Methods 120 (2017) 76–84

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

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

Katherine James [a,*], Simon J. Cockell [b], Nikolay Zenkin [a]

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

ARTICLE INFO

Article history:
Received 14 December 2016
Received in revised form 13 April 2017
Accepted 18 April 2017
Available online 21 April 2017

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

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.

Contents

1. Introduction . . . 77
2. The prokaryotic transcriptome . . . 77
3. RNA-Seq . . . 78
   3.1. Replicates . . . 78
   3.2. mRNA enrichment . . . 78
   3.3. Amplification . . . 79
   3.4. Sequencing . . . 79
4. Bioinformatic analysis . . . 79
   4.1. Quality control . . . 79
   4.2. Trimming and filtering . . . 79
   4.3. Alignment . . . 79
   4.4. Counting . . . 79
5. RNA-Seq for the study of prokaryotic transcription . . . 80
   5.1. Transcription start sites . . . 81
   5.2. Transcription termination sites . . . 81
   5.3. Polymerase dynamics and fidelity . . . 81
   5.4. RNA binding and modification . . . 82
6. Conclusions and perspectives . . . 82
   Funding . . . 82
   References . . . 82

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.

* Corresponding author.
E-mail address: katherine.james@newcastle.ac.uk (K. James).

http://dx.doi.org/10.1016/j.ymeth.2017.04.016
1046-2023/

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 (Fig. 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 (unassociated 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 (~50–500 bp), 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

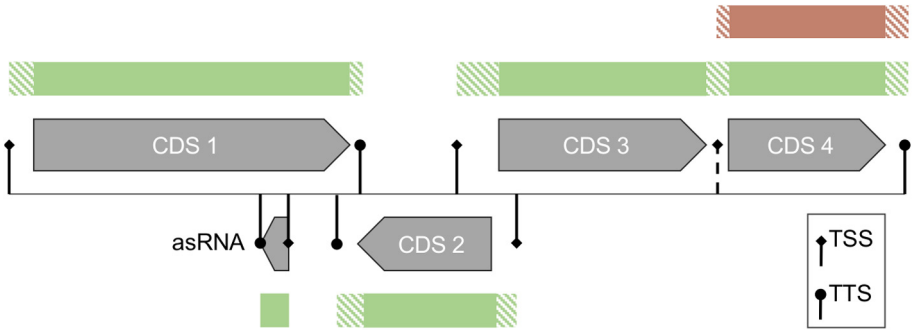

Fig. 1. Bacterial transcriptome complexity. In addition to the translated coding sequences (CDS), transcribed regions (green and red) include untranslated regions (UTRs – shaded) that are bordered by the transcription start (TSS) and termination sites (TTS). A transcript may contain a single CDS, an operon of multiple CDSs or another untranscribed 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). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)