Briefings in Bioinformatics, 00(00), 2021, 1–14

doi: 10.1093/bib/bbab162
Problem Solving Protocol

A novel computational framework for genome-scale alternative transcription units prediction

Qi Wang, Zhaoqian Liu, Bo Yan, Wen-Chi Chou, Laurence Ettwiller, Qin Ma and Bingqiang Liu

Corresponding authors: Bingqiang Liu, School of Mathematics, Shandong University, Jinan 250200, China. Tel: 86-531-88363455; E-mail: bingqiang@sdu.edu.cn and Qin Ma, Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA. Tel: 1-614-688-9857; E-mail: qin.ma@osumc.edu

Abstract

Alternative transcription units (ATUs) are dynamically encoded under different conditions and display overlapping patterns (sharing one or more genes) under a specific condition in bacterial genomes. Genome-scale identification of ATUs is essential for studying the emergence of human diseases caused by bacterial organisms. However, it is unrealistic to identify all ATUs using experimental techniques because of the complexity and dynamic nature of ATUs. Here, we present the first-of-its-kind computational framework, named SeqATU, for genome-scale ATU prediction based on next-generation RNA-Seq data. The framework utilizes a convex quadratic programming model to seek an optimum expression combination of all of the to-be-identified ATUs. The predicted ATUs in *Escherichia coli* reached a precision of 0.77/0.74 and a recall of 0.75/0.76 in the two RNA-Sequencing datasets compared with the benchmarked ATUs from third-generation RNA-Seq data. In addition, the proportion of 5′- or 3′-end genes of the predicted ATUs, having documented transcription factor binding sites and transcription termination sites, was three times greater than that of no 5′- or 3′-end genes. We further evaluated the predicted ATUs by Gene Ontology and Kyoto Encyclopedia of Genes and Genomes functional enrichment analyses. The results suggested that gene pairs frequently encoded in the same ATUs are more functionally related than those that can belong to two distinct ATUs. Overall, these results demonstrated the high reliability of predicted ATUs. We expect that the new insights derived by SeqATU will not only improve the understanding of the transcription mechanism of bacteria but also guide the reconstruction of a genome-scale transcriptional regulatory network.

Key words: bacterial transcription regulation; alternative transcription units; RNA-Seq; convex quadratic programming; non-uniform read distribution

Qi Wang is a PhD student at the School of Mathematics, Shandong University. Her research interest is microbiome studies.
Zhaoqian Liu is a PhD student at the School of Mathematics, Shandong University. Her research interest is microbiome studies.
Bo Yan, PhD, is a research scientist at New England Biolabs. Her primary research interests are molecular biology, next-generation sequencing and computational biology.
Wen-Chi Chou, PhD, is a postdoctoral associate at Broad Institute of MIT and Harvard. His primary research strength is data mining of OMICS datasets including human genomic, human gut metagenomic, (meta)transcriptomic and metabolomic data.
Laurence Ettwiller, PhD, is the head of Bioinformatics and Computational Biology at New England Biolabs. Her primary research interests are genomics, microbiome and computational biology.
Qin Ma, PhD, is an associate professor at the Department of Biomedical Informatics, the Ohio State University. Dr Ma has over 10 years of research experience in the development of enabling methods and studying how functional machinery is encoded in a (meta-)genome.
Bingqiang Liu, PhD, is a professor at the School of Mathematics, Shandong University. His primary research strength is regulatory network construction.
Submitted: 21 January 2021; **Received (in revised form):** 18 March 2021

Introduction

An operon in bacterial genomes is defined as a group of consecutive genes regulated by a common promoter that all share the same terminator [1]. Genes in the same operon generally encode proteins with relevant or similar biological functions; e.g. *lacZ*, *lacY* and *lacA* in the *lac* operon encode proteins that help cells use lactose [1, 2]. With decades of research on bacterial transcriptional regulation, the operon model has been found to have complex mechanisms that control expression [3–5]. Previous studies showcased that bacterial genes are dynamically transcribed into mRNA transcripts under different triggering conditions and different transcripts can share genes under a specific condition [6–8]. Such dynamic transcription units (TUs) with overlapping patterns can be defined as ATUs (a.k.a., ATUs) [3, 5], with more details in Supplementary Figure S1.

ATU identification is of fundamental importance for understanding the transcriptional regulatory mechanisms of bacteria, and these dynamic structures have been demonstrated to be associated with human diseases [9–12]. For example, Bhat *et al.* studied the *alr-groEL1* operon, which is essential for the survival or virulence of *M. tuberculosis* [9, 11], the causative agent of tuberculosis, and found that the regulation of the sub-operon is distinct from the main operon (*alr-groEL1* operon) under stress, especially during heat shock, pH and Sodium dodecyl sulfate stresses [9]. Another example is *Helicobacter pylori*, a gastric pathogen that is the primary known risk factor for gastric cancer [12]. Sharma *et al.* [10] found an acid-induced sub-operon *cag*22–18 transcribed from the primary *cag*25–18 operon in the *cag* pathogenicity island of the *H. pylori* genome under acid stress. The mechanism of the complex ATU structures in these pathogenic bacteria can help us to study the emergence of human diseases caused by bacterial organisms.

Recent experimental techniques have provided a comprehensive view of bacterial transcriptome by identifying full-length primary transcripts [13–17]. For example, PacBio SMRT (Single Molecule, Real-Time) sequencing (SMRT-Cappable-seq) [6] combines the isolation of the full-length bacterial primary transcriptome of *Escherichia coli* with PacBio SMRT third-generation sequencing [6]. Simultaneous 5′ and 3′ end sequencing (SEnd-seq) [7] captures both transcription start sites (TSSs) and transcription termination sites (TTSs) of *E. coli* via circularization of transcripts [17]. Despite the great progress in experimental techniques, there are still some deficiencies. On the one hand, the read depth and error rate of the third-generation sequencing used in SMRT-Cappable-seq have an impact on ATU prediction compared with Illumina-based RNA-Seq [7, 18]. On the other hand, the time-consuming, laborious and costly properties of these experimental techniques make them unrealistic to be generally applicable to ATU predictions in bacteria under specific conditions. Thus, novel, convenient and robust computational methods for ATU identification in bacterial genomes based on next-generation RNA-Seq are urgently needed.

Fortunately, many computational methods have been developed to identify TUs in bacteria, accumulating some preliminary studies for ATU prediction [19–25]. Several public databases, such as RegulonDB [26], DBTBS [27], MicrobesOnline [28], DOOR [29, 30], OperomeDB [31], DMINDA 2.0 [32] and ProOpDB [33], provide various levels of operon information and small amounts of ATU information. RegulonDB [26] provides the TU and operon information of *E. coli* K-12 from scientific publications as well as computational predictions. DOOR 2.0 [30] contains computationally predicted operons of more than 2000 prokaryotes with complete sequenced genomes. However, these databases cannot provide genome-scale ATU information under specific conditions. Rockhopper [34], SeqTU [4, 35], BAC-BROWSER [36], rSeqTU [5], TruHMM [37] and Operon-mapper [38] utilize machine learning, transcripts assembling and model integration methods to identify bacterial transcription architecture, based on genomic information and gene expression profiles. However, these works cannot identify the overlapping patterns of ATUs in a specific condition and has limited power to further elucidate their dynamic features among different conditions.

Here, we present SeqATU, a novel computational method for genome-scale ATU prediction based on next-generation RNA-Seq data (Figure 1 and Supplementary Table S1). SeqATU utilizes a convex quadratic programming (CQP) model and aims to provide the optimum expression combination of all of the to-be-identified ATUs. Specifically, CQP minimizes the squared error between the predicted expression level of ATUs and the actual expression levels in genic and intergenic regions. It is noteworthy that SeqATU also utilizes the information about the bias rate function in modeling non-uniform read distribution as the linear constraints of CQP to profile the complexity of the ATU architecture. It is noteworthy that SeqATU can be easily and robustly applied to any bacterial organism with accessible reference genome and genome annotations, in support of the identification of genome-scale ATUs and construction of a transcriptional regulatory network.

Methods

Data collection

Two next-generation RNA-Seq datasets used as input data for SeqATU, named Illumina 1 and Illumina 2, were obtained from condition 1 (M9 minimal medium) and condition 2 (Rich medium), respectively. The details about the generation of these data were shown in Supplementary Method S1. Two ATU datasets of *E. coli*, named SMRT 1 and SMRT 2, were used as the benchmark data to evaluate the predicted ATUs, which were generated by SMRT-Cappable-seq under condition 1 and condition 2, respectively [6]. All reads in Illumina 1 and 2 were mapped to the *E. coli* genome using Burrows–Wheeler Aligner (BWA) with the default parameters [39]. Read alignment and other computational analyses were carried out using the *E. coli* genome NC_000913.3, and the corresponding gene annotations (GCF_000005845.2_ASM584v2_genomic.gff) were downloaded from NCBI. In addition, the next-generation RNA-Seq datasets of *Bacteroides fragilis* were retrieved from NCBI's SRA database with project accession number PRJNA445716. The reference genome sequence and gene annotations of *B. fragilis* are GCF_000025985.1_ASM2598v1_genomic.fna and GCF_000025985.1_ASM2598v1_genomic.gff, respectively.

Calculation of the expression values of genic and intergenic regions

We mapped the RNA-Seq reads in Illumina 1 and Illumina 2 to the *E. coli* genome using BWA [39], and determined the number of reads $N(l)$ covering each genomic position l. Suppose that g_i and g_{i+1} are two consecutive genes on the same strand; we denote the expression value of g_i as c_i and the expression value of the intergenic region between genes g_i and g_{i+1} as $b_{i,i+1}$. Then, the calculation of c_i and $b_{i,i+1}$ is given by:

$$c_i = \frac{\sum_{k \in g_i} N(k)}{| g_i |} \quad (1)$$

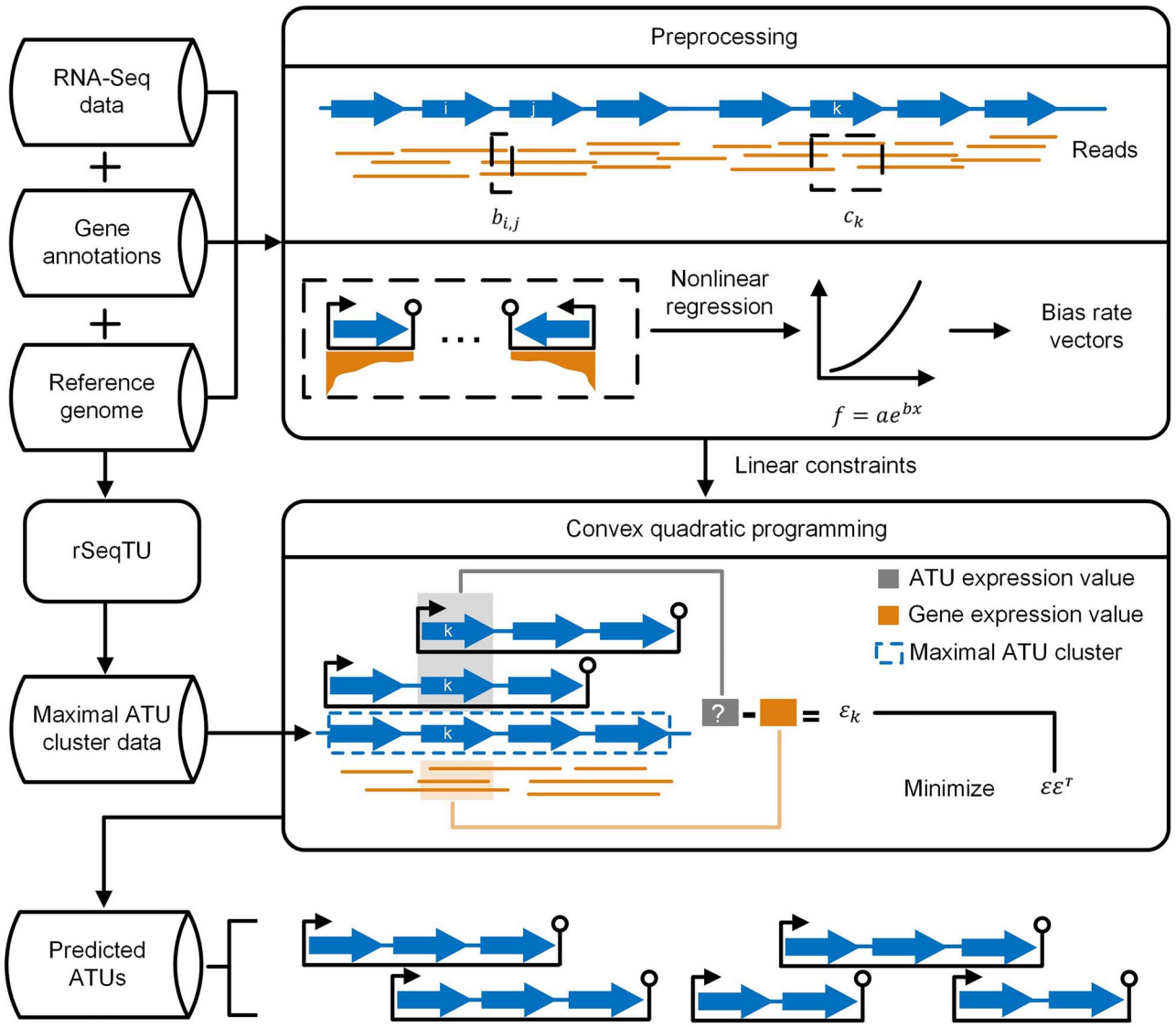

Figure 1. Schematic overview of SeqATU. The blue arrow and orange line denote gene and RNA-Seq read, respectively. The preprocessing stage requires next-generation RNA-Seq data in the FASTQ format, the reference genome sequence in the FASTA format, and gene annotations in the GFF format, generating linear constraints for the next CQP stage. There are two steps in the preprocessing stage: (i) calculating the expression value of the genic region c_k and intergenic region $b_{i,j}$ and (ii) modeling non-uniform read distribution along mRNA transcripts; specifically, we acquired a bias rate function $f(x) = ae^{bx}$ using nonlinear regression and then constructed genic or intergenic region bias rate vectors. The maximal ATU cluster data determined by rSeqTU and the linear constraints from preprocessing are both taken as inputs of CQP. CQP seeks the optimum expression combination of all of the to-be-identified ATUs to minimize the gap $\varepsilon\varepsilon^T$ between the predicted ATU expression profile and the genic and intergenic region expression profile. Finally, the output of CQP is the predicted ATUs.

$$b_{i,i+1} = \frac{\sum_{l \in g_{i,i+1}} N(l)}{| g_{i,i+1} |} \quad (2)$$

where $k \in g_i$ denotes that genomic position k is on the gene g_i and $| g_i |$ denotes the genomic length of g_i.

Modeling non-uniform read distribution along mRNA transcripts

We introduced the bias rate function, which is similar to the bias curves in the work of Wu *et al.* [40], to address the non-uniform distribution of the RNA-Seq reads along mRNA transcripts [40–43]. The bias function reflects the relative read distribution bias from the 3′ end to the 5′ end of an mRNA transcript. We assumed that the maximum read coverage of all the genomic positions of an mRNA transcript is the expression level without bias. It is noteworthy that a single gene mRNA transcript with no shared gene among different mRNA transcripts can serve as the ideal template for modeling non-uniform read distribution along mRNA transcripts. The specific steps of modeling non-uniform read distribution are detailed as follows:

Step 1: Single gene mRNA transcript selection

We selected single gene mRNA transcripts from the evaluation data and plotted their expression distributions. Specifically, 12 groups of single gene mRNA transcripts with lengths ranging from 300 to 1500 bp were selected from the evaluation data (more details are given in Supplementary Method S2), and each group had 10 randomly chosen mRNA transcripts. Apparent decline trends appeared in the single gene mRNA

transcripts with long lengths, ranging from 1100 to 1500 bp (Supplementary Figures S2–S5). The reason for this phenomenon may be that the incomplete transcription and 3′ end degradation or processing induce the enrichment of signal at 5′ end of the mRNA transcripts with long lengths [44, 45]. Finally, we plotted the expression distribution of single gene mRNA transcripts with lengths ranging from 1100 to 1500 bp.

Step 2: Acquiring the bias rate function

We applied nonlinear regression to the expression distribution of the selected single gene mRNA transcripts and acquired the hypothetical function $f(x)$. Specifically, the x axis and y axis of the expression distribution were converted to the distance from the 3′ end of an mRNA transcript and the bias rate of read distribution, respectively. To apply nonlinear regression to single gene mRNA transcripts with different lengths, normalization was also implemented on x. Here, $x = (x_1, x_2, \ldots, x_m)$ and $y = (y_1, y_2, \ldots, y_m)$ are defined by:

$$x_i = \begin{cases} \dfrac{l_m - l_{m-i+1}}{\max_l - l_1} \times 10^3, \text{forward} \\ \dfrac{l_i - l_1}{\max_l - l_1} \times 10^3, \text{reverse} \end{cases} \tag{3}$$

$$y_i = \begin{cases} \dfrac{N(l_{m-i+1})}{\max_y}, \text{forward} \\ \dfrac{N(l_i)}{\max_y}, \text{reverse} \end{cases} \tag{4}$$

where m denotes the number of genomic positions on an mRNA transcript; $l = (l_1, l_2, \ldots, l_m)$ denotes the genomic positions on an mRNA transcript; $\max_l = l_m$; $N(l_i)$ denotes the expression level of the genomic position l_i, i.e. the number of reads covering the genomic position l_i; and $\max_y$ denotes the expression level without bias in an mRNA transcript, which is calculated as $\max\{N(l_i)\}$, $1 \le i \le m$. We used the function *nls* in R to acquire the hypothetical function $f(x)$.

Step 3: Constructing bias rate vectors

We constructed a genic or intergenic region bias rate vector for each mRNA transcript by calculating the bias rate of all its component genic or intergenic regions. The bias rate of a genic or an intergenic region is the average bias rate of all the genomic positions that it contains. Considering an mRNA transcript T and its component gene set $\{g_1, g_2, \ldots, g_n\}$ (the details of the gene labels are described in Supplementary Method S3), we denoted the genic region bias rate vector as $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)$, which was calculated using the formula:

$$u_i = \begin{cases} \dfrac{\sum_{t=m-i_p+1}^{m-i_p+1} f(x_t)}{x_{m-i_p+1} - x_{m-i_q+1} + 1}, \text{forward} \\ \dfrac{\sum_{t=i_p}^{i_q} f(x_t)}{x_{i_q} - x_{i_p} + 1}, \text{reverse} \end{cases} \tag{5}$$

where m denotes the number of genomic positions on T; u_i denotes the bias rate of g_i for T; and $L_g = (l_{1_p}, l_{1_q}, l_{2_p}, l_{2_q}, \ldots, l_{n_p}, l_{n_q})$ is the range of the genomic positions of $\{g_1, g_2, \ldots, g_n\}$, while the range of the genomic positions of g_i is $[l_{i_p}, l_{i_q}]$, l_{i_p} is the left boundary position of gene g_i and l_{i_q} is the right boundary position of gene g_i, $1 \le i \le n$. Similarly, the calculation of the intergenic region bias rate vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_{n-1})$ is provided in Supplementary Method S4.

Modification of maximal ATU clusters

A maximal ATU cluster is defined as a maximal consecutive gene set such that each pair of its consecutive genes can be covered by at least one ATU. Similar to ATUs, maximal ATU clusters are also dynamically composed under different conditions or environmental stimuli in bacterial genomes [5, 46]. Such a maximal ATU cluster can be used as an independent genomic region for ATU prediction, which alleviates the difficulty in computationally predicting ATUs at the genome scale. The output of our in-house tool rSeqTU can serve as the maximal ATU cluster data, which lays a solid foundation for ATU prediction [5]. We modified the maximal ATU clusters from rSeqTU: (*i*) two maximal ATU clusters with distances less than 40 bp were combined into one cluster and (*ii*) a maximal ATU cluster was split at the intergenic region where the opposite-strand genes were located. In addition, we selected the maximal ATU clusters with expression values over 10 (see the details in Supplementary Method S5), according to the study of Etwiller *et al.* [13].

The mathematical programming model for ATU prediction

The predicted ATU expression profile should be consistent with the observed expression profiles of the genic and intergenic regions. Therefore, the prediction of the ATU profiles can be modeled as an optimization problem, which seeks an optimum expression combination of all of the to-be-identified ATUs to minimize the gap between the predicted ATUs and the observed genic and intergenic region expression profiles. Here, a CQP model was built to solve this optimization problem.

We denoted a maximal ATU cluster as G, assuming that it contains the consecutive genes $\{g_1, \ldots, g_n\}$, and the intergenic regions of these genes are $\{g_{1,2}, \ldots, g_{n-1,n}\}$. The size of G is defined as the number of its component genes n. Theoretically, there are $\frac{n\times(n+1)}{2}$ ATUs for G, and an ATU with consecutive genes $\{g_i, g_{i+1}, \ldots, g_j\}$ is denoted as $a^{i,j}$; the corresponding expression value is $x^{i,j}$, $1 \le i \le j \le n$.

For the component gene g_k of G, the gap between the gene expression value c_k and the sum of the expression level of the ATUs containing it is denoted as τ_k, which provides the first n equality constraints in our mathematical programming model, $k = 1, 2, \ldots, n$. Similarly, for the intergenic region $g_{l,l+1}$ of G, the gap between the intergenic region expression value $b_{l,l+1}$ and the sum of the expression level of the ATUs containing it is denoted as β_l, providing the last $n-1$ equality constraints in our mathematical programming model, $l = 1, 2, \ldots, n-1$.

The goal of our mathematical programming model is to minimize the square of $\boldsymbol{\varepsilon} = (\tau_1, \tau_2, \ldots, \tau_n, \beta_1, \ldots, \beta_{n-1})$, as the combination of $x^{i,j}$ with a minimal value of $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T$ is corresponding to an optimum expression combination of all ATUs $a^{i,j}$ for G, $1 \le i \le j \le n$. Additionally, to control the number of optimal solutions and reduce the false-positive errors, we added an L^1 regularization $\alpha\|\boldsymbol{x}\|_1$ to $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T$ with $x^{i,j} \ge 0$, which is a linear function. Because of the variant expression level of different maximal ATU clusters, we used the expression value of G as α. In total, the CQP model with unknown variables $(\boldsymbol{x}, \boldsymbol{\varepsilon})$ is shown as follows:

$$\min\ \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T + \alpha\|\boldsymbol{x}\|_1$$

$$\text{s.t.} \sum_{i=1}^{k}\sum_{j=k}^{n} u_{i,k} x^{i,j} = c_k + \tau_k \quad k = 1, 2, \ldots, n$$

$$\sum_{i=1}^{l}\sum_{j=l+1}^{n} v_{i,l+1}x^{i,j} = b_{l,l+1} + \beta_l \; l = 1, 2, \ldots, n-1$$

$$\boldsymbol{x} = \left(x^{i,j}\right), x^{i,j} \geq 0 \; 1 \leq i \leq j \leq n$$

$$\boldsymbol{\varepsilon} = (\tau_1, \tau_2, \ldots, \tau_n, \beta_1, \ldots, \beta_{n-1}) \tag{6}$$

where $\boldsymbol{u} = (u_{i,j})$ is the genic region bias rate vector for G, $u_{i,j}$ is the bias rate of gene g_j for ATU $a^{i,k}$, $1 \leq i \leq j \leq n, j \leq k \leq n$, $\boldsymbol{v} = (v_{i,j})$ is the intergenic region bias rate vector for G, and $v_{i,j}$ is the bias rate of the intergenic region $g_{j-1,j}$ for ATU $a^{i,l}$, $1 \leq i < j \leq n, j \leq l \leq n$ (see the details in Supplementary Method S6).

Two evaluation methods for ATU prediction

In the first evaluation method, precision and recall were defined based on perfect matching (Eq. 7), where perfect matching of two ATUs means that all of their component genes are the same. Here, the true positives (TP) are the number of predicted ATUs with the same component genes as an ATU in the evaluation data; the false positives (FP) are the number of predicted ATUs that do not exist in the evaluation data; the false negatives (FN) are the number of ATUs that appear in the evaluation data but not in the prediction results of SeqATU.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

In the second evaluation method, precision and recall were defined based on relaxed matching, which is measured by the similarity of two ATUs. Assuming that an ATU t is in one of two datasets (the predicted ATU dataset and evaluated ATU dataset), the definition and calculation of the similarity of t are shown in the following three cases:

Case 1: If t shares boundary genes at both ends of an ATU in the other dataset, i.e. all component genes of t are the same as one in the other dataset, then similarity(t) = 1.

Case 2: If t shares exactly one boundary gene of ATUs in the other dataset, then we denote U_a as the ATUs in the other dataset that share the 5′-end gene with t and denoted U_b as the ATUs in the other dataset that share the 3′-end gene with t, $U_a \cap U_b = \varnothing$, one of U_a and U_b can be empty. Then,

$$\text{similarity(t)} = \frac{1}{2}\max_{t' \in U_a}\frac{\alpha(t')}{\beta(t')} + \frac{1}{2}\max_{t' \in U_b}\frac{\alpha(t')}{\beta(t')} \tag{8}$$

where $\alpha(t')$ is the number of shared genes of t and t' and $\beta(t')$ is the maximal size of t and t'.

Case 3: If t shares no boundary genes at both ends of the ATUs in the other dataset, then similarity(t) = 0.

Finally, the precision and recall based on relaxed matching are calculated by the following formula:

$$\text{precision} = \frac{\sum_{t \in T_1}\text{similarity}(t)}{n_1}$$

$$\text{recall} = \frac{\sum_{t \in T_2}\text{similarity}(t)}{n_2} \tag{9}$$

where T_1 is the set of predicted ATUs, n_1 is the number of predicted ATUs, T_2 is the set of evaluated ATUs, and n_2 is the number of evaluated ATUs.

Results

A reliable bias rate function is acquired in modeling non-uniform read distribution along mRNA transcripts

To ensure the reliability of the bias rate function in modeling non-uniform read distribution, we selected four single gene mRNA transcript datasets randomly from the two evaluation datasets (SMRT 1 and 2), named Group 1–4. Four bias rate functions, which are exponential functions, were generated after conducting nonlinear regression on the mRNA transcripts across these four datasets (Figure 2). We found that these bias rate functions were similar ($R^2 > 0.998$) when we evaluated the R-square statistic (for more details, see Supplementary Method S7 and Supplementary Table S2). The similarity of the four bias rate functions indicated that the selection of the single gene mRNA transcript datasets had little impact on modeling non-uniform read distribution along mRNA transcripts, implying the universal common non-uniform read distribution of different mRNA transcripts of *E. coli*. Specifically, we used the average of these four coefficients as the final coefficients of the exponential function, which was $f(x) = ae^{bx}$ with $a = 0.256$ and $b = 0.00128$.

ATUs predicted by SeqATU achieve a satisfactory performance

The performance evaluation was conducted by comparing the predicted ATUs with the ATUs in SMRT 1 and 2, which were generated based on the third-generation sequencing and are not sensitive to transcripts with low expression levels. For a more accurate and fair evaluation, maximal ATU clusters after pre-selection were retained in the subsequent evaluations (more details about the pre-selection of maximal ATU clusters can be seen in Supplementary Method S8 and Supplementary Figure S6).

The precision and recall of the predicted ATUs were calculated for each maximal ATU cluster. By considering only perfect matching, the average precision and recall were 0.67 and 0.67 for Illumina 1 and 0.64 and 0.68 for Illumina 2, respectively. When using relaxed matching, the average precision and recall increased to 0.77 and 0.75 for Illumina 1 and 0.74 and 0.76 for Illumina 2, respectively. These results showed that the performance of SeqATU is satisfying, considering no existing tools for predicting ATUs with overlapping patterns. The statistics for precision and recall on maximal ATU clusters with different sizes, as shown in Figure 3A and Supplementary Figure S7A. These results showed that the average precision and recall were decreasing with the increasing size of maximal ATU clusters (other than several large size ones due to their small number of counts). The results also indicated that the evaluation results based on relaxed matching were significantly higher than those based on perfect matching across different sizes. This result implied that the incorrectly predicted ATUs by SeqATU based on perfect matching tended to have strong similarities with the ATUs in the evaluation data. In addition, we also found that more than a quarter of the incorrectly predicted ATUs (25%/29% for Illumina 1/Illumina 2) by SeqATU based on perfect matching matched with the TUs in RegulonDB [26].

The two evaluation datasets (SMRT 1 and 2) were both from SMRT-Cappable-seq, while one of the processing steps of the technique filtered RNA reads smaller than 1000 bp [6], which indicated that the ATUs in these two evaluation datasets were not comprehensive. To address this issue, we enriched the evaluation data by adding the ATUs defined by SEnd-seq [7], as SEnd-seq did not introduce any filtering based on RNA size.

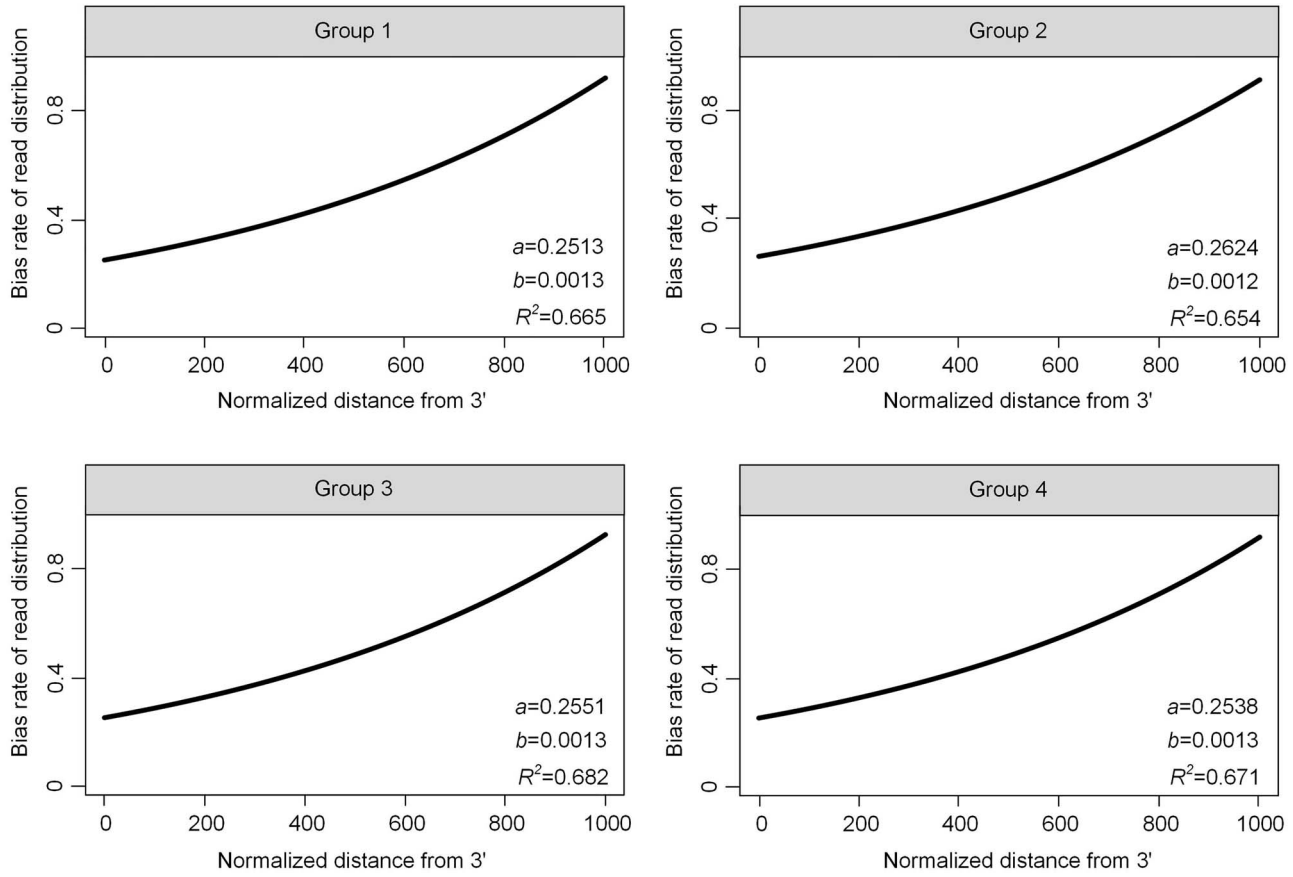

Figure 2. Results of modeling non-uniform read distribution along mRNA transcripts. The four bias rate functions ($y = ae^{bx}$) by nonlinear regression had similar coefficients (a and b) across the four datasets Group 1–4.

When we used the new evaluation data, the ATUs predicted by SeqATU improved by 15% (0.77) and 19% (0.76) in terms of the average precision based on perfect matching for Illumina 1 and Illumina 2, respectively, and by 9% (0.84) and 12% (0.83) based on relaxed matching. The statistics for precision across different sizes of the maximal ATU clusters are shown in Figure 3B and Supplementary Figure S7B, showing that the values of precision based on perfect matching were significantly improved across different sizes of maximal ATU clusters by using the evaluated ATUs from SMRT-Cappable-seq and SEnd-seq. This result suggested that the ATUs we predicted, which were not in SMRT 1 and 2, may be due to the RNA length selection of SMRT-Cappable-seq. We enriched the evaluation data by adding the ATUs in RegulonDB [26] and also found the improvement of precision across different sizes of maximal ATU clusters for Illumina 1 and Illumina 2 (Supplementary Figure S7C).

Furthermore, to facilitate the understanding of the performance of SeqATU and to measure the influence of the maximal ATU clusters from rSeqTU on our ATU prediction method, SMRT maximal ATU clusters collected from SMRT 1 and 2 (for more details, see Supplementary Method S9) were applied for the CQP in two conditions (M9 minimal medium and Rich medium). We found that precision and recall increased to 0.73 and 0.77 for Illumina 1, respectively, and 0.69 and 0.80 for Illumina 2 based on perfect matching (Supplementary Figure S7D). Additionally, when using relaxed matching, precision and recall significantly increased to 0.82 and 0.84 for Illumina 1, respectively, and 0.79 and 0.86 for Illumina 2 (Supplementary Figure S7D). The significantly improved results verified the ability of SeqATU to accurately predict ATU when giving more accurate maximal ATU clusters. In addition, we found that the number of predicted ATUs and the evaluated ATUs under the maximal ATU cluster with the same size were similar except for the maximal size (Figure 3C), and they were far less than the theoretical number, which indicated that SeqATU can effectively exclude most of the incorrect ATUs.

The bias rate constraints efficiently improve the ability of SeqATU to predict ATUs

We tried to use SeqATU without bias rate constraints to predict the ATUs of *E. coli* and found that its performance significantly decreased compared with SeqATU (Figure 4 and Supplementary Figure S8). Specifically, the F-score of SeqATU without bias rate constraints was 0.69/0.68 based on perfect matching for Illumina 1/Illumina 2, compared with 0.75/0.74 for SeqATU. When using relaxed matching, the F-score of SeqATU without bias rate constraints was 0.79/0.78 for Illumina 1/Illumina 2 compared with 0.83/0.83 for SeqATU. This result suggested that the bias rate constraints of SeqATU could capture useful information about the non-uniform distribution of the RNA-Seq reads along the mRNA transcripts [40–43] and then efficiently improve the ability of the model to predict complex ATUs.

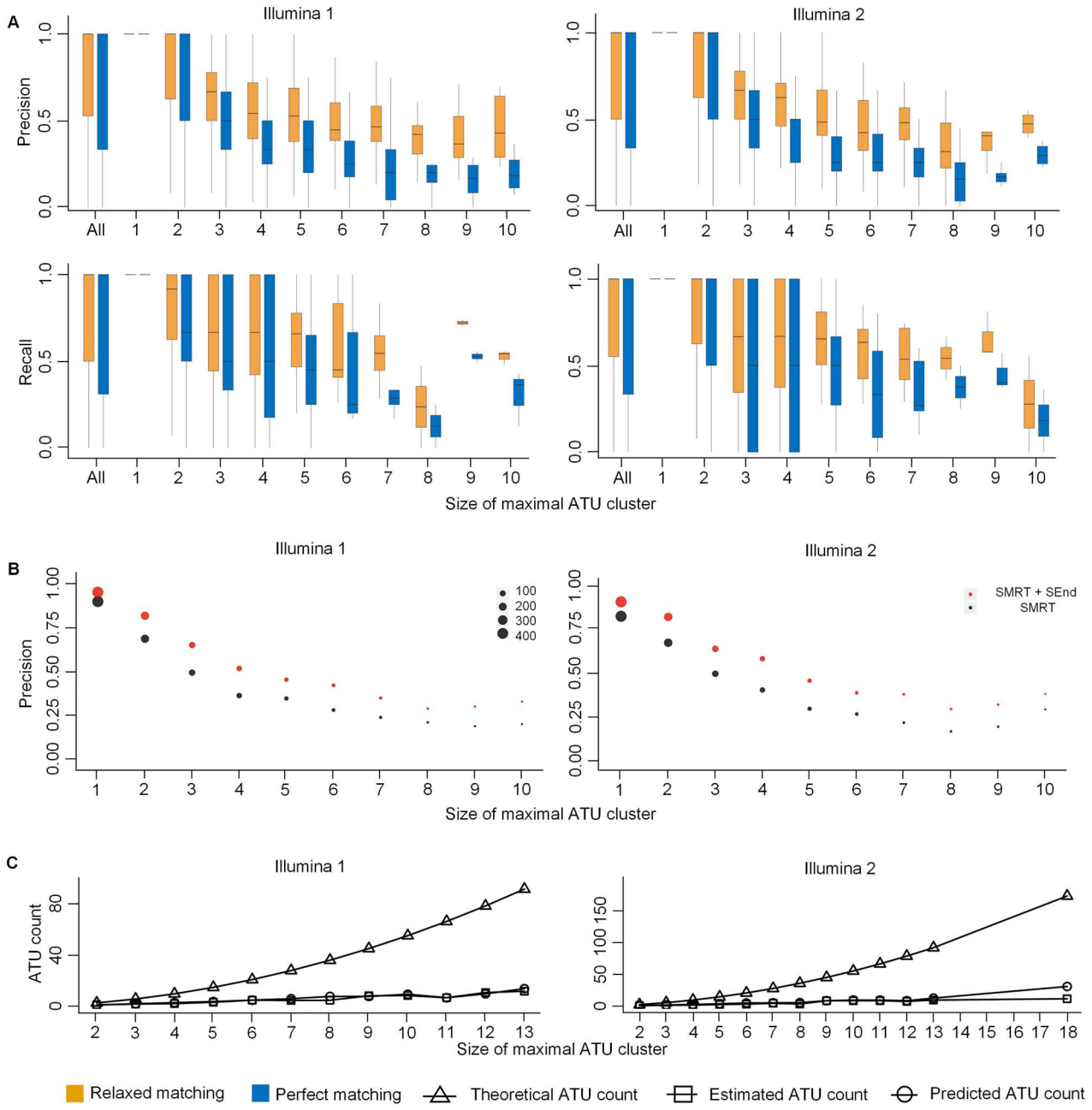

Figure 3. Overall evaluation results of SeqATU. (**A**) Precision and recall based on perfect matching and relaxed matching for Illumina 1 (left) and Illumina 2 (right) using evaluated ATUs from SMRT-Cappable-seq. (**B**) Average precision based on perfect matching for Illumina 1 (left) and Illumina 2 (right) using evaluated ATUs from SMRT-Cappable-seq (black) and evaluated ATUs from SMRT-Cappable-seq and SEnd-seq (red). The magnitude of the point denotes the number of maximal ATU clusters of the same size. (**C**) The average number of ATUs across different sizes of SMRT maximal ATU clusters for Illumina 1 (left) and Illumina 2 (right).

SeqATU outperforms Rockhopper in identifying complex transcription structures

Rockhopper is designed to identify TUs from the expression data under certain conditions (It is noteworthy that Rockhopper only provides TUs without overlapping genes). Intuitively, it should have the potential capability to elucidate dynamic transcription patterns. Hence, we applied Rockhopper [34] to Illumina 1 and 2 and compared the results with SeqATU's using the identification results from third-generation RNA-Seq as a benchmark. We found that the number of verified ATUs identified by SeqATU (764 and 782 for Illumina 1 and Illumina 2, respectively) is 3.4 times that of Rockhopper (222 and 226 for Illumina 1 and Illumina 2, respectively). The F-score of SeqATU is 0.67 and 0.66 for Illumina 1 and 2, respectively, whereas that of Rockhopper are 0.43 and 0.37. These results demonstrated that SeqATU has the extra power to identify complex transcription structures by solving the overlapping patterns of ATUs comparing with Rockhopper. Specifically, we showed an example of predicted transcription structures on the gene cluster containing four genes (*usr*, *dcm*, *yedJ* and *yedR* in Figure 5). SeqATU identified three ATUs, and all of them are consistent with the identification results from third-generation RNA-Seq. Note that an

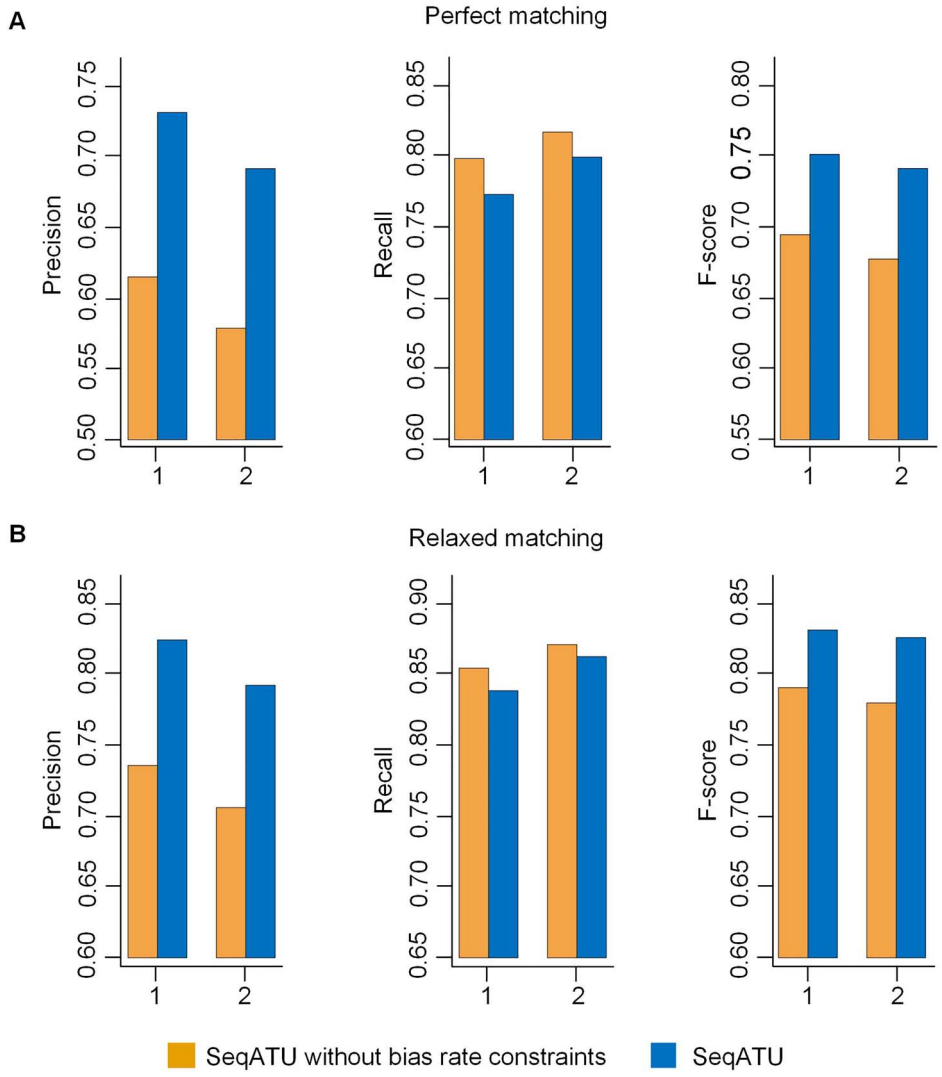

Figure 4. Comparative analysis of the performance between SeqATU and SeqATU without the bias rate constraints for SMRT maximal ATU clusters. (**A**) Precision, recall and F-score based on perfect matching for Illumina 1 and Illumina 2. (**B**) Precision, recall and F-score based on relaxed matching for Illumina 1 and Illumina 2.

alternative TU containing these four genes (*usr*, *dcm*, *yedJ* and *yedR*) was predicted by SeqATU. Rockhopper identified two TUs, and there is no overlapping between them. Only one of them is verified by third-generation RNA-Seq. More examples about the comparison of SeqATU and Rockhopper can be found in Supplementary Figure S9. Besides, we found the TUs predicted by Rockhopper are in the same structure between Illumina 1 and 2 while ATUs predicted by SeqATU display different structures (details in next subsection 'ATUs predicted by SeqATU display a dynamic composition and overlapping patterns'). The result suggests that the ATUs with complex structures play more important roles across different conditions and this important dynamic mechanism can be captured by the prediction results of SeqATU.

ATUs predicted by SeqATU display a dynamic composition and overlapping patterns

A total of 2973 distinct ATUs of *E. coli* were identified in M9 minimal medium, and 2767 were identified in Rich medium. Among them, there were 1423/1550 distinct ATUs on the forward strand and 1323/1444 on the reverse strand for Illumina 1/Illumina 2. Each of the predicted ATUs was comprised of an average of 2.59 genes, with the largest ATU containing 28 genes across the two conditions. The distribution of the size of the predicted ATUs is shown in Figure 6A, from which we can see that the majority of ATUs (more than 87%) contained fewer than five genes in M9 minimal medium and Rich medium. Approximately 41% of the genes in *E. coli* were contained in more than one ATU for Illumina 1, compared to 43% genes for Illumina 2, suggesting that the ATUs in a maximal ATU cluster generally overlapped with each other (Figure 6B). In addition, there were 1576 ATU maximal clusters for Illumina 1 and 1512 ATU maximal clusters for Illumina 2. SeqATU identified a total of 1977 identical ATUs under the two conditions, whereas there were 1786 distinct ATUs. Among the distinct ATUs across the two conditions, 394 ATUs were from the same maximal ATU clusters in the two maximal ATU cluster datasets, and the rest were from different maximal ATU clusters. The fact there were distinct ATUs under the two conditions suggests that ATUs are

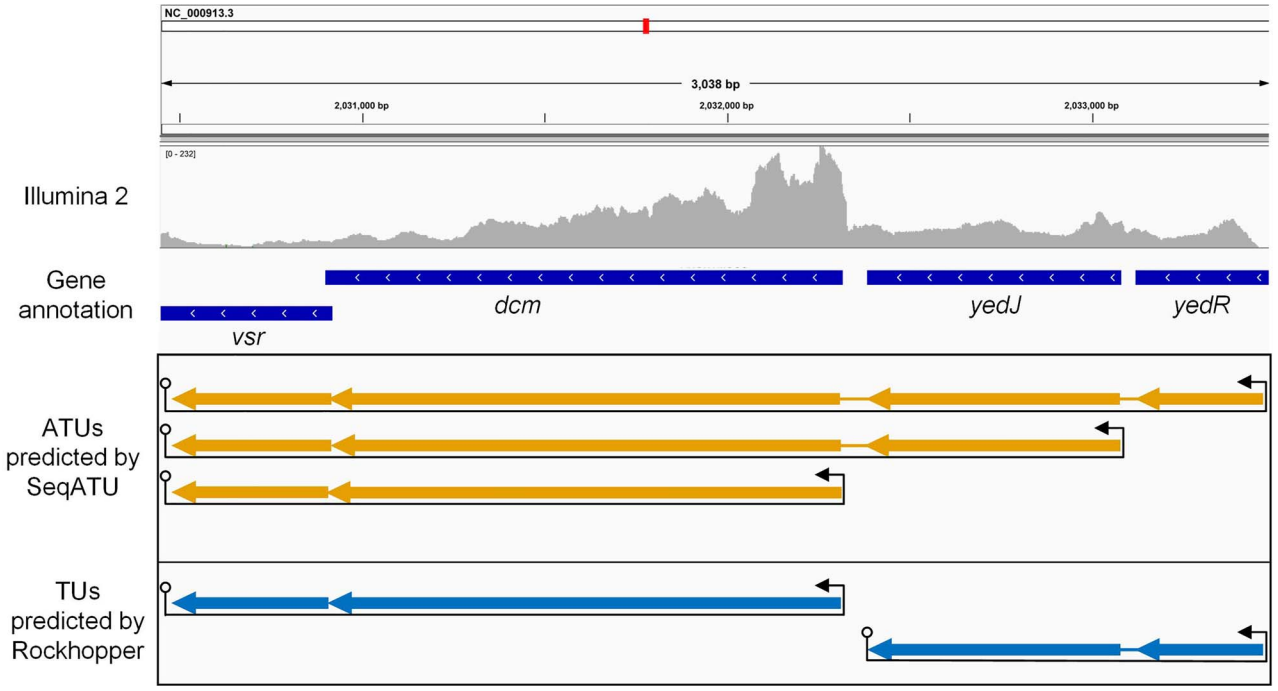

Figure 5. Integrative genomics viewer (IGV) representation of the prediction results from SeqATU and Rockhopper. IGV representation of ATUs predicted by SeqATU (orange) compared to TUs predicted by Rockhopper (blue) for the gene cluster of *E. coli* containing the four genes *vsr*, *dcm*, *yedJ* and *yedR*.

dynamically responsive to different conditions or environmental stimuli (for more real examples about the ATUs under different conditions, see Supplementary Figure S10).

Predicted ATUs by SeqATU are verified by experimental TSSs and TTSs

Two experimentally generated datasets were used to further verify the reliability of SeqATU (Table 1), which are the TSSs of *E. coli* from SEnd-seq [7] (named Data 1) and the transcription factor (TF) binding sites of *E. coli* (named Data 2) from RegulonDB [26]. We considered the 5′-end genes and no 5′-end genes of the predicted ATUs by SeqATU. A gene that is not the 5′-end gene of any predicted ATU is named a no 5′-end gene. We identified 2177/2005 5′-end genes and 1266/1160 no 5′-end genes of the predicted ATUs for Illumina 1/Illumina 2. A gene validated by experimental TSSs or TF binding sites means that it is the immediate downstream gene of an experimental TSS or TF binding site. As a result, the proportion of 5′-end genes (29%/30% for Illumina 1/Illumina 2) validated by experimental TF binding sites was over three times greater than the no 5′-end genes (9.2%/9.0% for Illumina 1/Illumina 2) (Table 1). The result further verified the reliability of the ATUs predicted by SeqATU in terms of the TSS level. In addition, four other experimental TSS or promoter datasets from RegulonDB [26], differential RNA sequencing (dRNA-seq) [14] and Cappable-seq [13] were also examined. The results are shown in Supplementary Table S3, and we also found a higher proportion of 5′-end genes of the predicted ATUs validated by experimental TSSs or promoters than that of no 5′-end genes.

We also used two experimental TTS datasets of *E. coli* from SEnd-seq [7] (named Data 3) and RegulonDB [26] (named Data 4) to verify the reliability of the predicted ATUs by SeqATU (Table 2). We considered the 3′-end genes and no 3′-end genes of the predicted ATUs by SeqATU. A gene that is not the 3′-end gene of any predicted ATU is named a no 3′-end gene. A gene validated by experimental TTSs means that it is the immediate upstream gene of an experimental TTS. As a result, the proportion of 3′-end genes (51%/53% for Illumina 1/Illumina 2) validated by experimental TTSs from SEnd-seq was over three times greater than that of no 3′-end genes (15%/14% for Illumina 1/Illumina 2) (Table 2). The result further verified the reliability of the ATUs predicted by SeqATU in terms of the TTS level. In addition, two other computationally predicted TTS datasets from the works by Nadiras *et al.* [47] and Kingsford *et al.* [48] were also examined. The results are shown in Supplementary Table S4, and we also found the proportion of 3′-end genes (63%/62% for Illumina 1/Illumina 2) validated by computationally predicted Rho-independent TTSs was over two times greater than that of no 3′-end genes (29%/29% for Illumina 1/Illumina 2).

The gene pairs frequently encoded in the same ATUs are more functionally related than those that can belong to two distinct ATUs

Functional analysis was conducted by integrating GO terms from the Gene Ontology database [49]. In detail, we measured the level of functional relatedness for two types of consecutive gene pairs, which is similar to the definition in the work by Mao *et al.* [46]. Two types of consecutive gene pairs were (*i*) gene pairs each consisting of a 5′-end gene of an ATU and the gene in its immediate upstream on the same strand and (*ii*) all the other gene pairs inside an ATU (Figure 7A). In addition, we used a scoring scheme to measure the GO-based functional similarity between a pair of genes by Wu *et al.* [50]. This study developed a GO similarity score and we can interpret that the larger the score, the more likely that two genes are functionally related (see the calculation of GO similarity score in Supplementary Method S10).

As a result, the mean GO similarity score was higher for type-*ii* gene pairs (5.97 versus 4.04 for Illumina 1 and 5.86

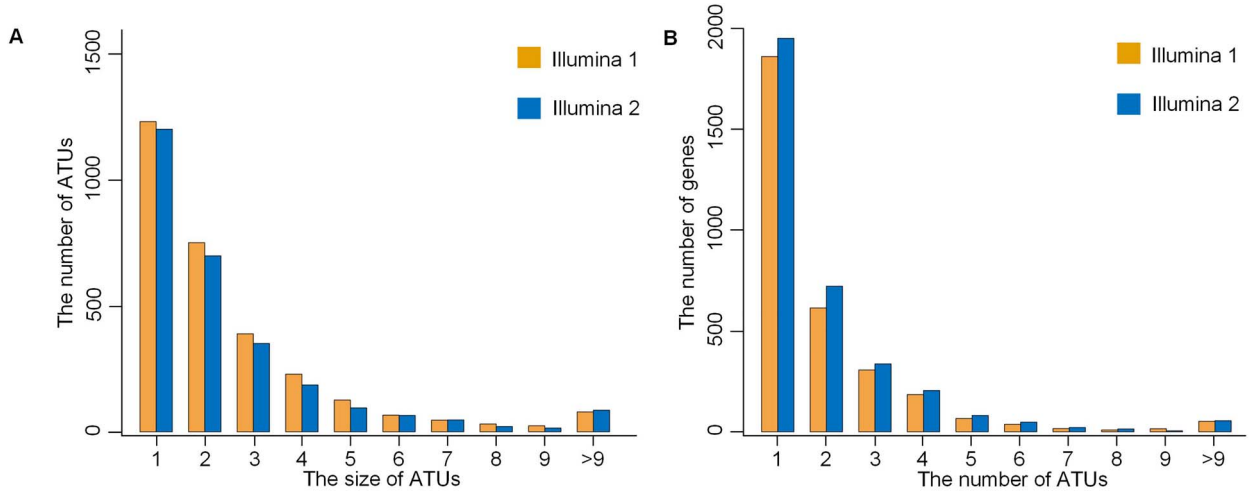

Figure 6. Comprehensive analysis of the predicted ATUs by SeqATU. (**A**) Number of ATUs across different sizes. The size of an ATU is the number of its component genes. (**B**) Distribution of the number of ATUs per gene.

Table 1. Results of predicted ATUs verified by experimental TSSs or TF binding sites. Overview of the experimental TSS and TF binding site datasets (Data 1 and Data 2) and the proportion of 5′-end genes and no 5′-end genes of the predicted ATUs by SeqATU for Illumina 1 and 2, which were validated by experimental TSSs or TF binding sites

		Data 1	Data 2
Source		Ju *et al.*	RegulonDB TF binding sites
Technique		SEnd-seq	Collection
TSSs/TF binding sites		5512	3220
Illumina 1	5′-end genes	83%	29%
	No 5′-end genes	47%	9.2%
Illumina 2	5′-end genes	89%	30%
	No 5′-end genes	44%	9.0%

Table 2. Results of predicted ATUs verified by experimental TTSs. Overview of the experimental TTS datasets (Data 3 and Data 4) and the proportion of 3′-end genes and no 3′-end genes of the predicted ATUs by SeqATU for Control 1 and 2, which were validated by experimental TTSs

		Data 3	Data 4
Source		Ju *et al.*	RegulonDB TTSs
Technique		SEnd-seq	Collection
TTSs		1540	367
Control 1	3′-end genes	51%	11%
	No 3′-end genes	15%	5.2%
Control 2	3′-end genes	53%	11%
	No 3′-end genes	14%	4.8%

versus 3.91 for Illumina 2) than for type-*i* gene pairs. A total of 574/524 type-*ii* gene pairs had GO similarity scores greater than four (64%/63% of a total of 899/834), while only 461/404 type-*i* gene pairs had GO similarity scores greater than four (36%/34% of a total of 1274/1179) for Illumina 1/Illumina 2. We also applied a χ^2-test [51] to determine whether the distribution of $S_{GO}(g_k, g_j)$ was different for the type-*i* gene pairs and type-*ii* gene pairs. The χ^2-statistics corresponded to a *P*-value less than 10^{-4}, which revealed that the distribution of $S_{GO}(g_k, g_j)$ for the type-*ii* gene pairs was significantly different from the type-*i* gene pairs. Figure 7B shows the distribution of $S_{GO}(g_k, g_j)$ for the type-*i* gene pairs and the type-*ii* gene pairs. These results strongly indicated that the type-*ii* gene pairs had a higher degree of GO similarity than the type-*i* gene pairs, suggesting that the gene pairs frequently encoded in the same ATUs (type-*ii* gene pairs) are more functionally related than those that can belong to two distinct ATUs (type-*i* gene pairs).

We also carried out a similar analysis of the two different gene pairs based on Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis [52] (see more details in Supplementary Method S10) and found that the proportion of type-*ii* gene pairs (59%/57% for Illumina 1/Illumina 2), whose two genes were contained in the same KEGG pathway, was higher than the proportion of type-*i* gene pairs (32%/28% for Illumina 1/Illumina 2) (Figure 7C). The distribution of the KEGG similarity scores of the two different types of gene pairs is shown in Figure 7D, suggesting that genes of type-*ii* gene pairs have a higher probability of participating in the same KEGG pathway than those of type-*i* gene pairs.

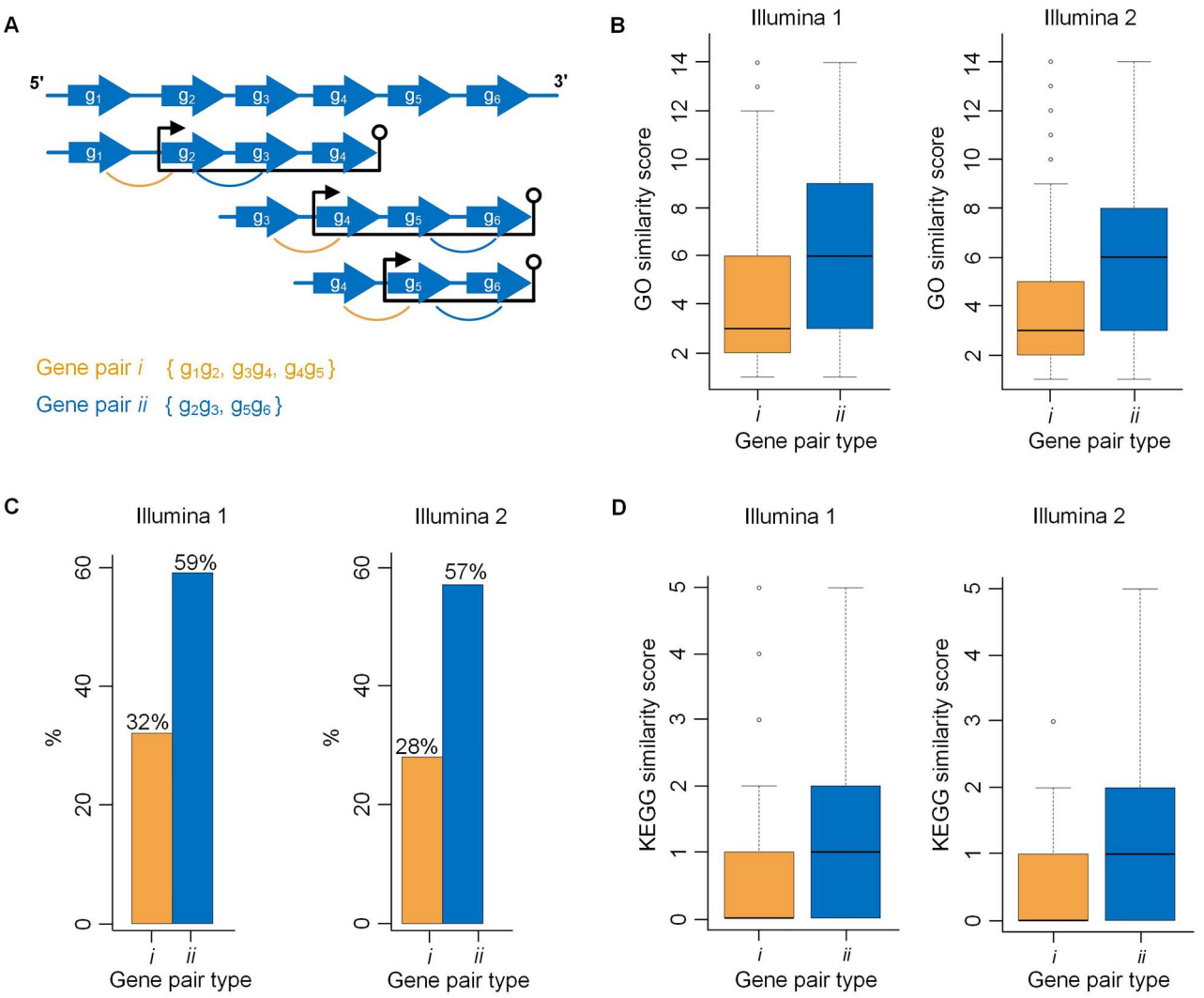

Figure 7. Interpretation and results of the functional relatedness of different gene pairs based on GO and KEGG enrichment analyses. (**A**) Illustration of two different gene pairs *i* and *ii*. (**B**) Functional relatedness results based on GO enrichment analysis for Illumina 1 (left) and Illumina 2 (right). (**C**) The proportion of two different gene pairs whose genes are contained in the same KEGG pathway for Illumina 1 (left) and Illumina 2 (right). (**D**) The functional relatedness results based on KEGG enrichment analysis for Illumina 1 (left) and Illumina 2 (right).

SeqATU can be applied to any bacterial organisms for ATU prediction

SeqATU is a generalized tool for the inference of ATUs based on next-generation RNA-Seq data, without any particular design for certain species, therefore can be applied to any bacterial organism. We run it on two *B. fragilis* datasets, which are derived from *ccfA* overexpression *B. fragilis* (SRR6899497) and wild-type *B. fragilis* (SRR6900706) by Donaldson *et al.* [53]. There were 2200 and 3040 identified ATUs for the ccfA overexpression *B. fragilis* dataset and the wild-type *B. fragilis* dataset, respectively. Specifically, 1548 ATUs existed in both conditions and 2144 ATUs existed in only one condition (more details can be found in Supplementary Method S11). We did functional annotation for the genes within these different ATUs by DAVID [54]. The results showed that these genes were enriched in six different functional domains (Supplementary Table S5). One of the domains, i.e. the outer membrane protein beta-barrel domain, has been demonstrated closely related to mucosal colonization [55–57]. Based on these, we can infer that the identified different ATUs are associated with mucosal colonization, which is consistent with the prior knowledge of the datasets (i.e. ccfA activates genes involved in mucosal colonization). These results suggested that SeqATU can provide insights about transcriptional regulatory differences of genes under distinct environments.

Discussion

We developed SeqATU, the first CQP model for the inference of dynamic ATUs with overlapping patterns based on next-generation RNA-Seq data. Linear constraints provided by the bias rate of read distribution were, for the first time, integrated into the CQP model. Positional bias refers to the non-uniform distribution of reads over different positions of a transcript [41, 43], which is handled by learning non-uniform read distributions from given RNA-Seq reads [40] or modeling the RNA degradation [58]. The bias rate function we proposed can address the non-uniform read distribution along mRNA transcripts and also be desirable for standard next-generation RNA-Seq data that

involves more degraded mRNAs, as the exponential function has been used to model the degradation of mRNA transcripts [58]. As a result, a total of 2973 distinct ATUs for Illumina 1 and 2767 distinct ATUs for Illumina 2 were identified by SeqATU. The precision and recall reached 0.67/0.64 and 0.67/0.68, respectively, based on perfect matching and 0.77/0.74 and 0.75/0.76, respectively, based on relaxed matching for Illumina 1/Illumina 2. In addition, the proportion of the 5′- or 3′-end genes of predicted ATUs that were validated by experimental TF binding sites and TTSs from RegulonDB and SEnd-seq was over three times greater than that of no 5′- or 3′-end genes, demonstrating the high reliability of predicted ATUs. Gene pairs frequently encoded in the same ATUs were more functionally related than those that can belong to two distinct ATUs according to GO and KEGG enrichment analyses.

The ATU architecture of bacteria is much more complex than that determined with currently used experimental techniques. We investigated the 5′-end genes and no 5′-end genes of the experimental ATUs identified by SMRT-Cappable-seq [6] using a combination of experimental TSSs from RegulonDB [26], dRNA-seq [14], Cappable-seq [13] and SEnd-seq [7]. As a result, we found that the proportion of 5′-end genes (99%) validated by experimental TSSs was not significantly different from that of no 5′-end genes (92%). The high percentage of no 5′-end genes validated by experimental TSSs implied that the ATUs identified by experimental techniques are only a small proportion of the comprehensive ATUs in bacterial organisms because of the dynamic mechanisms of ATUs. These results further verified the necessity of developing robust computational methods for ATU identification.

SeqATU not only provides a powerful tool to understand the transcription mechanism of bacteria but also provides a fundamental tool to guide the reconstruction of a genome-scale transcriptional regulatory network. First, the ATU structures can help us to make new functional predictions, as genes in an ATU tend to have related functions in a specific condition or environmental stimuli (Supplementary Method S12 and Supplementary Figure S11). Second, ATUs can elucidate condition-specific uses of alternative sigma factors [8, 59]. For example, the *thrLABC* operon is regulated by transcriptional attenuation. Totsuka *et al*. found that under the log phase growth condition, the *thrLABC* operon is the only transcript, whereas two transcripts are found under stationary phase growth condition, the *thrLABC* and *thrBC*. As validated experimentally, σ^S can regulate the additional promoter located in front of *thrB* under the stationary phase growth condition and then separately regulate *thrBC*, which elucidates the condition-specific uses of σ^S [8]. Third, understanding the ATU structures is of great help to construct transcriptional and translation regulatory networks, such as for the construction of the σ-TUG (σ-factor-TU gene) network [60]. The transcription regulatory network consists of nodes (ATU and regulatory proteins) and links (interactions) [61], and the comprehensive ATU structures can provide a nearly complete set of nodes, which can improve the accuracy of regulatory prediction.

Although SeqATU has obtained satisfactory predicted results, there are still several challenges regarding the computational prediction of ATUs. Firstly, because of the influence of the 3′ untranslated region (UTR) and 5′ UTR in the intergenic regions, the expression value of intergenic regions cannot be reproduced perfectly by the same calculation used for the expression value of genic regions. Without accurate reproduction, it is difficult to obtain the best expression combination of ATUs by the programming model based on the expression value of genic and intergenic regions. Secondly, because of the lack of strand-specific RNA-Seq data, it is difficult to distinguish the expression level of intergenic regions between two consecutive genes on the same strand derived from ATUs containing these two genes or antisense RNAs [6, 62]. All of these challenges and the great significance of ATU prediction inspire and encourage us to discover more information to determine the ATU structures in bacteria. For example, we plan to add high confidence TSSs and TTSs information to our programming model in the future. Additionally, since the microbiome is increasingly recognized as a critical component in human diseases, such as inflammatory bowel disease [63], antibiotic-associated diarrhea [64], neurological disorders [65] and cancer [66, 67], predicting new ATUs of uncultured species from metagenomic and metatranscriptomic data is of great significance in uncovering new regulatory pathway and metabolic products during the development of diseases [68]. However, because of a majority of species with unknown genomes or genome annotations within a microbial community, ATU prediction on metagenomics and metatranscriptomics is still a challenging task, which encourages us to pay more attention on it.

Key Points

- We developed SeqATU, the first CQP model for the inference of dynamic alternative transcription units (ATUs) with overlapping patterns based on next-generation RNA-Seq data.
- Non-uniform read distribution is integrated as linear constraints of the convex quadratic programming model utilized by SeqATU, which has significantly improved the prediction performance.
- The predicted ATUs of *E. coli* displayed a dynamic composition and overlapping patterns, and they reached a precision of 0.77/0.74 and a recall of 0.75/0.76 on two RNA-Seq datasets compared with the benchmarked ATUs from the corresponding third-generation RNA-Seq data.
- The proportion of 5′- or 3′-end genes of the predicted ATUs, having documented transcription factor binding sites and transcription termination sites, was three times greater than that of no 5′- or 3′-end genes.
- ATUs predicted by SeqATU were evaluated by GO and KEGG functional enrichment analyses and found that gene pairs frequently encoded in the same ATUs are more functionally related than those that can belong to two distinct ATUs.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Data access

The source code of SeqATU and a detailed tutorial can be found at https://github.com/OSU-BMBL/SeqATU. All raw data are available in https://bmbl.bmi.osumc.edu/downloadFiles/data.zip.

Authors' contributions

B.L., Q.M. and W.C. conceived the basic idea and designed the overall analyses. Q.W. carried out most of the

computational analysis and data interpretation. All the authors wrote the manuscript. The authors declare that they have no competing interests.

Acknowledgments

The authors would like to thank Yang Li for his assistance in language polishing.

Funding

This work was supported by the National Nature Science Foundation of China (61772313 to B.L., 11931008 to B.L.), the Interdisciplinary Science Innovation Group Project of Shandong University (2019) and the Innovation Method Fund of China (Ministry of Science and Technology of China) (2018IM020200 to B.L.).

References

1. Jacob F, Perrin D, Sanchez C, *et al*. Operon: a group of genes with the expression coordinated by an operator. *C R Hebd Seances Acad Sci* 1960;**250**:1727–9.
2. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 1961;**3**:318–56.
3. Liu Z, Feng J, Yu B, *et al*. The functional determinants in the organization of bacterial genomes. *Brief Bioinform* 2020. doi: 10.1093/bib/bbaa1172.
4. Chou W-C, Ma Q, Yang S, *et al*. Analysis of strand-specific RNA-seq data using machine learning reveals the structures of transcription units in Clostridium thermocellum. *Nucleic Acids Res* 2015;**43**:e67–7.
5. Niu S-Y, Liu B, Ma Q, *et al*. rSeqTU—a machine-learning based R package for prediction of bacterial transcription units. *Front Genet* 2019;**10**:374.
6. Yan B, Boitano M, Clark TA, *et al*. SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nat Commun* 2018;**9**:3676.
7. Ju X, Li D, Liu S. Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. *Nat Microbiol* 2019;**4**:1907–18.
8. Totsuka K, Totsuka K. The transcription unit architecture of the *Escherichia Coli* genome. *Nat Biotechnol* 2009;**27**:1043–9.
9. Bhat AH, Pathak D, Rao A. The alr-groEL1 operon in mycobacterium tuberculosis: an interplay of multiple regulatory elements. *Sci Rep* 2017;**7**:43772.
10. Sharma CM, Hoffmann S, Darfeuille F, *et al*. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* 2010;**464**:250–5.
11. Durand JM, Bjork GR. Putrescine or a combination of methionine and arginine restores virulence gene expression in a tRNA modification-deficient mutant of Shigella flexneri: a possible role in adaptation of virulence. *Mol Microbiol* 2010;**47**:519–27.
12. Wroblewski LE, Peek RM, Wilson KT. *Helicobacter pylori* and gastric cancer: factors that modulate disease risk. *Clin Microbiol Rev* 2010;**23**:713–39.
13. Ettwiller L, Buswell J, Yigit E, *et al*. A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. *BMC Genomics* 2016;**17**:199–9.
14. Thomason MK, Bischler T, Eisenbart SK, *et al*. Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in *Escherichia coli*. *J Bacteriol* 2015;**197**:18–28.
15. Bischler T, Tan HS, Nieselt K, *et al*. Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in *Helicobacter pylori*. *Methods* 2015;**86**:89–101.
16. Dar D, Shamir M, Mellin J, *et al*. Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science* 2016;**352**:6282.
17. Clauwaert J, Menschaert G, Waegeman W. An in-depth evaluation of annotated transcription start sites in E. coli using deep learning. In: *bioRxiv*, 2020.
18. Goodwin S, Mcpherson JD, Mccombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**:333–51.
19. Chen X, Su Z, Xu Y, *et al*. Computational prediction of operons in Synechococcus sp. WH8102. *Genome Inform* 2004;**15**:211–22.
20. Westover BP, Buhler JD, Sonnenburg JL, *et al*. Operon prediction without a training set. *Bioinformatics* 2005;**21**:880–8.
21. Price MN, Huang KH, Alm EJ, *et al*. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res* 2005;**33**:880–92.
22. Dam P, Olman V, Harris K, *et al*. Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Res* 2007;**35**:288–98.
23. Tran TT, Dam P, Su Z, *et al*. Operon prediction in Pyrococcus furiosus. *Nucleic Acids Res* 2007;**35**:11–20.
24. Bergman NH, Passalacqua KD, Hanna PC, *et al*. Operon prediction for sequenced bacterial genomes without experimental information. *Appl Environ Microbiol* 2007;**73**:846–54.
25. Taboada B, Verde C, Merino E. High accuracy operon prediction method based on STRING database scores. *Nucleic Acids Res* 2010;**38**:e130.
26. Santos-Zavaleta A, Salgado H, Gama-Castro S, *et al*. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Res* 2018;**47**:D212–20.
27. Sierro N, Makita Y, De Hoon MJL, *et al*. DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res* 2008;**36**:93–6.
28. Dehal PS, Joachimiak MP, Price MN, *et al*. MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res* 2010;**38**:D396–400.
29. Cao H, Ma Q, Chen X, *et al*. DOOR: a prokaryotic operon database for genome analyses and functional inference. *Brief Bioinform* 2019;**20**:1568–77.
30. Mao X, Ma Q, Zhou C, *et al*. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Res* 2013;**42**:D654–9.
31. Chetal K, Janga SC, Operome DB. A database of condition-specific transcription units in prokaryotic genomes. *Biomed Res Int* 2015;**2015**:1–10.
32. Yang J, Chen X, Mcdermaid A, *et al*. DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses. *Bioinformatics* 2017;**33**:2586–8.
33. Blanca T, Ricardo C, Martinez-Guerrero CE, *et al*. ProOpDB: prokaryotic operon DataBase. *Nucleic Acids Res* 2012;**40**:D627–31.
34. McClure R, Balasubramanian D, Sun Y, *et al*. Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res* 2013;**41**:e140–0.

35. Chen X, Chou W, Ma Q, *et al*. SeqTU: a web server for identification of bacterial transcription units. *Sci Rep* 2017;**7**:43925.
36. Garanina IA, Fisunov GY, Govorun VM. BAC-BROWSER: the tool for visualization and analysis of prokaryotic genomes. *Front Microbiol* 2018;**9**:2827.
37. Li S, Dong X, Directional SZ. RNA-seq reveals highly complex condition-dependent transcriptomes in E. coli K12 through accurate full-length transcripts assembling. *BMC Genomics* 2013;**14**:1–24.
38. Taboada B, Estrada K, Ciria R, *et al*. Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. *Bioinformatics* 2018;**34**:4118–20.
39. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* 2009;**25**:1754–60.
40. Wu Z, Wang X, Zhang X. Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq. *Bioinformatics* 2011;**27**:502–8.
41. Roberts A, Trapnell C, Donaghey J, *et al*. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* 2011;**12**:1–14.
42. Bohnert R, Rätsch G. rQuant. web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res* 2010;**38**:W348–51.
43. Li W, Jiang T. Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads. *Bioinformatics* 2012;**28**:2914–21.
44. Xiong B, Yang Y, Fineis FR, *et al*. DegNorm: normalization of generalized transcript degradation improves accuracy in RNA-seq analysis. *Genome Biol* 2019;**20**:75.
45. Chaitanya J. Degradation of mRNA in *Escherichia coli*. *IUBMB Life* 2010;**54**:315–21.
46. Mao X, Ma Q, Liu B, *et al*. Revisiting operons: an analysis of the landscape of transcriptional units in *E. Coli*. *BMC Bioinformatics* 2015;**16**:356.
47. Nadiras C, Eveno E, Schwartz A, *et al*. A multivariate prediction model for rho-dependent termination of transcription. *Nucleic Acids Res* 2018;**46**:8245–60.
48. Kingsford CL, Ayanbule K, Salzberg SL. Rapid, accurate, computational discovery of rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol* 2007;**8**:R22.
49. Ashburner M, Lewis S. On ontologies for biologists: the gene ontology—untangling the web. *Novartis Found Symp* 2002;**247**:66–80 discussion 80-63, 84-90, 244-252.
50. Wu H, Su Z, Mao F, *et al*. Prediction of functional modules based on comparative genome analysis and gene ontology application. *Nucleic Acids Res* 2005;**33**:2822–37.
51. Teukolsky SA, Flannery BP, Press W, *et al*. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge: Cambridge University Press, 1992.
52. Kanehisa M, Goto SKEGG. Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
53. Donaldson GP, Ladinsky MS, Yu KB, *et al*. Gut microbiota utilize immunoglobulin A for mucosal colonization. *Science* 2018;**360**:795–800.
54. Dennis G, Sherman BT, Hosack DA, *et al*. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 2003;**4**:1–11.
55. Hunter S, Apweiler R, Attwood TK, *et al*. InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009;**37**:D211–5.
56. De Jonge R, Durrani Z, Rijpkema SG, *et al*. Role of the *Helicobacter pylori* outer-membrane proteins AlpA and AlpB in colonization of the Guinea pig stomach. *J Med Microbiol* 2004;**53**:375–9.
57. Ottman N, Huuskonen L, Reunanen J, *et al*. Characterization of outer membrane proteome of Akkermansia muciniphila reveals sets of novel proteins exposed to the human intestine. *Front Microbiol* 2016;**7**:1157.
58. Wan L, Yan X, Chen T, *et al*. Modeling RNA degradation for RNA-Seq with applications. *Biostatistics* 2012;**13**:734–47.
59. Yanofsky C. Attenuation in the control of expression of bacterial operons. *Nature* 1981;**289**:751.
60. Cho BK, Kim D, Knight EM, *et al*. Genome-scale reconstruction of the sigma factor network in *Escherichia coli*: topology and functional states. *BMC Biol* 2014;**12**:4–4.
61. Cho B-K, Charusanti P, Herrgård MJ. Microbial regulatory and metabolic networks. *Curr Opin Biotechnol* 2007;**18**:360–4.
62. Toledo-Arana A, Dussurget O, Nikitas G, *et al*. The listeria transcriptional landscape from saprophytism to virulence. *Nature* 2009;**459**:950–6.
63. Yue B, Luo X, Yu Z, *et al*. Inflammatory bowel disease: a potential result from the collusion between gut microbiota and mucosal immune system. *Microorganisms* 2019;**7**:440.
64. Mullish BH, Williams HR. Clostridium difficile infection and antibiotic-associated diarrhoea. *Clin Med* 2018;**18**:237.
65. Maguire M, Maguire G. Gut dysbiosis, leaky gut, and intestinal epithelial proliferation in neurological disorders: towards the development of a new therapeutic using amino acids, prebiotics, probiotics, and postbiotics. *Rev Neurosci* 2019;**30**:179–201.
66. Vivarelli S, Salemi R, Candido S, *et al*. Gut microbiota and cancer: from pathogenesis to therapy. *Cancer* 2019;**11**:38.
67. Cammarota G, Ianiro G, Ahern A, *et al*. Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nat Rev Gastroenterol Hepatol* 2020;**17**:635–48.
68. Zaidi SSA, Zhang X. Computational operon prediction in whole-genomes and metagenomes. *Brief Funct Genomics* 2017;**16**:181–93.