Published in final edited form as:

Curr Opin Microbiol. 2015 February ; 0: 133–140. doi:10.1016/j.mib.2014.11.011.

Quantitative bacterial transcriptomics with RNA-seq

James P. Creecy^{1,2} and Tyrrell Conway^{1,*}

¹Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK 73019

²Department of Biology, University of Central Oklahoma, Edmond, OK 73034

Abstract

RNA sequencing has emerged as the premier approach to study bacterial transcriptomes. While the earliest published studies analyzed the data qualitatively, the data are readily digitized and lend themselves to quantitative analysis. High-resolution RNA sequence (RNA-seq) data allows transcriptional features (promoters, terminators, operons, etc.) to be pinpointed on any bacterial transcriptome. Once the transcriptome is mapped, the activity of transcriptional features can be quantified. Here we highlight how quantitative transcriptome analysis can reveal biological insights and briefly discuss some of the challenges to be faced by the field of bacterial transcriptomics in the near future.

Keywords

dRNA-seq; transcriptome; operon; promoter; terminator

RNA-seq comes of age

Advances in RNA sequencing technology have revolutionized the study of bacterial transcriptomes [1,2]. At its core, RNA-seq generates digital information that allows transcriptional features to be located with single-nucleotide precision in a strand specific manner. Since the data are digital, RNA-seq facilitates quantitative computational analysis of any selected region of the transcriptome, but the transcriptome must first be annotated properly. Since bacterial genomes are organized in operons, it is logical that RNA-seq data should be annotated with the operon architecture in mind. In practice, only three transcriptional features need to be defined: 5' transcript ends (promoters), 3' ends (terminators), and RNA sequence read coverage to connect the ends, which together define operons [3-5].

The true power of RNA-seq resides in its potential as an analytical tool for quantifying promoter activity, terminator efficiency, and differential expression of transcripts, including

© 2014 Elsevier Ltd. All rights reserved.

*Corresponding author: Department of Microbiology and Plant Biology University of Oklahoma Norman, OK 73019-0245
tconway@ou.edu 405 325 1683 .

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

operons, transcription units within operons (e.g. generated by promoters internal to operons), and antisense RNAs. As described in more detail below, RNA-seq datasets consist of tens of millions of sequence reads and typically the reads are 50 bases in length. The raw sequence reads are aligned to a reference genome and only high quality reads are retained and mapped. Conversion of sequence data into digital format is accomplished by employing freely available computer scripts that count the number of times each transcribed base was sequenced in a read-aligned dataset, thereby converting aligned sequence reads to base count data. Normalization of the base count data is necessary to quantify the differential expression (i.e., relative base counts) of each transcriptional feature within a sample or between different samples. The normalized base count data can be quantified by averaging the base count across a selected region of the genome. Since the average of the base counts is used, the relative expression of any given transcription feature, regardless of its length, can be expressed in this way. Here we focus on the analysis of an *E. coli* RNA-seq dataset to demonstrate the strategy we developed to quantify the expression of the transcriptional features that define operons in bacteria.

Single-nucleotide resolved RNA-seq dataset

To obtain an RNA-seq dataset suitable for quantitative analysis, we prepared RNA from a culture of *E. coli* K-12 strain BW38028 during logarithmic- and stationary-phase growth on glucose limited minimal medium, as described previously [4]. In addition, we starved *E. coli* BW38028 and its isogenic *rpoS* mutant BW39452 for nitrogen by decreasing by three-fold the amount of ammonium chloride in the growth medium [6]. The RNA was extracted by using the hot-phenol method [7] and DNase I treated to remove contaminating DNA. The RNA samples were not depleted for rRNA prior to sequencing, which tends to eliminate some experimental biases [8]. The RNA samples were shipped on dry ice to vertis Biotechnologie AG (Germany) for library preparation and Illumina HiSeq2000 sequencing, as described by others [7,9]. For library preparation the RNA samples were split and subjected to differential RNA-seq (dRNA-seq) as described [2,10]. Briefly, one portion of the RNA was fragmented by ultrasound and then the fragments were poly(A)-tailed and an RNA adapter was ligated to the 5' phosphate of the RNA. First strand cDNA synthesis was with a poly(dT) primer and reverse transcriptase. Second strand cDNA synthesis incorporated a barcoded 3' TruSeq adapter. The other portion of the RNA samples were fragmented and treated with terminator exonuclease (TEX), which enriches for 5' triphosphate containing transcripts that are generated by transcription initiation at promoters. The TEX treated samples then were tailed and ligated, and cDNA was prepared as described above. The cDNAs were sequenced on an Illumina HiSeq2000 system using 50 bp read length, with each library yielding approximately 20 million reads.

Datasets consisting of 10 million reads per sample are sufficient for transcriptional feature mapping and differential gene expression analysis without ribo-depletion for a transcriptome the size of *E. coli* [9,11]. For quantification the genome-aligned, strand-specific RNA-seq data should be converted from aligned reads to base counts. Our RNA-seq data analysis pipeline involves alignment of the raw data to the reference genome by using Bowtie2 to generate the sequence read alignment file (SAM) [12]. SAMTOOLS [13] were used to convert the SAM file to a binary alignment file (BAM). The BAM file was converted to a

BigWig file (base count file), which contains the count of the base at each base location and is the standard for visualization in genome browsers such as J-Browse [14]. Conversion of BAM to BigWig formatted files can be accomplished by using tools available in the Galaxy Toolshed [15] or at UCSC Genome Browser [16].

Alternatively, users can analyze their datasets by using pipelines such as Galaxy [17] or READemption [18], which outputs normalized wiggle files (base count files). A simple and straightforward way to normalize base count data is by using a strategy analogous to the total count approach [19] for normalizing gene-specific read alignments, which expresses each value as the base count per billion bases counted [4]. Because the BigWig file represents the base count at each nucleotide position, all downstream analysis begins with this file. The advantages of the base count approach are: a) the digital base count data are inherently computable because of their format and smaller size, b) the average base counts of individual transcriptional features can be computed and queried at any desired resolution, from a single nucleotide to an entire operon, to quantify the expression level or activity, c) normalization of base count data makes all samples directly comparable, and d) the use of average base count values eliminates the length bias when comparing transcriptional features of different length [19].

Identification of transcription start sites

Several published RNA-seq studies have focused on transcription start site (TSS) identification [7,9,10,20-28]. The annotation of TSSs is essential for analyzing promoters, 5' UTRs, operon architecture, and for discovering novel transcripts. To assure accuracy, a set of "best practices" for TSS identification is emerging. Enrichment of the 5' RNA ends that are generated by transcription initiation is critical for accurate TSS identification. The many advantages of dRNA-seq were recently reviewed [2]. The initiating nucleotide in bacteria is a nucleotide triphosphate, which can be distinguished from 5'-monophosphate and 5'-OH containing RNAs that are generated by RNA processing or RppH pyrophosphohydrolase activity [29]. The enrichment strategy preferred by many researchers makes use of 5' monophosphate-dependent terminator exonuclease (TEX), which degrades RNA with 5' monophosphate ends to enrich for primary transcripts that contain 5' triphosphate ends and hence represent the product of transcription initiation [10]. dRNA-seq works by enumerating differences in base counts between TEX-enriched and unenriched sequencing libraries. Experimental replication is critical for accurate TSS identification. Since dRNA-seq is remarkably reproducible, comparison of datasets generated by using the same protocols yet different growth conditions adds confidence to the process and the use of different growth conditions also increases the number of mapped TSSs. RNA samples from many growth conditions can be pooled for dRNA-seq identification of thousands of promoters [9]. For example, a recent dRNA-seq analysis of *Salmonella* using RNA pooled from 22 different growth conditions led to mapping of 96% of the TSSs that could be identified by independently analyzing the 22 samples [9].

When annotating transcriptome data, it is convenient to use widely available computer programs to search dRNA-seq datasets for TSSs [20,30,31]. The advantages of the computational process compared to manual annotation are the speed and precision of

recording transcription feature locations. However, like all bioinformatics approaches, some features will be missed and there will be false positives. In the end, human supervision of the results is critical and the state-of-the-art in transcriptome annotation remains a manual process [9]. Manual annotation of TSSs is made more efficient by plotting the count of only the first base at the 5' end of each TEX-enriched read (Fig. 1A) [32]. In practice this allows visualization of the 5' triphosphate nucleotide at the TSS.

Subsequent to identification of TSSs by dRNA-seq, bioinformatics and functional analyses can add weight to promoter identification. For example, the DNA sequences immediately upstream of putative TSSs can be analyzed by using a bioinformatics approach to score sigma factor specific RNA polymerase binding sequence motifs [4,33]. ChIP determination of RNA polymerase binding provides a robust and comprehensive validation of putative promoters [23]. When used in combination, dRNA-seq, consensus amongst experimental replicates, promoter sequence analysis, and RNA polymerase binding assays are a powerful set of tools for the identification of promoters.

Annotation of 3' ends

To obtain the full analytical value of RNA-seq data it is essential to map the 3' transcript ends. Annotating 3' ends is a notably more difficult endeavor than mapping TSSs because there currently is no method of enriching for them. The 3' ends are the primary sites of exonuclease-dependent RNA decay, which may be the reason that RNA base counts decline at the 3' ends of operons, and few reads extend into the stem loop structures of intrinsic terminators (Fig. 1C). Further complicating 3' end analysis is that termination is typically inefficient [34], which allows read-through transcription. Currently, the best method for annotating 3' ends is to search for correlation between replicates of the furthestmost downstream bases transcribed, keeping in mind that the base counts near the 3' end will be low even for highly expressed transcripts. Comparison of the 3' ends to terminator predictions adds confidence to the analysis. For example, the TransTermHP software package works very well for finding intrinsic terminators [35]. In addition, a ChIP-chip analysis of the distribution of RNA polymerase after treatment with the Rho-specific inhibitor bicyclomycin led to identification of 200 Rho-dependent terminators [36]. Once both the 5' and 3' transcript ends are mapped, it is possible to annotate operons.

Annotation of operons

The transcriptome is a map of the activities of promoters and terminators. These activities are located on both strands of the genome [37] and depending on their arrangement, can give rise to antisense transcription and overlapping, divergent [38,39] and convergent operons [40,41]. To accommodate this naturally occurring complexity it is necessary to annotate the operon architecture. Three transcriptional features are necessary to define operons: 5' ends (promoters), 3' ends (terminators), and sufficient RNA-seq read coverage to connect the ends. If sequence reads cover 90% of the bases, this is a sensible indicator that the operon is real [4,32]. While there are computer algorithms that can find operons [5,42,43], just as for TSS mapping, the state-of-the-art remains a manual process [9]. Once the operons have been mapped, it is a straightforward task to annotate additional promoters and terminators within

operons, which add complexity to the transcriptome. Mapping of internal promoters can be done manually or by bioinformatics analysis of mapped promoters that fall within the base locations of annotated operons. The transcriptional feature locations can be formatted as a GenBank feature file by using “promoter”, “terminator” and “operon” as feature keys (see for example, GSE52059 [4]). This format accommodates incremental annotation of condition specific regulatory information and is an accepted standard for disseminating genome annotation data [44]. Once the transcriptional feature locations are annotated, it is reasonably straightforward to calculate the average base count value for each feature, from each dataset, as described below.

Computing the activities of transcriptional features

Analysis of RNA-seq reads at the base count level permits normalized base counts to be readily averaged across any range of base locations to calculate the relative expression level, activity, or efficiency of individual transcriptional features [4]. We determined empirically that computing the average count of the first 10 transcribed bases accurately represents promoter activity and allows closely spaced promoters to be discriminated [4]. Likewise, the efficiency of transcription termination can be calculated as the relative decline in average base counts in 25-base windows before and after terminators (Fig. 1C). The relative transcript levels of operons can be calculated by averaging the base counts from the promoter to the terminator locations. Likewise, the expression levels of alternative transcripts generated by promoter and terminator activities within operons can be calculated. These applications of single-nucleotide-resolution analysis are exemplified in Fig. 1, for wild type *E. coli* K-12 during logarithmic growth on glucose minimal medium and during starvation for carbon (stationary phase) or nitrogen, as well as an *rpoS* mutant during nitrogen starvation.

The *cysK-ptsHI-crr* operon contains 4 genes and multiple transcription units (Fig. 1A). Conservatively, more than 40% of *E. coli* operons contain multiple transcription units that are differentially expressed, underscoring the need for an annotation system that accommodates operon architecture [4]. In addition to the primary promoter (P-1) and terminator (T-B) that define the operon, there are 8 additional promoters and one terminator within the operon (Fig. 1A). The activities of the promoters range from 12 to more than 10,000 average base counts (calculated from +1 to +10 at each promoter) and their relative activities under the four growth conditions are plotted in Fig. 1B.

There are two promoters (P-1 and P-2), separated by 33 base pairs, which drive transcription of *cysK* (Fig. 1B). Comparison of the average counts of the first 10 transcribed bases indicates that P-2 is greater than 30-fold more active than P-1. Inefficient termination (approximately 40% of *cysK* transcripts are not terminated, as indicated by the ratio of average base counts) at the internal terminator (T-A) suggests that *cysK* and *ptsHI-crr* are co-transcribed (Fig. 1C). Nevertheless, the T-A terminator segments the operon into *cysK* and *ptsHI-crr* specific transcripts, which makes sense because CysK is a cysteine biosynthetic enzyme and the remaining genes encode components of the phosphotransferase system (PTS) involved in sugar uptake [45]. In the current annotation these genes are thought to comprise two operons (*cysK* and *ptsHI-crr*) [46], but the data in Fig. 1 show a

low but significant number of RNA-seq reads across the terminator T-A, most clearly in the log phase sample. There is also a promoter (P-3) internal to *cysK* that under all four conditions is relatively active compared to the other promoters and could contribute to transcription across the *cysK-ptsH* intergenic region (Fig. 1B), yet P-3 activity does not appear to correlate with the base counts in the corresponding unenriched samples and therefore is unlikely to contribute to operon function (Fig. 1A). Given its location at the end of a transcript and immediately upstream of an inefficient terminator, this could be an example of a pervasive transcript, which is discussed below.

Two promoters, P-4 and P-5, which are located within the *cysK-ptsH* intergenic region, drive transcription of *ptsHI-crr*. P-4 is approximately 15 times more active in logarithmic phase than it is under the other three conditions (Fig. 1B). On the other hand, P-5 is induced (2.5-fold) in stationary phase and nitrogen-starved conditions by comparison to logarithmic phase and its activity is *rpoS*-dependent, as indicated by a 40-fold reduction in promoter activity by comparison to the wild type under the same conditions (Fig. 1B). The transcripts originating from these two promoters apparently are terminated at T-B, downstream of *crr* (Fig. 1C). The collective activities of P-4 and P-5 correlate well with the modest decline in average base counts of the P-4:T-B (*ptsHI-crr*) transcript upon entry into stationary phase (Fig. 1D). Within the *ptsI* gene are three closely spaced promoters (P-6, P-7, and P-8) that are of relatively low activity compared with the others (Fig. 1B). P-6 is expressed approximately equally in the four conditions, P-7 is induced in stationary phase and nitrogen-starved conditions and is RpoS-dependent, and the least active of the three, P-8, is also dependent RpoS. It does not appear that these three promoters contribute to transcription of the downstream *crr* gene, as indicated by a lack of change in the unenriched base counts visualized in Fig. 1A, and so these promoters could also generate pervasive transcripts. On the other hand, P-9 is highly active in stationary phase and nitrogen-starved conditions, is RpoS-dependent, and is located near the 3' end of *ptsI* (Fig. 1B), where it apparently drives expression of a *crr* specific transcript (Fig. 1A).

Time series analysis shows that the three major transcripts within the operon are differentially expressed during growth and entry into stationary phase (Fig. 1D). The *cysK*-specific transcript is expressed at high levels during logarithmic phase and its level declines rapidly during stationary phase. Hence expression of *cysK* reflects the decline in P-1 and P-2 promoter activity in stationary phase and nitrogen-starved conditions. The *ptsHI-crr* transcript level declines little during the first 30 min of stationary phase and then declines modestly 3 hours into stationary phase (Fig. 1D), probably because P-4 is less active and P-5 is induced upon entry into stationary phase (Fig. 1B). Expression of the *crr* transcript is partially dependent on read-through from promoters within *ptsH* and *ptsI*, and there is no evidence from the base counts to indicate that there is termination within the *ptsI-crr* intergenic region. The *crr*-specific transcript level increases upon entry into stationary phase in the wild type, yet declines in an RpoS-dependent manner in the *rpoS* mutant (Fig. 1D). Indeed, P-9 is RpoS dependent, as indicated by 16-fold higher expression in the wild type starved for nitrogen compared to the *rpoS* mutant, and it has a -10 promoter element with the base sequence (CTAnnnTTAA) that is characteristic of RpoS promoters [47].

The primary goal of many RNA-seq experiments is to determine differential gene expression between growth conditions and treatments [9,19,27,32,48-52]. Typically these experiments involve calculating for control and test conditions the number of reads that map to the genome between the start and stop codons of individual genes. Similarly, differential expression of operons can be determined by calculating the average base counts between the promoters and terminators. Since the average operon contains 2 genes, plus intragenic sequences, and 5' and 3' UTRs, there is significantly more information used (more bases) to compute the operon expression level than what is available to represent expression of individual genes. So, the statistical significance of differential expression can be greatly enhanced by using normalized base count data to measure relative operon or transcript expression levels. Differential transcription of operons is readily accomplished by employing algorithms such as DEseq [48] to compute the differential expression and statistics.

Challenges

Massive amounts of RNA sequencing data can now be readily obtained. Precise mapping of transcriptional features, logical organization of the annotated data, and meaningful feature quantitation are key to maximizing the value of the resulting transcriptomes. Critical analysis of dRNA-seq data is needed to minimize the number of false positive promoters annotated. Thus it is necessary not only to properly replicate dRNA-seq experiments, but also to augment the analysis with information to corroborate that a predicted TSS is indeed a functional promoter, such as by promoter motif analysis and RNA polymerase binding assays. It would be useful if future advances in TSS mapping technology include methods to directly label the nucleotides corresponding to TSSs, rather than simply enriching for them. Mapping of 3' transcript ends is an even larger issue and there is a real need for technology that directly labels the 3' ends generated by transcription termination. Perhaps in vitro poly(A) tailing of the 3' ends of RNA prior to fragmentation, followed by sequencing from that end would be helpful. However, it appears from existing RNA-seq data that termination is not a precise biological process and transcripts do not stop at a single nucleotide. For the time being, the state-of-the-art for 3' transcript end mapping remains consensus between replicates.

Lastly, it is important to determine whether “pervasive transcription”, defined as TSSs in non-canonical locations [53], is real and if such transcripts have a functional role. Pervasive transcription is seen in yeast, mammals, and fruit flies [54,55] and is frequently observed in viruses and bacteria [32,56,57]. So, there seems to be little doubt that pervasive transcription is real. As to whether pervasive transcripts are functional, that topic was recently reviewed, but it is too early to be sure [53]. The finding that some pervasive transcripts in herpesvirus decreased viral protein production [56] suggests that the functional role of such transcripts should be investigated in bacteria. It is becoming apparent that H-NS and NusG suppress some pervasive transcripts [57,58]. Several potential examples of pervasive transcription can be seen in Fig. 1. Using a conservative approach we previously mapped 4 promoters to the *cysK-ptsHI-crr* operon [4]. However, dRNA-seq revealed 9 promoters that map to the operon (Fig. 1A), only 4 of which appear to drive transcription of the corresponding genes (P-2, P-4, P-5, and P-9). The other 5 include a weak promoter upstream of the major

promoter in front of *cysK* and a relatively strong promoter located within the *cysK* coding region and just upstream of the terminator that is intergenic to *cysK-ptsH*. Neither of these promoters appears to contribute to transcript expression levels. The remaining 3 putative pervasive promoters are located within the *ptsI* gene, have relatively low activity levels, and yet all have reasonably well conserved -10 promoter sequence elements, including two that have RpoS promoter motifs and appear to be RpoS-dependent. If these turn out to be real promoters, and there is no reason to think they are not, then the number of promoters on bacterial genomes is being underestimated by perhaps two-fold [9,32].

Acknowledgements

Research in the authors' laboratory was funded by the NIH (GM095370).

References

1. Croucher NJ, Thomson NR. Studying bacterial transcriptomes using RNA-seq. *Curr Opin Microbiol.* 2010; 13:619–624. [PubMed: 20888288]
- 2 **. Sharma CM, Vogel J. Differential RNA-seq: the approach behind and the biological insight gained. *Curr Opin Microbiol.* 2014; 19:97–105. [PubMed: 25024085]
- 3 *. Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BO. The transcription unit architecture of the *Escherichia coli* genome. *Nat Biotechnol.* 2009; 27:1043–1049. [PubMed: 19881496]
- 4 **. Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, Teramoto J, San Miguel P, Shimada T, Ishihama A, et al. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. *MBio.* 2014; 5:e01442–01414. [PubMed: 25006232]
5. Li S, Dong X, Su Z. Directional RNA-seq reveals highly complex condition-dependent transcriptomes in *E. coli* K12 through accurate full-length transcripts assembling. *BMC Genomics.* 2013; 14:520. [PubMed: 23899370]
6. Neidhardt FC, Bloch PL, Smith DF. Culture medium for enterobacteria. *J Bacteriol.* 1974; 119:736–747. [PubMed: 4604283]
- 7 *. Thomason MK, Bischler T, Eisenbart SK, Forstner KU, Zhang A, Herbig A, Nieselt K, Sharma CM, Storz G. Global transcriptional start site mapping using dRNA-seq reveals novel antisense RNAs in *Escherichia coli*. *J Bacteriol.* 2014
8. Lahens NF, Kavakli IH, Zhang R, Hayer K, Black MB, Dueck H, Pizarro A, Kim J, Irizarry R, Thomas RS, et al. IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol.* 2014; 15:R86. [PubMed: 24981968]
- 9 **. Kroger C, Colgan A, Srikumar S, Handler K, Sivasankaran SK, Hammarlof DL, Canals R, Grissom JE, Conway T, Hokamp K, et al. An infection-relevant transcriptomic compendium for *Salmonella enterica* Serovar Typhimurium. *Cell Host Microbe.* 2013; 14:683–695. [PubMed: 24331466]
- 10 **. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R, et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature.* 2010; 464:250–255. [PubMed: 20164839]
11. Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics.* 2012; 13:734. [PubMed: 23270466]
12. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012; 9:357–359. [PubMed: 22388286]
13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009; 25:2078–2079. [PubMed: 19505943]
14. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-generation genome browser. *Genome Res.* 2009; 19:1630–1638. [PubMed: 19570905]

15. Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, Galaxy T, Taylor J, Nekrutenko A. Dissemination of scientific software with Galaxy ToolShed. *Genome Biol.* 2014; 15:403. [PubMed: 25001293]
16. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics.* 2010; 26:2204–2207. [PubMed: 20639541]
17. Goecks J, Nekrutenko A, Taylor J, Galaxy T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 2010; 11:R86. [PubMed: 20738864]
- 18 *. Forstner KU, Vogel J, Sharma CM. READemption-a tool for the computational analysis of deep-sequencing-based transcriptome data. *Bioinformatics.* 2014
19. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics.* 2012
- 20 *. Dugar G, Herbig A, Forstner KU, Heidrich N, Reinhardt R, Nieselt K, Sharma CM. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. *PLoS Genet.* 2013; 9:e1003495. [PubMed: 23696746]
21. Jager D, Forstner KU, Sharma CM, Santangelo TJ, Reeve JN. Primary transcriptome map of the hyperthermophilic archaeon *Thermococcus kodakarensis*. *BMC Genomics.* 2014; 15:684. [PubMed: 25127548]
22. Kim D, Hong JS, Qiu Y, Nagarajan H, Seo JH, Cho BK, Tsai SF, Palsson BO. Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide transcription start site profiling. *PLoS Genet.* 2012; 8:e1002867. [PubMed: 22912590]
- 23 *. Kroger C, Dillon SC, Cameron AD, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hebrard N, Handler K, et al. The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. *Proc Natl Acad Sci U S A.* 2012; 109:E1277–1286. [PubMed: 22538806]
- 24 *. Shao W, Price MN, Deutschbauer AM, Romine MF, Arkin AP. Conservation of transcription start sites within genes across a bacterial genus. *MBio.* 2014; 5:e01398–01314. [PubMed: 24987095]
25. Behrens S, Widder S, Mannala GK, Qing X, Madhugiri R, Kefer N, Mraheil MA, Rattei T, Hain T. Ultra Deep Sequencing of *Listeria monocytogenes* sRNA Transcriptome Revealed New Antisense RNAs. *PLoS One.* 2014; 9:e83979. [PubMed: 24498259]
- 26 *. Passalacqua KD, Varadarajan A, Weist C, Ondov BD, Byrd B, Read TD, Bergman NH. Strand-specific RNA-seq reveals ordered patterns of sense and antisense transcription in *Bacillus anthracis*. *PLoS One.* 2012; 7:e43350. [PubMed: 22937038]
- 27 *. Soutourina OA, Monot M, Boudry P, Saujet L, Pichon C, Sismeiro O, Semenova E, Severinov K, Le Bouguenec C, Coppee JY, et al. Genome-wide identification of regulatory RNAs in the human pathogen *Clostridium difficile*. *PLoS Genet.* 2013; 9:e1003493. [PubMed: 23675309]
- 28 *. Wiegand S, Dietrich S, Hertel R, Bongaerts J, Evers S, Volland S, Daniel R, Liesegang H. RNA-Seq of *Bacillus licheniformis*: active regulatory RNA features expressed within a productive fermentation. *BMC Genomics.* 2013; 14:667. [PubMed: 24079885]
29. Deana A, Celesnik H, Belasco JG. The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal. *Nature.* 2008; 451:355–358. [PubMed: 18202662]
30. Bischler T, Kopf M, Voss B. Transcript mapping based on dRNA-seq data. *BMC Bioinformatics.* 2014; 15:122. [PubMed: 24780064]
31. Jorjani H, Zavolan M. TSSer: an automated method to identify transcription start sites in prokaryotic genomes from differential RNA sequencing data. *Bioinformatics.* 2014; 30:971–974. [PubMed: 24371151]
- 32 **. Lin YF, A DR, Guan S, Mamanova L, McDowall KJ. A combination of improved differential and global RNA-seq reveals pervasive transcription initiation and events in all stages of the life-cycle of functional RNAs in *Propionibacterium acnes*, a major contributor to wide-spread human disease. *BMC Genomics.* 2013; 14:620. [PubMed: 24034785]
33. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics.* 2011; 27:1017–1018. [PubMed: 21330290]

34. Chen YJ, Liu P, Nielsen AA, Brophy JA, Clancy K, Peterson T, Voigt CA. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nat Methods*. 2013; 10:659–664. [PubMed: 23727987]
35. 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. [PubMed: 17313685]
36. Peters JM, Mooney RA, Kuan PF, Rowland JL, Keles S, Landick R. Rho directs widespread termination of intragenic and stable RNA transcription. *Proc Natl Acad Sci U S A*. 2009; 106:15406–15411. [PubMed: 19706412]
37. Taylor K, Hradecna Z, Szybalski W. Asymmetric distribution of the transcribing regions on the complementary strands of coliphage lambda DNA. *Proc Natl Acad Sci U S A*. 1967; 57:1618–1625. [PubMed: 5231395]
38. Piette J, Cunin R, Boyen A, Charlier D, Crabeel M, Van Vliet F, Glansdorff N, Squires C, Squires CL. The regulatory region of the divergent *argECBH* operon in *Escherichia coli* K-12. *Nucleic Acids Res*. 1982; 10:8031–8048. [PubMed: 6761650]
39. Wek RC, Hatfield GW. Nucleotide sequence and in vivo expression of the *ilvY* and *ilvC* genes in *Escherichia coli* K12. Transcription from divergent overlapping promoters. *J Biol Chem*. 1986; 261:2441–2450. [PubMed: 3003115]
40. Nomura T, Aiba H, Ishihama A. Transcriptional organization of the convergent overlapping *dnaQ-rnh* genes of *Escherichia coli*. *J Biol Chem*. 1985; 260:7122–7125. [PubMed: 2987244]
41. Sameshima JH, Wek RC, Hatfield GW. Overlapping transcription and termination of the convergent *ilvA* and *ilvY* genes of *Escherichia coli*. *J Biol Chem*. 1989; 264:1224–1231. [PubMed: 2642900]
42. Fortino V, Smolander OP, Auvinen P, Tagliaferri R, Greco D. Transcriptome dynamics-based operon prediction in prokaryotes. *BMC Bioinformatics*. 2014; 15:145. [PubMed: 24884724]
- 43 *. McClure R, Balasubramanian D, Sun Y, Bobrovskyy M, Sumby P, Genco CA, Vanderpool CK, Tjaden B. Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res*. 2013; 41:e140. [PubMed: 23716638]
44. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res*. 2014; 42:D32–37. [PubMed: 24217914]
45. De Reuse H, Danchin A. The *ptsH*, *ptsI*, and *crr* genes of the *Escherichia coli* phosphoenolpyruvate-dependent phosphotransferase system: a complex operon with several modes of transcription. *J Bacteriol*. 1988; 170:3827–3837. [PubMed: 2457575]
46. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A, et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res*. 2013; 41:D203–D213. [PubMed: 23203884]
47. Weber H, Polen T, Heuveling J, Wendisch VF, Hengge R. Genome-wide analysis of the general stress response network in *Escherichia coli*: sigmaS-dependent genes, promoters, and sigma factor selectivity. *J Bacteriol*. 2005; 187:1591–1603. [PubMed: 15716429]
48. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome biology*. 2010; 11:R106. [PubMed: 20979621]
49. Balasubramanian D, Kumari H, Jaric M, Fernandez M, Turner KH, Dove SL, Narasimhan G, Lory S, Mathee K. Deep sequencing analyses expands the *Pseudomonas aeruginosa* AmpR regulon to include small RNA-mediated regulation of iron acquisition, heat shock and oxidative stress response. *Nucleic Acids Res*. 2014; 42:979–998. [PubMed: 24157832]
50. Frazee AC, Sabunciyan S, Hansen KD, Irizarry RA, Leek JT. Differential expression analysis of RNA-seq data at single-base resolution. *Biostatistics*. 2014
51. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*. 2012
52. Wagner GP, Kin K, Lynch VJ. A model based criterion for gene expression calls using RNA-seq data. *Theory Biosci*. 2013; 132:159–164. [PubMed: 23615947]
- 53 *. Wade JT, Grainger DC. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat Rev Microbiol*. 2014; 12:647–653. [PubMed: 25069631]

54. Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al. Diversity and dynamics of the *Drosophila* transcriptome. *Nature*. 2014
55. Jensen TH, Jacquier A, Libri D. Dealing with pervasive transcription. *Mol Cell*. 2013; 52:473–484. [PubMed: 24267449]
56. Canny SP, Reese TA, Johnson LS, Zhang X, Kambal A, Duan E, Liu CY, Virgin HW. Pervasive transcription of a herpesvirus genome generates functionally important RNAs. *MBio*. 2014; 5:e01033–01013. [PubMed: 24618256]
57. Singh SS, Singh N, Bonocora RP, Fitzgerald DM, Wade JT, Grainger DC. Widespread suppression of intragenic transcription initiation by HNS. *Genes Dev*. 2014; 28:214–219. [PubMed: 24449106]
58. Peters JM, Mooney RA, Grass JA, Jessen ED, Tran F, Landick R. Rho and NusG suppress pervasive antisense transcription in *Escherichia coli*. *Genes Dev*. 2012; 26:2621–2633. [PubMed: 23207917]

Highlights

- An annotation schema that accommodates operon architecture is ideal for genome-wide bacterial transcriptome analysis.
- RNA sequencing data can pinpoint transcriptional features with single-nucleotide-resolution.
- Single-nucleotide-resolved data essentially are digital and facilitate quantitative analysis of promoter activity and terminator efficiency, as well as differential expression of transcripts.
- Challenges facing the transcriptomics field are the lack of technologies to directly label transcript ends, need to validate promoters identified by dRNA-seq, and resolving the issue of “pervasive transcription”.

Figure 1.

Transcriptional feature map and analysis of the *cysK-ptsHI-crr* operon. The dRNA-seq data are available at GEO, GSE58556. (A) The genes and feature locations are drawn to scale and annotated to the positive strand of the *E. coli* MG1655 U00096.3 reference genome. Promoters (P) are indicated by an arrow and are numbered in order from left to right on the positive strand. Terminators (T) are indicated by a diamond. The base count data, consisting of TEX-treated samples pointing up and unenriched coverage data (fragmented RNA not treated with TEX) pointing down, are visualized in J-Browse [14], as described previously [4]. Only positive strand data are shown. Tracks: wild type (WT), glucose-grown *E. coli* K-12 in logarithmic phase (blue track); WT in stationary phase, 30 min after exhaustion of glucose (red track); WT starved for nitrogen (green track); and an isogenic *rpoS* mutant starved for nitrogen (tan track). The base count scale (on the left) is from 0 to 100, with values exceeding 100 indicated by dark red. (B) The relative activities of the nine promoters is plotted in the graphs as log2 average counts of the first 10 transcribed bases under the four different growth conditions, which are colorized as above. (c) The decrease in average counts of the 25 bases before and after the terminator T-A are shown by light green and pink

arrows. (D) Time series analysis of the relative expression levels of three transcripts within the complex *cysK-ptsHI-crr* operon is plotted as the log₂ average counts of bases from the indicated promoters to terminators, as described previously [4]. Time point 1 is during middle logarithmic phase, time point 2 is immediately prior to entry into stationary phase, time point 3 is 15 min after entry into stationary phase, time point 4 is 30 min after entry into stationary phase, and time point 5 is 180 min after entry into stationary phase.

Additional details of the analysis are described in the text.