RNA SEQ DESCRIPTION DATA FILE

1. GENERAL INFORMATION.

Title:

Global transcriptional start site mapping of *E. coli* MG1655 using dRNA-seq, version 1.0

Reference: Thomason MK*, Bischler T*, Eisenbart SK, Förstner KU, Zhang A, Herbig A, Nieselt K, Sharma CM, Storz G (201$) *Global transcriptional start site mapping using dRNA-seq reveals novel antisense RNAs in Escherichia coli.* Submitted to NAR
*Equally contributing authors

Contact person for this dataset:

Questions concerning the content of the dataset by users of RegulonDB will be forwarded to this person. We appreciate to receive copy of the responses to users, so we can keep track of taking care of user requests.

Person: Dr. Gisela Storz

Email address: storz@helix.nih.gov

Person: Dr. Cynthia M. Sharma

Email address: cynthia.sharma@uni-wuerzburg.de

2. DATASET DESCRIPTION.

Summary:

This dataset consists of annotations for 14,868 transcriptional start sites (TSS) in the *E. coli* K12 MG1655 genome. The TSS were detected by combining a differential RNA-seq (dRNA-seq) approach, which distinguishes between primary and processed transcripts (Sharma et al., 2010, Nature), with an automated transcriptional start site (TSS) prediction algorithm (Dugar et al., 2013, PloS Genetics). With the criteria of expression in at least one of three growth conditions examined, the TSS predictions from this RNA-seq data set converged on a core set of 14,868 TSS, including 5,495 TSS corresponding to potential antisense RNAs (asRNAs).

Experiment:

The following information is essential to facilitate curation, availability and classification of your dataset within RegulonDB:

- Version of *E. coli´* s Genome used in the experiment: NC_000913.2
- Sequence Identifier: 49175990
- Experimental conditions: Samples originate from three different biological conditions: MG1655 wild type strain grown to exponential (OD$_{600}$ ~0.4) or stationary phase (OD$_{600}$ ~2.0) in LB medium (samples LB 0.4 and LB 2.0, respectively) as well as grown to exponential phase (OD$_{600}$ ~0.4) in M63 minimal glucose medium (sample M63 0.4).
- Replicates: The following table gives an overview of the different biological and technical replicates that were generated for the three biological conditions M63 0.4, LB 0.4 and LB 2.0. The libraries were analyzed by Illumina sequencing in three independent sequencing runs on either a GAIIx (GA sample) or a HiSeq 2000 machine (HS1 and HS2 samples). The biological replicates B1 and B2 originate from two independent culture sample preparations. The library replicates, L1 and L2, describe repeated generation of a cDNA library from the same biological sample. From all replicates two distinct dRNA-seq cDNA libraries were generated, one from RNA that was treated with terminator exonuclease (+TEX) and one from total RNA without treatment (□TEX) resulting in a total of 22 libraries. Treatment with TEX leads to a characteristic enrichment in the cDNA coverage pattern at the 5′ end of transcripts compared to the –TEX library. This enrichment pattern can be used to globally map transcriptional start stites (TSS). Note that Libraries were generated form RNAs without any fragmentation step.

Biological replicate	Library replicate	Sequencing run	Biological condition			Sequencing technique
			M63 0.4	LB 0.4	LB 2.0	
B1	L1	GA	x	x	x	GAIIx
B2	L1	HS1	x	x	x	HiSeq 2000
B2	L1	HS2	x	x	x	HiSeq 2000
B1	L2	HS2			x	HiSeq 2000
B2	L2	HS2			x	HiSeq 2000

- TSS annotation: All replicates were used to globally annotate TSS based on "detection" of sharp flanks in the expression graphs of the +TEX libraries and the characteristic "enrichment" between +TEX and –TEX libraries. For this we applied an automated approach (TSSpredator) which annotates TSS in a comparative way by directly considering different biological conditions including replicates. In order to take into account the varying number of replicates we required a TSS to exceed the "flank detection" thresholds in at least 2 out of 3 replicates of M63 0.4 and LB 0.4 and in at least 3 out of 5 replicates of LB 2.0 ("matching replicates" parameter) to mark it as "detected" in the respective condition. Furthermore, a TSS was marked as "enriched" in a specific condition if it exceeded the "enrichment" threshold in at least one replicate. A TSS was annotated if it was "detected" as well as "enriched" in at least one of the three conditions but can be possibly only "detected" in the other conditions (for details see Methods).

Methods:

RNA isolation

Growth conditions. To harvest total RNA, overnight cultures of wild type MG1655 grown in LB at 37°C were diluted 1:500 in either fresh LB or M63 minimal glucose medium (supplemented with final concentrations of 0.001% vitamin B1 and 0.2% glucose) and allowed to grow until the cultures reached an OD_{600} of ~ 0.4 and 2.0 for cultures grown in LB and an OD_{600} of ~0.4 for cultures grown in M63. For samples grown to OD_{600} of ~0.4 a total volume of 25 ml of cells was harvested and combined with 5 ml of stop solution (95% Ethanol, 5% acid phenol). For samples grown to OD_{600} of 2.0, a total volume of 5 ml of cells was harvested and combined with 1 ml of stop solution. All samples were mixed, incubated on ice for 10 min after which cells were collected by centrifugation at 4150 rpm at 4°C for 10 min. Cell pellets were snap frozen in an ethanol/dry ice slurry and stored at -80°C until RNA could be extracted.

RNA extraction. Cell pellets were thawed on ice and resuspended in 880 □l of lysis buffer (0.5 mg/ml lysozyme dissolved in TE pH 8.0, 1% SDS), mixed by inversion and incubated at 65°C for 2 min or until the samples cleared. The samples were cooled and 88 □l of 1M sodium acetate, pH 5.2 was added along with 1 ml of acid phenol:chloroform (Ambion). Samples were incubated at 65°C for 6 min with mixing and spun 10 min at 13,000 rpm, 4°C. The aqueous layer was extracted a second time with chloroform using Phase Lock Gel 2.0 tubes (5Prime) after which the aqueous layer was ethanol precipitated, washed and resuspended in 100 □l of DEPC-H_2O. RNA concentration was measured by reading the absorbance at OD_{260} and RNA integrity was checked by running ~2 □g aliquots of each sample on a denaturing 1% agarose 1X TBE gel followed by ethidium bromide staining.

Deep sequencing sample preparation

DNase I treatment. Samples were treated with DNase I to remove residual genomic DNA. 40 □g of total RNA was denatured at 65°C for 5 min. The RNA was then combined with 1X DNase I buffer + $MgCl_2$ (Fermentas), 20 U of RNase Inhibitor (Invitrogen), and 10 U of DNase I (Fermentas) in a final volume of 100 □l. The mixture was incubated for 45 min at 37°C and then extracted with phenol:chloroform:isoamylalcohol (Invitrogen) in 2 ml Phase

Lock Gel Heavy tubes (5Prime). Samples were precipitated, washed, and resuspended in 40 ☐l of DEPC-H_2O. RNA concentration and integrity of ~100 ng aliquots were checked as above. Genomic DNA contamination was tested by a PCR reaction with oligonucleotides MK0095 and MK0096. Samples free of gDNA contamination were used in subsequent steps.

Terminator exonuclease treatment. 7 µg of DNase I-treated RNA was denatured for 2 min at 90˚C, cooled on ice for 5 min and combined with 10 U RNase Inhibitor (Invitrogen), 1X Terminator Exonuclease Buffer A (Epicentre), and 7 U of Terminator Exonuclease (Epicentre) in a final reaction volume of 50 ☐l. Control reactions lacking terminator exonuclease were run in parallel for each sample. Reactions were incubated at 30˚C for 1 h and stopped by the addition of 0.5 ☐l of 0.5 M EDTA, 50 ☐l DEPC-H_2O and extraction with phenol:chloroform:isoamylalcohol with Phase Lock Gel 2.0 Tubes. The supernatant was precipitated, washed and resuspended in 11 ☐l of DEPC-H_2O. RNA concentration was determined by reading the absorbance at OD_{260}.

Tobacco acid pyrophosphatase treatment. Samples were prepared for cDNA library construction by treating the TEX-treated and untreated control samples with tobacco acid pyrophosphatase. Briefly, samples were incubated for 1 h at 37˚C with 1X TAP Buffer (Invitrogen), 10 U RNase Inhibitor and 5 U Tobacco Acid Pyrophosphatase (Invitrogen) in a final reaction volume of 20 µl. The samples were extracted with phenol:chloroform:isoamylalcohol, precipitated, washed, and resuspended in 20 µl of DEPC-H_2O. RNA concentration was determined by reading the absorbance at OD_{260}, and RNA integrity was checked on a denaturing 4% acrylamide-7M urea gel in 1X TBE and visualized with Stains-all nucleic acid stain (Sigma).

Illumina sample preparation and sequencing. Equal amounts of RNA samples were poly(A)-tailed using poly(A) polymerase. Then, the 5'-PPP structures were removed using tobacco acid pyrophosphatase (TAP). Afterwards, an RNA adapter was ligated to the 5´-phosphate of the RNA. In the case of the GAIIx-libraries, the 5′ linker contained the barcode sequence at its 3′ end. For HiSeq 2000 libraries, the barcode was introduced in a later step during PCR-amplification of the cDNA library. First-strand cDNA was synthesized by using an oligo(dT)-adapter primer and the M-MLV reverse transcriptase. In a PCR-based amplification step using a high fidelity DNA polymerase the cDNA concentration was increased to 20-30 ng/☐l. For all libraries the Agencourt AMPure XP kit (Beckman Coulter Genomics) was used to purify the DNA, which was subsequently analyzed by capillary electrophoresis.

GAIIx-libraries. For the GAIIx-libraries, PCR products for sequencing were generated using the following primers designed for amplicon sequencing according to the instructions of Illumina/Solexa:

5′-end_primer

5’-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGACGATCNNNN-3’

3’-end_primer

5’-CAAGCAGAAGACGGCATACGATTTTTTTTTTTTTTTTTTTTTTTT-3’

The samples were run on an Illumina GAIIx instrument with 120 cycles in single-read mode.

HiSeq 2000-libraries. For the HiSeq 2000 libraries, a library-specific barcode for multiplex sequencing was included as part of a 3'-sequencing adapter. The following adapter sequences flank the cDNA inserts:

TrueSeq_Sense_primer

5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3’

TrueSeq_Antisense_NNNNNN_primer (NNNNNN = 6n barcode for multiplexing)

5’-CAAGCAGAAGACGGCATACGAGAT-NNNNNN-
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC(dT25)-3’

The samples were run on an Illumina HiSeq 2000 instrument with 100 cycles in single-read mode.

Raw sequence reads were uploaded at GEO with accession number GSE55199.

Analysis of deep sequencing data

Read mapping and coverage plot construction. To assure high quality of sequencing reads, the Illumina reads in FASTQ format were trimmed with a cut-off phred score of 20 by the program fastq_quality_trimmer from FASTX toolkit version 0.0.13. After trimming, poly(A)-tail sequences were removed and a size filtering step was applied in which sequences shorter than 12 nt were eliminated. The collections of remaining reads were mapped to the *E. coli* MG1655 genome (NCBI Acc.-No: NC_000913.2; Jun 24, 2004) using *segemehl* 0.1.4 with an accuracy cut-off of 95%. Coverage plots representing the numbers of mapped reads per nucleotide were generated. Reads that mapped to multiple locations contributed a fraction to the coverage value. For example, reads mapping to three positions contributed only 1/3 to the coverage values. Each graph was normalized to the number of reads that could be mapped from the respective library. To restore the original data range, each graph was then multiplied by the minimum number of mapped reads calculated over all libraries.

Normalization of expression graphs. Prior to the comparative analysis, the expression graphs with the cDNA coverage that resulted from the read mapping were further normalized. A percentile normalization step was applied to normalize the +TEX graphs. To this end, the 90th percentile of all data values was calculated for each +TEX graph. This value was then used to normalize the +TEX graph as well as the respective □TEX graph. Thus, the relative differences between each +TEX and □TEX graph were not changed in this normalization step. Again, all graphs were multiplied with the overall lowest value to restore the original data range. To account for different enrichment rates, a third normalization step was applied. During this step, prediction of TSS candidates was performed for each replicate of each strain. These candidates were then used to determine the median enrichment factor for each +/□TEX library pair. Using these medians all □TEX libraries were then normalized against the library with the strongest enrichment. Besides annotation of transcriptional start sites, the resulting graphs were also used for visualization in the Integrated Genome Browser.

Transcriptional start site prediction. Based on the normalized expression graphs we conducted an automated TSS prediction utilizing TSSpredator (v1.04-beta; http://it.inf.uni-tuebingen.de/TSSpredator). In brief, for each position (i) in the expression graph corresponding to the TEX treated libraries, the algorithm calculates an expression height, e(i), and compares that expression height to the preceding position by calculating e(i) – e(i-1), which is termed the flank height. Additionally, the algorithm calculates a factor of height change e(i)/e(i-1). To determine if a TSS is a primary TSS and not a processed transcript end an enrichment factor is calculated as $e_{+TEX}(i)/e_{-TEX}(i)$, where $e_{+TEX}(i)$ is the expression height for the terminator exonuclease treated sample and $e_{-TEX}(i)$ is the expression height for the untreated sample. For all positions where these parameters exceed the predefined thresholds a TSS is annotated.

We set the thresholds for the "minimum flank height" and the "minimum factor of height change" which are used to determine if a TSS is "detected" to 0.3 and 2.0, respectively. Here, the value for the "minimum flank height" is a factor to the minimum 90th percentile over all libraries resulting in an absolute value of 1.62. If the TSS candidate reaches these thresholds in at least one replicate of one condition, the thresholds are decreased for the other replicates to 0.1 (0.54 absolute) and 1.5, respectively. Furthermore, we set the "matching replicates" parameter which determines the number of replicates in which a TSS must exceed these thresholds in order to be marked as "detected" within a condition to 2 for M63 0.4 and LB 0.4 and to 3 for LB 2.0. If a TSS was detected in a certain condition, the lowered thresholds also apply to all remaining libraries of the other conditions. Furthermore, we consider a TSS candidate to be enriched in a condition if the respective enrichment factor for at least one replicate is not less than 2.0. A TSS candidate has to be enriched in at least one condition and is discarded otherwise. If a TSS candidate is not enriched in a condition but still reaches the other thresholds it is only indicated as "detected". However, a TSS candidate can only be labeled as detected in a condition if its enrichment factor is above 0.66. Otherwise we consider it to be a processing site. In order to take into account slight variations between TSS positions the respective parameters for clustering between replicates and conditions were set to a value of 1. In doing so a consensus TSS position in a three nucleotide window is determined based on the maximum "flank height" among the respective libraries.

Predicted TSS are further classified as primary, secondary, antisense, internal or orphan TSS based on the location of the TSS relative to gene annotations. Primary and secondary TSS are identified as being within 300 nucleotides upstream of a gene, with primary TSS being those designated as having the highest expression values. All other TSS associated with the gene are considered secondary. Internal TSS are those that are calculated to be internal to a gene on the sense strand. asTSS are internal or within 100 nucleotides of a gene on the opposite strand of the gene annotation. Orphan TSS do not meet any of the above requirements.

3. Content of the data file

The data file contain the next columns:

1) TSS position *

2) RPKM expression value for a 50 nt window downstream of TSS (maximum of all libraries for the respective condition) *

3) Sequence of the promoter region with TSS in upper case (-50 nt upstream + TSS)

4) Orientation (+ or - strand) *

5) Relative position of TSS to the start of the gene

6) Gene name

7) Blattner number

8) Left end position of gene

9) Right end position of gene

10) Relative orientation to gene (upstream/sense, intragenic/sense, upstream/antisense, intragenic/antisense, downstream/antisense)

11) TSS class (primary, secondary, internal, antisense, orphan)

12) Enrichment: contains 1 if TSS is enriched in the condition and 0 if not