LEXOGEN
The RNA Experts

Accurate Analysis of RNA-Seq Data

RNA-Seq Data Analysis Software

User Guide

023UG050V0220

FOR RESEARCH USE ONLY. NOT INTENDED FOR DIAGNOSTIC OR THERAPEUTIC USE.
INFORMATION IN THIS DOCUMENT IS SUBJECT TO CHANGE WITHOUT NOTICE.
Lexogen does not assume any responsibility for errors that may appear in this document.

PATENTS AND TRADEMARKS
The Mix-Square algorithm is covered by issued and/or pending patents. Mix-Square is a trademark of Lexogen. Lexogen is a registered trademark (EU, CH, US, CN, AU, NO, BR).

All other brands and names contained in this user guide are the property of their respective owners.

Lexogen does not assume responsibility for violations or patent infringements that may occur with the use of its products.

LIABILITY AND LIMITED USE LABEL LICENSE: FOR RESEARCH USE ONLY
This document is proprietary to Lexogen. The Mix-Square software is intended for use in research and development only. It needs to be handled by qualified and experienced personnel to ensure proper use. Lexogen does not assume liability for any damage caused by the improper use or the failure to read and explicitly follow this user guide. Furthermore, Lexogen does not assume warranty for merchant-ability or suitability of the product for a particular purpose.

Download of the software does not convey the right to resell, distribute, further sublicense, repackage, or modify the product or any of its components. This document and its contents shall not be used or distributed for any other purpose and/or otherwise communicated, disclosed, or reproduced in any way without the prior written consent of Lexogen.

For information on purchasing additional rights or a license for use other than research, please contact Lexogen.

WARRANTY
Lexogen is committed to providing excellent products. Lexogen warrants that the product performs to the standards described in this user guide. Should this product fail to meet these standards due to any reason other than misuse or improper handling Lexogen will replace the product free of charge or issue a credit for the purchase price. Lexogen does not provide any warranty if product components are replaced with substitutes. Under no circumstances shall the liability of this warranty exceed the purchase price of this product.

LITERATURE CITATION
When describing a procedure for publication using this product, please refer to it as Lexogen's Mix-Square software.

We reserve the right to change, alter, or modify any product without notice to enhance its performance.

CONTACT INFORMATION

Lexogen GmbH
Campus Vienna Biocenter 5
1030 Vienna, Austria
www.lexogen.com
E-mail: support@lexogen.com

Support
E-mail: support@lexogen.com
Tel. +43 (0) 1 3451212-41
Fax. +43 (0) 1 3451212-99

Table of Contents

LEXOGEN
The RNA Experts

1. Introduction

This manual describes the system requirements of the Mix^2 software. In addition, command line options of the software are discussed as well as its input and output format. The Mix^2 version is updated yearly and the current version is applicable only for the current year.

For further questions related to the Mix^2 software please contact support@lexogen.com.

2. Requirements

The Mix^2 software runs on Linux x64 distributions.

The Mix^2 software has been tested on:

- Ubuntu 18.04
- Ubuntu 20.04

If you encounter any problems when running the Mix^2 software, please contact us at support@lexogen.com.

3. Running Mix²

The Mix² software can be run from the command line as follows:
./mix-square [options] <arguments>

Options

General Options:	
-h [--help]	Describe options.
-G [--GTF] arg	Directory of the reference annotation file. Please refer at Mix ² Input section.
-B [--BAM] arg	Directory of the RNA-Seq read alignments in BAM format. SAM file format is not supported. The alignments need to be sorted by their leftmost coordinates.
-o [--output-dir] arg	Sets the output directory which the results will be saved to. The default is a directory called "output" in the current working directory. If the path to output-dir is relative it will be generated within the current working directory.
-p [--threads] arg	Number of threads to be used for the estimation process.
Clustering Options:	
-R [--results-dir] arg	Directory which contains mix-square run results.
-n [--nr-clusters] arg	Maximum number of clusters to be produced. Default: 5
--min-trans-len arg	Transcripts shorter than minimum transcript length are excluded from the clustering process. Default: 200bp
--min-trans-frags arg	Transcripts which has less fragments than min-trans-frags are excluded from the clustering process. Default: 100
Advanced Abundance Estimation Options:	
-x [--max-total-frags] arg	Sets the maximum number of fragments in a locus. A locus which has more fragments than the maximum number is skipped. Genes skipped can be found in genes_skipped.list. Default: 5000000
-M [--max-comp-frags] arg	Sets the maximum number of valid fragments in a locus. A locus which has more valid fragments than the maximum number is skipped. Genes skipped can be found in genes_skipped.list. Default: 5000000
-m [--min-comp-frags] arg	Sets the minimum number of valid fragments in a locus. A locus with less valid fragments than the minimum number is skipped. Genes skipped can be found in genes_skipped.list. Default: 1
-q [--min-param-diff] arg	Sets the minimum parameter difference between 2 iterations. Default: 1e-5
-i [--nr-iterations] arg	If the minimum Log Likelihood condition is not reached then the EM algorithm will terminate if the maximum number of iterations is reached. Default: 500
-T [--likelihood-threshold] arg	Sets the minimum log likelihood difference between 2 iterations. If the log likelihood difference between two iterations is below this value, then the EM algorithm terminates. Default: 0.5
-L [--genes-list] arg	A file containing gene IDs which are included or excluded in the experiment.

-b [--blocks] arg	This number defines how many mixture components are used to model the bias of fragment startsites. This number can be understood as the 'resolution' of the pdf. Accepted values are natural numbers from 1 to 10. The default is 3.
-e [--exclude-genes]	With this option, mix-square model excludes the genes which are specified in the genes list file via the –L option.
-t [--global-tying]	With this option, global tying is turned on which means that all the isoforms of a gene share the same parameters for the fragment start distributions This options should only be used if the relative fragment start distributions of the isoforms within a gene can be expected to have a similar shape, or in case of data sparsity.
-l [--log-files]	Turns on estimation process logging. An individual file is created for each gene.
Advanced Program Behavior Options:	
-s [--EULA]	With this option, you can view the EULA.
-r [--ignore]	With this option, the warnings, which may be shown while using the max-frags-locus option, are turned off.
-d [--debug]	This option turns on the debugging mode. This should only be used to obtain diagnostic information when facing problems with mix-square.
-D [--library-strandedness]	The strandedness of the library can be fwd/rev/none. Default: none
--quiet	Uses a simple progress bar.

4. Mix² Input

GTF (gene transfer) format and a file which contains the alignments in BAM (binary SAM) format.

The structure of the annotation file should be like:
<seqname> <feature> <start> <end> <strand> [attributes]

Field number	Field name	Example	Description
1	seqname	19	The name of the sequence. Chromosome ID or contig ID.
2	feature	Exon	Record type which can be "CDS", "start codon", "stop codon", "intron", "exon", "transcript" etc. All the record types are ignored except "exon".
3	start	51456206	Start coordinate of the feature, in this case the start coordinate of the exon.
4	end	51456321	End coordinate of the feature, in this case the end coordinate of the exon.
5	strand	+	The strand which exon comes from. Should be "-" or "+".

Attribute number	Attribute name	Example	Description
1	gene_id	ENSG00000167754	A globally unique identifier for the genomic locus of the transcript.
2	transcript_id	ENST00000391809	A globally unique identifier for the transcript.
3	gene_name	KLK5	The name of the gene.
4	end	51456321	End coordinate of the feature, in this case the end coordinate of the exon.

If one of the above fields/attributes is missing, the entry is skipped.

If an experiment needed to be done on a specific list of genes, then -L option could be used. That option expects a file which includes the gene IDs (one gene ID per line).

A typical list should be as below:

ENSG00000167754
ENSG00000187999
ENSG00000123437
ENSG00000145310

Optionally, the -e flag can be used to exclude the genes specified in the genes-list.

5. Mix² Output

5.1. BAM Index File

Mix² will produce an index file for the input BAM file if no such index file is present.

5.2. Genes_summary file

Field number	Field name	Example	Description
1	gene_ID	ENSG00000167754	A globally unique identifier for the genomic locus of the transcript.
2	gene_name	KLK5	The name of the gene.
3	locus	19:51446559-51456349	The locus which the gene is referenced to. Chromosome ID:start coordinate - end coordinate.
4	frags_locus	20000	Number of fragments in the specified locus.
5	frags_expt	200000000	Total number of fragments in the experiment.
6	FPKM_THN	452420.36	FPKM total hits norm. FPKM_THN is calculated counting all fragments including those which are not compatible with any reference transcript. FPKM_THN is calculated continuously during the experiment.
7	comp_frags_locus	10000	Number of fragments in the specified locus, which are compatible with a reference transcript. comp_frags_locus should be used to calculate isoform row counts for differential expression analysis. comp_frags_locus should be used to calculate isoform row counts for differential expression analysis.
8	comp_frags_expt	100000000	Total number of fragments in the experiment, which are compatible with a reference transcript.
9	FPKM_CHN	904840.73	FPKM compatible hits norm. FPKM_CHN is calculated counting only the fragments, which are compatible with a reference transcript. FPKM_CHN is calculated at the end of the experiment. FPKM_CHN should be used for differential expression analysis.
10	status	OK	Whether the estimation process was successful or not.

5.3. Transcripts_summary file

Field number	Field name	Example	Description
1	tracking_ID	ENST00000391809	A unique identifier for the transcript.
2	gene_ID	ENSG00000167754	A globally unique identifier for the genomic locus of the transcript.
3	gene_name	KLK5	The name of the gene.
4	locus	19:51446559-51456349	The locus which the gene is referenced to. Chromosome ID:start coordinate - end coordinate.
5	length	1405	Transcript length in basepairs.
6	fragment_validity_coverage	0.93	Validity coverage for the specified transcript.
7	abundance	0.23416	Estimated relative abundance.
8	frags_locus	20000	Number of fragments in the specified locus.
9	frags_expt	200000000	Total number of fragments in the experiment.
10	FPKM_THN	452420.36	FPKM total hits norm. FPKM_THN is calculated counting all fragments including those, which are not compatible with any reference transcript. FPKM_THN is calculated continuously during the experiment.
11	comp_frags_locus	10000	Number of fragments in the specified locus, which are compatible with a reference transcript.
12	comp_frags_expt	100000000	Total number of fragments in the experiment, which are compatible with a reference transcript.
13	FPKM_CHN	904840.73	FPKM compatible hits norm. FPKM_CHN is calculated counting only the fragments, which are compatible with a reference transcript. FPKM_CHN is calculated at the end of the experiment.
14	nr_mixture_comp	3	Number of the mixture components used in the experiment.
15	mean_N	376	These values are used for recalculation of the estimated fragment distributions for later use with the clustering algorithm.
16	beta_N	0.359	These values are used for recalculation of the estimated fragment distributions for later use with the clustering algorithm.

6. Test Case

The distribution of the Mix^2 software contains a small test set of artificial data, which enables the user to try out the basic functionality of the software. The example directory contains a GTF file for gene KLK5 and a sorted BAM file.

Here are two examples for how Mix^2 can be run from the command line on the test data:

- **./mix-square -G example/KLK5.gtf -B example/KLK5.sorted.bam**
  In order to run Mix^2 the above parameters are required at least. Since no output directory is specified, the results are saved in the current working directory under a directory called output.

- **./mix-square -G example/KLK5.gtf -B example/KLK5.sorted.bam -b 3 -t -o test-example-data**
  In this example the output directory has been specified as well as the number of blocks. In addition, the global tying option has been switched on, which means that the fragment start distributions of all isoforms within a gene share the same set of parameters.

- **./mix-square -R /home/user/mix2_results_dir/ -n 5 --min-trans-len 250 --min-trans-frags 50**
  This command starts the clustering procedure. The clustering algorithm restricts the numbers of the clusters to 5 and filters out the transcripts which are shorter than 250 bp as well as the transcripts which have less fragments than 50.

7. Appendix A: Revision History

Publication No. / Revision Date	Change	Page
023UG050V0220 Nov. 14, 2024	Legal terms and conditions statements updated.	2
	Tested operating systems updated.	4
	Advanced program behavior options updated.	6
	Chapters: Mix² License and Investigation of the Positional Bias were deleted.	
023UG050V0210 Dec. 28, 2016	Section "Clustering Options" added.	8
	Chapters "6.4. Error_log file" and "6.5. Genes_skipped file" were deleted.	12
	New Chapter "7. Investigation of the Positional Bias" added.	13
	Additional command line added.	17
023UG050V0100 Jan. 26, 2015	Initial Release.	

LEXOGEN
The RNA Experts

Associated Products:

025, 050, 051, 141 (SIRVs Spike-in RNA Variant Control Mixes)
144 (RiboCop rRNA Depletion Kits for Human/Mouse/Rat)
145 (RiboCop rRNA Depletion Kits for Human/Mouse/Rat Plus Globin)
157 (Poly(A) RNA Selection Kit)
171-176 (CORALL RNA-Seq V2 Library Prep Kits)
190 (RiboCop rRNA Depletion Kits for Yeast)
237 (RiboCop rRNA Depletion Kits for Plants)
241 (RiboCop rRNA Depletion Kits for Fish)

Mix^2 User Guide

Lexogen GmbH
Campus Vienna Biocenter 5
1030 Vienna, Austria
Telephone: +43 (0) 1 345 1212-41
Fax: +43 (0) 1 345 1212-99
E-mail: support@lexogen.com

Lexogen, Inc.
51 Autumn Pond Park
Greenland, NH 03840, USA
Telephone: +1-603-431-4300
Fax: +1-603-431-4333
www.lexogen.com