SeqTU: A Web Server for Identification of Bacterial Transcription Units

Xin Chen 1,2,3,4,*, Wen-Chi Chou 5,*, Qin Ma 6,7 & Ying Xu 1,2,3

A transcription unit (TU) consists of $K \geq 1$ consecutive genes on the same strand of a bacterial genome that are transcribed into a single mRNA molecule under certain conditions. Their identification is an essential step in elucidation of transcriptional regulatory networks. We have recently developed a machine-learning method to accurately identify TUs from RNA-seq data, based on two features of the assembled RNA reads: the continuity and stability of RNA-seq coverage across a genomic region. While good performance was achieved by the method on *Escherichia coli* and *Clostridium thermocellum*, substantial work is needed to make the program generally applicable to all bacteria, knowing that the program requires organism specific information. A web server, named SeqTU, was developed to automatically identify TUs with given RNA-seq data of any bacterium using a machine-learning approach. The server consists of a number of utility tools, in addition to TU identification, such as data preparation, data quality check and RNA-read mapping. SeqTU provides a user-friendly interface and automated prediction of TUs from given RNA-seq data. The predicted TUs are displayed intuitively using HTML format along with a graphic visualization of the prediction.
computational results will be saved on the server for at least six months. The back-end programs, implemented using PERL and R, along with test examples and documents are freely available online at http://csbl.bmb.uga.edu/SeqTU_dev/index.php.

It is noteworthy that we have previously developed an integrative operon database, DOOR2 (http://csbl.bmb.uga.edu/DOOR/), covering 2,072 bacteria genomes. For each genome, the identified operons and TUs are shown in a genome browser. By integrating the prediction results of SeqTU and DOOR2, a user can see the dynamic changes of expressions under different conditions.

Results
The predicted TUs are summarized in an HTML table, containing the following information: gene name, genomic location, strand orientation, and a read-abundance plot for each nucleotide in a TU. All the results are accessible and more detailed information can be found in the following. The RNA-seq dataset, SRR400619, is used as an example to illustrate how the server can be used to solve a TU identification problem in a bacterial genome.

Step 1: Data submission. On the homepage of SeqTU server (Fig. 2a), a user needs to enter a SRA ID (e.g., the non-strand-specific data SRR400619) to the search box; and click on the next button below the box. A new page will pop up to collect two pieces of information: (i) the email address for forwarding the prediction results (Fig. 2b), and (ii) the host genome for the given RNA-seq dataset. In this example, only one genome of E. coli str. K-12 substr. MG1655 is listed in the genome table (Fig. 2c). The user can select this genome, and specify this data-set is not a strain specific RNA-seq data in Fig. 2d. The NC number of the host genome may be required in case the desired host genome is not listed in the genome table (Fig. 2e). Then, the user can click on the submit button in Fig. 2e. The server will pop up a page with warning information if an invalid email address is entered. SeqTU can automatically download the RNA-seq dataset and the relevant genome sequence and annotation data from the NCBI database based on the submitted information.

To upload a user's private RNA-seq data from a local machine, one needs to click on the Submit your own data button blow the submit button (Fig. 2a). The user can upload their RNA-seq data, the host genome along with the annotation files (gff file, ptt file) and the user's email address in the next page (Fig. 2f).

Step 2: Job status checking. The user will receive an email immediately after the submission, containing a link to a job-status page. Clicking on this link, the user can find out the job status with the following information: (i) download progress of the RNA-seq data download from NCBI; (ii) quality check of the RNA-seq data using FastQC. The example SRR400619 has short reads with a length of 36 bps and the sequencing quality is high based on the quality report (Fig. 3a); (iii) mapping progress of the RNA reads to the host genome; and (iv) the final TU prediction (Fig. 3b). It is noteworthy that if this dataset has been submitted and analyzed previously, this step will be skipped and directly to the next step.

Step 3: Result page interpretation. When a job is done, the user will be notified by an email with a link to the final TU prediction. In total, 2,329 predicted TUs are listed in the result table of SRR400619. In addition, the start and end positions of each TU on the reference genome, their strand information, gene functional annotations and a download link to read abundance plots are also provided (Fig. 3c). Accuracy scores of each TU prediction are also provided. For example, the TU0005 with a TU prediction accuracy of 0.7813 is composed of talB and mog genes ranging from 8,238 bp to 9,893 bp on the reverse strand. For the strand-specific example (SRR578142), users can follow the same procedure provided above; and the result table and the TU plot are shown in Fig. 4.

Discussion
The SeqTU server provides a useful and free service for TU prediction from two provided RNA-seq datasets along with the host genomes. A unique characteristic is that the system has a built-in training program that can train an organism-specific classifier for TU prediction. This property makes the prediction highly reliable and desirable. We anticipate that the usefulness of this server will increase over the next few years as more genome-scale transcriptomic data become available and the needs for TU prediction continue to increase.
Methods

The representation and the logic layers of SeqTU are implemented using Web 2.0 technologies (HTML5, CSS3 and Javascript along with jQuery library) and PHP server-side scripting language. It runs on a Red Hat Enterprise Linux 6 box (32 Intel Xeon CPUs with 2.4 GHz and 64GB memory). The input RNA-seq data and the computing results will be stored on a local 10TB hard drive. Considering that a job may take hours up to days to complete, it will not be realistic for a user to sit in front of the computer, waiting for the results to return; hence an email address is collected via the input page, to which the final results will be sent.

SeqTU consists of four main functions: (i) data preparation, (ii) RNA-read mapping, (iii) TU prediction, and (iv) result visualization (Fig. 1), as detailed below.

Data preparation. A user can submit his/her RNA-seq data in one of the two ways: (i) select a dataset from the NCBI SRA database using the ID of the dataset, e.g. SRR400619; or (ii) upload a RNA-seq dataset from the local computer. For now, we restrict the maximum data size for upload to 500MB. In cases where a user’s data size is larger, special arrangements can be made with the developer to allow uploading the dataset in multiple fragments, which will be integrated into one and stored in a temporary location at the server side. For an organism needed for TU prediction, the user can upload the reference genome to the server or select it from the genome list provided by the server.

Read mapping. Before mapping the RNA reads to the target genome, quality check on the uploaded RNA-seq reads will be performed using FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/), with the check results shown on the result page. The Bowtie2 program is used to map RNA-seq reads to the relevant genome, and bedtools will be used to provide the genome coverage information by the uploaded RNA-seq reads, where the genome coverage is normalized by total-sum per million for cross-sample comparisons.
TU prediction. TU prediction has two steps: (i) training of an organism-specific classifier for TU prediction; and (ii) TU prediction by the trained classifier based on the given RNA-seq dataset. A highly effective method was developed to prepare for positive and negative datasets for training a TU predictor. Considering that experimentally validated TU data are generally lacking, the expressed genes of at least certain lengths and with certain characteristics in their expression pattern are used to mimic two consecutive genes in the same TU and are used as the positive training data. The negative training data consist of consecutive gene pairs deemed to be not in the same TU based on the following criteria: (a) the gap region(s) accounting for more than 50% of the intergenic region between the two genes; and (b) the average expression levels of the two genes are at least 10 fold different, where a gap is defined as a continuous genomic region without any expressions.

A TU classifier is trained based on the given RNA-seq data as follows: an expressed TU should have all its genes expressed at a relatively stable level continuously across the whole TU. The detailed characteristics in

Figure 3. Screenshots of the SeqTU result pages for SRR400619. (a) The progress report table; (b) The final TU prediction table; and (c) An example of computed expression levels over an identified TU in non-strand-specific dataset with default 800 bps upstream and downstream, where the blue histogram represents the read depth over a TU, with the lower part showing the genes in a TU.
stability and continuity of gene-expression patterns of true TUs are learned by the classifier through training on both positive and negative data. A support vector machine-based classifier is trained to distinguish the positive training data from the negative ones. A five-fold cross-validation is used to ensure the classifier is reliably trained. All these are done in an automated manner after the input data file is uploaded. After the training is done, the classifier is applied on the input RNA-seq dataset to generate the final result.

Currently, there are no other programs or web servers can do the same thing as SeqTU, so we have evaluated its performance using multiple RNA-seq datasets from different species. The test results of this strategy for calling TUs in *E. coli* and *Clostridium thermocellum* suggest that this strategy works very well. Specifically, on three sets of RNA-seq data of *E. coli* str. K-12 substr. MG1655, SRX315217, SRX315218, and SRX315219, the trained TU predictors have the prediction accuracy at 0.95, 0.94, and 0.93, respectively. In all the *Clostridium thermocellum* RNA-seq datasets, the trained predictors achieved prediction sensitivities over 0.9. The detailed results can be found in the original algorithm paper.

### References

1. Pierce, B. *Genetics: A Conceptual Approach*, 2nd ed., (W. H. Freeman, 2004).
2. Tjaden, B. De novo assembly of bacterial transcriptomes from RNA-seq data. *Genome Biol* 16, 1 (2015).
3. Jacob, F., Perrin, D., Sanchez, C. & Monod, J. Operon: a group of genes with the expression coordinated by an operator. *C R Hebdo Seances Acad Sci* 250, 1727–9 (1960).
4. Mao, X. *et al.* DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Res* 42, D654–9 (2014).
5. Siervo, N., Makita, Y., de Hoon, M. & Nakai, K. DDBTB: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res* 36, D93–6 (2008).
6. Pertea, M., Ayanbule, K., Smedinghoff, M. & Salzberg, S. L. OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Res* 37, D479–82 (2009).
7. Mao, F., Dam, P., Chou, J., Oliman, V. & Xu, Y. DOOR: a database for prokaryotic operons. *Nucleic Acids Res* 37, D459–63 (2009).
8. Okuda, S. *et al.* Characterization of relationships between transcriptional units and operon structures in Bacillus subtilis and Escherichia coli. *BMC Genomics* 8, 48 (2007).
9. Oliver, H. F. *et al.* Deep RNA sequencing of L. monocytogenes reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs. *BMC Genomics* 10, 641 (2009).
10. Cho, B. K. et al. The transcription unit architecture of the Escherichia coli genome. *Nat Biotechnol* **27**, 1043–9 (2009).
11. McClure, R. et al. Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res* **41**, e140 (2013).
12. Chou, W. C. et al. Analysis of strand-specific RNA-seq data using machine learning reveals the structures of transcription units in *Clostridium thermocellum*. *Nucleic Acids Res* **43**, e67 (2015).
13. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–9 (2012).
14. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **4711**, 12 1–34 (2014).
15. Bordbar, A. et al. Minimal metabolic pathway structure is consistent with associated biomolecular interactions. *Mol Syst Biol* **10**, 737 (2014).

**Acknowledgements**
This work was supported by a grant from the U.S. Department of Energy (#DE-PS02-06ER64304). The BioEnergy Science Center (BESC) is supported by the Office of Biological and Environmental Research in the DOE Office of Science. This work was also supported in part by the State of South Dakota Research Innovation Center and the Agriculture Experiment Station of South Dakota State University and the Natural Science Foundation of China (NSFC). We want to thank Mr. Jinyu Yang from South Dakota State University for his efforts in testing our example job and polishing all the three figures.

**Author Contributions**
Y.X. and Q.M. conceived the basic idea and planned the project. X.C. and W.-C.C. designed the web server and carried out the experiments. All authors edited the manuscript and approved the final version.

**Additional Information**

**Competing Interests:** The authors declare no competing financial interests.

**How to cite this article:** Chen, X. et al. SeqTU: A Web Server for Identification of Bacterial Transcription Units. *Sci. Rep.* **7**, 43925; doi: 10.1038/srep43925 (2017).

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2017