Genome-wide mapping of the distribution of CarD, RNAP $\sigma_A$, and RNAP $\beta$ on the Mycobacterium smegmatis chromosome using chromatin immunoprecipitation sequencing

Robert Landick  
*University of Wisconsin-Madison*

Azra Krek  
*Memorial Sloan-Kettering Cancer Center*

Michael S. Glickman  
*Memorial Sloan-Kettering Cancer Center*

Nicholas D. Socci  
*Memorial Sloan-Kettering Cancer Center*

Christina L. Stallings  
*Washington University School of Medicine in St. Louis*

Follow this and additional works at: [https://digitalcommons.wustl.edu/open_access_pubs](https://digitalcommons.wustl.edu/open_access_pubs)

**Recommended Citation**

Landick, Robert; Krek, Azra; Glickman, Michael S.; Socci, Nicholas D.; and Stallings, Christina L., "Genome-wide mapping of the distribution of CarD, RNAP $\sigma_A$, and RNAP $\beta$ on the Mycobacterium smegmatis chromosome using chromatin immunoprecipitation sequencing." Genomics Data. 2., 110-113. (2014). [https://digitalcommons.wustl.edu/open_access_pubs/4754](https://digitalcommons.wustl.edu/open_access_pubs/4754)

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
Data in Brief

Genome-wide mapping of the distribution of CarD, RNAP $\sigma^A$, and RNAP $\beta$ 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

**A R T I C L E  I N F O**

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

**A B S T R A C T**

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 $\beta$ and $\sigma^A$ were also profiled. We expected that RNAP $\beta$ would be present throughout transcribed regions and RNAP $\sigma^A$ 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 $\sigma^A$. The colocalization of $\sigma^A$ 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).

© 2014 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

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 mc2155 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 $\beta$, and RNAP $\sigma^A$, a carD merodiploid strain was produced by integrating pMSG430smcarD-HA (constitutively expresses M. smegmatis C-terminal HA tagged CarD, kanamycin resistant) into the attB site of M. smegmatis mc2155. Allelic exchange experiments were performed...

---

* Corresponding author.
E-mail address: stallings@borcim.wustl.edu (C.L. Stallings).
with the *carD* merodiploid strain using a DNA donor sequence with homology to mc²155 nucleotides 6141480 to 6142268 and 6140266 to 614010 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 posts*::tetcarD-HA [6]. For immunoprecipitation of unfused HA peptide as a control, mc²155 was transformed with pmsg431, which integrates into the attB site of the genome and constitutively expresses HA peptide. This strain was called mc²155 attB::pmsg431.

**Chromatin immunoprecipitation**

Cultures of *M. smegmatis* Δ*carD posts*::tetcarD-HA and mc²155 attB::pmsg431 strains were grown to late log phase (OD₆₀₀ = –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 (±2.5 × 10⁸ 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-Ultrasoundicator 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-Ultrasoundicator 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* mc²155 Δ*carD posts*::tetcarD-HA strain cell lysate by adding 50 μl of anti-HA agarose (Sigma). Complexes containing unfused HA were immunoprecipitated from the mc²155 attB::pmsg431 strain with the same anti-HA agarose. RNAP β and σ⁶ were immunoprecipitated from Δ*carD posts*::tetcarD-HA with monoclonal antibodies specific for these subunits (Neoclonie; 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 σ⁶ samples was lost during library preparation and, therefore, there is only data for one RNAP σ⁶ 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 log₂ 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* *rrmA* 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 CardD correlated to the distributions of RNAP β and RNAP σ⁶ (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 CardD-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 CardD-HA and the distribution of RNAP σ⁶ (bolded in Table 3) was almost as high as the correlation between the two CardD-HA replicates, indicating that the distribution of CardD-HA is very similar to that of RNAP σ⁶.

To directly compare the genome distributions of CardD-HA, RNAP β, and RNAP σ⁶, 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 non-specific crosslinking to the DNA. The normalized, background-corrected log₂ reads per base pair were then smoothed over a 20-bp window and RNAP σ⁶ and CardD-HA peaks were identified as described previously [3,4]. Briefly, maxima and minima were assigned as infection

**Table 1**

| 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   | 12,799,226 | 9,690,422         | 76.01%        |
| RNAP σ⁶-2  | 15,596,174 | 14,688,445        | 72.87%        |
| Unfused HA-2| 11,641,903 | 8,015,559         | 68.85%        |
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 log2 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 × 106 bp around the transcriptional start site, since most transcriptional start sites are not from the center of the associated operon). An arrangement would indicate the gene is an internal member of an operon.

Average Pearson correlations of the genomic coverage profiles for each immunoprecipitation condition examined. Each sample was done in duplicate, except αA 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 αA. Levels of both CarD-HA and RNAP αA 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 αA 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.

**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 αA, suggesting that CarD is targeted to the genome through its interaction with RNAP. The ChIP-seq data for the distribution of RNAP αA also serves as a map of potential promoter elements and RNAP αA were primarily associated with promoter regions. These data matched the high correlation calculated for the distribution of CarD-HA and RNAP αA (Tables 2 and 3). Levels of both CarD-HA and RNAP αA 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 αA 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 2**

|            | CarD-HA-1 | RNAP β-1 | Unfused HA-1 | CarD-HA-2 | RNAP β-2 | RNAP αA-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      | 0.87       | 0.76         |
| Unfused HA-1| 1.00      | 0.41      | 0.65         | 0.36      | 0.70      | 0.95       | 0.52         |
| CarD-HA-2  | 1.00      | 0.71      | 1.00         | 0.66      | 0.70      | 0.95       | 0.80         |
| RNAP αA-1  | 1.00      | 0.71      | 1.00         | 0.66      | 0.70      | 0.95       | 0.80         |
| Unfused HA-2| 1.00      | 0.41      | 1.00         | 0.66      | 0.70      | 0.95       | 0.80         |

**Table 3**

|            | CarD-HA | RNAP β | RNAP αA |
|------------|---------|--------|---------|
| HA         | 0.934   | 0.530  | 0.730   | 0.417   |
| CarD-HA    | 0.962   | 0.667  | 0.919   | 0.629   |
| RNAP β     | 0.954   | 0.682  | 1.000   |
| RNAP αA    |         |        |         |
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 $\sigma^A$ 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 $\sigma^A$ 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.