Tuberculosis xxx (2016) 1–10

Contents lists available at ScienceDirect

Tuberculosis

journal homepage: http://intl.elsevierhealth.com/journals/tube

MOLECULAR ASPECTS

REMap: Operon map of *M. tuberculosis* based on RNA sequence data

Shaaretha Pelly [a, 1], Kathryn Winglee [a, 1], Fang Fang Xia [b], Rick L. Stevens [b], William R. Bishai [a, c], Gyanu Lamichhane [a, *]

[a] Center for Tuberculosis Research, School of Medicine, Johns Hopkins University, 1550 Orleans St, Baltimore, MD 21287, USA
[b] Argonne National Laboratory, Argonne, IL 60439, USA
[c] Howard Hughes Medical Institute, Center for Tuberculosis Research, Johns Hopkins University School of Medicine, 1550 Orleans St, Baltimore, MD 21231, USA

ARTICLE INFO

Article history:
Received 17 February 2016
Received in revised form
19 April 2016
Accepted 24 April 2016

Keywords:
Operon
Mycobacterium tuberculosis
RNAseq

SUMMARY

A map of the transcriptional organization of genes of an organism is a basic tool that is necessary to understand and facilitate a more accurate genetic manipulation of the organism. Operon maps are largely generated by computational prediction programs that rely on gene conservation and genome architecture and may not be physiologically relevant. With the widespread use of RNA sequencing (RNAseq), the prediction of operons based on actual transcriptome sequencing rather than computational genomics alone is much needed. Here, we report a validated operon map of *Mycobacterium tuberculosis*, developed using RNAseq data from both the exponential and stationary phases of growth. At least 58.4% of *M. tuberculosis* genes are organized into 749 operons. Our prediction algorithm, REMap (RNA Expression Mapping of operons), considers the many cases of transcription coverage of intergenic regions, and avoids dependencies on functional annotation and arbitrary assumptions about gene structure. As a result, we demonstrate that REMap is able to more accurately predict operons, especially those that contain long intergenic regions or functionally unrelated genes, than previous operon prediction programs. The REMap algorithm is publicly available as a user-friendly tool that can be readily modified to predict operons in other bacteria.

1. Introduction

The success of *Mycobacterium tuberculosis* (*M. tuberculosis*) as the etiological agent of tuberculosis disease is largely due to its ability to adapt to the wide range of stresses it faces upon entering its host [1]. These include hypoxia in granulomatous lesions [2], low nutrient conditions [3] and nitric oxide exposure upon macrophage activation [4]. In addition to inhabiting both intracellular environments within host macrophages and extracellular environments within lesions, *M. tuberculosis* bacilli are able to switch from periods of active growth to dormancy, giving rise to latent tuberculosis disease [5]. The ability of *M. tuberculosis* to thrive in diverse microenvironments and modulate growth accordingly is due to complex transcriptional regulation that is often unique to mycobacterial species [6].

Although transcriptional studies are widely used in *M. tuberculosis* research, a map of operon structure based on the *M. tuberculosis* transcriptome remains unavailable. To date, the *M. tuberculosis* operon databases in use have been computationally generated based on statistical modeling. For example, predictions of *M. tuberculosis* operons have been made by the Database for prOkaryotic OpeRons (DOOR) [7]. However, the algorithm used in DOOR is optimized for *Escherichia coli* (*E. coli*) and *Bacillus subtilis* data. Transcription in mycobacteria has been shown to differ from these models as *M. tuberculosis* does not consistently harbor consensus −35 element in 5′ UTRs and also shows a heavy reliance on alternative sigma factors, suggesting the use of alternative promoters for ORFs [8,9]. In addition, a recent study demonstrated that a quarter of genes in *M. tuberculosis* are expressed as leaderless transcripts which lack a 5′UTR and a classical ribosomal binding site [10]. These findings suggest that the use of computational methods built on classical transcription mechanisms may not be accurate for determining operon structure in *M. tuberculosis*.

Abbreviations: REMap, RNA Expression Mapping of operons.

* Corresponding author. Tel.: +1 410 502 8162; fax: +1 410 614 8173.
E-mail address: lamichhane@jhu.edu (G. Lamichhane).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.tube.2016.04.010
1472-9792/

Four operon maps have been predicted specifically for *M. tuberculosis*. Each map used a different prediction approach, including an annotation-based method using gene pathways and length of intergenic region (IGR) (BioCyc Pathway/Genome Database, MtbRvCyc) [11], a comparative method using cross-species conservation of gene proximity (The Institute of Genomic Research, TIGR) [12], a method that combined IGR length and *M. tuberculosis* microarray coexpression data [13] and a method that predicted operons based on IGR length, conserved gene clusters and transcriptional termination predictions (MycoperonDB) [14]. A recent review on the transcriptome of *M. tuberculosis* revealed significant discrepancies and little overlap between these four databases, suggesting that these methods of predicting operons are not optimal for *M. tuberculosis* [15]. Furthermore, the MtbRvCyc and TIGR databases, both based on methods in *E. coli*, showed the least overlap, indicating the importance of species-specific transcriptome analysis in the determination of transcriptional units of *M. tuberculosis*.

Experimental methods, such as RNase protection assays, have been used to delineate operon boundaries [16]. Despite the advantages of specificity and physiological relevance of these methods compared to computational predictions, they are not feasible for the prediction of transcriptional units on a genome-wide scale. RNA-sequencing technology (RNAseq) provides researchers with the ability to analyze the transcriptome at the resolution of a single nucleotide. This has revolutionized the mapping of ORFs and transcriptional units as well as the study of gene expression across all fields. A recent publication updated annotations in the TubercuList database, resulting in the correction of multiple ORF annotations in *M. tuberculosis* based on alternative start codon usage in RNAseq data [17]. RNAseq studies in *Helicobacter pylori* and *M. tuberculosis* have also revealed a wide prevalence of alternative transcriptional start sites within operons, suggesting the uncoupling of polycistrons under different environmental conditions [10,18]. Another currently available transcriptome-based operon prediction algorithm is included in the Rockhopper RNAseq analysis package, which bases its classification on intergenic distances and expression level correlation across experiments [19]. These studies highlight the importance of transcriptional unit mapping based on transcriptome analysis and suggest that an operon map cannot be generalized for expression of an organism under all conditions, but is highly specific for the conditions during which the transcriptome is analyzed.

However, given the widespread use of RNAseq by researchers who are not trained bioinformaticians, and the limits of current operon prediction programs for *M. tuberculosis*, we developed a user-friendly program, REMap (RNA Expression Mapping of operons), to predict operons based on RNAseq data. A similar method has also been used to study the transcriptome of *Mycobacterium marinum* [20]. In this study, we report an operon map of *M. tuberculosis* during both exponential and stationary phases of growth and validate the REMap predictions through comparisons to previously published operons as well as confirmation of newly predicted operons. This study reveals extensive co-transcription in the *M. tuberculosis* genome, with operons up to 14 genes in length and 58.4% of genes being transcribed in 749 operons during the exponential phase of growth. This map is also presented in the form of a heat map that provides researchers with a whole genome view of both operonic structure and expression within operons in *M. tuberculosis*. REMap and the heat map output enable researchers to analyze RNAseq data for transcriptional units and compare expression under the different conditions being tested.

2. Methods

2.1. Strains and growth conditions

A *M. tuberculosis* clinical isolate CDC1551 [21] carrying a plasmid, pMH94 [22], integrated at the attB site, was used in this study. By carrying an empty vector that is often used for introducing DNA into *M. tuberculosis*, this strain would be an ideal parent strain for gene deletion and complemented strains. *M. tuberculosis* was grown in Middlebrook 7H9 broth (Difco) supplemented with 0.5% glycerol, 10% oleic acid–albumin–dextrose–catalase (OADC), 0.05% Tween 80 under constant shaking at 37 °C and total RNA was isolated from exponential (optical density of culture $A_{600nm} = 0.8$) and stationary phases of growth (two days following peak optical density).

2.2. RNA isolation and RNA sequencing

Total RNA was extracted from duplicate 150 ml cultures of *M. tuberculosis* at both exponential and stationary phases. *M. tuberculosis* cultures were centrifuged and the bacterial pellet was resuspended in Trizol (Invitrogen). This mixture was transferred to 1.8 ml O-ring tubes containing 0.5 ml of 0.1 mm zirconia beads (BioSpec Products). Cells were incubated at 25 °C for 10 min, lyzed by six cycles of bead-beating for 30 s and cooling on ice for 1 min, using a mini-beadbeater at 4800 RPM. Lysed cells were centrifuged for 5 min at 13,000 RPM, the supernatant was transferred to a fresh microfuge tube and RNA was then extracted as described [23], followed by column purification with the RNeasy Mini Kit (Qiagen). 16S and 23S rRNA were removed from the sample with MICROBExpress Bacterial mRNA Enrichment Kit (Invitrogen). The quality of RNA was assessed using a Nanodrop (ND-1000, Labtech) and Agilent 2100 Bioanalyzer (Agilent Technologies) after each step of processing. cDNA library preparations, fragment library protocols, and all parameters followed standard SOLiD Applied Biosystems protocols. Sequence reads were aligned to *M. tuberculosis* CDC1551 reference sequence (EMBL Accession number AE000516) using the Bioscope v1.3 Whole Transcriptome plugin. Sequencing data was visualized using the Integrated Genome Viewer V2.2. RNA-sequencing and analysis was carried out at the Next Generation Sequencing Center, Johns Hopkins University.

2.3. Reverse-transcription and polymerase chain reaction

For reverse-transcription PCR (RT-PCR), 5 μg total RNA was treated with 2 U of TURBO™ DNA-free DNase (Applied Biosystems) according to manufacturer's instructions for 30 min, followed by addition of 2 U of DNase for another 30 min. Reverse transcription was carried out using SuperScriptIII Reverse Transcriptase (Invitrogen) and a gene specific reverse primer complementary to the gene at the 3′ end of the IGR. Approximately 350 ng of DNase-treated RNA, dNTPs and the reverse primer were incubated at 65 °C for 5 min, then on ice for 1 min. This was followed by the addition of FS buffer, DTT and either reverse transcriptase or water (negative control) to each tube according to manufacturer's instructions. The final concentration of dNTPs was 1 mM and the final concentration of the reverse primer was 1 uM. Reverse transcription was conducted in an isothermal cycler at 55 °C for 30 min, followed by 70 °C for 15 min to generate cDNA. cDNA was amplified using Taq DNA Polymerase, Recombinant (Invitrogen). PCR samples were run on a 1.5% agarose gel for analysis. PCR amplification from gDNA and a cDNA sample without reverse transcriptase were used as a positive and negative control respectively. Gene specific primers were designed to amplify intergenic region between two

Please cite this article in press as: Pelly S, et al., REMap: Operon map of *M. tuberculosis* based on RNA sequence data, Tuberculosis (2016), http://dx.doi.org/10.1016/j.tube.2016.04.010