PIPETS: A statistically robust, gene-annotation agnostic analysis method to study bacterial termination using 3'-end sequencing.

Quinlan Furumo[1] and Michelle Meyer[1]
[1]Department of Biology, Boston College, Chestnut Hill, MA, 02135, United States.

Abstract:
We present PIPETS, an R package available on Bioconductor that provides a novel analysis approach to study 3'-end sequencing data. PIPETS statistically robust, gene-annotation agnostic methodology identified more significant 3'-end termination signal across a wider range of genomic contexts than existing analysis tools. PIPETS provides a broadly applicable platform to compare data sets across different organisms.

Availability and Implementation:
PIPETS is available on the Bioconductor Repository at:
https://bioconductor.org/packages/devel/bioc/html/PIPETS.html
and on Github at: http://https://github.com/qfurumo/PIPETS.

Introduction:

The study of alternative and premature bacterial transcription termination is a rapidly growing field that aims to understand bacterial responses to changes in environmental conditions. The current standard for studying this is 3-prime-end-sequencing (3'-seq)/Term-seq which non-specifically tags the 3' ends of RNA, allowing for short-read sequencing to map these ends and quantify their relative frequencies. Initial studies using 3'-seq focused on identifying regulatory elements in the 5'UTRs of annotated genes by focusing exclusively on signal in the 5'- and 3'-UTRs of these genes (Dar *et al.,* 2016, Dar and Sorek, 2018; Konikkat *et al.,* 2021. This analysis methodology was used or closely mirrored by several studies afterwards, many of which noted an abundance of 3'-seq signal in other genomic regions of the data (Warrier *et al.,* 2018; Millman *et al.,* 2017; Petroni *et al.,* 2023). Other studies created their own custom analysis methods to study 3'-seq data in order to study specific regulatory signals or to add additional sequencing information (Ju *et al.,* 2019; Mandell *et al.,* 2021). Although many different studies have identified that there are biologically relevant 3'-seq signal levels in regions outside of the 3'-UTR of genes, there is not a consensus on an analysis method to best analyze 3'-seq data. Without a standardized 3'-seq analysis method, findings from different studies are impossible to compare short of reanalyzing many large-scale data sets, and it is challenging to contextualize whether findings from diverse organisms are due to differences in the biology, the data collection method, or the data analysis performed.

In this work, we introduce **PIPETS** (**P**oisson **I**dentification of **PE**aks from **T**erm-**S**eq data). PIPETS is an R package available on Bioconductor that analyzes mapped data from 3'-end sequencing, which nonspecifically captures transcript termination signal, whether that be from transcript termination, regulatory cleavage, truncation, or other biological processes (Saberi *et al.,* 2016; Chao and Vogel, 2016; Ruiz de los Mozos *et al.,* 2013; Vanderpool *et al.,* 2011; Rosinski-Chupin *et al.,* 2015; Papenfort and Vogel, 2010). In contrast to many existing approaches for 3'-seq analysis, PIPETS analyzes all 3'-seq signal, regardless of genomic context, and uses statistically robust methods to identify biologically relevant 3'-end signal. We demonstrate the use of PIPETS on two existing 3'-seq datasets.

Methods and Features:

The inputs for PIPETS are mapped 3'-seq Bed files that contain at least chromosome, read start, read stop, and strand information for each read. Alternatively PIPETS can use Genomic Ranges objects from the GenomicRanges package so long as the input data contains the same information as above (Lawrence *et al.,* 2013). With these inputs it calculates the 3'-end read coverage of every genomic position using the terminal position of each read corresponding to the terminal site in the transcript. PIPETS subsequently analyzes this data in three main steps: 1) A sliding window function which uses a Poisson Distribution Test to identify genomic coordinates with significant 3'-end termination read coverage and 2) A condensing step that combines the signal of consecutive significant genomic positions into one termination "peak" and 3) A peak condensing step that combines proximal significant peaks into one signal. These

steps ensure that technical or biological error does not conflate noise as significant signal while pruning the total number of individual peaks reported to make analysis manageable.

One of the inherent challenges associated with analysis of RNA-seq data generally is the large range of read values obtained across different transcripts. Low read coverage does not necessarily indicate a lack of biological significance, as some transcripts have a low basal expression level. However, in the case of 3'-end sequencing, the "noise" created from cleavage of transcripts (either during *in vitro* RNA processing or within the cell) will also be higher for more highly expressed transcripts, potentially much higher than true signal for a lowly expressed transcript. To account for this, PIPETS uses a sliding window approach to ensure that significant termination read coverage is assigned locally. By default, the sliding window spans 51 genomic positions and uses a Poisson Distribution Test to determine if the read coverage of any position is significantly higher than expected from the surrounding average. This ensures that genomic positions with read coverage dwarfed by highly expressed regions are not overlooked, increasing PIPETS sensitivity. Even for input data with millions of reads, a vast majority of genomic coordinates will have very few or no termination reads mapped to them. Because every single genomic coordinate cannot be the site of a termination event, we associate most low read coverage events to technical or biological noise that are unlikely to be biologically significant. PIPETS uses a global minimum read cutoff based on the total read coverage of each file. Positions with 3'-end read depth below this cutoff are not considered, speeding analysis and preventing false positive results from noise. While the total number of tests that occur is high with input data of large genomes, the concern of type 1 errors remains low as each genomic coordinate is tested at most twice. In addition, PIPETS' default p-value for significance is set to 0.0005 and can be adjusted to increase stringency.

The structure of a true bacterial termination for 3'-seq data resembles a peak, with a position of highest read coverage and flanking positions of lesser read coverage that are still larger than surrounding noise. This imperfect signal is the result of several factors, namely imperfect bacterial termination, RNA degradation, and technical noise from the sequencing itself. Each genomic position from the significant results of step one is unlikely to be an independent termination signal; many of these positions are adjacent to others and together they form one site of termination. To better reflect the biological context of these data, the significant genomic positions identified in step one undergo two condensing steps: 1) Combine consecutive significant positions into single termination peaks, and 2) Combine proximal termination peaks to abbreviate total results volume. The first round of condensing performed by PIPETS identifies significant positions that are within 2 nucleotides of each other. Consecutive significant positions are then combined into a single termination "peak" which contains the full range of consecutive positions and uses the highest read coverage position within for downstream analysis. PIPETS then combines proximal termination peaks within 20 nucleotides of each other as termination signals so close to each other are likely not independent sites of transcript (Conway *et al.,* 2014; Stringer *et al.,* 2020; Mitra *et al.,* 2017).

<u>Default Parameters:</u>

To establish default values for the parameters in PIPETs, we assessed robustness across all run parameters (Supplementary Data 1). The most impactful parameters for adjusting the analysis are threshAdjust and topEndPercentage. These two parameters are used to generate the global read coverage cutoff used to differentiate noise from biological signal, and testing revealed that minor adjustments to these values can greatly affect the distribution of read coverage values that are assessed. PIPETS sorts the genomic coordinates from highest read coverage to lowest and identifies the genomic positions that account for a percentage of the total read coverage of the file as defined by threshAdjust (75% by default). These highest read coverage values of each file, which are most likely to be derived from biologically significant signal, are used to create a minimum read coverage cutoff for that input file. This results in a global cutoff that is informed by biologically relevant signal that can be used to better filter out noise. To account for possible high read coverage outliers, topEndPercentage defines the top percentage (1% by default) of these highest read coverage positions that are excluded from the calculation of the global read coverage cutoff. All positions removed this way will still be analyzed by PIPETS, but this helps to ensure that PIPETS is not heavily biased by either high or low read coverage outliers.

In datasets with very high read depths or datasets from very long genomes, we recommend increasing the strictness of PIPETS' noise filtering by lowering the value of threshAdjust (from 75% to values from 50% - 60%). This will reduce the effect of global increases to basal noise levels and ensure that PIPETS remains most sensitive to biologically relevant signal. Alternatively, increasing the value of threshAdjust (from 75% to values from 80% - 95%) will make PIPETS less strict and is useful when studying data sets with low read coverage or data from smaller genomes. In the event where a small portion of the genomic positions in the data account for a disproportionately large percentage of the total read coverage for a data set, increasing the value of topEndPercentage (from 1% to values from 1.5% - 2.5%) can reduce the bias imparted by these outliers. topEndPercentage is a more sensitive parameter and its values may be more variable from dataset to dataset. Ranges of values were tested for all other parameters, however none were identified as having a large impact on the final output. Aside from pValue, these parameters do not affect the identification of significant results, and are used to attenuate the data scanning process.

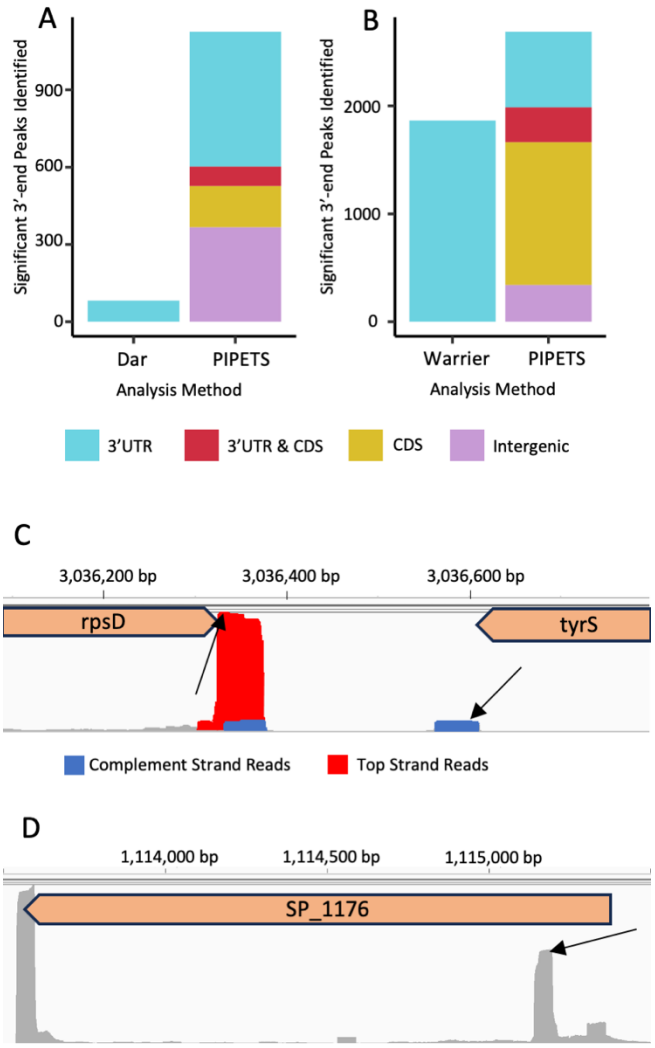

Figure 1: (A,B) PIPETS identified substantially more significant peaks when compared to the original analysis methods for both *B. subtilis* (A) (Dar *et al.,* 2016) and *S. pneumoniae* (B) (Warrier *et al.,* 2018) data sets. At least half of the significant peaks identified by PIPETS in either data set were found in gene coding regions (CDS), intergenic regions (Intergenic), or in the 3'-UTR of one gene and the coding region of a different gene (3'-UTR & CDS); all of which were completely unanalyzed by the original method. (C) Example of proximal termination peaks on opposite strands that would be missed by many existing analysis methods from the *B. subtilis* data. (D) Example of a termination peak within the CDS of the SP_1176 gene that could cause premature termination from the *S. pneumoniae* data.

Results:

We first tested PIPETS on *Bacillus subtilis* data from the first paper to propose implement 3'-seq as a method of studying bacterial termination (Dar *et al.,* 2016). The sequencing depth of the original data was low relative to modern sequencing standards, with no replicate sequencing file with more than 3 million aligned reads and some with as few as 4000 aligned reads, and we were interested to see how more robust analysis methods would alter the conclusions. Their analysis methodology focused only on the position of highest termination signal in the 3'-UTR of annotated genes, did not consider reads strand specifically, and excluded all other data from

analysis. They identified 82 total termination positions that passed their thresholds and were near genes of interest. PIPETS identified 1125 significant termination peaks from the same data (See Figure 1A). 523 of these peaks were in the 3'-UTR of genes, but the remaining 602 were not (368 in intergenic regions, 159 inside of coding regions, and 75 that were present in the 3'-UTR of one gene and inside the coding region of a different adjacent gene)(Supplementary Data 2). One example of two significant peaks identified by PIPETS are shown in Figure 1C, where termination signals from two genes on opposite strands are close enough together that previous analysis methods would combine all signal into one termination event. By properly using strand information, PIPETS identified the strand specific termination events for each gene as well as some antisense termination signal for the rpsD gene.

PIPETS not only identified more than 6x as many termination peaks in 3'-UTRs of genes, it also identified 602 significant termination peaks that existing methods would have filtered out. While it cannot be determined if all the significant peaks are sites of altered or premature termination, they all had read coverage that was significantly higher than the surrounding data, indicating that they are all confident sites of transcript ends, which may be due to either RNA-cleavage or bona fide transcription termination. There were 37 termination positions noted in the Dar data that were not identified by PIPETS. However, only two of these termination positions had an average read coverage across all data sets of over 100. While these termination signals were proximal to gene coding regions, they lacked the significant read coverage to reach the global cutoff to be identified and thus, their biological relevance is difficult to assess.

We next tested PIPETS on *Streptococcus pneumoniae* data generated in our own lab (Warrier *et al.,* 2018). The analysis procedure for this work mirrored the Dar paper (Dar *et al.,* 2016), but significantly more reads were collected. Here we were interested how the PIPETS results would compare to an analysis tool designed for lower read depth data. The original method identified 1864 termination sites across three conditions, and as a result of the analysis parameters many of these identified sites had very low read coverages that were difficult to differentiate from surrounding noise. PIPETS identified 2686 unique significant peaks between just two of these conditions (no-drug-control samples at 0 and 30 minutes). Of these 2686 peaks, 342 were in intergenic regions, 1323 were inside coding regions, 700 were in the 3'-UTR of a gene, and 321 that were present in the 3'-UTR of one gene and inside the coding region of a different adjacent gene (See Figure 1B, Supplementary Data 2). Another example of a significant peak identified by PIPETS is seen in Figure 1D where there is a robust 3'-end signal inside of the SP_1176 gene that is comparable in magnitude to the full termination signal identified at the end of the gene. Of the original 1864 termination sites identified, PIPETS identified 904. Of the remaining 960, only 32 had read coverage over 500, and only 7 had over 1000 reads. With such relatively low read coverage at most of the missed positions, we were confident that PIPETS was consistently identifying biologically relevant termination signal without identifying noise as false positives.

Discussion:

The number of studies utilizing 3'-seq is rapidly increasing, yet there is not an easy to download, standardized analysis method currently available. Here we present PIPETS, a new 3'-seq analysis

method that identifies statistically significant 3'-seq signal in a gene annotation agnostic fashion. When compared to the existing 3'-seq analysis methodology, PIPETS identified almost all 3'-seq signal with high read coverages in the 3'-UTR of annotated genes while also identifying many more significant results in other genomic regions. PIPETS is easily available for use via Bioconductor and provides easy to parse results files. We anticipate that the annotation agnostic approach of PIPETS will enable more streamlined and consistent analysis of 3'-end sequencing data to drive the field of bacterial transcriptomics forward.

Competing Interests:
No competing interest is declared.

Author Contributions Statement:
Q.F. and M.M. conceptualized the software, Q.F. developed the software and published it, Q.F. and M.M analysed the results. Q.F. and M.M. wrote and reviewed the manuscript.

Acknowledgements:
We would like to thank Irem Ozkan and Leo Ahumada Manuel for their invaluable assistance as early testers of PIPETS. We would also like to thank the Bioconductor team for their support during the submission process.

Funding:
This work was supported by National Institute of Health (NIH) grants awarded to M.M.: R21AI148895 and R01GM134259.

Citations:

Chao Y, Vogel J. A 3'-UTR-Derived Small RNA Provides the Regulatory Noncoding Arm of the Inner Membrane Stress Response. *Molecular Cell* 2016;**61**:352-363.

Conway T, Creecy JP, Maddox SM *et al.* Unprecedented High-Resolution View of Bacterial Operon Architecture Revealed by RNA Sequencing. *mBio* 2014;**5** (4):e01442-14.

Dar D, Sorek R. High-resolution RNA 3'-ends mapping of bacterial Rho-dependent transcripts. *Nucleic Acids Res.* 2018;**46**:6797-6805.

Dar D, Shamir M, Mellin JR *et al.* Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. Science 2016;**352** (6282):aad9822.

Ju X, Li D, Liu S. Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. *Nat. Microbiol.* 2019;**4**:1907-1918.

Konikkat S, Scribner MR, Eutsey R *et al.* Quantitative mapping of mRNA 3' ends in *Pseudomonas aeruginosa* reveals a pervasive role for premature 3' end formation in response to azithromycin. *PLoS Genet.* 2021;**17**(7):e1009634.

Lawrence M, Huber W, Pagès H, *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Comput. Biol.* 2013;**9**(8):e1003118.

Mandell ZF, Oshiro RT, Yakhnin AV *et al.* NusG is an intrinsic transcription termination factor that stimulates motility and coordinates gene expression with NusA. *eLife* 2021;**10**:e61880.

Millman A, Dar D, Shamir M *et al.* Computational prediction of regulatory, premature transcription termination in bacteria. *Nucleic Acids Res.* 2017; **45**:886–893.

Mitra P, Ghosh G, Hafeezunnisa M *et al.* Rho Protein: Roles and Mechanisms. *Annu. Rev. Microbiol.* 2017;**71**:687-709.

Papenfort K, Vogel J. Regulatory RNA in bacterial pathogens. *Cell Host & Microbe* 2010;**8**:116-127.

Petroni E, Esnault C, Tetreault D *et al.* Extensive diversity in RNA termination and regulation revealed by transcriptome mapping for the Lyme pathogen *B. burgdorferi*. *bioRxiv* 2023.01.04.522626

Rosinski-Chupin I, Sauvage E, Sismeiro O *et al.* Single nucleotide resolution RNA-seq uncovers new regulatory mechanisms in the opportunistic pathogen *Streptococcus agalactiae*. *BMC Genomics* 2015; **16**:419.

Ruiz de los Mozos I, Vergara-Irigaray M, Segura V *et al.* Base pairing interaction between 5'- and 3'-UTRs controls icaR mRNA translation in *Staphylococcus aureus*. *PLoS Genet.* 2013; **9**(12):e1004001.

Saberi F, Kamali M, Najafi A *et al.* Natural antisense RNAs as mRNA regulatory elements in bacteria: a review on function and applications. *Cellular & Molecular Biology Letters* 2016;**21**:6.

Stringer AM, Baniulyte G, Lasek-Nesselquist E *et al.* Transcription termination and antitermination of bacterial CRISPR arrays. *eLife* 2020;**9**:e58182.

Vanderpool CK, Balasubramanian D, Lloyd CR. Dual-function RNA regulators in bacteria. *Biochimie.* 2011;**93** (11):1943-1949.

Warrier I, Ram-Mohan N, Zhu Z *et al.* The Transcriptional landscape of *Streptococcus pneumoniae* TIGR4 reveals a complex operon architecture and abundant riboregulation critical for growth and virulence. *PLOS Pathog.* 2018;**14** (12):e1007461.