LOX: inferring level of expression from diverse methods of census sequencing

Version 1.8, February 2012

Zhang Zhang, Francesc Lopez-Giraldez, and

Jeffrey P. Townsend

Content

1 Introduction

LOX is a software package that employs Markov Chain Monte Carlo to estimate the Level Of gene eXpression from high-throughput expressed sequence data sets with multiple treatments or samples. Unlike most analyses, LOX incorporates a gene bias model that facilitates integration of diversetranscriptomic sequencing data that arises when transcriptomic data have been produced using diverse experimental methodologies. LOX integrates over all sequence count tallies normalized by total expressed sequence count to provide expression levels for each gene relative to all treatments as well as Bayesian credible intervals.

2 Copyright & License

3 Installation

For high efficiency and compatibility with more platforms, LOX is written in standard C++. You can download from the LOX webpage at http://www.yale.edu/townsend/software.html the newest package, normally named LOXXXX.tar.gz (XXX stands for the version).

3.1 Compiled Executables

Executables have been precompiled for Linux/Unix/Mac/Windows. Please unpack the package of LOXXXX.tar.gz (see below) and then you will find compiled executables in the folder of ××LOXXXX/bin/××.

3.2 Linux/Unix/Mac/Windows

For compilation on your specific platform, please follow the steps below.

☒☒ Unpack the package of LOXXXX.tar.gz by the following commands.

```
tar –zxf LOXXXX.tar.gz
```

☒☒ If you use other Linux/Unix OS, you have to compile the program in the source codes folder with the help of g++/gcc compiler.

```
cd LOXXXX/src
make
```

That►►s it. Then you can find an executable named ►►LOX►► in this folder.

4 Setting Parameters

LOX allows the user to customize parameters. The following are the parameters►► settings, which can also be found by typing ►►LOX -h►►.

►► -i►►►►►►►	input file name [string, required]
►► -m►►►►►►►	whether to output inferred methodology biases [integer, optional], {0:No || 1:Yes}, default = 0
►► -r►►►►►►►	whether to show maximum likelihood estimates as well as posterior means [integer, optional], {0:No || 1:Yes], default = 0
►► -g►►►►►►►	whether to output posterior distribution for each gene [integer, optional], {0:No || 1:Yes], default = 0
►► -s►►►►►►►	set chain sample size [integer, optional], default = 1000
►► -b►►►►►►►	set burnin size [integer, optional], default = 2000
►► -p►►►►►►►	set sampling period [integer, optional], default = 20
►► -d►►►►►►►	set tuning depth [integer, optional], default = 8
►► -l►►►►►►►►►	set lower acceptance rate [double, optional], default = 0.15
►► -u►►►►►►►	set upper acceptance rate [double, optional], default = 0.50

5 Format of the Input File

The input data for LOX are expression counts of multiple genes, under one or more treatments and with one or more methodologies. To ease data input, LOX accepts tab-delimited text file with three header rows. Input row one is set aside for user-customized information, row two contains text codes designating the methodology applied, and row three includes text codes designating the treatment type. The subsequent rows contain gene ID, gene name, and expression counts under corresponding treatments and methodologies. Please note that while the first row is for user use only, the second two header rows are used by LOX to calculate the expression levels. Thus, the methodology/treatment abbreviation used must be identical for each identical methodology/treatment type.

An example data file (named ►►LOXinputfile.txt►►) containing 5525 genes and its results file accompanies the LOX package. For instance, the input information for gene Q0085 is listed below.

		x1	x2	x3	x4
		oligodT	oligodT	MTP	MTP
ID	Common Name	MM	SC	MM	SC
Q0085	ATP6	4327	2026	53	18

The first row ××x1××, ××x2××, ××x3××, ××x4×× can be set to contain information (usually sample numerals or labels) customized by each user. From the second and third rows, it can be seen that the gene ID and common name are ××Q0085×× and ××ATP6××, respectively, and that two experiments (with different priming strategies, oligodT and MTP) were cells grown in Minimal Media (denoted as MM) and Synthetic Complete (denoted as SC). ××Therefore, the treatments are ××MM×× and ××SC××, and the methodologies are ××oligodT×× and ××MTP××. The fourth row contains expression counts under a combination of treatments and methodologies. For example, 4327 under ××MM×× and ××oligodT××. More information can be found in the folder ××LOXXXX/example/××.

Additionally, in order to facilitate use of LOX, a basic pipeline for generating the LOX input file from raw sequence reads and genome features of interest is provided in the LOX package. More details can be found in the folder of ××LOXXXX/how to generate input file/××.

6 Output

LOX output is in the form of a tab-delimited text file with one header row. Each row thereafter displays the results for a single gene, including columns with gene ID and gene name, the estimate of expression level for each treatment (the median of the posterior distribution), ninety-five percent Bayesian credible intervals (the additions and subtractions to make upper and lower bounds) for that estimate, the stationary acceptance rates for the MCMC steps, a boolean value indicating whether those rates are within an acceptable range (by default, 0.15 to 0.50), and the best log posterior probability. Bayesian P-values for differential expression are also reported regarding all pairs of treatments, and may be used in conjunction with effect sizes and credible intervals to rank genes by their differential expression. Lastly, optional columns can be output that report the methodological effects and the parameter estimates at the peak of maximum likelihood.

Taking gene Q0085 as an example, the following are extracted from the example output file ××LOXinputfile.txt.lox×× in the folder of ××LOXXXX/example/××.

ID	Common Name	MM	SC	(-)97.5%[MM]	(-)97.5%[SC]	(+)97.5%[MM]	(+)97.5%[SC]	Acceptance Ratio	Density of Best Log-Likelihood
Q0085	ATP6	2.13815	1	0.0344703	0.0344273	0.0344644	0.0346123	0.158959 (TRUE)	-71704.7

The description for each column is listed as follows.

- ID, Common Name: Gene ID and its common name
××××
- MM, SC: Relative expression levels of the treatment (the median of the posterior distribution). In the model, the value 2.13815 under column title ××MM×× represents the relative expression level of treatment ××MM×× to treatment ××SC××, as determined by the ratios of the p_{ik}, treatment i for a given gene k,

where *k* = ��Q0085��, and where, for this experiment, *i* can take on values ��MM�� or ��SC��.

- (-)97.5%[MM], (-)97.5%[SC]: Subtractions to make 95% lower bounds of the Bayesian credible intervals on the relative expression level for the ��MM�� and ��SC�� treatments. For gene Q0085, the expression level of treatment MM, 2.13815, is the median of the posterior distribution, and (-)97.5%[MM], 0.0344703, indicates the difference between the median and the lower bound. It is not a log transformation.

- (+)97.5%[MM], (+)97.5%[SC]: Additions to make 95% upper bounds of the Bayesian credible intervals. For gene Q0085, (+)97.5%[MM] indicates the difference between the median and the upper bound. It is not a log transformation.

- Acceptance Ratio: Stationary acceptance rates for the Monte Carlo steps and a boolean value (true or false) indicating whether those rates are within an acceptable range (by default 0.15 to 0.50; which can be changeable by the user).

- Density of Best Log-Likelihood: Best log posterior probability.

Similar to treatments, the output data for methodologies can be generated by setting the parameter ��-m 1��. The filename will be same as the original input filename with the characters ��.met�� appended.

Bayesian P-values for differential expression are also reported regarding all pairs of treatments, and may be used in conjunction with effect sizes and credible intervals to rank genes by their differential expression. For each treatment, there are P-values for the treatment against the rest treatments, and thus, the P-values filename will be same as the original input file name with the treatment name plus the characters ��.pvalue�� appended.

Posterior distributions for all examined genes can be outputted by setting the parameter ��-g 1��. Each gene will have a file containing the posterior distribution data for each treatment and each method and the corresponding file name will be same as the gene ID. As a result, all these files will be stored into a folder named ��Posterior_�� plus the original input file name appended.

Optional columns can be output that report the methodological effects and the parameter estimates at the peak of maximum likelihood.

Please see details in the folder of ��LOXXXX/example/��. In this folder, a LOX input file is provided (LOXinputfile.txt) containing biological data of 5525 yeast genes for the combination of 2 methodologies to obtain the transcriptome (oligodT and MTP) and the two treatments (MM and SC) mentioned earlier. The file ��command.txt�� contains the command line to execute the software with the example file. The corresponding output files are also provided (LOXinputfile.txt.lox and LOXinputfile.txt.met).

7 Acknowledgements

We thank reviewers for their constructive comments to further improve this work. We also thank Zheng Wang and Andrea Hodgins-Davis for valuable discussions and many users as well as members of the Townsend lab for reporting bugs and sending comments.

8 Contact

Please send bugs or advice to Dr. Zhang Zhang (Zhang.Zhang@Kaust.edu.sa) or Dr. Jeffrey Townsend (Jeffrey.Townsend@Yale.edu).