User Manual of ERVcaller v1.1

July 15, 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.html

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 tool using both simulated and benchmark whole-genome sequencing (WGS) datasets. ERVcaller is capable of accurately detecting various TE insertions of any lengths, particularly ERVs. It can be applied to both paired-end and single-end WGS, RNA-Seq or targeted DNA sequencing data. In addition, ERVcaller is capable of detecting the breakpoints at single-nucleotide resolution. It allows for the use of a TE reference library regardless of sequence complexity, such as the entire RepBase database, suggesting that it may be used to detect highly divergent or novel TE insertions. It is easy to install and use from the command line.

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 Installation

2.1 Unzip ERVcaller installer

$ *tar vxzf ERVcaller_v.1.1.tar.gz*

2.2 Installing dependent software

Users need to successfully install the following software separately and make them available in the default path (such as by 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 included in the Scripts folder of ERVcaller installer)
- R-3.3.2 (or higher)

2.3 Databases

- The human reference genome (hg38 by default; http://hgdownload.soe.ucsc.edu/downloads.html#human) (if you use BAM file as the input, the chromosome IDs of the input BAM file should be identical to those of the human reference genome used, such as “Chr1”, “chr1”, and “1”).
- One of the following TE reference libraries: 1) a TE reference provided by ERVcaller installer (i.e., the HERV-K reference sequence, the ERV library or the human TE database), or 2) a user-defined TE reference library.

2.4 Databases indexing

- The human reference genome and TE reference library should be indexed separately using BWA “*bwa index*”.

3 Running ERVcaller

3.1 Print help page

$ *perl user_installed_path/ERVcaller.pl*

3.2 TE detection

For detecting and genotyping all ERV insertions:

$ *perl user_installed_path/ERVcaller.pl -i sample_ID -f .bam -H human_genome_hg38.fa -T ERV_library.fa -S -V -G*

More examples of ERVcaller running command lines are shown in the help page.

3.3 Parameters

All available parameters are listed below. Four parameters are required, including input_sampleID (-i), file_suffix (-f), Human_reference_genome (-H), and TE_reference_genomes (-T).

Table 1 List of parameters and interpretation

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 (*required;* 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 (*required*)
-T	TE_reference_genomes	The TE library (FASTA) used for screening (*required*)
-l	length_insertsize	Insert size length (bp) (default: *500*)
-S	Split	Is the split reads used for detection
-V	Validation	Validation function (input bam file need to be indexed)
-G	Genotyping	Genotyping function (input bam file need to be indexed)
-w	window_size	Window size of selected genomic locations for genotyping (bp) (default: *10,000*)

(see next page)

4 Output file

The output file with the suffix of ".output" (without validation or genotyping functions performed) or ".output2" (with either validation or genotyping function performed) will be generated after the running. All the columns are listed below.

Table 2 Headers of the final output file

Column	Header	Description
Col 1	Sample_ID	Sample ID
Col 2	Is_Split_mode	Were split reads used
Col 3	Is_genotyped	Was genotyping function performed
Col 4	Is_validated	Was validation function performed
Col 5	Human_ref.	Human reference genome
Col 6	TE_reference	TE reference genome
Col 7	TE_sequence_name	TE sequence ID (name)
Col 8	Chr.	Human chromosome ID
Col 9	Start	Start position of the span genomic region of all chimeric and split reads in the human reference genome Total count of chimeric reads of the integration
Col 10	End	End position of the span genomic region of all chimeric and split reads in the human reference genome 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	Insertion_site	Insertion size on human
Col 17	Group	If the TE insertions were located in repeat regions or false positives; If the TE insertions were successfully validated (confident level)
Col 18	Average_AS_for_chimeric_and_improper_reads_on_human	Average value of (AS flag value) for chimeric reads on human
Col 19	Average_XS_for_chimeric_and_improper_reads_on_human	Average value of (XS flag value) for chimeric reads on human
Col 20	Maximum_AS_for_chimeric_and_improper_reads_on_human	Maximum value of (AS flag value) for chimeric reads on human
Col 21	No._supporting_reads	No. reads support TE insertion, including chimeric and split reads
Col 22	Average_AS_supporting_reads	Average alignment score (AS) of reads support TE insertion, including chimeric and split reads
Col 23	No._chimeric_and_improper_reads	No. chimeric reads
Col 24	V_True_chimeric_and_improper_reads	No. truly validated chimeric reads
Col 25	V_False_chimeric_and_improper_reads_PE	No. false chimeric reads (mapped in proper paired) within the candidate genomic region
Col 26	V_False_chimeric_and_improper_reads_others	No. false chimeric reads (both ends mapped but not in proper paired) within the candidate genomic region
Col 27	No._split_reads_(≥20bp)	No. split reads (⩾ 20 bp)

Col 28	V_True_split_reads	No. truly validated split reads
Col 29	No._split_reads_(<20bp)	No. split reads (< 20 bp)
Col 30	Geno_No._reads_supporting_non_TE_insertions	No. reads support no TE insertion
Col 31	Geno_Read_depth_of_the_genomic_window_(Mean)	Mean value of the read count of randomly-selected genomic locations interspersed every 50 bp within a window
Col 32	Geno_Read_depth_of_the_genomic_window_(SD)	SD value of the read count of randomly-selected genomic locations interspersed every 50 bp within a window
Col 33	Geno_No._random_locations_of_the_genomic_window	The number of randomly-selected genomic locations interspersed every 50 bp within a window
Col 34	Geno_Read_depth_of_the_genomic_window_(Quantile=0.1)	The quantile value of (threshold =0.1)
Col 35	Genotype	Genotypes of the TE insertion