RNA SEQ DESCRIPTION DATA FILE

1. GENERAL INFORMATION.

Title:

High-throughput transcription initiation mapping, version 3.0

Reference:

Salgado, H. et al. (2013). "RegulonDB v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more". Nucleic Acids Research 2012 Nov; doi: 10.1093/nar/gks1201.

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 users requests.

Person: Dr. Enrique Morett

Email address: emorett@ibt.unam.mx

2. DATASET DESCRIPTION.

Summary:

This dataset describes 5197 positions of the E. coli MG1655 genome considered highly likely TSSs. These positions were detected in a series of 6 independent Illumina directional RNA-seq experiments were total RNA received different treatments to enrich for 5' monophosphate or 5' triphosphate ends (the later are characteristic of primary mRNA), leading to the generation of 22 different libraries. The selected data appeared in at least half of the libraries expected to represent 5' triphosphate RNA species. Importantly, our TSSs are described either as a single position or as a cluster (a group of consecutive positions).

Experiment:

- Version of *E. coli*´s Genome used in the experiment: NC_000913.2
- Sequence Identifier: 49175990
- Experimental conditions: LB (1 library) or MOPS media supplemented with 0.2% Glucose (20 libraries) or 0.2% acetate (1 library). In all cases shaking speed was 300 rpm and temperature 37°C. MG1655 WT was used in all but four experiments where an *rppH* null mutant derivative of this strain was employed.
- Number of independent experiments: 6 independent library sets each containing at least 1 and maximum 4 libraries corresponding to different total RNA treatments (M, MT, TA and TE, as explained below).

Methods:

Using an Illumina Genome Analyzer IIx (GAIIx), we sequenced the 5' ends (36bp reads) of all transcripts from E. coli MG1655 using four experimental methods. Samples were treated with DNase I, and ribosomal RNA was reduced with the RiboMinus Transcriptome Isolation Kit (Invitrogen, Carlsbad, CA). Then the RNA was either (1) enriched for 5' monophosphate transcripts (M), (2) enriched for 5' triphosphate transcripts by degrading 5' monophosphate transcripts with a specific exonuclease (TE), (3) enriched for 5' triphosphate transcripts by ligating an adapter only to these transcripts (TA), or (4) left untreated (MT). Finally, an adapter was ligated to the 5' end; all the above enables the identification of TSSs by filtering out products of degradation.

Reads were mapped to the genome using Bowtie -v mode allowing maximum 3 mismatches. We used the start position of the reads as the putative TSSs position and filtered out reads mapping to ribosomal operons. A total of 821,789 5' ends where detected in the sum of all libraries. We then selected 5197 positions as high confidence TSS because they represent expressed genes in the conditions tested and appeared in at least half of the MT, TA and TE libraries that contain or were enriched for 5' tryphosphorylated RNA species. Importantly, our TSSs are described either as a single position or as a cluster (a group of consecutive positions).

Finally, each cluster or single position TSS was classified into one of seven possible regions with the following priority: a) Upsense (from -150bp to 50bp up to the start of the gene); b) InSense (into the gene); c) UpSenseExten (upstream to gene over 150bp if no coding region was reached); d) UpAntiSense (in the upstream region of the gene but in the opposite orientation); e) InAntiSense (into the gene but in the opposite orientation); f) Convergent (in a region flanked by two genes with opposite direction); g) UpAntiSenseExten (upstream to gene over 150bp and in the opposite direction).

3. Column format of the data file

1) TSS left position
2) TSS right position: This value is the same as in column one for single position TSSs and higher that one for cluster TSSs
3) TSS position Max frequency: This value is the Position within the cluster with the highest frequency.
4) TSS length: 1 for single position TSSs and more than one for cluster TSSs.
5) Number of reads: For cluster TSSs the accumulated number of reads for all its positions is given.
6) Orientation (strand)
7) Gene name
8) Blattner number
9) Left end position for gene
10) Right end position for gene
11) Relative orientation to gene: Upsense, InSense, UpAntiSense, InAntiSense, Convergent, UpSenseExten or UpAntiSenseExten.