Northumbria
University
NEWCASTLE

Deep sequencing approaches for the analysis of prokaryotic transcriptional boundaries and dynamics

Katherine James[1,*], Simon J. Cockell[2], Nikolay Zenkin[1]

[1] Centre for Bacterial Cell Biology, Institute for Cell and Molecular Bioscience, Newcastle University, Baddiley-Clark Building, Richardson Road, Newcastle Upon Tyne, NE2 4AX, UK.
[2] Bioinformatics Support Unit, Newcastle University, William Leech Building, Framlington Place, Newcastle Upon Tyne, NE2 4HH, UK.

*Correspondence to:

Dr Katherine James,
The Centre for Bacterial Cell Biology,
Baddiley-Clark Building,
Newcastle University,
Newcastle upon Tyne,
NE2 4AX,
United Kingdom.

Email: katherine.james@newcastle.ac.uk.

Abbreviations: CDS coding sequence; DOOR Database of prOkaryotic OpeRons; EMOTE Exact Mapping Of Transcription Ends; HMM hidden Markov model; IP immunoprecipitation; NN neural network; RACE Rapid-Amplification of cDNA Ends; RBP RNA-binding protein; RF random forest; RNAP RNA polymerase; RT reverse transcription; SVM support vector machine; TEX terminator exonuclease; TAP tobacco acid pyrophosphatase; TSS transcription start site; TTS transcription termination site; UTR untranslated region.

Abstract:

The identification of the protein-coding regions of a genome is straightforward due to the universality of start and stop codons. However, the boundaries of the transcribed regions, conditional operon structures, non-coding RNAs and the dynamics of transcription, such as pausing of elongation, are non-trivial to identify, even in the comparatively simple genomes of prokaryotes. Traditional methods for the study of these areas, such as tiling arrays, are noisy, labour-intensive and lack the resolution required for densely-packed bacterial genomes. Recently, deep sequencing has become increasingly popular for the study of the transcriptome due to its lower costs, higher accuracy and single nucleotide resolution. These methods have revolutionised our understanding of prokaryotic transcriptional dynamics. Here, we review the deep sequencing and data analysis techniques that are available for the study of transcription in prokaryotes, and discuss the bioinformatic considerations of these analyses.

Keywords: Deep-sequencing; Prokaryotic transcription; Transcription start sites; Transcription termination sites; Transcriptional dynamics; Bioinformatics

Funding: This work was supported by Grants from UK Biotechnology and Biological Sciences Research Council, Wellcome Trust and Leverhulme Trust to N.Z.

1. Introduction

Deep sequencing techniques have provided the opportunity to gain a more detailed and accurate understanding of the bacterial transcriptome [1-7]. These techniques were originally designed for the study of eukaryotes, and have traditionally been used for the analysis of differential gene expression [8, 9]. The development of experimental techniques and analysis resources for prokaryotic transcription has therefore lagged behind. This deficiency was due in part to technical difficulties involved in enriching bacterial mRNAs, which lack the poly(A) tail utilised in eukaryotic RNA-Seq; alternative priming approaches, such as artificial polyadenylation and random hexamers are used for bacterial RNA-Seq [5]. It was also generally assumed that bacterial genomes are very simple and do not require such in-depth analysis [4]. However, bacterial transcriptomes have been found to be far more complex and dynamic than previously thought [10], and a number of prokaryote-specific deep sequencing methods have been developed to accurately investigate this complexity [4].

2. The prokaryotic transcriptome

Prokaryotic transcriptional units often overlap (Figure 1) [11]. In addition to the translated coding sequences (CDS), which produce the final protein products, a bacterial transcriptional unit can contain untranslated regions (UTRs) that are bordered by the transcription start and termination sites (TSS and TTS, respectively), and which can contain regulatory regions [12, 13]. The DNA sequences downstream of the TSS (5' UTRs) are often essential to transcription, since they may contain regulatory factors such as secondary structures [14]. However, leaderless mRNAs are also found in prokaryotes that have no 5’ UTR; the ribosome binds directly to the start AUG without the need for additional regulatory structures [15, 16]. TSS can be classified as primary (upstream of a CDS), secondary (upstream but weaker than a CDS's primary TSS), internal (within a sequence feature on the sense strand), antisense (within a sequence feature on the antisense strand), or orphan (un-associated with annotated regions) [17-20]. The bioinformatic identification of promoters and binding sites can be non-trivial from genome sequence alone. However, since promoter binding occurs ~6-8 nucleotides from the TSS, experimental identification of the TSS aids in the identification of promoters, binding sites and other regulatory structures [21, 22].

Traditionally, TSS have been identified for specific genes of interest by small scale methods such as primer extension [23] or the PCR-based 5' RACE (Rapid-Amplification of cDNA Ends) [24], which are accurate for TSS identification but inefficient and time-consuming [7]. Tiling arrays consisting of high density oligonucleotide probes can be used to identify TSS with accuracy varying from ~30 to 5 nucleotide resolution and, therefore lack precision [4, 12, 25-29]. Furthermore, the signal for some genes can be close to the level of background noise [30, 31]. Finally, ChIP-chip array methods have been used to identify promoters by capturing the transcription machinery following immobilisation of the RNA polymerase [32, 33]. These methods, however, provide even lower resolution TSS determination.

The DNA sequences upstream of the TTS (3' UTR) also contain regulatory regions, such as conditional terminators, and have been linked to translational regulation in archaea [34]. There are two types of terminator in prokaryotes: Rho-dependent and intrinsic (reviewed in [35] and [36]). Intrinsic terminators consist of a thymine-rich stretch of DNA preceded by a GC-rich hairpin [37]. While terminators can be identified from sequence to a certain extent [38], their identification is greatly aided by identification of the TTS. However, the identification of the TTS is non-trivial, due to the inefficiency of termination [39] and exonuclease degradation [1] making the boundary less clear than that of the TSS, particularly where transcripts overlap.

Once the TSS and TTS have been identified, the continuous expressed sequence in between them defines the transcriptional unit [12], which may contain a single CDS, an operon of multiple CDSs or other untranslated elements such as tRNAs, rRNAs and regulatory small RNAs [5, 6]. Computational methods can use sequence data and features of known operons to predict transcriptional units, but these methods lack sensitivity [40-42].

Non-coding RNAs are widespread in bacterial genomes, both intergenically (sRNAs) and on the anti-sense strand (asRNAs) [5, 18, 29, 43-48]. Many of these RNAs can be difficult to identify due to their

small size (~50-500bp), location and short half-life [4, 49]. Small non-coding RNAs (sRNAs) have been linked to several aspects of gene expression control including mRNA stability, transcriptional termination, and the RNA-based regulation of diverse cellular processes [50-55]. While several asRNAs have been functionally characterised (reviewed by Georg and Hess [56]), it remains uncertain whether most asRNAs have a biological role or are artefacts produced by spurious promoters and are mostly transcriptional noise [57-60].

The complexity of the prokaryotic genome is further increased by its conditional nature. TSS can change depending on condition [11, 12, 21, 61] and can be cell cycle dependent [25]. Consequently, the transcriptome identified in one condition can differ greatly from that in another [62, 63]. Internal promoters can produce sub-operons, making operons modular and giving flexibility to gene expression [4, 40, 64]. For instance, the glpEGR operon of *E, coli* has three internal promoters potentially producing three suboperons of different lengths [65]. Detection of these operon dynamics requires specific experimental design and analysis, for example the use of differential RNA-Seq (discussed below).

The process of transcription is itself dynamic and non-continuous. RNAP has been observed to pause approximately once per every 100 bp in the *E. coli* genome [66, 67]. Pausing is thought to be involved in the regulation of initiation at the promoter [68]. Pauses are over-represented at TSSs, and are enriched within the first 100 nt of expressed genes [66]. Pausing is also involved in Rho-dependent and intrinsic termination [36, 69, 70], where pausing allows Rho factor to catch up with RNAP or GC hairpin to form, respectively [35, 71]. Finally, pausing has been associated with misincorporation events [72-74].

3. RNA-Seq

Deep sequencing has quickly taken over from array-based methods for the study of bacterial transcriptomics, since it allows direct sequencing of the entire transcriptome in a high-throughput manner [4, 6, 19, 45, 48, 64, 75-84]. Sequencing techniques are also far more efficient and cost-effective, as well as being far more accurate, since arrays have high levels of noise due to non-specific cross-hybridisation of the probes [6, 12]. A basic RNA-Seq protocol involves extraction of total RNA, which is converted into a cDNA library by reverse transcription (RT), fragmented and then sequenced. However, there are several possible variations to this protocol (Figure 2 A).

The sequencing depth required for transcriptional analysis is dependent upon the focus of the study varies (discussed in [85]). Different types of transcripts vary in abundance by several orders of magnitude, therefore, the number of reads required to detect these transcripts also varies. Given that sequencing projects often have constrained budgets, several techniques can be used to optimise read depth which are discussed in the following sections.

3.1 Replicates

Several studies have addressed the need for experimental replicates in eukaryotic RNA-Seq data [86-90]. Although sequencing is highly reproducible [89], replicates are essential when looking at single base resolution data [87], and provide more power than increased sequencing depth [88]. Technical replicates allow for technical noise in the data and are needed when evaluating methodologies [87], although it should be noted that at very low coverage technical replicates can vary [91]. Biological replicates are more meaningful when studying bacterial transcriptomics since they take into account both technical and biological variation between samples [87]. The number of replicates required for a study is dependent upon the biological question being asked of the data [92], although in *Saccharomyces cerevisiae* a minimum of six biological replicates has been recommended for differential expression analysis [90]. A recent study of replicate variation in bacterial RNA-Seq highlighted the need to minimise experimental variation between replicates, in particular by using consistent media lots [93]. Batch effects caused by experimental variation, such as media lots, personnel or laboratory conditions, can be a major problem in high-throughput deep sequencing, particularly if the batches correlate with the measured outcome [94]. For instance, if samples from two conditions are prepared on separate days by different individuals, there will likely be a batch effect in addition to a biological effect between the two samples. Therefore, variation should be minimised

during experimental design, and batch correction should be carried out during the normalisation stage where required (see Section 4.4).

3.2 mRNA enrichment

Since mRNAs can comprise just 1-20% of the total RNA extract [12, 95, 96], enrichment of mRNAs, for instance by rRNA depletion, is often carried out prior to RT in order to improve coverage of protein coding regions (methods for enrichment are reviewed in [5] and evaluated in [97]). Enrichment may also be necessary when using host-bacterial mixed samples in which the host RNAs will also contaminate the sample [85, 98]. However, this enrichment step is not always necessary; for instance, dRNA-Seq (described in Section 5.1) is often performed without enrichment in order to study the whole transcriptome [99]. However, a far greater read depth is required if mRNA enrichment is not used [85]. Alternatively, fragment selection based on strand size may also be used to improve sequencing results; for example to remove larger RNAs where short RNAs (sRNAs) are of interest [100], or fractionation to divide the RNAs into different groups prior to sequencing (see Section 5).

3.3 Amplification

Where the original sample is very small a PCR amplification step is often required prior to sequencing. For example, some RNA-Seq variations, such as NET-seq discussed in Section 5.3, only use a subset of the cellular RNAs and require amplification. In addition, single cell samples also require amplification prior to sequencing [101].

3.4 Sequencing

There are various options for sequencing which will each require the addition of the correct adaptors during library preparation. Paired end sequencing (sequencing from both ends of the strand) is not generally needed for prokaryotic genomes, since they lack the splicing variants common to eukaryotes [6], but can be useful to increase coverage, when *de novo* transcript assembly is required, and when genomic rearrangements are of interest. However, in most cases, single-ended 50bp RNA-Seq is adequate for most prokaryotic transcriptional studies to provide unique mappings to the genome, and is more cost effective than paired end sequencing.

Strand specificity is important for the detection of overlapping UTRs and antisense transcription [5, 102]. These methods, termed ssRNA-Seq, use directional adaptors to allow the distinction between strands during the alignment stage [12, 82, 103]. A combination of paired end and stranded sequencing has been used to improve the identification of transcript boundaries [104]. The incorporation of barcoded adapters, termed multiplexing, allows reads from different samples to be sequenced in the same lane [85], which may be preferable when sequencing studies are budget-constrained.

4. Bioinformatic analysis

RNA-Seq produces millions of short sequence reads ranging from ~20 bp to 200 bp depending on the sequencing platform used. To analyse these data they must first be mapped to a reference genome and counted. However, there are several optional steps to the analysis (Figure 2 B). The handling of the data is dependent on the question being asked and the data type [105]; many existing RNA-Seq analysis pipelines are designed for differential expression analysis in eukaryotes and are not suitable for other analyses or data types in prokaryotes.

4.1 Quality control

An essential first step is to assess the quality of the data; for instance, by using an automated program such as FastQC[1] which assesses quality, content and sequence duplication levels. These tests identify possible artefacts which can be removed by pre-processing. However, several RNA-Seq variations have unusual sequence distributions, and will fail most of the sequence content and duplication tests used for standard RNA-Seq data. For instance, NET-seq data (discussed in Section

[1] http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc

5.3) is repetitive due to pause consensus sequences and will fail the sequence GC content test, which assumes an even GC distribution across all the reads. Therefore, interpretation of the quality control reports is dependent on the data type.

4.2 Trimming and filtering

Some pre-processing of the reads may be required prior to alignment. Many RNA-Seq analysis pipelines often heavily trim and filter reads based on sequencing quality [92]. Trimming of low quality bases is useful to improve alignment for differential expression analysis, where read counts are per genomic feature (such as CDS), but is often inappropriate for single nucleotide resolution data, where counts are per nucleotide and preserving the ends of the reads is often essential. For example, trimming at the 3' ends of reads will add noise when identifying TTS from RNA-Seq data (see Section 5). Therefore, while adaptor and primer sequences should be clipped from the reads prior to alignment [106], further trimming may be detrimental to the results. Where the 5' end of the read is of interest, for example NET-seq and Term-seq data, where it corresponds to the 3' end of the transcript (Sections 5.2, 5.3 and Figure 3), truncation of the error-prone 3' section of the read will aid alignment, although some read aligners, for example segemehl [107], will ignore this section during mapping anyway. Alternately, rather than hard trimming, filtering based on sequence quality may be applied at each position individually to preserve 5' positions [74]. Filtering can also be applied to remove whole reads that fall beneath a minimum length and those with low overall quality.

4.3 Alignment

Once pre-processed, the reads are aligned to a reference genome. There are many alignment tools for RNA-Seq reads (reviewed in [108]). Due to the simplicity of the prokaryotic genome, many of the features of these tools are not required for bacterial data, and the simple un-gapped alignment tool Bowtie can be used for short reads ≤ 50bp [109], or Bowtie2 for longer reads [110]. Some pipeline tools, such as Rockhopper [111], include an alignment step prior to analysis.

The parameter selection for alignment is highly dependent on the data type and intended results. The three parameters of greatest importance are the seed length, the number of mismatches, and the number of times a read aligns; however, there are several other parameters which may be changed to optimize results. The seed is the region at the high quality 5' end of the read and is the basis of the alignment, since sequencing quality falls significantly towards the 3' of the read (Figure 3). The number of mismatched bases allowed within this region can be altered to allow for sequencing errors. The alignment algorithm will then identify genomic matches to the seed region given this parameter, and then extend the alignment along the rest of the read using dynamic programming [110]. The mismatched bases outside the seed region are not counted; however, minimums can be set for both mismatches and sequence quality in the non-seed region. Reducing the length of the seed can optimise an alignment, particularly where the 5' position is of interest [74]. The number of times a read maps to the genome may also be altered. Generally, most transcriptome analyses accept only unique matches. However, if tRNAs, rRNAs or other multi-copy genes are of interest then multiple mappings must be allowed, due to the repetitive nature of these sequences. Ideally, sequential alignments can be carried out to first align the uniquely-matching reads, and then to align the remainder of the reads that match in multiple positions, therefore, allowing assessment of all the read data.

4.4 Counting

Following alignment most RNA-Seq analyses will perform counts of the reads using a tool such as bedtools [112] or featureCounts [113]. While differential expression analysis counts reads or fragments by gene feature, the analysis of transcriptional features is usually based on per base counts, or in some cases 5' position counts [114, 115]. Following counting, normalisation between replicates (and between conditions and batches, if applicable) must be carried out prior to the analysis. Several normalisation and batch correction algorithms are available [94, 116]. Analysis tools such as Rockhopper, Condop and Parseq (discussed in Section 5) automatically include a normalisation step [111, 117, 118].

5. RNA-Seq for the study of prokaryotic transcription

RNA-Seq data can be used to identify TSS, TTS, and operons, and to study transcriptional dynamics. Identification of TSS and TTS involves identifying breaks in the transcribed regions (Figure 4 A) [6, 81]. There are many confounding factors that make it difficult to pinpoint the borders between transcriptional units, including their dense nature, overlapping areas and lack of uniform expression [81], so *a priori* CDS knowledge is often used as a starting point [119]. Strand-specific sequencing can aid greatly in the identification of TSS and TTS. For instance, White and colleagues developed an empirical methodology to automatically identify transcript boundaries from strand-specific data to a resolution of 10-20 nucleotides [101]. O'Shea and co-workers used a probabilistic approach based on *a priori* knowledge [119], Borodovsky and co-workers trained hidden Markov models (HMMs) using high confidence transcriptional units as a gold standard datasets [79, 120], and Xu and co-workers applied support vector machines (SVM) to identify transcriptional units [121]. Integration of RNA-Seq with other data types can aid also TSS identification [12, 21, 64, 122].

Although sRNAs are hard to identify, they can be enriched by combining depletion of tRNAs and rRNAs with size selection prior to sequencing [100], or using a low molecular weight RNA enrichment step [123]. Gradient fractionation of the RNAs prior to sequencing can also be used [54]. In this case the cellular RNAs are separated into 20 fractions which are sequenced in turn. This technique has the advantage of grouping the RNAs that have similar function, so mRNA enrichment (Section 3.2) is not required for this technique.

Several algorithms and software tools have been developed for identification of transcriptional units from deep sequencing data. The open source Rockhopper pipeline can identify transcript boundaries, operons and sRNAs from single stranded RNA-Seq reads or alignments using a Bayesian approach [111]. Rockhopper can performs alignment of sequencing reads using bowtie2 [110]. However, since bowtie2 is designed for use with longer reads, a separate alignment may be preferred as input where reads are <50bp. Rockhopper handles replicates from multiple conditions, performing normalisation, aggregation and comparison in one step, and also outputs differentially expressed genes where more than one condition is supplied. Rockhopper also has a *de novo* transcript assembly option for organisms without a reference genome [124].

The Parseq algorithm which identifies TSS and TTS from alignment data based on state space modelling, and accounts for variations in the transcriptional signal and technical artefacts in the data [118]. Parseq is available as a free command line tool.

Condop is an algorithm which identifies condition-dependent operon patterns from RNA-Seq by comparison with the Database of prOkaryotic OpeRons (DOOR) [125]. The algorithm uses an ensemble classifier based on Neural Networks (NN), Support Vector Machines (SVMs) and Random Forests (RFs) which is trained on the DOOR operons. The algorithm then identifies operon differences between conditions. Condop is freely available as a bioconductor package [117]. In addition to traditional RNA-Seq, several variations have been developed for the analysis of specific areas of prokaryotic transcription, which we discuss in sections 5.1 to 5.4.

5.1 Transcription start sites

Differential RNA-Seq (dRNA-Seq) exploits the fact that the RNAs extracted and sequenced from the cell are of two types: primary nascent transcripts with a 5' triphosphate and processed full length transcripts with a 5' monophosphate (broken RNAs with a 5' OH cannot ligate to the linker prior to sequencing). By distinguishing between the primary and processed reads a more accurate estimation of TSS positions is possible [19, 126-128]. The primary transcripts are enriched via the degradation of the processed transcripts, using 5′ mono-phosphate-dependent terminator exonuclease (TEX). Comparison of TEX+ and TEX- sequences allows accurate identification of TSS (Figure 4 B). The detection of small RNAs has been significantly improved by this method [20, 129] and, dRNA-Seq has also been applied to the study of condition-dependent operons [130], leaderless mRNAs [131] and sub-operon expression [99].

Initially, TSS were manually annotated, however this method is labour-intensive, time-consuming, and subjective, so prone to variation between individual researchers. Therefore, several computational

approaches have been developed for TSS identification from dRNA-Seq data. TSSPredator is an open source program which generates TSS maps from the single base read counts of dRNA-Seq replicates [17]. TSSAR is both a standalone package and a Web server for the automatic *de novo* identification of TSS from dRNA-Seq data [132]. The TSSAR algorithm assumes that the differences between the dRNA-Seq libraries follow a Skellam distribution, allowing the identification of primary transcript enrichment at TSSs. The TruHMM algorithm uses HMMs to identify transcription units from dRNA-Seq data and is available as C++ code [130]. Evguenieva-Hackenberg and co-workers developed a machine learning algorithm for TSS recognition from dRNA-Seq data, which requires an expert-curated training dataset [133]. Finally, the TSSer software identifies TSSs from dRNA-Seq data using a probabilistic framework for identifying read enrichment [134].

A variation of dRNA-Seq utilises tobacco acid pyrophosphatase (TAP), an enzyme that converts a 5′ triphosphate to monophosphate, prior to constructing and sequencing cDNA libraries of native 5′-end segments [45-47, 122]. This treatment enriches the 5' read depth at TSS sites which are identified by comparison of the TAP+ and TAP- alignments in a number of ways. Kenneth and co-workers used M-A (ratio-intensity) scatterplots in a similar fashion as that used for microarrays [122]. Sorek and colleagues developed a machine learning approach that used random forest machine learning from a heuristic-derived training dataset [135].

A further variation of dRNA-Seq is Transcription Start Site-Exact Mapping Of Transcription Ends (TSS-EMOTE). This experimental technique uses the EMOTE protocol in order to precisely identify the mono-phosphorylated 5'-ends of the processed mRNAs [14, 136, 137]. Finally, several studies have used a modified 5′ RACE protocol (see Section 2) combined with high-throughput sequencing, to identify TSS with high precision [12, 21, 138, 139].

5.2 Transcription termination sites

Although dRNA-Seq improves the accuracy of TSS identification, it cannot provide accurate information about TTS. Finding 3' ends of reads is hard as they decay very quickly, meaning that there are few reads extending as far as the terminator [1]. The inefficiency of termination means that there is also considerable read-through of the terminator [39]. Furthermore, the absence of a poly(A) tail makes experimental identification of TTS challenging [4, 5].

Term-seq is a recently developed deep sequencing based technique which allows the direct sequencing of the RNA 3' ends to single base resolution [140]. Terminators are then identified as the largest peak in the 5’ end read counts on the opposite strand (Figure 4 C). A recent study using this technique has revealed that 30% of archaeal genes are controlled by multiple consecutive terminators [115]. An alternative method for TTS identification combines 3’ RACE with high-throughput sequencing in a similar fashion to the modified 5’ RACE described in Section 5.1 [138].

5.3 Polymerase dynamics and fidelity

RNAP pauses are involved in initiation, regulation and termination of transcription [68, 70, 141, 142], and have been linked to transcriptional fidelity [73, 74]. Nascent elongating transcript sequencing (NET-seq) was developed to enable characterisation of RNAP positions at the single nucleotide level. The technique, developed by Churchman and Weissman [114], exploits the stability of the transcription elongation complex in order to isolate nascent RNA prior to library preparation by immunoprecipitation (IP) of the RNAP. The 5’ ends of the reads may then be mapped to the 3’ base of nascent RNA positions and counted following genome alignment (Figure 4 D). Importantly, the transcripts are captured directly from live cells without any cross-linking, allowing the study of transcriptional dynamics under physiological conditions.

Weissman and co-workers used NET-seq to identify the consensus sequence linked to polymerase pausing in *E. coli* and observed similar results in *B. subtilis* [66]. Kashlev and colleagues used a RNase-based NET-seq variant, read-length-specific NET-seq (RNET-seq), to investigate pause positions and backtracking [73]. While RNAP fidelity can be measured in RNA-Seq data [143, 144], NET-seq allows measurement of transcriptional fidelity prior to proofreading [73, 74]. Due to the

sequence-biased nature of these data types, the sequence reads will fail most standard quality control pathways (Section 3.1) and require distinct analysis protocols from standard RNA-Seq.

5.4 RNA binding and modification

RNA Immunoprecipitation sequencing (RIP-seq) is another IP-based method, which is used to identify sRNAs and which has revealed their abundance in several bacterial genomes [83]. This technique has been used to investigate RNAs that are bound to an RNA-binding protein (RBP) of interest, for example the RNA chaperon, Hfq [145], by exploiting the specificity of the RNA-RBP binding. However, the basic protocol does not identify the RBP binding site itself. A later digestion optimized variation, DO-RIP-seq, combines RIP-seq and cross-linking to successfully identify the binding sites [146, 147]. The RIPSeeker R package uses HMMs to identify peaks from RIP-seq alignment files [148], while the free Piranha package identifies peaks based on a negative binomial regression model [149]. However, both of these tools are designed for eukaryotic data and to date no prokaryote-specific software has been created.

Bisulfite treatment of DNA can be used prior to sequencing in order to determine its pattern of methylation in eukaryotes (Bis-seq). In prokaryotes, methylation of rRNAs and tRNAs is linked to translation fidelity and ribosome assembly. A prokaryotic variation of Bis-seq has been used to identify RNA m^5C modification in bacteria and archaea [150]. Methylation sites are identified as cytosine residues that are not converted to uridine by the bisulfite treatment (since they are protected by the methylation). Since RNA modification is most common in tRNAs and rRNAs, no mRNA enrichment is required for the Bis-seq protocol. Flexible alignment parameters are required due to the bisulfite-induced mismatches and multiple copies of rRNA and tRNA genes, and several mappers have been developed for eukaryotic data (reviewed in [151]). However, software for prokaryotic Bis-seq data is yet to be developed.

6. Conclusions and perspectives

Deep sequencing has become the method of choice for the high-throughput study of bacterial transcriptomics. While the majority of deep sequencing techniques and analysis tools are designed for eukaryotes, prokaryotic-specific variations are increasing, and have revealed a far more complex bacterial transcriptome than previously thought. In addition to standard RNA-Seq, new variations such as dRNA-Seq, NET-seq and Term-seq allow the analysis of specific areas of prokaryotic transcription. The analysis of deep sequencing data for transcriptome analysis requires different bioinformatic processing than RNA-Seq gene expression studies. While automated pipelines for prokaryotic analysis are being developed, these resources are still lacking in some areas. However, as these deep sequencing technologies continue to be developed and improved, prokaryote-specific bioinformatic resources will also continue to be produced, increasing the utility of deep sequencing for the analysis of bacterial transcriptomes.

References

[1] J.P. Creecy, T. Conway, Curr. Opin. Microbiol., 23 (2015) 133-140.
[2] N.J. Croucher, N.R. Thomson, Curr. Opin. Microbiol., 13 (2010) 619-624.
[3] M.J. Filiatrault, Curr. Opin. Microbiol., 14 (2011) 579-586.
[4] M. Güell, E. Yus, M. Lluch-Senar, L. Serrano, Nat. Rev. Microbiol., 9 (2011) 658-669.
[5] R. Sorek, P. Cossart, Nat. Rev. Genet., 11 (2010) 9-16.
[6] V. van, M. Arnoud H, FEMS Microb. Lett., 302 (2010) 1-7.
[7] J. Wang, L. Chen, Z. Chen, W. Zhang, Integr. Biol. (Camb), 7 (2015) 1466-1476.
[8] S. Pepke, B. Wold, A. Mortazavi, Nat. Methods, 6 (2009) S22-S32.
[9] Z. Wang, M. Gerstein, M. Snyder, Nat. Rev. Genet., 10 (2009) 57-63.
[10] D. Medini, D. Serruto, J. Parkhill, D.A. Relman, C. Donati, R. Moxon, S. Falkow, R. Rappuoli, Nat. Rev. Microbiol., 6 (2008) 419-430.
[11] S. Okuda, S. Kawashima, K. Kobayashi, N. Ogasawara, M. Kanehisa, S. Goto, BMC Genomics, 8 (2007) 48-48.

[12] B.-K. Cho, K. Zengler, Y. Qiu, Y.S. Park, E.M. Knight, C.L. Barrett, Y. Gao, B.Ø. Palsson, Nat. Biotechnol., 27 (2009) 1043-1049.
[13] A. Millman, D. Dar, M. Shamir, R. Sorek, Nuc. Acids Res., (2016) gkw749 [Epub ahead of print].
[14] J. Prados, P. Linder, P. Redder, BMC Genomics, 17 (2016) 849-849.
[15] J.E. Brock, S. Pourshahian, J. Giliberti, P.A. Limbach, G.R. Janssen, Rna, 14 (2008) 2159-2169.
[16] I. Moll, S. Grill, C.O. Gualerzi, U. Blasi, Mol Microbiol, 43 (2002) 239-246.
[17] G. Dugar, A. Herbig, K.U. Förstner, N. Heidrich, R. Reinhardt, K. Nieselt, C.M. Sharma, PLoS Genet., 9 (2013) e1003495-e1003495.
[18] J. Mitschke, J. Georg, I. Scholz, C.M. Sharma, D. Dienst, J. Bantscheff, B. Voss, C. Steglich, A. Wilde, J. Vogel, W.R. Hess, Proc. Natl. Acad. Sci. U.S.A., 108 (2011) 2124-2129.
[19] C.M. Sharma, S. Hoffmann, F. Darfeuille, J. Reignier, S. Findeiss, A. Sittka, S. Chabas, K. Reiche, J. Hackermüller, R. Reinhardt, P.F. Stadler, J. Vogel, Nature, 464 (2010) 250-255.
[20] M.K. Thomason, T. Bischler, S.K. Eisenbart, K.U. Förstner, A. Zhang, A. Herbig, K. Nieselt, C.M. Sharma, G. Storz, J. Bacteriol., 197 (2015) 18-28.
[21] A. Mendoza-Vargas, L. Olvera, M. Olvera, R. Grande, L. Vega-Alvarado, B. Taboada, V. Jimenez-Jacinto, H. Salgado, K. Juárez, B. Contreras-Moreira, A.M. Huerta, J. Collado-Vides, E. Morett, PLoS One, 4 (2009) e7526-e7526.
[22] G. Oliva, T. Sahr, C. Buchrieser, FEMS Microbiol. Lett., 39 (2015) 331-349.
[23] J.A. Thompson, M.F. Radonovich, N.P. Salzman, J. Virol., 31 (1979) 437-446.
[24] M.A. Frohman, M.K. Dush, G.R. Martin, Proc. Natl. Acad. Sci. U.S.A., 85 (1988) 8998-9002.
[25] P.T. McGrath, H. Lee, L. Zhang, A.A. Iniesta, A.K. Hottes, M.H. Tan, N.J. Hillson, P. Hu, L. Shapiro, H.H. McAdams, Nat. Biotech., 25 (2007) 584-592.
[26] D.W. Selinger, K.J. Cheung, R. Mei, E.M. Johansson, C.S. Richmond, F.R. Blattner, D.J. Lockhart, G.M. Church, Nat. Biotech., 18 (2000) 1262-1268.
[27] B. Tjaden, R.M. Saxena, S. Stolyar, D.R. Haynor, E. Kolker, C. Rosenow, Nuc. Acids Res., 30 (2002) 3732-3738.
[28] A. Toledo-Arana, O. Dussurget, G. Nikitas, N. Sesto, H. Guet-Revillet, D. Balestrino, E. Loh, J. Gripenland, T. Tiensuu, K. Vaitkevicius, M. Barthelemy, M. Vergassola, M.-A. Nahori, G. Soubigou, B. Régnault, J.-Y. Coppée, M. Lecuit, J. Johansson, P. Cossart, Nature, 459 (2009) 950-956.
[29] S. Rasmussen, H.B. Nielsen, H. Jarmer, Mol. Microbiol., 73 (2009) 1043-1057.
[30] J.M. Johnson, S. Edwards, D. Shoemaker, E.E. Schadt, Trends Genet., 21 (2005) 93-102.
[31] T.E. Royce, J.S. Rozowsky, P. Bertone, M. Samanta, V. Stolc, S. Weissman, M. Snyder, M. Gerstein, Trends Genet., 21 (2005) 466-475.
[32] C.D. Herring, M. Raffaelle, T.E. Allen, E.I. Kanin, R. Landick, A.Z. Ansari, B.Ø. Palsson, J. Bacteriol., 187 (2005) 6166-6174.
[33] J.T. Wade, K. Struhl, S.J.W. Busby, D.C. Grainger, Mol. Microbiol., 65 (2007) 21-26.
[34] M. Brenneis, J. Soppa, PloS One, 4 (2009) e4484-e4484.
[35] A. Ray-Soni, M.J. Bellecourt, R. Landick, Annu. Rev. Biochem., 85 (2016) 319-347.
[36] J.M. Peters, A.D. Vangeloff, R. Landick, J. Mol. Biol., 412 (2011) 793-813.
[37] K.S. Wilson, P.H. von Hippel, Proc. Natl. Acad. Sci. U.S.A., 92 (1995) 8793-8797.
[38] C.L. Kingsford, K. Ayanbule, S.L. Salzberg, Genome Biol., 8 (2007) R22-R22.
[39] Y.-J. Chen, P. Liu, A.A.K. Nielsen, J.A.N. Brophy, K. Clancy, T. Peterson, C.A. Voigt, Nat. Meth., 10 (2013) 659-664.
[40] R.W.W. Brouwer, O.P. Kuipers, S.A.F.T. van Hijum, Brief. Bioinform., 9 (2008) 367-375.
[41] P. Dam, V. Olman, K. Harris, Z. Su, Y. Xu, Nucleic Acids Res., 35 (2007) 288-298.
[42] P.R. Romero, P.D. Karp, Bioinformatics, 20 (2004) 709-717.
[43] L. Barquist, J. Vogel, Ann. Rev. Genet., 49 (2015) 367-394.
[44] J.E. Dornenburg, A.M. Devita, M.J. Palumbo, J.T. Wade, mBio, 1 (2010) e00024-00010.
[45] O. Wurtzel, R. Sapra, F. Chen, Y. Zhu, B.A. Simmons, R. Sorek, Genome Res., 20 (2010) 133-141.
[46] O. Wurtzel, N. Sesto, J.R. Mellin, I. Karunker, S. Edelheit, C. Bécavin, C. Archambaud, P. Cossart, R. Sorek, Mol. Syst. Biol., 8 (2012) 583-583.

[47] O. Wurtzel, D.R. Yoder-Himes, K. Han, A.A. Dandekar, S. Edelheit, E.P. Greenberg, R. Sorek, S. Lory, PLoS Pathog., 8 (2012) e1002945-e1002945.
[48] D.R. Yoder-Himes, P.S.G. Chain, Y. Zhu, O. Wurtzel, E.M. Rubin, J.M. Tiedje, R. Sorek, Proc. Natl. Acad. Sci. U.S.A., 106 (2009) 3976-3981.
[49] J. Vogel, V. Bartels, T.H. Tang, G. Churakov, J.G. Slagter-Jäger, A. Hüttenhofer, E.G.H. Wagner, Nuc. Acids Res., 31 (2003) 6435-6443.
[50] J.P. Bardill, B.K. Hammer, RNA Biology, 9 (2012) 392-401.
[51] M. Bobrovskyy, C.K. Vanderpool, Ann. Rev. Genet., 47 (2013) 209-232.
[52] I. Caldelari, Y. Chao, P. Romby, J. Vogel, Cold Spring Harb. Perspect. Med., 3 (2013) a010298-a010298.
[53] J. Georg, B. Voss, I. Scholz, J. Mitschke, A. Wilde, W.R. Hess, Mol. Syst. Biol., 5 (2009) 305-305.
[54] M. Lybecker, B. Zimmermann, I. Bilusic, N. Tukhtubaeva, R. Schroeder, Proc. Natl. Acad. Sci. U.S.A., 111 (2014) 3134-3139.
[55] M.K. Thomason, G. Storz, Ann. Rev. Genet., 44 (2010) 167-188.
[56] J. Georg, W.R. Hess, Microbiol Mol Biol Rev., 75 (2011) 286-300.
[57] V. Lloréns-Rico, J. Cano, T. Kamminga, R. Gil, A. Latorre, W.-H. Chen, P. Bork, J.I. Glass, L. Serrano, M. Lluch-Senar, Sci. Adv., 2 (2016) e1501363-e1501363.
[58] J.T. Wade, D.C. Grainger, Nat. Rev. Microbiol., 12 (2014) 647-653.
[59] R. Robinson, PLoS Biol., 8 (2010) e1000370.
[60] J.L. Slonczewski, Mbio, 1 (2010).
[61] J. Mitschke, A. Vioque, F. Haas, W.R. Hess, A.M. Muro-Pastor, Proc. Natl. Acad. Sci. U.S.A., 108 (2011) 20130-20135.
[62] P. Nicolas, U. Mäder, E. Dervyn, T. Rochat, A. Leduc, N. Pigeonneau, E. Bidnenko, E. Marchadier, M. Hoebeke, S. Aymerich, D. Becher, P. Bisicchia, E. Botella, O. Delumeau, G. Doherty, E.L. Denham, M.J. Fogg, V. Fromion, A. Goelzer, A. Hansen, E. Härtig, C.R. Harwood, G. Homuth, H. Jarmer, M. Jules, E. Klipp, L. Le Chat, F. Lecointe, P. Lewis, W. Liebermeister, A. March, R.A.T. Mars, P. Nannapaneni, D. Noone, S. Pohl, B. Rinn, F. Rügheimer, P.K. Sappa, F. Samson, M. Schaffer, B. Schwikowski, L. Steil, J. Stülke, T. Wiegert, K.M. Devine, A.J. Wilkinson, J.M. van Dijl, M. Hecker, U. Völker, P. Bessières, P. Noirot, Science, 335 (2012) 1103-1106.
[63] V. Fortino, O.-P. Smolander, P. Auvinen, R. Tagliaferri, D. Greco, BMC Bioinform., 15 (2014) 145-145.
[64] M. Güell, V. van Noort, E. Yus, W.-H. Chen, J. Leigh-Bell, K. Michalodimitrakis, T. Yamada, M. Arumugam, T. Doerks, S. Kühner, M. Rode, M. Suyama, S. Schmidt, A.-C. Gavin, P. Bork, L. Serrano, Science, 326 (2009) 1268-1271.
[65] B. Yang, T.J. Larson, Biochim Biophys Acta, 1396 (1998) 114-126.
[66] M.H. Larson, R.A. Mooney, J.M. Peters, T. Windgassen, D. Nayak, C.A. Gross, S.M. Block, W.J. Greenleaf, R. Landick, J.S. Weissman, Science, 344 (2014) 1042-1047.
[67] I.O. Vvedenskaya, H. Vahedian-Movahed, J.G. Bird, J.G. Knoblauch, S.R. Goldman, Y. Zhang, R.H. Ebright, B.E. Nickels, Science, 344 (2014) 1285-1289.
[68] D. Duchi, D.L.V. Bauer, L. Fernandez, G. Evans, N. Robb, L.C. Hwang, K. Gryte, A. Tomescu, P. Zawadzki, Z. Morichaud, K. Brodolin, A.N. Kapanidis, Mol. Cell., 63 (2016) 939-950.
[69] G.A. Kassavetis, M.J. Chamberlin, J. Biol. Chem., 256 (1981) 2777-2786.
[70] R. Landick, Biochem. Soc. Trans., 34 (2006) 1062-1066.
[71] I. Gusarov, E. Nudler, Mol. Cell, 3 (1999) 495-504.
[72] P. Gamba, K. James, N. Zenkin, Transcription, [in press] (2016).
[73] M. Imashimizu, H. Takahashi, T. Oshima, C. McIntosh, M. Bubunenko, D.L. Court, M. Kashlev, Genome Biol., 16 (2015) 98-98.
[74] K. James, P. Gamba, S.J. Cockell, N. Zenkin, Nuc. Acids Res., (2016) gkw969 [Epub ahead of print].
[75] L. Camarena, V. Bruno, G. Euskirchen, S. Poggio, M. Snyder, PLoS Pathog., 6 (2010) e1000834-e1000834.

[76] M.J. Filiatrault, P.V. Stodghill, C.R. Myers, P.A. Bronstein, B.G. Butcher, H. Lam, G. Grills, P. Schweitzer, W. Wang, D.J. Schneider, S.W. Cartinhour, PloS One, 6 (2011) e29335-e29335.
[77] J. Frias-Lopez, Y. Shi, G.W. Tyson, M.L. Coleman, S.C. Schuster, S.W. Chisholm, E.F. Delong, Proc. Natl. Acad. Sci. U.S.A., 105 (2008) 3805-3810.
[78] J.A. Gilbert, D. Field, Y. Huang, R. Edwards, W. Li, P. Gilna, I. Joint, PloS One, 3 (2008) e3042-e3042.
[79] J. Martin, W. Zhu, K.D. Passalacqua, N. Bergman, M. Borodovsky, BMC Bioinform., 11 Suppl 3 (2010) S10-S10.
[80] P.R. McAdam, E.J. Richardson, J.R. Fitzgerald, Curr. Opin. Microbiol., 19 (2014) 106-113.
[81] K.D. Passalacqua, A. Varadarajan, B.D. Ondov, D.T. Okou, M.E. Zwick, N.H. Bergman, J. Bacteriol. , 191 (2009) 3203-3211.
[82] T.T. Perkins, R.A. Kingsley, M.C. Fookes, P.P. Gardner, K.D. James, L. Yu, S.A. Assefa, M. He, N.J. Croucher, D.J. Pickard, D.J. Maskell, J. Parkhill, J. Choudhary, N.R. Thomson, G. Dougan, PLoS Genet., 5 (2009) e1000569-e1000569.
[83] A. Sittka, S. Lucchini, K. Papenfort, C.M. Sharma, K. Rolle, T.T. Binnewies, J.C.D. Hinton, J. Vogel, PLoS Genet., 4 (2008) e1000163-e1000163.
[84] P.J. Turnbaugh, C. Quince, J.J. Faith, A.C. McHardy, T. Yatsunenko, F. Niazi, J. Affourtit, M. Egholm, B. Henrissat, R. Knight, J.I. Gordon, Proc. Natl. Acad. Sci. U.S.A., 107 (2010) 7503-7508.
[85] B.J. Haas, M. Chin, C. Nusbaum, B.W. Birren, J. Livny, BMC Genomics, 13 (2012) 734-734.
[86] M.A. Busby, C. Stewart, C.A. Miller, K.R. Grzeda, G.T. Marth, Bioinformatics, 29 (2013) 656-657.
[87] Z. Fang, X. Cui, Brief. Bioinform., 12 (2011) 280-287.
[88] Y. Liu, J. Zhou, K.P. White, Bioinformatics, 30 (2014) 301-304.
[89] J.C. Marioni, C.E. Mason, S.M. Mane, M. Stephens, Y. Gilad, Genome Res., 18 (2008) 1509-1517.
[90] N.J. Schurch, P. Schofield, M. Gierliniski, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G.G. Simpson, T. Owen-Hughes, M. Blaxter, G.J. Barton, Rna, 22 (2016) 1641-1641.
[91] L.M. McIntyre, K.K. Lopiano, A.M. Morse, V. Amin, A.L. Oberg, L.J. Young, S.V. Nuzhdin, BMC Genomics, 12 (2011) 293.
[92] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M.W. Szcześniak, D.J. Gaffney, L.L. Elo, X. Zhang, A. Mortazavi, Genome Biol., 17 (2016) 13-13.
[93] P. Manga, D.M. Klingeman, T.Y.S. Lu, T.L. Mehlhorn, D.A. Pelletier, L.J. Hauser, C.M. Wilson, S.D. Brown, Front. Microbiol., 7 (2016) 794.
[94] J.T. Leek, R.B. Scharpf, H.C. Bravo, D. Simcha, B. Langmead, W.E. Johnson, D. Geman, K. Baggerly, R.A. Irizarry, Nature Reviews Genetics, 11 (2010) 733-739.
[95] C. Condon, Curr. Opin. Microbiol. , 10 (2007) 271-278.
[96] M.P. Deutscher, J. Biol. Chem., 278 (2003) 45041-45044.
[97] G. Giannoukos, D.M. Ciulla, K. Huang, B.J. Haas, J. Izard, J.Z. Levin, J. Livny, A.M. Earl, D. Gevers, D.V. Ward, C. Nusbaum, B.W. Birren, A. Gnirke, Genome Biol., 13 (2012) R23-R23.
[98] N. Kumar, M. Lin, X. Zhao, S. Ott, I. Santana-Cruz, S. Daugherty, Y. Rikihisa, L. Sadzewicz, L.J. Tallon, C.M. Fraser, J.C. Dunning Hotopp, Sci Rep, 6 (2016) 34850.
[99] C.M. Sharma, J. Vogel, Curr. Opin. Microbiol., 19 (2014) 97-105.
[100] J.M. Liu, J. Livny, M.S. Lawrence, M.D. Kimball, M.K. Waldor, A. Camilli, Nuc. Acids Res., 37 (2009) e46-e46.
[101] Y. Wang, K.D. MacKenzie, A.P. White, BMC Genomics, 16 (2015) 359-359.
[102] N.J. Croucher, M.C. Fookes, T.T. Perkins, D.J. Turner, S.B. Marguerat, T. Keane, M.A. Quail, M. He, S. Assefa, J. Bähler, R.A. Kingsley, J. Parkhill, S.D. Bentley, G. Dougan, N.R. Thomson, Nucleic Acids Res., 37 (2009) e148-e148.
[103] A.P. Vivancos, M. Guell, J.C. Dohm, L. Serrano, H. Himmelbauer, Genome Res., 20 (2010) 989-999.
[104] T. Conway, J.P. Creecy, S.M. Maddox, J.E. Grissom, T.L. Conkle, T.M. Shadid, J. Teramoto, P. San Miguel, T. Shimada, A. Ishihama, H. Mori, B.L. Wanner, mBio, 5 (2014) e01442-e01414.
[105] D. MacLean, J.D.G. Jones, D.J. Studholme, Nat. Rev. Microbiol., 7 (2009) 287-296.
[106] A.M. Bolger, M. Lohse, B. Usadel, Bioinformatics, 30 (2014) 2114-2120.

[107] S. Hoffmann, C. Otto, S. Kurtz, C.M. Sharma, P. Khaitovich, J. Vogel, P.F. Stadler, J. Hackermüller, PLoS Comp. Biol., 5 (2009) e1000502-e1000502.
[108] H. Li, N. Homer, Brief. Bioinform., 11 (2010) 473-483.
[109] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Genome Biol., 10 (2009) R25-R25.
[110] B. Langmead, S.L. Salzberg, Nat. Meth., 9 (2012) 357-359.
[111] R. McClure, D. Balasubramanian, Y. Sun, M. Bobrovskyy, P. Sumby, C.A. Genco, C.K. Vanderpool, B. Tjaden, Nuc. Acids Res., 41 (2013) e140-e140.
[112] A.R. Quinlan, I.M. Hall, Bioinformatics, 26 (2010) 841-842.
[113] Y. Liao, G.K. Smyth, W. Shi, Bioinformatics, 30 (2014) 923-930.
[114] L.S. Churchman, J.S. Weissman, Nature, 469 (2011) 368-373.
[115] D. Dar, D. Prasse, R.A. Schmitz, R. Sorek, Nat. Microbiol., 1 (2016) 16143-16143.
[116] J.H. Bullard, E. Purdom, K.D. Hansen, S. Dudoit, BMC Bioinformatics, 11 (2010).
[117] V. Fortino, R. Tagliaferri, D. Greco, Bioinformatics, 32 (2016) 3199-3200.
[118] B. Mirauta, P. Nicolas, H. Richard, Bioinformatics, 30 (2014) 1409-1416.
[119] V. Vijayan, I.H. Jain, E.K. O'Shea, Genome Biol., 12 (2011) R47.
[120] J. Martin, W.H. Zhu, N. Bergman, M. Borodovsky, Proceedings IEEE Int. Conf. Bioinform. Biomed. , (2009) 54-59.
[121] W.-C. Chou, Q. Ma, S. Yang, S. Cao, D.M. Klingeman, S.D. Brown, Y. Xu, Nucleic Acids Res. , 43 (2015) e67-e67.
[122] Y.-f. Lin, A. David Romero, S. Guan, L. Mamanova, K.J. McDowall, BMC Genomics, 14 (2013) 620-620.
[123] A. Shinhara, M. Matsui, K. Hiraoka, W. Nomura, R. Hirano, K. Nakahigashi, M. Tomita, H. Mori, A. Kanai, BMC Genomics, 12 (2011).
[124] B. Tjaden, Genome Biol., 16 (2015) 1.
[125] X. Mao, Q. Ma, C. Zhou, X. Chen, H. Zhang, J. Yang, F. Mao, W. Lai, Y. Xu, Nuc. Acids Res., 42 (2014) D654-D659.
[126] C. Kroger, S.C. Dillon, A.D.S. Cameron, K. Papenfort, S.K. Sivasankaran, K. Hokamp, Y.J. Chao, A. Sittka, M. Hebrard, K. Handler, A. Colgan, P. Leekitcharoenphon, G.C. Langridge, A.J. Lohan, B. Loftus, S. Lucchini, D.W. Ussery, C.J. Dorman, N.R. Thomson, J. Vogel, J.C.D. Hinton, Proc. Natl. Acad. Sci. U.S.A., 109 (2012) E1277-E1286.
[127] J.P. Schluter, J. Reinkensmeier, M.J. Barnett, C. Lang, E. Krol, R. Giegerich, S.R. Long, A. Becker, BMC Genomics, 14 (2013) 156.
[128] T. Bischler, H.S. Tan, K. Nieselt, C.M. Sharma, Methods, 86 (2015) 89-101.
[129] I. Irnov, C.M. Sharma, J. Vogel, W.C. Winkler, Nuc. Acids Res., 38 (2010) 6637-6651.
[130] S. Li, X. Dong, Z. Su, BMC Genomics, 14 (2013) 520-520.
[131] J. Babski, K.A. Haas, D. Nather-Schindler, F. Pfeiffer, K.U. Forstner, M. Hammelmann, R. Hilker, A. Becker, C.M. Sharma, A. Marchfelder, J. Soppa, BMC Genomics, 17 (2016) 629.
[132] F. Amman, M.T. Wolfinger, R. Lorenz, I.L. Hofacker, P.F. Stadler, S. Findeiß, BMC Bioinform., 15 (2014) 89-89.
[133] J. Čuklina, J. Hahn, M. Imakaev, U. Omasits, K.U. Förstner, N. Ljubimov, M. Goebel, G. Pessi, H.-M. Fischer, C.H. Ahrens, M.S. Gelfand, E. Evguenieva-Hackenberg, BMC Genomics, 17 (2016) 302-302.
[134] H. Jorjani, M. Zavolan, Bioinformatics, 30 (2014) 971-974.
[135] O. Cohen, S. Doron, O. Wurtzel, D. Dar, S. Edelheit, I. Karunker, E. Mick, R. Sorek, Nucleic Acids Res., 44 (2016) W46-W53.
[136] P. Redder, Methods in molecular biology (Clifton, N.J.), 1259 (2015) 69-85.
[137] P. Linder, S. Lemeille, P. Redder, PLoS Genetics, 10 (2014).
[138] D. Matteau, S. Rodrigue, Methods Mol. Biol., 1334 (2015) 143-159.
[139] D. Kim, J.S.J. Hong, Y. Qiu, H. Nagarajan, J.H. Seo, B.K. Cho, S.F. Tsai, B.O. Palsson, PLoS Genet., 8 (2012).
[140] D. Dar, M. Shamir, J.R. Mellin, M. Koutero, N. Stern-Ginossar, P. Cossart, R. Sorek, Science, 352 (2016) aad9822-aad9822.

[141] G.A. Belogurov, I. Artsimovitch, Ann. Rev. Microbiol., 69 (2015) 49-69.
[142] P. Gollnick, P. Babitzke, Biochim Biophys Acta., 1577 (2002) 240-250.
[143] L.B. Carey, eLife, 4 (2015) e09945.
[144] M. Imashimizu, T. Oshima, L. Lubkowska, M. Kashlev, Nuc. Acids Res., 41 (2013) 9090-9104.
[145] P. Möller, A. Overlöper, K.U. Förstner, T.-N. Wen, C.M. Sharma, E.-M. Lai, F. Narberhaus, PloS One, 9 (2014) e110427-e110427.
[146] C.O. Nicholson, M.B. Friedersdorf, L.S. Bisogno, J.D. Keene, Methods, (2016) S1046-2023(1016)30437-30436 [Epub ahead of print].
[147] C.O. Nicholson, M.B. Friedersdorf, J.D. Keene, Rna, (2016) rna.058115.058116. [Epub ahead of print].
[148] Y. Li, D.Y. Zhao, J.F. Greenblatt, Z. Zhang, Nuc. Acids Res., 41 (2013) e94-e94.
[149] P.J. Uren, E. Bahrami-Samani, S.C. Burns, M. Qiao, F.V. Karginov, E. Hodges, G.J. Hannon, J.R. Sanford, L.O.F. Penalva, A.D. Smith, Bioinformatics, 28 (2012) 3013-3020.
[150] S. Edelheit, S. Schwartz, M.R. Mumbach, O. Wurtzel, R. Sorek, PLoS Genet., 9 (2013) e1003602-e1003602.
[151] G. Kunde-Ramamoorthy, C. Coarfa, E. Laritsky, N.J. Kessler, R.A. Harris, M. Xu, R. Chen, L. Shen, A. Milosavljevic, R.A. Waterland, Nuc. Acids Res., 42 (2014) e43-e43.

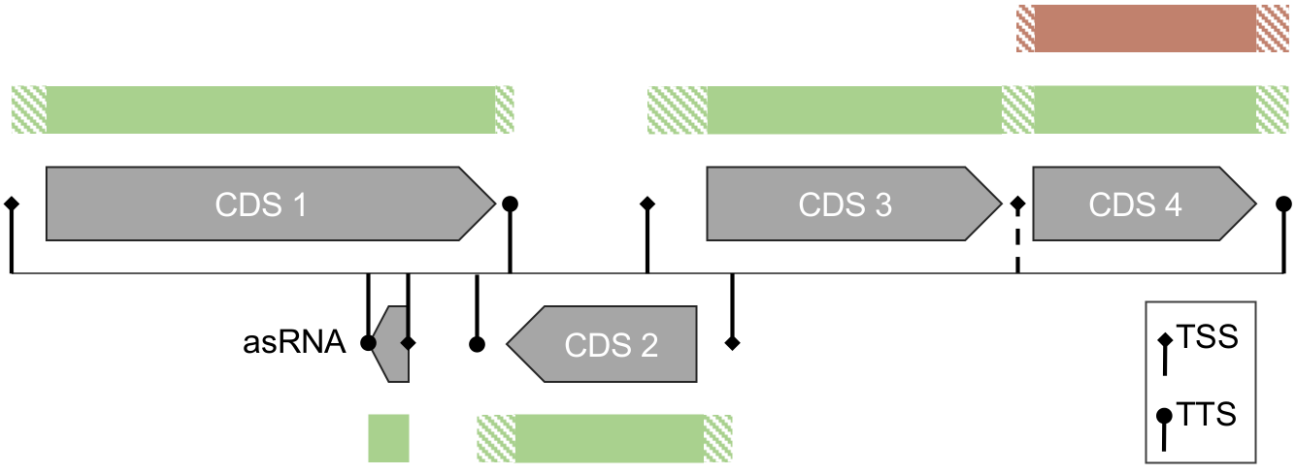
Figure Legends

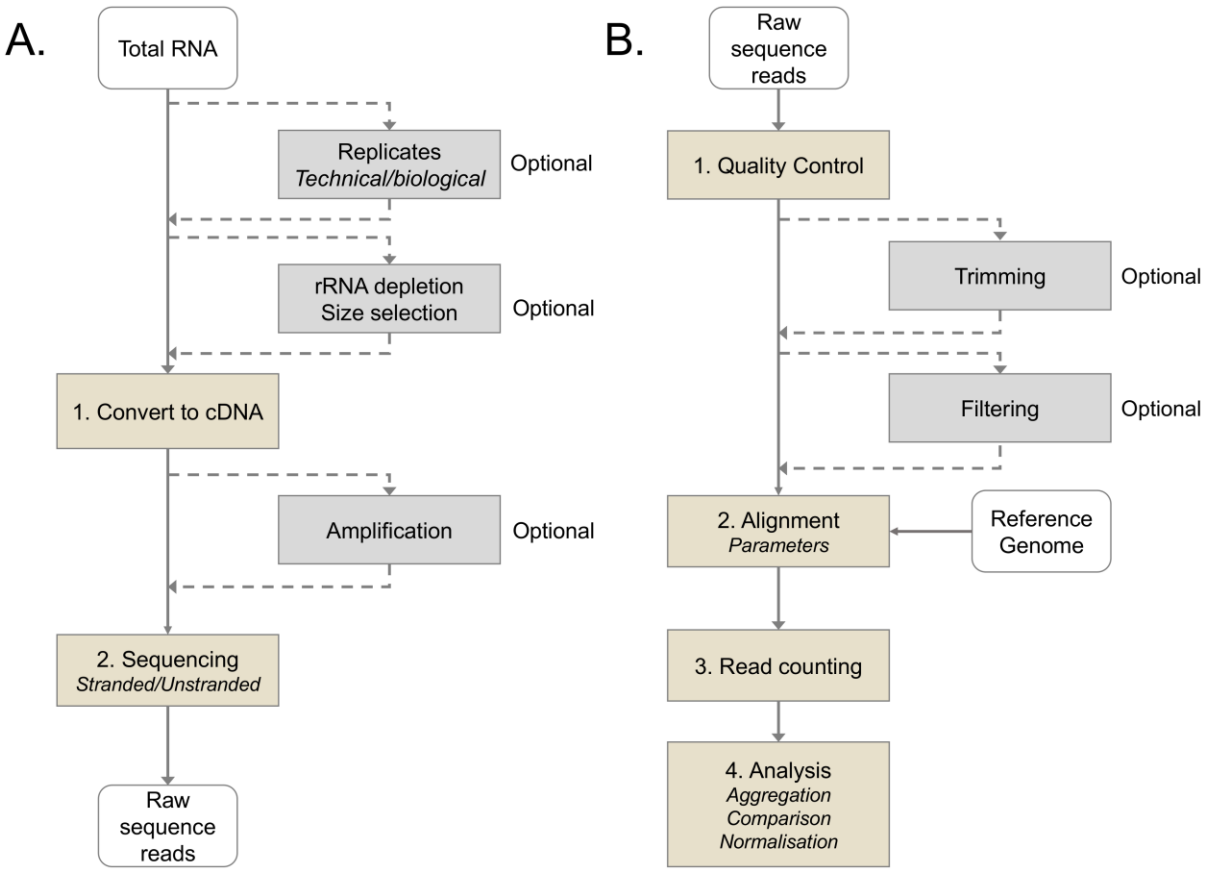
Figure 1. Bacterial transcriptome complexity. In addition to the translated coding sequences (CDS), transcribed regions (green and red) include untranslated regions (UTRs - shaded) that are bordered by the transcription start (TSS) and termination sites (TTS). A transcript may contain a single CDS, an operon of multiple CDSs or another un-transcribed elements such as the antisense RNA (asRNA) shown here. TSS can change depending on condition. Here a two-CDS transcript is produced under condition 1 (green), while a single CDS transcript is produced by the alternate TSS (dashed) under condition 2 (red).

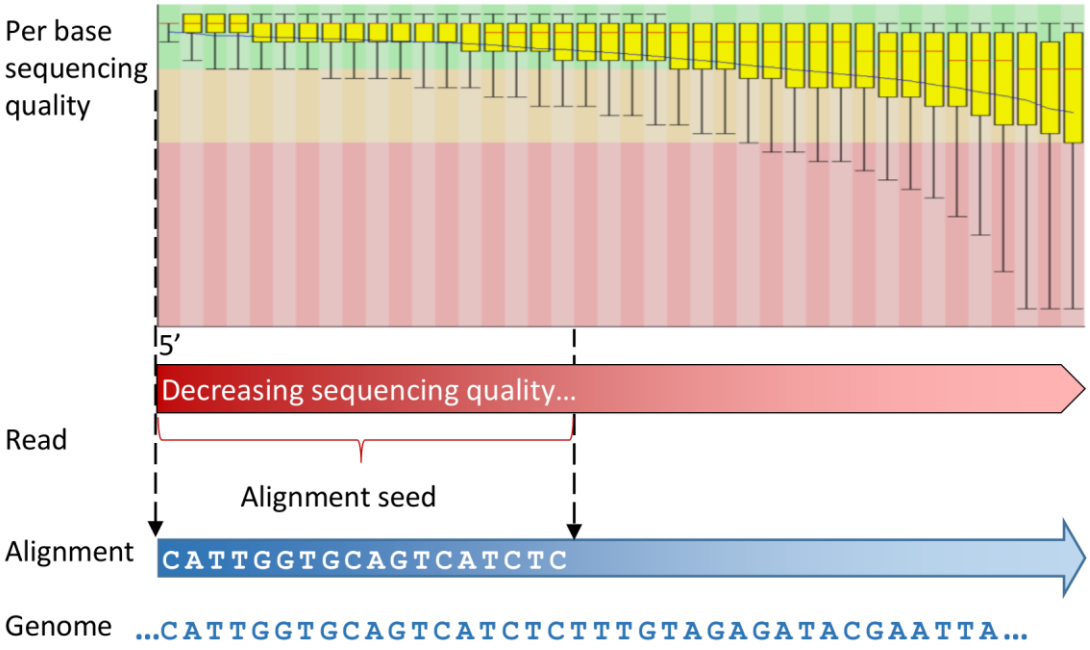
Figure 2. RNA-Seq for transcriptomics and its analysis. A. The experimental pipeline. B. The analysis pipeline.

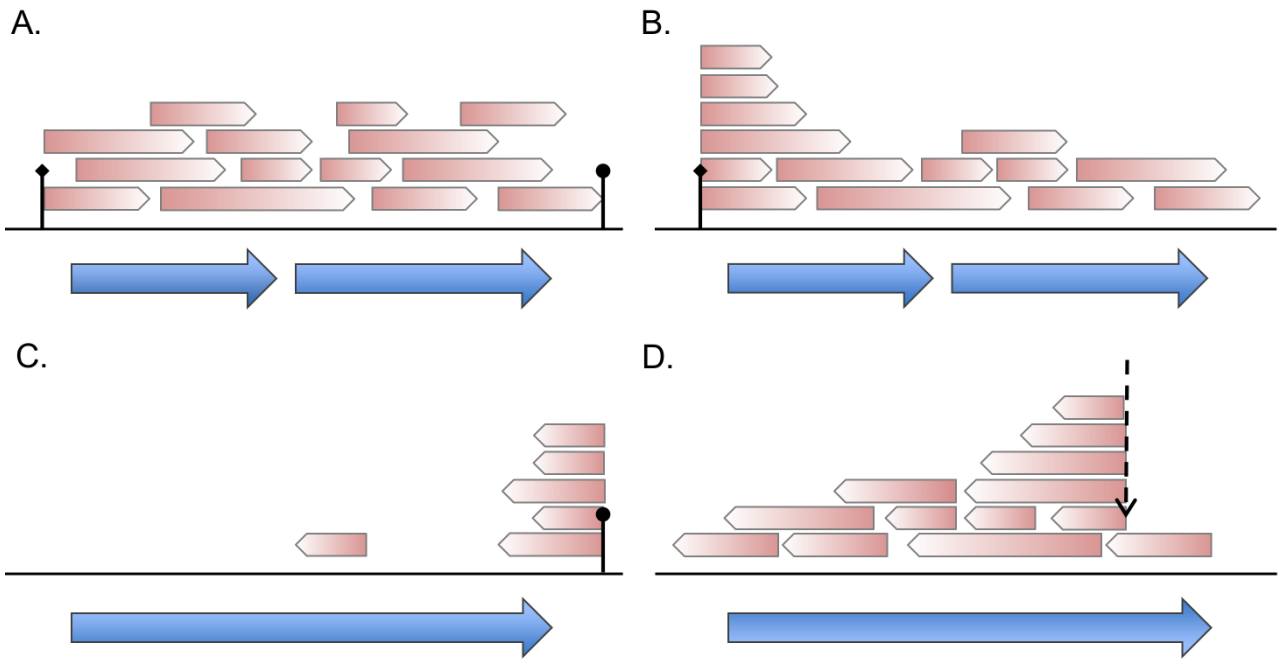
Figure 3 Seeded alignment. The per base sequencing quality of an RNA-Seq library decreases from 5' to 3' (measured here using FastQC). Therefore, for each individual read a seed region at the high quality 5' end is first aligned to the genome using strict parameters (here a perfect alignment with no mismatches), before the alignment is extended along the length of the read using less strict parameters.

Figure 4. RNA-Seq technologies. A. In standard RNA-Seq the reads align as a solid block across the transcriptional unit; here a two-gene operon. Transcription start sites (diamond) and termination sites (circle) can then be identified from breaks in read depth. B. In dRNA-Seq reads next to the start sites are enriched in the TEX+ reads, allowing far more accurate TSS identification. C. In Term-seq data reads next to the termination site are enriched, with reads aligning to the opposite strand. D. NET-seq reads also align on the opposite strand. Pause sites (dashed arrow) are identified as enrichment of read 5' ends (equivalent to the 3' of the nascent RNA).


CDS 1
CDS 3
CDS 4
asRNA
CDS 2
TSS
TTS


A.
Total RNA
Replicates
Technical/biological
Optional
rRNA depletion
Size selection
Optional
1. Convert to cDNA
Amplification
Optional
2. Sequencing
Stranded/Unstranded
Raw sequence reads
B.
Raw sequence reads
1. Quality Control
Trimming
Optional
Filtering
Optional
2. Alignment
Parameters
Reference Genome
3. Read counting
4. Analysis
Aggregation
Comparison
Normalisation


Per base sequencing quality
5’
Decreasing sequencing quality…
Read
Alignment seed
Alignment
CATTGGTGCAGTCATCTC
Genome
…CATTGGTGCAGTCATCTCTTTGTAGAGATACGAATTA…


A.
B.
C.
D.