**User Manual of ERVcaller v1.0**

**May 27, 2018**

**Citation:** Chen X and Li D. ERVcaller: Identifying and genotyping non-reference unfixed endogenous retroviruses (ERVs) and other transposable elements (TEs) using next-generation sequencing data. *Manuscript in submission*.

**Download:** [www.uvm.edu/genomics/software/ERVcaller](http://www.uvm.edu/genomics/software/ERVcaller).html

**Copyright:** ERVcaller is licensed under the Creative Commons Attribution-NonCommercial 4.0 International license. It may be used for non-commercial use only. For inquiries about a commercial license, please contact the corresponding author at [dawei.li@uvm.edu](mailto:dawei.li@uvm.edu) or The University of Vermont Innovations at [innovate@uvm.edu](mailto:innovate@uvm.edu).

**1. Introduction**

ERVcaller is a tool designed to accurately detect and genotype non-reference unfixed endogenous retroviruses (ERVs) and other transposon elements (TEs) in the human genome using next-generation sequencing (NGS) data. We evaluated the tools using both simulated and real benchmark whole-genome sequencing (WGS) datasets. ERVcaller is capable to accurately detect various TE insertions of any lengths, particularly ERVs. It allows for use of a TE reference library regardless of sequence complexity, such as use of the entire RepBase database. It is easy to install and use with command lines.

Complementary to ERVcaller, other bioinformatics tools designed to detect large deletions, such as Breakdancer, may be used to detect TEs that are present in the human reference genome but not in testing samples.

**2. Obtaining and Compiling**

**2.1. Users need to successfully install the following software separately, and make them available in the default path, such as using the Linux command “export”.**

* BWA-0.7.10
* Bowtie2
* Tophat-2.1.1
* Samtools-1.6 (or later than 1.2)
* Hydra-0.5.3
* SE\_MEI (Modified version)

**2.2. Databases**

* The human reference genome (hg38 by default)
* One TE reference database from these options: the human TE database, the HERVK reference sequence, ERV library or the entire RepBase ([www.girinst.org/repbase/](http://www.girinst.org/repbase/))

**3. Running ERVcaller (more examples are shown in the help print page)**

**3.1. Print help page**

$ *perl user\_installed\_path/ERVcaller.pl*

**3.2 TE detection**

$ *perl user\_installed\_path/ERVcaller.pl -i sample\_ID -f .bam -S -r -g -w 5000*

**3.3. Parameters**

**Table 1** List of parameters and explanation

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Full name** | **Description** |
| -i | input\_sampleID | Sample ID (*required*) |
| -h | help | Print this help |
| -t | threads | The number of threads (default: *1*) |
| -f | file\_suffix | The suffix of the input data (default: *.fq.gz*) |
| -d | data\_type | Data type, including WGS, and RNA-seq (default: *WGS*) |
| -s | sequencing\_type | Type of sequencing data, including paired-end, and single-end (default: *paired-end*) |
| -H | Human\_reference\_genome | The FASTA file of the human reference genome |
| -T | TE\_reference\_genomes | The TE library (FASTA) used for screening |
| -l | length\_insertsize | Insert size length (bp) (default: *500*) |
| -S | Split | If the split reads is used for detection |
| -r | Reciprocal\_alignment | Reciprocal align the supporting reads against the candidate genomic regions |
| -g | genotyping | Genotyping function |
| -w | window\_size | Window size of selected genomic locations for genotyping (bp) (default: *5,000*) |

**4 Output file**

**Table 2** Header of the final output file

|  |  |  |
| --- | --- | --- |
| **Column** | **Header** | **Description** |
| Col 1 | Sample\_ID | Sample ID |
| Col 2 | Split\_reads | Use split reads or not |
| Col 3 | Genotyping | Use genotyping function or not |
| Col 4 | Reciprocal\_alignment | If the reads were reciprocal aligned back to the candidate genomic regions |
| Col 5 | TE\_sequence\_name | Sequence name of the detected TE |
| Col 6 | Chr. | Human chromosome ID |
| Col 7 | Start | Start position of the span genomic region of all chimeric and split reads in the human reference genome |
| Col 8 | End | End position of the span genomic region of all chimeric and split reads in the human reference genome |
| Col 9 | No.\_chimeric\_reads | Total count of chimeric reads of the integration |
| Col 10 | No.\_split\_reads | Total count of split reads of the integration |
| Col 11 | Upstream\_breakpoint\_on\_human | Upstream breakpoint detected in the human reference genome |
| Col 12 | Downstream\_breakpoint\_on\_human | Downstream breakpoint detected in the human reference genome |
| Col 13 | Upstream\_breakpoint\_on\_TE | Upstream breakpoint detected in the TE sequence |
| Col 14 | Downstream\_breakpoint\_on\_TE | Downstream breakpoint detected in the TE sequence |
| Col 15 | Information\_both\_breakpoints | Upstream and downstream breakpoint information. Upstream and downstream breakpoints were separated by semicolon; "D" and "E" represent if this breakpoint is detected by split reads (D), or estimated by chimeric reads separately (E); "+" and "-", represent the forward and reverse direction for both human (left) and TE (right) genome in the square per breakpoint; "na" represent this breakpoint is not covered by any chimeric and split reads |
| Col 16 | Genotype | Genotypes of the TE insertion |
| Col 17 | No.\_reads\_supporting\_nonTE | No. reads support no TE insertion |
| Col 18 | No.\_reads\_supporting\_TE | No. reads support TE insertion, including chimeric and split reads |
| Col 19 | Average\_alignment\_score | Average alignment score (AS) of reads support TE insertion, including chimeric and split reads |
| Col 20 | Read\_depth\_(Mean)\_of\_the\_window | Mean value of the read count of randomly-selected genomic locations interspersed every 50 bp within a window |
| Col 21 | Read\_depth\_(SD)\_of\_the\_window | SD value of the read count of randomly-selected genomic locations interspersed every 50 bp within a window |
| Col 22 | No.\_sampled\_genomic\_locations | The number of randomly-selected genomic locations interspersed every 50 bp within a window |
| Col 23 | Read\_depth\_(Quantile =0.05) | The quantile value of (threshold =0.05) |
| Col 24 | Paired-end\_false\_Chimeric\_reads | The number of chimeric reads (mapped in proper paired) within the candidate genomic region |
| Col 25 | Other\_false\_Chimeric\_reads | The number of chimeric reads (both ends mapped but not in proper paired) within the candidate genomic region |
| Col 26 | Average\_(AS-XS)\_for\_chimeric\_reads\_on\_human | Average value of (AS – XS flag value) for chimeric reads on human |
| Col 27 | Maximum\_(AS-XS)\_for\_chimeric\_reads\_on\_human | Maximum value of (AS – XS flag value) for chimeric reads on human |