LTRdigest User's Manual

Sascha Steinbiss
Ute Willhoeft
Gordon Gremme
Stefan Kurtz

Research Group for Genome Informatics
Center for Bioinformatics
University of Hamburg
Bundesstrasse 43
20146 Hamburg
Germany

`steinbiss@zbh.uni-hamburg.de`

In any documentation or publication about research using *LTRdigest* please cite the following paper:

S. Steinbiss, U. Willhoeft, G. Gremme and S. Kurtz. Fine-grained annotation and classification of *de novo* predicted LTR retrotransposons. *Nucleic Acids Research* 2009, 37(21):7002–7013
`http://nar.oxfordjournals.org/cgi/content/full/37/21/7002`

26/08/2013

1 Introduction

This document describes *LTRdigest*, a software tool for identification and annotation of characteristic sequence features of LTR retrotransposons in predicted candidates, like those reported by*LTRharvest* [1]. In particular, *LTRdigest* can be used to find

- protein domains,
- polypurine tracts (PPT) and
- primer binding sites (PBS)

inside a sequence region predicted to be a LTR retrotransposon.

For this identification, *LTRdigest* utilises a number of algorithms to create an annotation based on user-supplied constraints. For example, length and position values for possible features can be extensively parameterised, as can be algorithmic parameters like alignment scores and cut-off values.

LTRdigest computes the boundaries and attributes of the features that fit the user-supplied model and outputs them in GFF3 format [2] (in addition to the existing LTR retrotransposon annotation), as well as the corresponding sequences in multiple FASTA format. In addition, a tab-separated summary file is created that can conveniently and quickly browsed for results.

LTRdigest is written in C and it is based on the *GenomeTools* library [3]. It is called as part of the single binary named `gt`.

The source code can be compiled on 32-bit and 64-bit platforms without making any changes to the sources. It incorporates HMMER [5], a popular and widely used profile hidden Markov model package that is used for identification of protein domains, for example using pHMMs taken from the Pfam [4] database. The protein domain search is implemented to be run in a multi-threaded fashion, thus making use of modern multi-core computer systems.

2 Building *LTRdigest*

As *LTRdigest* is part of the *GenomeTools* software suite, a source distribution of *GenomeTools* must be obtained, e.g. via the *GenomeTools* home page (`http://genometools.org`), and decompressed into a source directory:

```
$ tar -xzvf genometools-X.X.X.tar.gz
$ cd genometools-X.X.X
```

Then, it suffices to call `make` to compile the source. To enable protein domain search functionality, please also append the `with-hmmer=yes` option to the `make` call. This option will make sure that a HMMER source package is downloaded and compiled along with the `gt` binary. Note that the `wget` executable must be available in the current PATH to do so (alternatively, you can download HMMER manually from `ftp://selab.janelia.org/pub/software/hmmer/CURRENT/` and untar it in the `src/external/` subdirectory).

```
$ make with-hmmer=yes
```

If the `with-hmmer` option is not specified or set to `no`, protein domain finding functionality will be missing from the *LTRdigest* program. If the build process reports an error due to an unavailable Cairo library, also append the `cairo=no` option to build the `gt` binary without Cairo support. This has no influence on the function of *LTRdigest*.

To enable multithreading support (that is, to speed up protein domain search by searching for multiple pHMMs at once), please also specify the `threads=yes` option to the `make` call:

```
$ make with-hmmer=yes threads=yes
```

After successful compilation, the *GenomeTools* executable containing *LTRdigest* can then be installed for system-wide use as follows:

```
$ make install
```

If a `prefix=<path>` option is appended to this line, a custom directory can be specified as the installation target directory, e.g.

```
$ make install prefix=/home/user/gt
```

will install the `gt` binary in the `/home/user/gt/bin` directory. Please also consult the `README` and `INSTALL` files in the root directory of the uncompressed source tree for more information and troubleshooting advice.

3 Usage

Some text is highlighted by different fonts according to the following rules.

- `Typewriter font` is used for the names of software tools.
- `Small typewriter font` is used for file names.
- `Footnote sized typewriter font` with a leading `'-'` is used for program options.
- *small italic font* is used for the argument(s) of an option.

3.1 *LTRdigest* command line options

Since *LTRdigest* is part of `gt`, *LTRdigest* is called as follows.

`gt ltrdigest` [*options*] *GFF3_file indexname*

where *GFF3_file* denotes the GFF3 input file and *indexname* the name of an encoded sequence as created by the `gt suffixerator` tool. An overview of all possible options with a short one-line description of each option is given in Table 1. They can also be displayed when invoking *LTRdigest* with the option `-help` or `-help+`. All options can be specified only once.

To run the protein domain search in a parallel fashion, use the `-j` parameter to `gt` to specify the number of concurrent threads to use:

```
gt -j 3 ltrdigest [options] GFF3_file indexname
```

Table 1: Overview of the *LTRdigest* options sorted by categories.

	Input options
GFF3_file	specify the path to the GFF3 input file
indexname	specify the path to the input sequences
	Output options
`-outfileprefix`	specify prefix for sequence and tabular output files
`-o`	specify file to output result GFF3 into
`-gzip`	gzip-compress GFF3 output file specified by `-o`
`-bzip2`	bzip2-compress GFF3 output file specified by `-o`
`-force`	force output file to be overwritten
`-aaout`	output amino acid sequences for protein domain hits
`-aliout`	output HMMER amino acid alignments
`-seqnamelen`	set maximal length of sequence names in output FASTA headers (e.g. for clustalw or similar tools)
	PPT options
`-pptlen`	specify a range of acceptable PPT lengths
`-uboxlen`	specify a range of acceptable U-box lengths
`-pptradius`	specify region around 3' LTR beginning to search for PPT
`-pptrprob`	purine emission probability inside PPT
`-pptyprob`	pyrimidine emission probability inside PPT
`-pptaprob`	background A emission probability outside PPT
`-pptcprob`	background C emission probability outside PPT
`-pptgprob`	background G emission probability outside PPT
`-ppttprob`	background T emission probability outside PPT
`-pptuprob`	U/T emission probability inside U-box
	PBS options
`-trnas`	tRNA library in multiple FASTA format
`-pbsalilen`	specify a range of acceptable PBS lengths
`-pbsoffset`	specify a range of acceptable PBS start distances from 3' end of 5' LTR
`-pbstrnaoffset`	specify a range of acceptable tRNA/PBS alignment offsets from tRNA 3' end
`-pbsmaxedist`	specify the maximal allowed unit edit distance in tRNA/PBS alignment
`-pbsradius`	specify region around 5' LTR end to search for PBS
	Protein domain search options
`-hmms`	specify a list of pHMMs for domain search in HMMER2 format
`-pdomevalcutoff`	specify an E-value cutoff for pHMM search
`-maxgaplen`	maximum allowed chaining gap size between fragments (in amino acids)
	Alignment options
`-pbsmatchscore`	specify matchscore for PBS/tRNA Smith-Waterman alignment
`-pbsmismatchscore`	specify mismatchscore for PBS/tRNA Smith-Waterman alignment
`-pbsinsertionscore`	specify insertionscore for PBS/tRNA Smith-Waterman alignment
`-pbsdeletionscore`	specify deletionscore for PBS/tRNA Smith-Waterman alignment
	Miscellaneous options
`-v`	verbose mode
`-help`	show basic options
`-help+`	show basic and extended options

3.2 Input parameters

GFF3_file
specifies the path to the GFF3 input file. It has to include at least position annotation for the LTR retrotransposon itself (`LTR_retrotransposon` type) and the predicted LTRs (`long_terminal_repeat` type) because this information is needed to locate the favored positions of the features in question. The GFF3 file must also be sorted by position, which can be done using *GenomeTools*:

```
gt gff3 -sort unsorted_gff3_file > sorted_gff3_file
```

indexname
specifies the index name of an encoded sequence file containing the sequences the GFF3 coordinates refer to. For each sequence in the encoded sequence file given as the second parameter, there must exist a sequence region in the GFF3 file named '`seq`*i*' where i is the (zero-based) index number of the corresponding sequence in the encoded sequence. For instance, all GFF3 feature coordinates for features on the first sequence in the index file must be on sequence region `seq0`, and so on. An encoded sequence can be created from a FASTA, GenBank or EMBL format file using the `gt suffixerator` command:

```
gt suffixerator -tis -des -dna -ssp -db sequencefile -indexname indexname
```

3.3 Output options

Results are reported in GFF3 format on stdout and can easily be written to a file using the notation > *GFF3_resultfile* as in the following example:

```
gt ltrdigest [options] GFF3_file indexname > GFF3_resultfile
```

`-outfileprefix` *prefix*

If this option is given, a number of files containing further information will be created during the *LTRdigest* run:

- `<prefix>_tabout.csv` contains a tab-separated summary of the results that can, for example, be opened in a spreadsheet software or processed by a script. Each column is described in the file's header line and each row describes exactly one LTR retrotransposon candidate.
- `<prefix>_conditions.csv` contains information about the parameters used in the current run for documentation purposes.
- `<prefix>_pbs.fas` contains the PBS sequences identified in the current run in multiple FASTA format.
- `<prefix>_ppt.fas` contains the PPT sequences identified in the current run in multiple FASTA format.
- The files `<prefix>_5ltr.fas` and `<prefix>_3ltr.fas` contain the 5' and 3' LTR sequences identified in the current run in multiple FASTA format. Please note: If the direction of the retrotransposon could be predicted, the files will contain the corresponding 3' and 5' LTR sequences. If no direction could be predicted, forward direction with regard to the original sequence will be assumed, i.e. the 'left' LTR will be considered the 5' LTR.

- Additionally, one `<prefix>_pdom_<domainname>.fas` file will be created per protein domain model given. This file contains the FASTA DNA sequences of the HMM matches to the LTR retrotransposon candidates.

In FASTA output files, each FASTA header contains position and sequence region information to match the hit to the corresponding LTR retrotransposon.

`-aaout` *yes/no*

If this option is set to *yes*, one `<prefix>_pdom_<domainname>_aa.fas` file will be created per protein domain model given. This file contains the (concatenated) FASTA amino acid sequences of the HMM matches to the LTR retrotransposon candidates.

`-aliout` *yes/no*

If this option is set to *yes*, one `<prefix>_pdom_<domainname>.ali` file will be created per protein domain model given. This file contains alignment information for all matches to of the given protein domain model to the translations of all candidate.

3.4 PPT options

`-pptlen` L_{min} L_{max}

Specify the minimum and maximum allowed lengths for PPT predictions. If a purine-rich region shorter than L_{min} or longer than L_{max} is found, it will be skipped.
L_{min} and L_{min} have to be positive integers. If this option is not selected, then L_{min} is set to 8, L_{max} to 30.

`-uboxlen` L_{min} L_{max}

Specify the minimum and maximum allowed lengths for U-box predictions. If a T-rich region preceding a PPT shorter than L_{min} or longer than L_{max} is found, it will be skipped.
L_{min} and L_{min} have to be positive integers. If this option is not selected, then L_{min} is set to 3, L_{max} to 30.

`-pptradius` r

Specify the area around the 3' LTR beginning (l_s) to be searched for PPTs, in other words, define the search interval $[l_s - r, l_s + r]$.
r has to be a positive integer. If this option is not selected, then r is set to 30.

`-pptrprob` p_R

Specify the emission probability of a purine base (`A`/`G`) inside a PPT. This value must be a valid probability value ($0 \leq p_R \leq 1$). If this option is not set, then p_R is set to 0.97.

`-pptyprob` p_Y

Specify the emission probability of a pyrimidine base (`T`/`C`) inside a PPT. This value must be a valid probability value ($0 \leq p_Y \leq 1$). If this option is not set, then p_Y is set to 0.03.

`-pptaprob` p_A

Specify the background emission probability of an `A` base outside of PPT and U-box regions. This value must be a valid probability value ($0 \leq p_A \leq 1$). If this option is not set, then p_A is set to 0.25.

`-pptcprob` p_C

Specify the background emission probability of a `C` base outside of PPT and U-box regions. This value must be a valid probability value ($0 \leq p_C \leq 1$). If this option is not set, then p_C is set to 0.25.

`-pptgprob` p_G

Specify the background emission probability of a `G` base outside of PPT and U-box regions. This value must be a valid probability value ($0 \leq p_G \leq 1$). If this option is not set, then p_G is set to 0.25.

`-ppttprob` p_T

Specify the background emission probability of a `T` base outside of PPT and U-box regions. This value must be a valid probability value ($0 \leq p_T \leq 1$). If this option is not set, then p_T is set to 0.25.

Note that $\sum_{x \in \{A,C,G,T\}} p_x = 1$ must hold if the `-ppt{a,c,g,t}prob` options are used.

`-pptuprob` p_U

Specify the emission probability of a thymine base (`T`) inside a U-box. This value must be a valid probability value ($0 < p_U \leq 1$). All other emission probabilities are calculated as uniform probabilities $p_x = \frac{1-p_U}{3}$ for all $x \in \{A, C, G\}$. If this option is not set, then p_U is set to 0.91.

3.5 PBS options

`-trnas` *trnafile*

Specify a file in multiple FASTA format to be used as a tRNA library that is aligned to the area around the end of the 5' LTR to find a putative PBS. The header of each sequence in this file should reflect the encoded amino acid and codon. If this option is not selected, then PBS searching is skipped altogether.

`-pbsalilen` L_{min} L_{max}

Specify the minimum and maximum allowed lengths for PBS/tRNA alignments. If a local alignment shorter than L_{min} or longer than L_{max} is found, it will be skipped.

L_{min} and L_{min} have to be positive integers. If this option is not selected, then L_{min} is set to 11, L_{max} to 30.

`-pbsoffset` L_{min} L_{max}

Specify the minimum and maximum allowed distance between the start of the PBS and the 3' end of the 5' LTR. If a local alignment with such a distance smaller than L_{min} or greater than L_{max} is found, it will be skipped.

L_{min} and L_{min} have to be positive integers. If this option is not selected, then L_{min} is set to 0, L_{max} to 5.

`-pbstrnaoffset` L_{min} L_{max}

Specify the minimum and maximum allowed PBS/tRNA alignment offsets from the 3' end of the tRNA. If a local alignment with an offset smaller than L_{min} or greater than L_{max} is found, it will be skipped.

L_{min} and L_{min} have to be positive integers. If this option is not selected, then L_{min} is set to 0, L_{max} to 5.

`-pptmaxedist` d

Specify the maximal allowed unit edit distance in a local PBS/tRNA alignment. All optimal local alignments with a unit edit distance $> d$ will be skipped. Set this to 0 to accept exact matches only. It is also possible to fine-tune the results by adjusting the match/mismatch/indelscores used in the Smith-Waterman alignment (see below).
d has to be a positive integer. If this option is not selected, then d is set to 1.

`-pbsradius` r

Specify the area around the 5' LTR end (l_e) to be searched for a PBS, in other words, define the search interval $[l_e - r, l_e + r]$.
r has to be a positive integer. If this option is not selected, then r is set to 30.

`-pbsmatchscore` $score_m$

Specify the match score used in the PBS/tRNA Smith-Waterman alignment. Lower this value to discourage matches, increase this value to prefer matches.
$score_m$ has to be an integer. If this option is not selected, then $score_m$ is set to 5.

`-pbsmismatchscore` $score_{mm}$

Specify the mismatch score used in the PBS/tRNA Smith-Waterman alignment. Lower this value to discourage mismatches, increase this value to prefer mismatches.
$score_{mm}$ has to be an integer. If this option is not selected, then $score_{mm}$ is set to -10.

`-pbsdeletionscore` $score_d$

Specify the deletion score used in the PBS/tRNA Smith-Waterman alignment. Lower this value to discourage deletions, increase this value to prefer deletions.
$score_d$ has to be an integer. If this option is not selected, then $score_d$ is set to -20.

`-pbsinsertionscore` $score_i$

Specify the insertion score used in the PBS/tRNA Smith-Waterman alignment. Lower this value to discourage insertions, increase this value to prefer insertions.
$score_i$ has to be an integer. If this option is not selected, then $score_i$ is set to -20.

3.6 Protein domain search options

`-hmms` $hmmfile_1, hmmfile_2, \ldots, hmmfile_n$

Specify a list of pHMM files in HMMER2 format. The pHMMs must be defined for the amino acid alphabet and follow the Plan7 specification. For example, pHMMs defining protein domains taken from the Pfam database can be used here. Every file must exist and be readable, otherwise an error is reported. If this option is not given, protein domain searching is skipped altogether. Please note that shell globbing can be used here to specify large numbers of files, e.g. by using wildcards.

`-pdomevalcutoff` c

Specify the E-value cutoff for HMMER searches. All hits that fail to meet this maximal e-value requirement are discarded.
c has to be a probability ($0 \leq c \leq 1$). If this option is not selected, then c is set to 10^{-6}.

4 Example

This section describes an example session with *LTRdigest*. For simplicity, we assume that a *LTRharvest* run on a genome has already been performed and produced a `ltrs.gff3` file containing the basic GFF3 annotation describing LTR positions. The enhanced suffix array (ESA) index, any sequence output or the tabular standard output from *LTRharvest* will not be needed. Instead, we require an encoded sequence representation of the original input sequence, in this example called `genome.fas`. This can be created using the *GenomeTools* `gt suffixerator` tool as follows:

```
gt suffixerator -tis -des -dna -ssp -db genome.fas -indexname genome.fas
```

This step creates additional files with the suffixes `.al1`, `.des`, `.esq`, `.prj` and `.ssp` in the sequence file directory.

We will also assume that the HMM files to be used are called `HMM1.hmm`, `HMM2.hmm` and `HMM3.hmm`, we will be using a tRNA library called `tRNA.fas`, and we are running *LTRdigest* on a dual-core system. We also want to restrict the PBS offset from the LTR end to a maximum of 3 nucleotides, while we want the PPT length to be at least 10 nucleotides. Finally, we want all output such as sequences written to files beginning with “mygenome-ltrs”.

First, sort the GFF3 output by position:

```
gt gff3 -sort ltrs.gff3 > ltrs_sorted.gff3
```

Then, the *LTRdigest* run can be started with the following command line using the parameters above:

```
gt -j 2 ltrdigest -pptlen 10 30 -pbsoffset 0 3 -trnas tRNA.fas
-hmms HMM*.hmm -outfileprefix mygenome-ltrs ltrs_sorted.gff3 genome.fas
> ltrs_after_ltrdigest.gff3
```

No screen output (except possible error messages) is produced since the GFF3 output on stdout is redirected to a file. Additionally, the files `mygenome-ltrs_conditions.csv`, `mygenome--ltrs_3ltr.fas`, `mygenome-ltrs_5ltr.fas`, `mygenome-ltrs_ppt.fas`, `mygenome--ltrs_pbs.fas`, `mygenome-ltrs_tabout.csv` and one FASTA file for each of the HMM models will be created and updated during the computation. As the files are buffered, it may take a while before first output to these files can be observed.

The calculation may be restarted with the same or changed parameters afterwards, overwriting the output files in the process. If it is desired to keep sequences etc. from each run, keep in mind to assign specific `-outfileprefix` values to each run.

References

[1] D. Ellinghaus, S. Kurtz, and U. Willhoeft. *LTRharvest*, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9:18, 2008.

[2] L. Stein. Generic Feature Format Version 3. `http://www.sequenceontology.org/gff3.shtml`.

[3] G. Gremme. GenomeTools. `http://genometools.org`.

[4] R.D. Finn, J. Mistry, B. Schuster-Boeckler, S. Griffiths-Jones, V. Hollich, T. Lassmann,S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L.L. Sonnhammer and A. Bateman. Pfam: clans, web tools and services. *Nucleic Acids Research* (Database Issue), 34:D247-D251, 2006.

[5] S.R. Eddy. HMMER: Biosequence analysis using profile hidden Markov models. `http://hmmer.janelia.org`.