LEXOGEN

Enabling complete transcriptome sequencing

Accurate Analysis of RNA-Seq Data

RNA-Seq data analysis software

Scientific Evaluation Release

User Guide

023UG133V0100

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, USA).

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.

The purchase of the product 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 and cite Tuerk A, Wiktorin G, Güler S (2017) Mixture models reveal multiple positional bias types in RNA-Seq data and lead to accurate transcript concentration estimates. PLOS Computational Biology 13(5): e1005515. https://doi.org/10.1371/journal.pcbi.1005515

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: info@lexogen.com

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

Table of Contents

LEXOGEN

Enabling complete transcriptome sequencing

1. Introduction

This manual describes system requirements, command line options as well as input and output format of the Mix^2 software. For further questions related to the Mix^2 software please contact bioinfo@lexogen.com.

2. Requirements

The Mix² software runs on Linux x64 distributions, as well as on Mac OSX and Windows. Please note that the current version of our software will expire on 31.12.2017. We will, however, release regular updates of our software valid for at least 9 months from the date of issue before the previous version expires.

The Mix² software has been tested on:

- Linux distributions
  - Ubuntu 12.04+ Desktop x64
  - Ubuntu 12.04 Server x64
  - openSUSE 13.2 Desktop x64
  - openSUSE 12 Server x64
  - Linux Mint 17.1 Desktop x64
  - Fedora Live 20 Desktop x64
  - CentOS 7.0 Desktop x64
- Mac OSX 10.10 (Yosemite), Mac OSX 10.12 (Sierra)
- Windows 8, Windows 10

If you encounter any problems when running the Mix² software, please contact us at bioinfo@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	Reference annotation file in GTF format.
-B [--BAM] arg	Alignment file in BAM format. SAM file format is not supported. The alignments need to be sorted by their leftmost coordinate.
-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.
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	Sets the number of mixture components in the positional bias model of fragment start sites. Accepted values are natural numbers from 1 to 10. The default is 3.
-e [--exclude-genes]	Exclude genes specified via the -L option in the gene list file.
-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 option 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.

Other Options:	
-s [--EULA]	Prints the end-user license agreement.
-r [--ignore]	With this option, the warnings, which may be shown while using the max-frags-locus option, are turned off.
--quiet	Minimize diagnostic output.

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 not the complete genome is to be processed, the -L option can be used to select a subset of genes. The -L option expects a file containing one gene ID per line.

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

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.

6. Test Case

This distribution of the Mix² 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.

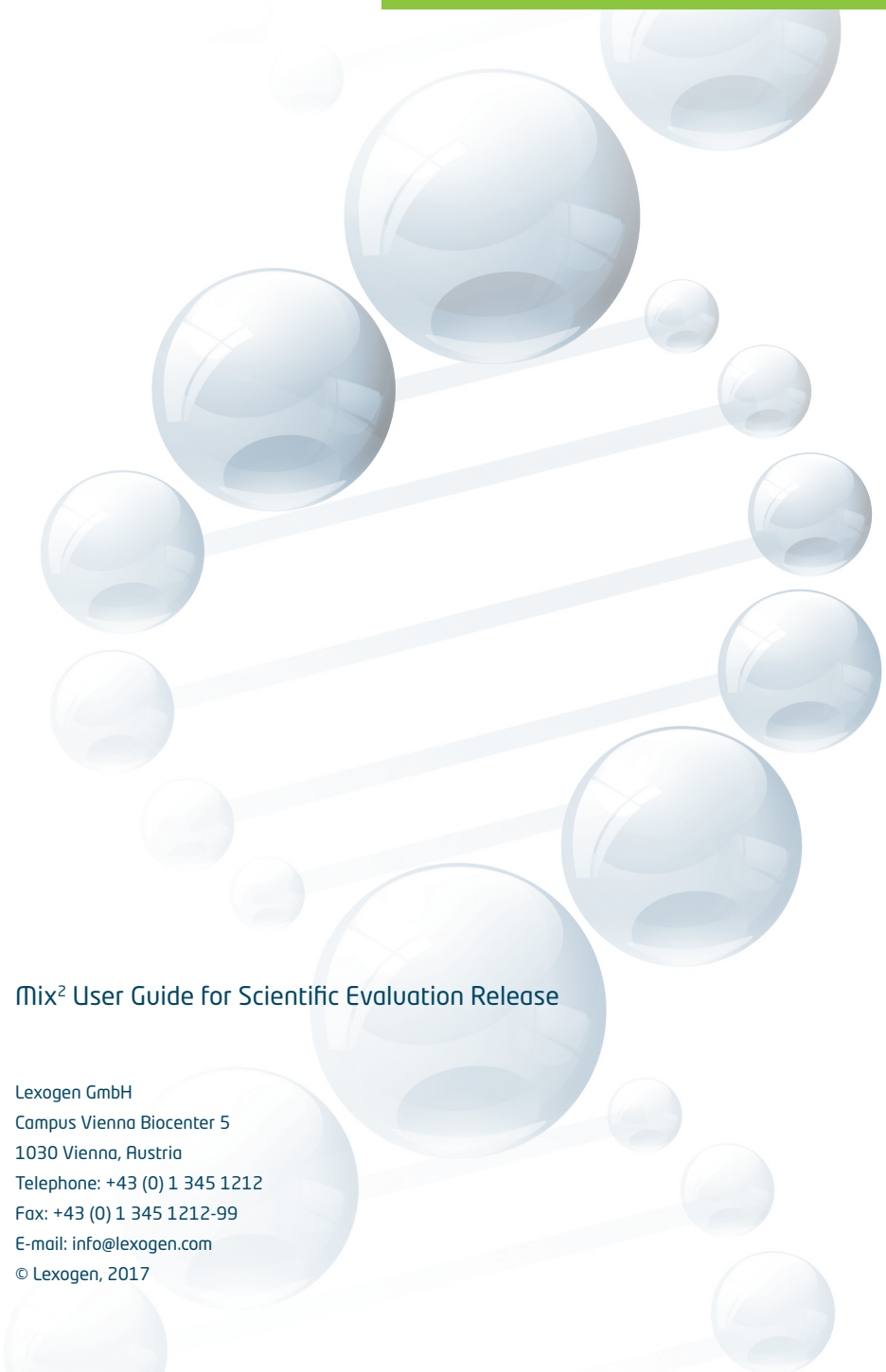
Here are two examples for how Mix² can be run from the command line on the test data:

- ./mix-square -G example/KLK5.gtf -B example/KLK5.sorted.bam
  - In order to run Mix² 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.

LEXOGEN

Enabling complete transcriptome sequencing



Mix^2 User Guide for Scientific Evaluation Release

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