Genomics Data 2 (2014) 110–113

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Genome-wide mapping of the distribution of CarD, RNAP σ^A, and RNAP β on the *Mycobacterium smegmatis* chromosome using chromatin immunoprecipitation sequencing

Robert Landick [a,b], Azra Krek [c], Michael S. Glickman [d], Nicholas D. Socci [c], Christina L. Stallings [e,*]

[a] Department of Biochemistry, University of Wisconsin, Madison, WI 53706, USA
[b] Department of Bacteriology, University of Wisconsin, Madison, WI 53706, USA
[c] Bioinformatics Core, Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA
[d] Immunology Program, Sloan Kettering Institute and Division of Infectious Diseases, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[e] Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110, USA

ARTICLE INFO

Article history:
Received 9 May 2014
Received in revised form 20 May 2014
Accepted 20 May 2014
Available online 11 June 2014

Keywords:
Mycobacteria
RNA polymerase
Transcription
Tuberculosis
CarD

ABSTRACT

CarD is an essential mycobacterial protein that binds the RNA polymerase (RNAP) and affects the transcriptional profile of *Mycobacterium smegmatis* and *Mycobacterium tuberculosis* [6]. We predicted that CarD was directly regulating RNAP function but our prior experiments had not determined at what stage of transcription CarD was functioning and at which genes CarD interacted with the RNAP. To begin to address these open questions, we performed chromatin immunoprecipitation sequencing (ChIP-seq) to survey the distribution of CarD throughout the *M. smegmatis* chromosome. The distribution of RNAP subunits β and σ^A were also profiled. We expected that RNAP β would be present throughout transcribed regions and RNAP σ^A would be predominantly enriched at promoters based on work in *Escherichia coli* [3], however this had yet to be determined in mycobacteria. The ChIP-seq analyses revealed that CarD was never present on the genome in the absence of RNAP, was primarily associated with promoter regions, and was highly correlated with the distribution of RNAP σ^A. The colocalization of σ^A and CarD led us to propose that *in vivo*, CarD associates with RNAP initiation complexes at most promoters and is therefore a global regulator of transcription initiation. Here we describe in detail the data from the ChIP-seq experiments associated with the study published by Srivastava and colleagues in the Proceedings of the National Academy of Science in 2013 [5] as well as discuss the findings from this dataset in relation to both CarD and mycobacterial transcription as a whole.
The ChIP-seq data have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE48164).

Specifications	
Sample and organism	Genomic DNA from *Mycobacterium smegmatis* mc^{2}155 derived strains
Sequencer	AB SOLiD 4 system high-throughput genome sequencer
Data format	Raw data: sra files, normalized data: wig, SOFT, MINiML, and TXT files
Experimental factors	In the *M. smegmatis* strain that was used, the *carD* gene had been deleted from the native chromosomal locus and the strain instead constitutively expressed a functional C-terminal HA tagged version of CarD. The exception was the control strain that expressed an untagged HA peptide and retained the *carD* gene at its endogenous locus.
Experimental features	All *M. smegmatis* strains were isogenic to mc^{2}155 and were grown at 37 °C in LB supplemented with 0.5% dextrose, 0.5% glycerol, and 0.05% Tween 80 to late log phase ($OD\lambda_{600}$ of 1.0) before crosslinking the protein–nucleic acid complexes.

* Corresponding author.
E-mail address: stallings@borcim.wustl.edu (C.L. Stallings).

Direct link to deposited data

The direct link for the ChIP-seq data is: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE48164.

Experimental design, materials and methods

Bacterial strains and culture conditions

All *M. smegmatis* strains were isogenic to mc^{2}155 and were grown at 37 °C in LB supplemented with 0.5% dextrose, 0.5% glycerol, and 0.05% Tween 80 (broth). For immunoprecipitation of CarD, RNAP β, and RNAP σ^A, a *carD* merodiploid strain was produced by integrating pMSG430sm*carD*-HA (constitutively expresses *M. smegmatis* C-terminal HA tagged CarD, kanamycin resistant) into the *attB* site of *M. smegmatis* mc^{2}155. Allelic exchange experiments were performed

http://dx.doi.org/10.1016/j.gdata.2014.05.012

with the *carD* merodiploid strain using a DNA donor sequence with homology to mc^{2}155 nucleotides 6141480 to 6142268 and 6140266 to 6141010 to delete all of the *carD* gene except the nucleotides encoding the first 10 and last 3 amino acids from the endogenous locus, generating Δ*carD attB*::tet*carD*-HA [6]. For immunoprecipitation of unfused HA peptide as a control, mc^{2}155 was transformed with pmsg431, which integrates into the *attB* site of the genome and constitutively expresses HA peptide. This strain was called mc^{2}155 *attB*::pmsg431.

Chromatin immunoprecipitation

Cultures of *M. smegmatis* Δ*carD attB*::tet*carD*-HA and mc^{2}155 *attB*::pmsg431 strains were grown to late log phase ($OD\lambda_{600} = \sim 1$) before adding a final concentration of 2% formaldehyde and shaking at room temperature for 30 min to crosslink DNA and proteins. The crosslinking was quenched by the addition of 0.25 ml of 2.5 M glycine per 5 ml of culture and incubated 5 min at 25 °C with shaking. 5 ml ($\sim 2.5 \times 10^9$ mycobacterial cells) of each culture was then collected by centrifugation. The cells were washed once with TE and resuspended in 100 μl of TE supplemented with Roche Complete protease inhibitor cocktail. The cell suspension was lysed using a Covaris Focused-Ultrasonicator so that the genomic DNA was sheared into ~100 base pair (bp) fragments, as assessed by DNA gel electrophoresis. The use of the Covaris Focused-Ultrasonicator was critical for this step and other sonicator systems were unable to yield a comparable consistency and homogeneity of DNA fragment distribution. The cell debris was spun down and the lysate was added to 400 μl ChIP lysis buffer (50 mM HEPES-KOH [pH 7.5], 140 mM NaCl, 1 mM EDTA, 1% Triton X-100) plus Roche Complete protease inhibitor cocktail.

Protein–nucleic acid complexes containing CarD-HA were immunoprecipitated from the *M. smegmatis* mc^{2}155 Δ*carD attB*::tet*carD*-HA strain cell lysate by adding 50 μl of anti-HA agarose (Sigma). Complexes containing unfused HA were immunoprecipitated from the mc^{2}155 *attB*::pmsg431 strain with the same anti-HA agarose. RNAP β and σ^A were immunoprecipitated from Δ*carD attB*::tet*carD*-HA with monoclonal antibodies specific for these subunits (Neoclone; 8RB13 for β, 2G10 for σ) immobilized on GammaBind G Sepharose (GE Healthcare Life Sciences). Each immunoprecipitation was performed in duplicate from two separate cultures, thus comprising two biological replicates. However, one of the RNAP σ^A samples was lost during library preparation and, therefore, there is only data for one RNAP σ^A replicate.

The lysates and antibodies were incubated overnight by rotating at 4 °C. The antibody matrix was washed 2× with ChIP lysis buffer, 2× with ChIP lysis buffer plus an additional 360 mM NaCl, 2× with ChIP wash buffer (10 mM Tris-HCl pH 8.0, 250 mM LiCl, 0.5% NP-40, 0.5% sodium deoxycholate, 1 mM EDTA), and 2× with TE, each time by rotating for 10 min at 4 °C. Complexes that co-precipitated with the respective antibody matrix were eluted twice by adding 100 μl of ChIP elution buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% SDS), incubating for 10 min at 65 °C with agitation, spinning down the antibody matrix, and transferring the eluate to a new tube. Wash and elution buffers were all supplemented with Roche Complete protease inhibitor cocktail. To reverse the crosslinks, the eluates were incubated overnight at 65 °C. 15 μl of each sample was removed for Western blot analysis of proteins, while 100 μg/ml of proteinase K was added to the rest of each sample and incubated at 37 °C for 2 h before isolating nucleic acid by chloroform phenol extracting 2 times, ethanol precipitating and resuspending the DNA pellet in 34 μl of water.

Sequencing

Co-precipitated DNA was sequenced using an AB SOLiD 4 high-throughput genome sequencer (Life Technologies) and a 50 bp read length, which provided sufficient reads for over 100-fold coverage of the genome in each sample, wherein the *M. smegmatis* genome is 6,988,209 bp in length and the coverage of each sequencing reaction was over 800 Mbp. Table 1 shows the total number of reads and number of mapped reads for each sample.

Normalization

The DESeq method [1] was used to normalize the raw data sequence reads. Specifically, the normalized coverage (or counts) was determined by multiplying the raw (sequenced) coverage (or counts) in each sample by that sample's size factor. The size factors are determined by taking the median of the ratios of observed counts. The normalized number of sequence reads per base pair was then expressed as a $\log_2$ value. If a read mapped with equal quality at multiple loci (but not more than 3), its contribution was distributed evenly among them. For example, the sequences of the 16S, 23S, and 5S ribosomal RNA are identical in the *M. smegmatis rrnA* and *rrnB* operons. Therefore, the total number of reads for those sequences was split equally between the operons. If the number of mapping loci was higher than 3, the read was discarded. The normalized number of reads for each base pair was saved as a wig file for each sample.

Data analysis

We first determined how well replicate samples of the distribution of a given protein correlated to each other and how well the distribution of CarD correlated to the distributions of RNAP β and RNAP σ^A (Tables 2 and 3). The correlations were obtained by computing the Pearson correlation of the genomic coverage profiles of each pair of samples. The coverage profiles were computed by summing the contributions of all mapped fragments, assuming they were 100 bp long, and then, in 20-bp steps along the entire genome, computing the average coverage of the surrounding 100-bp window. Table 2 shows the correlations between the individual replicates. These data showed that individual replicates for a single immunoprecipitation condition correlated highly with one another (bolded in Table 2) and indicated that the distribution of CarD-HA or RNAP β was consistent between biological replicates. This consistency between replicates allowed us to average the Pearson correlation values for each comparison to simplify the comparisons between immunoprecipitation conditions (Table 3). The correlation between the distribution of CarD-HA and the distribution of RNAP σ^A (bolded in Table 3) was almost as high as the correlation between the two CarD-HA replicates, indicating that the distribution of CarD-HA is very similar to that of RNAP σ^A.

To directly compare the genome distributions of CarD-HA, RNAP β, and RNAP σ^A, the reads per base pair from the unfused HA peptide sample served as the background control and were subtracted from the other datasets. The rationale for this control was that as a non-DNA binding protein, the HA peptide should be diffusely localized throughout the cell and serve as a readout for the background levels of nonspecific crosslinking to the DNA. The normalized, background-corrected $\log_2$ reads per base pair were then smoothed over a 20-bp window and RNAP σ^A and CarD-HA peaks were identified as described previously [3,4]. Briefly, maxima and minima were assigned as inflection

Table 1
Number of sequencing reads for each sample from the AB SOLiD 4 high-throughput genome sequencer set to a 50 bp read length.

Sample	# of reads	# of mapped reads	% mapped reads
CarD-HA-1	24,988,001	16,452,015	65.84%
RNAP β-1	24,145,461	16,249,329	67.30%
Unfused HA-1	27,153,580	17,194,808	63.32%
CarD-HA-2	9,323,217	7,097,095	76.12%
RNAP β-2	12,709,226	9,660,422	76.01%
RNAP σ^A-2	19,596,174	14,868,445	75.87%
Unfused HA-2	11,641,903	8,015,559	68.85%

Table 2
Pearson correlations of the genomic coverage profiles of each pair of samples. The bolded numbers show the correlation between the distributions of individual replicates for a single immunoprecipitation condition.

	CarD-HA-1	RNAP β-1	Unfused HA-1	CarD-HA-2	RNAP β-2	RNAP σ^A-1	Unfused HA-2
CarD-HA-1	1.00	0.71	0.57	**0.92**	0.71	0.89	0.62
RNAP β-1		1.00	0.71	0.63	**0.91**	0.60	0.76
Unfused HA-1			1.00	0.41	0.65	0.36	**0.87**
CarD-HA-2				1.00	0.70	0.95	0.52
RNAP β-2					1.00	0.66	0.80
RNAP σ^A-2						1.00	0.47
Unfused HA-2							1.00

points where the values ± 10 bp were both lower or both higher, respectively. Maxima within 20 bp were merged with the peak location assigned to the maximum with the highest absolute signal value. Adjacent minima were merged analogously. To assess the statistical significance, peaks were divided into 0.1 interval bins of peak heights with a lower cutoff of peak height of 0.4 $\log_2$ reads per base pair. Starting with the lowest bin, we then calculated the distance of each peak to the nearest gene start and compared these distances to those computed using genome coordinates arbitrarily rotated 1×10^6 bp around the *M. smegmatis* genome. Using the Wilcoxon–Mann–Whitney ranksum test for nonsimilarity of distributions [3,4], RNAP σ^A peaks in the 1.1–1.2 peak-height bins and CarD-HA peaks in the 0.5–0.6 peak-height bins were statistically significant (P values for similarity of the distributions <0.0001). In other words, for each peak-height bin, the two lists of peak-to-gene start distances (actual and rotated by 1×10^6 bp) were tested for whether they were from different populations using the Wilcoxon–Mann–Whitney ranksum test. Peaks in the lowest peak-height bin for which peak-to-gene start distances differed from random with a P value of <0.0001 plus all peaks with greater heights were then used as the statistically significant peaks. We then identified RNAP σ^A peaks associated with each gene as the closest RNAP σ^A peak upstream from the gene start and CarD-HA peaks associated with the RNAP σ^A peaks as the closest CarD-HA peak to each selected RNAP σ^A peak. To calculate average ChIP signals for the aggregate profiles (Fig. 1), we selected a subset of 62 genes meeting the following criteria: (i) ≥300 bp in gene length, (ii) average RNAP $\log_2$ ChIP signal ≥1.6/bp, (iii) associated with an RNAP σ^A peak with $\log_2$ ChIP signal ≥3/bp, (iv) absence of other RNAP σ^A peaks within 500 bp upstream or 1000 bp downstream of the associated RNAP σ^A peak, (v) absence of an oppositely oriented gene with an average RNAP β $\log_2$ ChIP signal ≥1 upstream from the gene (because an oppositely oriented gene could create a divergent promoter region with potential for overlapping RNAP σ^A and CarD-HA ChIP signals), and (vi) absence of an upstream gene with average RNAP $\log_2$ ChIP signal >0 within 100 bp upstream from the gene (because such an arrangement would indicate the gene is an internal member of an operon).

The RNAP β, CarD-HA, and RNAP σ^A signals from the 62 genes were then averaged using the distance from the center of the associated σ^A peaks to align the genes (Fig. 1). For the gene alignments, the distance from the center of the associated σ^A peak served as a proxy for the transcriptional start site, since most transcriptional start sites are not mapped in *M. smegmatis*. This analysis showed that whereas RNAP β was found throughout transcribed regions of the genome, CarD-HA and RNAP σ^A were primarily associated with promoter regions. These data matched the high correlation calculated for the distribution of CarD-HA and RNAP σ^A (Tables 2 and 3). Levels of both CarD-HA and RNAP σ^A dropped off immediately following the promoter sequences, suggesting that these proteins are lost from the RNAP elongating complex after transcription initiation. The colocalization of RNAP σ^A and CarD-HA led us to propose that *in vivo*, CarD associates with RNAP initiation complexes at most promoters and is therefore a global regulator of transcription initiation. Further analysis of the dataset also revealed that CarD was never present on the genome in the absence of RNAP, suggesting that it may be targeted to the genome through its interaction with RNAP.

Table 3
Average Pearson correlations of the genomic coverage profiles for each immunoprecipitation condition examined. Each sample was done in duplicate, except σ^A was done once. Correlations are the average of each duplicate to one another. The bolded number shows the correlation between the distribution of CarD-HA and the distribution of RNAP σ^A.

	HA	CarD-HA	RNAP β	RNAP σ^A
HA	0.934	0.530	0.730	0.417
CarD-HA		0.962	0.687	**0.919**
RNAP β			0.954	0.629
RNAP σ^A				1.000

Discussion

CarD modulates transcription through its direct interaction with RNAP [6,7]. To determine at which stage of the transcription cycle (initiation, elongation, or termination) CarD acts, we used ChIP-seq [2] to survey the distribution of CarD throughout the *M. smegmatis* chromosome. Our data shows that CarD is localized to promoters throughout the *M. smegmatis* genome, indicating that CarD functions during transcription initiation. Despite the previous finding that CarD has sequence non-specific DNA binding activity [5], the ChIP-seq experiments also revealed that CarD was never present on the genome in the absence of RNAP β or RNAP σ^A, suggesting that CarD is targeted to the genome through its interaction with RNAP. The ChIP-seq data for the distribution of RNAP σ^A also serves as a map of potential promoter elements

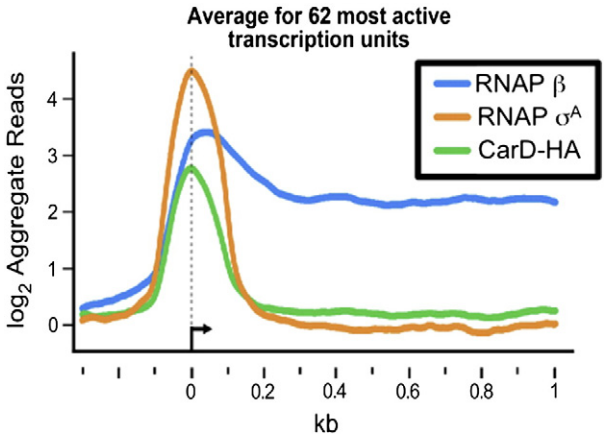

Fig. 1. Normalized $\log_2$ of ChIP-seq reads from *M. smegmatis* DNA co-immunoprecipitated with RNAP β, RNAP σ^A, or CarD-HA. Protein–DNA complexes containing CarD-HA, RNAP β, and RNAP σ^A were immunoprecipitated from *M. smegmatis* lysates. The co-precipitated DNA was sequenced, and the number of sequence reads per bp was normalized to total reads per sample and expressed as a $\log_2$ value. Normalized reads per base pair from DNA precipitated from cells expressing only the HA epitope were used as background and subtracted from the other samples. Shown are the aggregate profiles averaged over 62 highly active transcription units with the 0 designating the estimated transcriptional start sites. The 62 transcription units were selected on the basis of high signal and isolation from surrounding transcription units.

throughout the *M. smegmatis* genome, which has never before been experimentally examined.

The ChIP-seq experimental dataset has also raised a number of new questions. Compilation of the ChIP-seq data and previous microarray expression profiling analyses [6] indicates that CarD is broadly distributed on promoters of most transcription units regardless of whether they were deregulated during CarD depletion. This brings into question whether CarD activity exhibits promoter specificity. There is also the striking correlation between the distributions of CarD and RNAP σ^A on the genome, despite the fact that no direct interaction between these proteins has been reported. The factors contributing to the enrichment of CarD at RNAP σ containing holoenzymes as opposed to elongating RNAP core complexes remain unknown and will be a topic of future study. All together, results from these experiments have provided invaluable information that will help direct the ongoing efforts in determining the mechanism of transcription regulation by CarD. In addition, this work serves as a framework for further investigations into RNAP function in mycobacteria.

Acknowledgments

The authors thank the Genomics Core Laboratory (GCL) at Memorial Sloan Kettering Cancer Center (MSKCC) for performing the next-generation sequencing for ChIP-seq experiments. The GCL is supported by the cancer center core grant P30 CA008748.

References

[1] S. Anders, W. Huber, Differential expression analysis for sequence count data. Genome Biol. 11 (2010) R106.
[2] D.S. Johnson, A. Mortazavi, R.M. Myers, B. Wold, Genome-wide mapping of *in vivo* protein-DNA interactions. Science 316 (2007) 1497–1502.
[3] R.A. Mooney, S.E. Davis, J.M. Peters, J.L. Rowland, A.Z. Ansari, R. Landick, Regulator trafficking on bacterial transcription units *in vivo*. Mol. Cell 33 (2009) 97–108.
[4] N.B. Reppas, J.T. Wade, G.M. Church, K. Struhl, The transition between transcriptional initiation and elongation in *E. coli* is highly variable and often rate limiting. Mol. Cell 24 (2006) 747–757.
[5] D.B. Srivastava, K. Leon, J. Osmundson, A.L. Garner, L.A. Weiss, L.F. Westblade, M.S. Glickman, R. Landick, S.A. Darst, C.L. Stallings, E.A. Campbell, Structure and function of CarD, an essential mycobacterial transcription factor. Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 12619–12624.
[6] C.L. Stallings, N.C. Stephanou, L. Chu, A. Hochschild, B.E. Nickels, M.S. Glickman, CarD is an essential regulator of rRNA transcription required for *Mycobacterium tuberculosis* persistence. Cell 138 (2009) 146–159.
[7] L.A. Weiss, P.G. Harrison, B.E. Nickels, M.S. Glickman, E.A. Campbell, S.A. Darst, C.L. Stallings, Interaction of CarD with RNA polymerase mediates *Mycobacterium tuberculosis* viability, rifampin resistance, and pathogenesis. J. Bacteriol. 194 (2012) 5621–5631.