUs used were Intel(R) Xeon(R) CPU E5-2660 v3 @ 2.60GHz. Full performance results for single sample processing are in Table 2. These statistics refer to the mean of three different samples.

### DISCUSSION

RetroSnake is an efficient and comprehensive computational pipeline for the detection of transposable elements. The main advantages of RetroSnake over the other available tools are that it is end-to-end, based on an extensively wet-lab validated protocol for HERV-K detection, and developed within the Snakemake framework.
framework which provides computational efficiency, scalability, customizability, crash recovery features, and it is easy to use.

End-to-end

RetroSnake starts with an alignment file and encompasses all steps commonly needed for the characterization of mobile elements. It eliminates the need for postprocessing of predicted insertions that are commonly necessary for the other existing tools. These post processing steps include not only comparing the predictions with previously known ones but also annotation of the results with biologically relevant information from various databases.

Based on a wet-lab validated protocol

RetroSnake is based on a previously described and experimentally validated protocol (Keane et al., 2013; Wildschutte et al., 2016). The authors have used PCR and capillary sequencing and have validated the presence of 34 of the 36 candidate insertions in at least one individual predicted to have the insertion. Using the same protocol gives us confidence in the quality of predicted insertions. Furthermore, we have verified the presence or absence of one novel, high frequency insertion in one sample predicted to carry the insertion and two samples predicted not to carry it.

Efficiency

The Snakemake framework allows the analysis to scale from single samples on personal computers to multiple samples on HPC clusters. When run on a cluster, Snakemake takes care of submitting jobs in parallel in an optimized way. The full pipeline including CRAM to BAM conversion takes around 6 hours of CPU time.
CRAM to BAM conversion is only performed if the alignment files in CRAM format are used. If data in BAM format are provided, the pipeline can process one sample in less than 4 hours, including verification and annotation; for the same task, other tools such as STEAK (Santander et al., 2017) and Ervcaller (Chen and Li, 2019) would take 7.5 and 14.5 CPU hours, respectively (Bowles et al., 2022).

Customizability
RetroSnake is a highly modular pipeline that allows for a fast integration of additional steps and the replacement of existing ones. For example, the HERV-K insertion detection step could be accomplished using another tool without affecting the rest of the pipeline. Other annotation rules can be added to complement the existing AnnotSV annotations. RetroSnake can be used for the detection of any transposable elements by providing their reference sequences.

Ease of use
RetroSnake can be installed and used even by users with limited computational experience. RetroSnake uses the package manager Conda and its Bioconda channel, which enables automatic installation of all tools provided by these channels. For example, in order to run Retroseq, filter the results and classify insertions into known and novel, no additional tools need to be manually installed as they are automatically obtained by Conda. For the verification and annotation steps, additional tools need to be installed and detailed instructions are available on GitHub. In order to run RetroSnake, only basic knowledge of the terminal usage is necessary. Detailed examples and instructions are provided on our GitHub.

Crash recovery
When Snakemake determines which rules are to be executed by building a DAG of jobs, it checks time-stamps of input and output files. As long as a rule was properly executed and its input file is older than its output file, Snakemake will not attempt to regenerate the output. The rule will run only if the timestamp of the input file has changed, or if the output file is missing. If an error occurs during execution, the output up to the point of failure is preserved, and the user will not need to rerun the whole pipeline. Similarly, if the parameters need to be changed, only particular rules need to be rerun. Another advantage provided by the Snakemake framework is that in case of a failed rule, intermediate files created by that rule are removed by Snakemake before exiting. This prevents truncated files from being mistakenly used in downstream analysis, which is a problem occasionally encountered in bioinformatics pipelines.

Conclusions
In conclusion, we presented RetroSnake, a computationally efficient, customizable, easy to use, and comprehensive pipeline for detecting HERVs in short-read genome NGS data. RetroSnake presents important advantages with respect to other available tools. For example, it is the only pipeline based on an extensively wet-lab validated protocol (Wildschutte et al., 2016), and it is the most complete HERV detection pipeline, producing annotated insertions presented as an html file with hyperlink, easy enough to be used by life scientists without substantial computational training.

Limitations of the study
RetroSnake is able to predict insertion points of HERVs. However, the short-read sequencing technology does not enable it to characterise they sequence beyond the presence of the (LTR). HERV-K LTRs alone are almost 1kb long and a complete provirus can be up to 10 kb, while reads are 100–250 bp long. Therefore, unmapped parts of the insertion (e.g. viral protein coding genes) cannot be mapped uniquely, and our tool only detects the insertion point using the LTR reference. However, with the increased availability of
long-read sequencing data (read length >10,000 bp), it is possible to not only predict the insertion point but to also determine the entire sequence of the insertion. Another limitation is that RetroSnake does not provide the zygosity of predicted insertions.

STAR METHODS
Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Subject details
- METHOD DETAILS
  - Pipeline steps
  - Whole-genome sequencing
  - Lab validation

SUPPLEMENTAL INFORMATION
Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105289.

ACKNOWLEDGMENTS
UK Research and Innovation; Medical Research Council; South London and Maudsley NHS Foundation Trust; MND Scotland; Motor Neurone Disease Association; National Institute for Health Research; Spastic Paraplegia Foundation; Rosetrees Trust; Darby Rimmer MND Foundation. Funding for open access charge: UKRI. R.K. is funded by the MND Scotland. H.M. is supported by a GSK studentship and the KCL funded Centre for Doctoral Training (CDT) in Data-Driven Health. A.I. is funded by the Motor Neurone Disease Association and South London and Maudsley NHS Foundation Trust. 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-http://www.neurodegenerationresearch.eu/ (United Kingdom, Medical Research Council MR/L501529/1 to A.A.-C., principal investigator [PI] and MR/R024804/1 to A.A.-C., PI); Economic and Social Research Council ES/L008238/1 to A.A.-C. [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 Center 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). A.A.K. is funded by ALS Association Milton Safenowitz Research Fellowship (grant number22-PDF-609.DOI:10.52546/pc.gr.150909.* title = "doi:DOI:10.52546/pc.gr.150909.*">DOI:10.52546/pc.gr.150909.), The Motor Neurone Disease Association (MND Association) Fellowship (AI Khleifat/Oct21/975-799), The Darby Rimmer Foundation, and The NIHR Maudsley Biomedical Research Centre. H.B. his is funded by the National Institute for Health and Care Research (NIHR) Biomedical Research Center based at Guy’s and St Thomas’ NHS Foundation Trust and King’s College London. C.R.B. is funded by MRC (MR/S000844/1). 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 Centers at South London & Maudsley and Guy’s & 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). 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.

AUTHOR CONTRIBUTIONS
R.K., H.B., and A.I. contributed to concept and design the study. R.K. implemented the software, run the analyses, and drafted the manuscript. H.M. and H.B. have tested the pipeline. H.B., C.R.B., and C.M.S. have designed and preformed the wet lab validation. A.I. supervised the study. A.A.-C. and A.I. raised the study funding. All authors contributed to the write-up.

DECLARATION OF INTERESTS
The authors declare no competing interests.

Received: March 3, 2022
Revised: August 8, 2022
Accepted: October 4, 2022
Published: November 18, 2022

REFERENCES
Bowles, H., Kabiljo, R., Jones, A., Al Khleifat, A., Quinn, J.P., Dobson, R.J., Swanson, C.M., Al-Chalabi, A., and Iacoangeli, A. (2022). An assessment of bioinformatics tools for the detection of human endogenous retroviral insertions in short-read genome sequencing data. Preprint at bioRxiv. https://doi.org/10.1101/2022.02.18.481042.

Chen, J., Forozesh, M., and Qin, Z. (2019). Transactivation of human endogenous retroviruses by tumor viruses and their functions in virus-associated malignancies. Oncogene 8, 6.

Chen, X., and Li, D. (2019). ERVcaller: identifying polymorphic endogenous retrovirus and other transposable element insertions using whole-genome sequencing data. Bioinformatics 35, 3913–3922.

Dembny, P., Newman, A.G., Singh, M., Hinz, M., Szczepak, M., Kruger, C., Adalbert, R., Deyoe, O., Trimbuch, T., Wallach, T., et al. (2020). Human endogenous retrovirus HERV-K(HML-2) RNA causes neurodegeneration through Toll-like receptors. JCI Insight 5, 131093.

Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E.C., Pittard, W.S., Mills, R.E., and Devine, S.E. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. 27, 1916–1929.

Geoffroy, V., Guignard, T., Kress, A., Gaillard, J.B., Solli-Nowlan, T., Schalk, A., Gatinov, N., Dollfus, H., Scheidecker, S., and Muller, J. (2021). AnnotSV and knobAnnotSV: a web server for human structural variations annotations, ranking and analysis. Nucleic Acids Res. 49, W21–W28.

Geoffroy, V., Herengy, Y., Kress, A., Stoetzel, C., Pison, A., Dollfus, H., and Muller, J. (2018). AnnotSV: an integrated tool for structural variations annotation. Bioinformatics 34, 3572–3574.

Gifford, R., and Tristem, M. (2003). The evolution, distribution and diversity of endogenous retroviruses. Virus Gene. 26, 291–315.

Grandi, N., and Tramontano, E. (2018). Human endogenous retroviruses are ancient acquired elements still shaping innate immune responses. Front. Immunol. 9, 2039.

Gruning, B., Dale, R., Spödi, A., Chapman, B.A., Rowe, J., Tomkins-Tinch, C.H., Valieris, R., and Köster, J. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat. Methods 15, 475–476.

Huang, X., and Madan, A. (1999). CAP3: a DNA sequence assembly program. Genome Res. 9, 868–877.

Iacoangeli, A., Al Khleifat, A., Jones, A.R., Sproviero, W., Shatunov, A., Opie-Martin, S., Alzheimer’s Disease Neuroimaging Initiative, Morrison, K.E., Shaw, P.J., Shaw, C.E., et al. (2019). C9orf72 intermediate expansions of 24–30 repeats are associated with ALS. Acta Neuropathol. Commun. 7, 115.

Iacoangeli, A., Fogh, I., Selvackadunco, S., Topp, S.D., Shatunov, A., van Rheenen, W., Al-Khleifat, A., Opie-Martin, S., Ratti, A., Calvo, A., et al. (2021). SCFD1 expression quantitative trait loci in amyotrophic lateral sclerosis are differentially expressed. Brain Commun. 3, fca236.

Keane, T.M., Wong, K., and Adams, D.J. (2013). RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics 29, 389–390.

Koster, J., and Rahmann, S. (2018). Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 34, 3600.

Lee, A., Huntley, D., Aiewsakun, P., Kanda, R.K., Lynn, C., and Tristem, M. (2014). Novel denisovan and neandertal retroviruses. J. Virol. 88, 12907–12909.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.

Marchi, E., Kanapin, A., Magiorkinis, G., and Belshaw, R. (2014). Unfixed endogenous retroviral insertions in the human population. J. Virol. 88, 9529–9537.

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, 1537–1546.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.

Santander, C.G., Gambroń, P., Marchi, E., Karamitros, T., Katsourakis, A., and Magiorkinis, G. (2017). STEAK: a specific tool for transposable elements and retrovirus detection in high-throughput sequencing data. Virus Evol. 3, vex023.

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, 90.

Wildschutte, J.H., Williams, Z.H., Montesion, M., Subramanian, R.P., Kidd, J.M., and Coffin, J.M. (2016). Discovery of unfixed endogenous retroviral insertions in diverse human populations. Proc. Natl. Acad. Sci. USA 113, E2326–E2334.
STAR METHODS

KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| **Software and algorithms** | | |
| Snakemake | Koster and Rahmann (2018) | https://github.com/snakemake/snakemake |
| RetroSnake | This paper | https://doi.org/10.5281/zenodo.7050012, https://github.com/KHP-Informatics/RetroSnake |
| **Oligonucleotides** | | |
| PCR primers | This paper | N/A |
| CTCATTCTCCCCCCTTGTG | | |
| TAACCAAATGCGCCGCTGCT | | |
| TTTCAAGGAGCGGGTGGTGG | | |
| **Critical commercial assays** | | |
| Q5 High-fidelity 2X master-mix | New England Biolab | M0492S |
| pCR™ Blunt II-TOPO™ | Invitrogen | 451,245 |
| QIAprep Spin Miniprep | QIAGEN | 27106X4 |

RESOURCE AVAILABILITY

Lead contact
Requests for further information should be directed to the lead contact, Alfredo Iacoangeli (alfredo.iacoangeli@kcl.ac.uk).

Materials availability
This study did not generate any new materials.

Data and code availability
- Data
Sequencing data from ProjectMinE is available upon reasonable request to the lead contact.
- Code
All original code has been deposited at https://github.com/KHP-Informatics/RetroSnake and is publicly available as of the date of publication. https://doi.org/10.5281/zenodo.7050012.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Subject details
RetroSnake has been applied for the detection of HERV-Ks in a dataset of 162 human whole genome sequencing samples from the MRC London Neurodegenerative Diseases Brain Bank based at the Institute of Psychiatry, Psychology & Neuroscience, King’s College London. The dataset comprises 107 ALS cases (60 male, 47 female; age at death 64.49 ± 12.31 years), 54 controls (22 male and 32 female, age at death 76.31 ± 14.56 years), and one sample with missing phenotype information.

METHOD DETAILS

Pipeline steps
CRAMToBam
This step calls Samtools (Li et al., 2009) in order to convert CRAM files to BAM files. CRAM files are a compressed version of BAM files, and frequently, due to the large size of the files and memory constraints, files
are stored as CRAMs. This first, optional step does the conversion to BAM format which is necessary for the next step. If the BAM files are already present, this step is skipped.

**Retroseq**

Retroseq is executed with default parameters – 80 identity threshold which can be adjusted. The fasta file containing the sequence of the repetitive element needed for the execution of Retroseq has been downloaded from RepeatMasker with query ‘LTR5_Hs’ and is available on our GitHub PAGE. Both Retroseq discover and call options have been used.

**Filtering**

Output of Retroseq call is a vcf file with predicted insertions and various quality flags characterising each prediction. For filtering of Retroseq results we have used the following combination of flags \( fl \) and \( gq \) from Retroseq output: For \( fl \) of 8, \( gq \) has to be 10 or more; for \( fl \) of 7 \( gq \) has to be 20 or more and for \( fl \) of 6, \( gq \) needs to be as high as 29. All predictions with \( fl \) lower than 6 are rejected.

**Marking known and novel insertions**

The compiled list we used as known non reference HERV-K insertions is in Table S2. This list has been assembled from literature (Lee et al., 2014; Marchi et al., 2014; Subramanian et al., 2011; Wildschutte et al., 2016).

We used BEDTools (Quinlan and Hall, 2010) to mark as known insertions all predictions within 500 bp of a known prediction. Novel insertions were defined as predicted insertions not found within 5000bp of a previously reported insertion. All novel predicted insertions in all subjects were merged using BEDTools; the presence/absence of an insertion in each subject was then obtained by using BEDTools to check the overlap with any of the merged insertions.

**Verification**

The verification step is based on our interpretation of the verification described by (Wildschutte et al., 2016). All reads from the aligned BAM that overlap each predicted insertion point until 1000bp downstream of it are collected and assembled into contigs using CAP3 (Huang and Madan, 1999). The assembled contigs are then run through RepeatMasker using DFAM libraries of repetitive elements. Only hits with less than 10% substitutions in the matching region compared to the consensus, and with less than 3% of bases opposite a gap in either a query or a repeat sequence pass the initial filter. For medium level verification, we further request at least one contig to contain LTR5_Hs. For strict verification, we require the presence of multiple contigs with LTR5_Hs. If only a single contig has LTR5_Hs, we require that the contig has to be in the first half of list of contigs, as the contigs assembled by CAP3 have been sorted by quality. The verification level is easily controlled by parameter verificationLevel.

**Annotation**

AnnotSV and knotAnnotSV are run with default parameters, on either filtered or verified predictions. A custom annotation track with known mobile element insertions and their frequencies has been added and is automatically included. When invoked, the annotation step executes both AnnotSV which searches for elements overlapping the predicted insertions, and KnotAnnotSV which takes the annotations predicted by AnnotSV and renders them into an HTML annotation report with clickable links.

**Whole-genome sequencing**

Whole-genome sequencing (WGS) data were obtained from frozen human postmortem tissue taken from primary motor cortex of 162 individuals from Project MinE described in detail in (Iacoangeli et al., 2019; Project MinE ALS Sequencing Consortium, 2018). Samples were sequenced on the Illumina Hiseq X platform, using PCR-free library preparation, 150 bp paired-end reads, with 30x coverage depth. The full list of IDs of subjects from Project MinE included in this study is in Table S4.

**Lab validation**

Nested PCR was used to validate one novel, high frequency HERV-K detected by RetroSnake located approximately at Chr6:153429801. The protocol was applied to one Project MinE sample predicted to contain the HERV-K insertion and two samples in which it was predicted to be absent. 50ng genomic
DNA was used for all PCR reactions with Q5 High-fidelity 2X master-mix (NEB) in the presence of 10pM of primers for 30 amplification cycles. The first PCR step used flanking external chromosome 6 primers which were designed using PrimerBlast (forward: CTCACTTCTCCCCCTTGTG, reverse: TAACCAAATGTGCGGCTGCT). The PCR product was purified using the QIAquick protocol (Qiagen). This was followed by a second PCR reaction using the forward flanking primer and an internal primer derived from the HERV-K LTR5_Hs sequence (TTTCAGAGAGCACGGGGTTG). This protocol should therefore amplify the 5' junction between genomic DNA (chromosome 6) and the predicted HERV-K LTR. The final PCR product was TOPO cloned in the pCR™Blunt II-TOPO™ vector (Invitrogen) for 30 min at 22°C, and the TOPO cloning product was introduced into DH10B E. coli by chemical transformation. Transformed bacteria were selected in presence of kanamycin and plasmid DNA was isolated from single colonies by miniprep (QiaRep Spin Miniprep, Qiagen). The insert size in the plasmids was verified by EcoRI digest, and plasmids with the expected insert size (~350bp) were sequenced by Sanger sequencing (GENEWIZ). A chromosome 6 contig (NCBI nucleotide accession AL356963) and the HERV-K LTR sequence (Dfam DF0000558.4) were aligned to the sequencing read using MacVector 17 (MacVector). The sequencing read is representative of two independent bacterial clones from the patient’s PCR reaction.