G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2018;16(4):e36
https://doi.org/10.5808/GI.2018.16.4.e36

APPLICATION NOTE

COEX-Seq: Convert a Variety of Measurements of Gene Expression in RNA-Seq

Sang Cheol Kim[1]*, Donghyeon Yu[2], Seong Beom Cho[1]

[1]Division of Bio-Medical Informatics, Center for Genome Science, National Institute of Health, Korea Centers for Disease Control and Prevention, Cheongju 28159, Korea, [2]Department of Statistics, Inha University, Incheon 22212, Korea

Next generation sequencing (NGS), a high-throughput DNA sequencing technology, is widely used for molecular biological studies. In NGS, RNA-sequencing (RNA-Seq), which is a short-read massively parallel sequencing, is a major quantitative transcriptome tool for different transcriptome studies. To utilize the RNA-Seq data, various quantification and analysis methods have been developed to solve specific research goals, including identification of differentially expressed genes and detection of novel transcripts. Because of the accumulation of RNA-Seq data in the public databases, there is a demand for integrative analysis. However, the available RNA-Seq data are stored in different formats such as read count, transcripts per million, and fragments per kilobase million. This hinders the integrative analysis of the RNA-Seq data. To solve this problem, we have developed a web-based application using Shiny, COEX-seq (Convert a Variety of Measurements of Gene Expression in RNA-Seq) that easily converts data in a variety of measurement formats of gene expression used in most bioinformatic tools for RNA-Seq. It provides a workflow that includes loading data set, selecting measurement formats of gene expression, and identifying gene names. COEX-seq is freely available for academic purposes and can be run on Windows, Mac OS, and Linux operating systems. Source code, sample data sets, and supplementary documentation are available as well.

Keywords: integrative analysis, measurements of gene expression, RNA-Seq, web-based application using Shiny

Availability: https://github.com/kimsc77/COEX-seq.

Introduction

Next generation sequencing (NGS), a high-throughput DNA sequencing technology, is widely used for molecular biological studies. In NGS, the RNA-sequencing (RNA-Seq), which is a short-read massively parallel sequencing, is a major quantitative transcriptome tool for many types of transcriptome studies, such as mRNA and miRNA. To utilize the RNA-Seq data, various quantification and analysis methods have been developed to solve specific research goals, including identifying differentially expressed genes and detection of novel transcripts. With the accumulation of RNA-Seq data in the public databases, there is a demand for integrative analysis, and its development is an ongoing challenge [1]. However, the available RNA-Seq data are stored in different formats. In particular, in the public databases such as GEO (Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/), ArrayEXpress (https://www.ebi.ac.uk/arrayexpress/), The Cancer Genome Atlas (TCGA), the quantitative measurements of the processed RNA-Seq data sets are available in various formats, which are not unified. The following are different formats of measurements provided by the public databases. Read count routinely refers to the number of reads that align to a particular region. Counts per million mapped reads are counts scaled by the number of sequenced fragments multiplied by one million. Transcripts per million (TPM) is a measurement of the proportion of transcripts in a pool of RNA. Reads per kilobase of exon per million reads mapped (RPKM) and the more generic fragments per kilobase million (FPKM), which substitutes reads in RPKM with fragments, are essentially the same measurements [2-5]. Utilizing different quantitative measurements provided by the public databases hinders the integrative analysis. Like recount2, a project is underway to provide results from

Received October 15, 2018; Accepted November 14, 2018; Published online December 28, 2018

***Corresponding author:** Tel: +82-43-719-8861, Fax: +82-43-719-8869, E-mail: sckim.knih@korea.kr

various databases through a single analytical pipeline [6]. To solve this problem, we aimed to develop a web-based application using Shiny, COEX-seq (COnvert a variety of measurements of gene EXpression in RNA-Seq) that easily converts data in a variety of measurement formats of gene expression used in most bioinformatic tools for RNA-Seq.

Results

Fig. 1 shows the graphical user interface for COEX-seq, which consists of two parts. The first is handling data and formats (loading data set, selecting measurements of gene expression, and identification of gene names). The second part reports the converted data and depicts their boxplots.

COEX-seq

COEX-seq is a web-based application using Shiny (Shiny; a web application framework for R) [7, 8] that converts a variety of measurements of gene expression in RNA-Seq experiments. It provides a workflow that includes loading data set, selecting measurements of gene expression, and identifying gene names. COEX-seq is freely available for academic purposes and can be run on Windows, Mac OS and Linux operating systems. Source code, sample data sets, and supplementary documentation are available at https://github.com/kimsc77/COEX-seq.

Measurements of gene expression

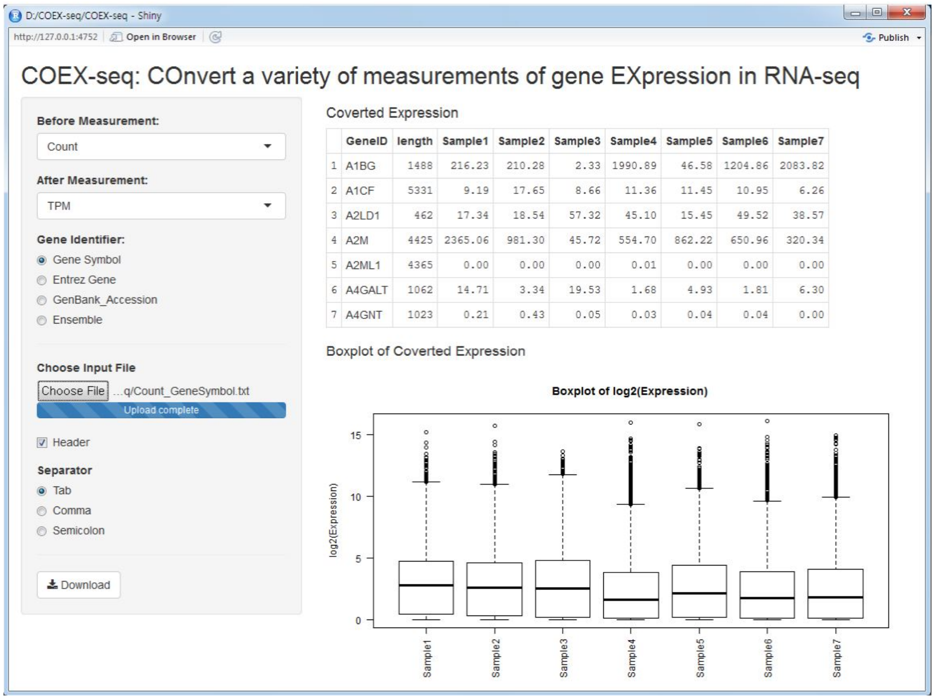
The following are the measurements of gene expression used in the public databases. Read counts are simply the number of reads overlapping a given feature such as a gene. Counts are often used by the methods identifying differentially expressed genes as a counting model, such as a Poisson or negative binomial, which naturally represents them. Fragments per kilobase of exon per million reads are much more complicated. Fragment means fragment of DNA; therefore, the two reads that comprise a paired-end read count as one. Per kilobase of exon means the count of fragments is then normalized by dividing by the total length of all exons in the gene (or transcript).

$$FPKM_g = \frac{RC_g}{\frac{RC_{pc}}{10^3} * \frac{L}{10^6}} = \frac{RC_g * 10^9}{RC_{pc} * L},$$

where, RC_g is number of reads mapped to the gene, RC_{pc} is number of reads mapped to all protein-coding (exon) genes, and L is length of the gene in base pairs. TPM is a measurement of the proportion of transcripts in mRNA. TPM is probably the most stable unit across experiments, although you still should not compare it across experiments.

$$TPM_g = \frac{RC_g}{L_g} * \frac{1}{\Sigma_j \frac{RC_j}{L_k}} * 10^6,$$

where, RC_g is the number of reads mapped for each gene and



D:/COEX-seq/COEX-seq - Shiny

http://127.0.0.1:4752 | Open in Browser

COEX-seq: COnvert a variety of measurements of gene EXpression in RNA-seq

Before Measurement: Count

After Measurement: TPM

Gene Identifier:
● Gene Symbol
○ Entrez Gene
○ GenBank_Accession
○ Ensemble

Choose Input File
Choose File ...q/Count_GeneSymbol.txt
Upload complete

☑ Header

Separator
● Tab
○ Comma
○ Semicolon

Download

Coverted Expression

	GeneID	length	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7
1	A1BG	1488	216.23	210.28	2.33	1990.89	46.58	1204.86	2083.82
2	A1CF	5331	9.19	17.65	8.66	11.36	11.45	10.95	6.26
3	A2LD1	462	17.34	18.54	57.32	45.10	15.45	49.52	38.57
4	A2M	4425	2365.06	981.30	45.72	554.70	862.22	650.96	320.34
5	A2ML1	4365	0.00	0.00	0.00	0.01	0.00	0.00	0.00
6	A4GALT	1062	14.71	3.34	19.53	1.68	4.93	1.81	6.30
7	A4GNT	1023	0.21	0.43	0.05	0.03	0.04	0.04	0.00

Boxplot of Coverted Expression

Fig. 1. Graphical user interface of COEX-seq (COnvert a variety of measurements of gene EXpression in RNA-Seq).

L_g is the length of the gene.

Relationship between TPM and FPKM

The relationship between TPM and FPKM is derived by Pachter (2011) [9] in review of transcript quantification method, using Eq. (10)–(13) in Pachter's study [9].

$$TPM_g =$$

$$\frac{RC_g}{L_g} * \frac{1}{\sum_j \frac{RC_j}{L_k}} * 10^6 \propto \frac{RC_g}{L_g * N} * \frac{1}{\sum_j \frac{RC_j}{L_k * N}} \propto \frac{RC_g}{L_g * N} * 10^9,$$

where $N = \sum_t RC_t$ is the total number of mapped reads. If FPKM is available, then TPM can be easily computed as

$$TPM_g = \left(\frac{FPKM_g}{\sum_j FPKM_g}\right) \times 10^6.$$

Discussion

Recently, based on the advances in NGS technologies, various quantification and analysis methods have been developed for the transcriptome studies. In addition, with the accumulation of RNA-Seq data sets in the public databases, there is a demand for integrative analysis; therefore, it has become an active research field. However, the available RNA-Seq data are stored in different formats such as read count, TPM, and FPKM. This hinders the integrative analysis of the RNA-Seq data. To solve this problem, we have developed a web-based application using Shiny, COEX-seq that easily converts data in a variety of measurement formats of gene expression used in most bioinformatic tools for RNA-Seq. Thus, COEX-seq is very useful to use with other analysis tools developed using R.

ORCID: Sang Cheol Kim: https://orcid.org/0000-0003-4389-9178; Donghyeon Yu: https://orcid.org/0000-0003-4519-8500; Seong Beom Cho: https://orcid.org/0000-0002-5734-1645

Author's contribution

Conceptualization: SCK, SBC
Funding acquisition: SCK
Methodology: SCK, DY
Writing – original draft: SCK
Writing – review & edition: SCK, DY, SBC

Conflicts of Interest

No potential conflicts of interest relevant to this article was reported.

Acknowledgments

This research was supported by a fund (2017-NI72002-00) by the Research of Korea Centers for Disease Control and Prevention.

References

1. Goncalves A, Tikhonov A, Brazma A, Kapushesky M. A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics* 2011;27:867-869.
2. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;15:R29.
3. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* 2012;131:281-285.
4. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 2010;26:493-500.
5. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;5:621-628.
6. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, *et al*. Reproducible RNA-seq analysis using recount2. *Nat Biotechnol* 2017;35:319-321.
7. RStudio Team. RStudio: Integrated Development for R. Boston: RStudio, Inc., 2015. Accessed 2018 Oct 1. Available from: http://www.rstudio.com/.
8. RStudio Team. shiny: Web Application Framework for R. Boston: RStudio, Inc., 2013. Accessed 2018 Oct 1. Available from: http://shiny.rstudio.com/.
9. Pachter L. Models for transcript quantification from RNA-Seq. Ithaca: Cornell University Library, 2011. Accessed 2018 Oct 1. Available from: http://arxiv.org/abs/1104.3889.