Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in *Escherichia coli*

Maureen K. Thomason,* Thorsten Bischler,¹ Sara K. Eisenbart,¹ Konrad U. Förster,¹ Aixia Zhang,¹ Alexander Herbig,² Kay Nieselt,² Cynthia M. Sharma,²* Gisela Storz*²

Cell Biology and Metabolism Program, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland, USA; Research Center for Infectious Diseases (ZINF), University of Würzburg, Würzburg, Germany; Center for Bioinformatics Tübingen (ZBIT), University of Tübingen, Tübingen, Germany

While the model organism *Escherichia coli* has been the subject of intense study for decades, the full complement of its RNAs is only now being examined. Here we describe a survey of the *E. coli* transcriptome carried out using a differential RNA sequencing (dRNA-seq) approach, which can distinguish between primary and processed transcripts, and an automated prediction algorithm for transcriptional start sites (TSS). With the criterion of expression under at least one of three growth conditions examined, we predicted 14,868 TSS candidates, including 5,574 internal to annotated genes (iTSS) and 5,495 TSS corresponding to potential antisense RNAs (asRNAs). We examined expression of 14 candidate asRNAs by Northern analysis using RNA from wild-type *E. coli* and from strains defective for RNases III and E, two RNases reported to be involved in asRNA processing. Interestingly, nine asRNAs detected as distinct bands by Northern analysis were differentially affected by the *rnc* and *rnt* mutations. We also compared our asRNA candidates with previously published asRNA annotations from RNA-seq data and discuss the challenges associated with these cross-comparisons. Our global transcriptional start site map represents a valuable resource for identification of transcription start sites, promoters, and novel transcripts in *E. coli* and is easily accessible, together with the cDNA coverage plots, in an online genome browser.

Despite the hundreds of asRNAs reported, even using the most conservative estimates, it is surprising how few functions have been elucidated for these RNAs. A limited number of asRNAs have been shown to modulate transcription, stability, or translation of the corresponding sense transcripts (reviewed in references 23 and 24). Other recent genome-wide studies have proposed more general functions for asRNAs. These include asRNA-directed digestion of sense transcripts by RNase III in Gram-positive but not Gram-negative organisms (25) and reciprocal effects on the expression of sense RNAs in a so-called “excludon” model (reviewed in reference 26). Still other studies conclude most asRNAs lack function and result from pervasive transcription (16, 27), collisions between replication and transcription machinery (28), or inefficient transcription termination, particularly in the absence of a transcription stop signal.
of the Rho protein (9, 17, 18), or correspond to contaminating genomic DNA (22).

To further explore the *Escherichia coli* transcriptome on a genome-wide scale, particularly the subset of asRNAs, we carried out differential RNA sequencing (dRNA-seq) analysis (reviewed in reference 29), which we analyzed by an automated TSS prediction algorithm (30). This approach led us to identify, across three growth conditions, >5,500 potential TSS within genes, 212 divergently transcribed gene pairs with overlapping 5′ untranslated regions (UTRs), and >5,400 potential asRNA loci. We examined expression of 14 candidate asRNAs by Northern analysis and found 9 to be differentially degraded by RNase III and RNase E, two RNases implicated in asRNA-based regulation. Our global TSS map is one of the best and most sensitive data sets for promoter and transcript identification in the widely used model organism *E. coli* and is easily accessible at RegulonDB (21) and via an online browser at http://cbmp.nichd.nih.gov/segr/ecoli/.

**MATERIALS AND METHODS**

**Strain construction.** The strains and oligonucleotides used for this study are listed in Tables S1 and S2, respectively, in the supplemental material. The asRNA deletion control strains were constructed using λ Red-mediated recombinational (31) to replace the region encompassing the asRNA signal along with 300 nucleotides (nt) on either side with a kanamycin cassette. Deletion constructs were confirmed by sequencing and moved into new wild-type or mutant backgrounds by P1 transduction.

**Growth conditions.** Cells were grown at 37°C in LB (10 g of tryptone, 5 g of yeast extract, 10 g of NaCl per liter) or M63 minimal glucose medium (supplemented with final concentrations of 0.001% vitamin B1 and 0.2% glucose) to an optical density at 600 nm (OD_{600}) of ~0.4 and 2.0 for LB and an OD_{600} of ~0.4 for M63. At the indicated OD_{600}, 25 ml of cells (OD_{600} of 0.4) or 5 ml of cells (OD_{600} of 2.0) was combined in a 5:1 ratio of cells to stop solution (95% ethanol, 5% acid phenol [pH 4.5]), vortexed, incubated on ice for 10 min, and collected by centrifugation. Cell pellets were snap-frozen in an ethanol-dry ice slurry and stored at −80°C.

**Deep-sequencing sample preparation.** Details for sample preparation for deep sequencing can be found in Materials and Methods in the supplemental material. Briefly, RNA extraction for RNA-seq analysis was performed as described previously using hot-acid phenol chloroform (3, 32). RNA samples were treated with DNase I to remove contaminating genomic DNA. RNA samples free of genomic DNA were treated with terminator 5′-phosphate-dependent exonuclease (TEX) (Epicentre) followed by tobacco acid pyrophosphatase (TAP) treatment (Invitrogen) as described previously (3). Control reactions lacking terminator exonuclease were run in parallel for each sample. Unfractionated total RNA was used to construct cDNA libraries for sequencing on GAIIx and HiSeq 2000 machines.

**Analysis of deep-sequencing data.** For a detailed description of the read mapping, expression graph construction, normalization of expression graphs, correlation analysis, TSS prediction, comparison to other data sets, and other computational analyses, see Materials and Methods in the supplemental material.

(i) **Read mapping.** Between 1.8 and 9.8 million reads for each of the cDNA libraries were mapped to the *E. coli* MG1655 genome (NCBI accession no. NC_000913.2 [24 June 2004]) using our RNA-seq pipeline READemption (33) and segemehl, with an accuracy cutoff of 95% (34).

(ii) **Correlation analysis.** Nucleotide- and gene-wise Spearman and Pearson correlation coefficients were calculated based on concatenated values of forward and reverse strand position-wise coverage files and visualized using the R package corrplot. Gene-wise correlation values utilized read overlap counts based on NCBI annotations (accession no. NC_000913.2).

(iii) **TSS prediction.** Transcriptional start site (TSS) prediction was performed using the program TSSpredator (http://it.inf.uni-tuebingen.de/TSSpredator) (30). TSS were classified as primary TSS (pTSS), secondary TSS (sTSS), internal TSS (iTSS), or orphan TSS (oTSS) based on the location relative to gene annotations. pTSS and sTSS are within 300 nucleotides upstream of a gene, with pTSS having the highest expression values. All other TSS associated with the gene are considered secondary. iTSS are internal to a gene on the sense strand, while asTSS are internal or within 100 nucleotides of a gene on the opposite strand of the annotation. oTSS do not meet any of the above requirements.

(iv) **Comparison to DOOR.** A table containing all operon annotations (1,526 single-gene operons and 851 operons consisting of multiple genes) was downloaded from the Database of Prokaryotic Operons (DOOR) 2.0 website (35) and compared to a final set of 2,441 TSS.

(v) **Comparison of pTSS and sTSS to RegulonDB promoters.** We extracted 6,406 TSS annotated based on the “strong evidence” classification (21) from the RegulonDB promoter table (version RegulonDB 8.3, 11-28-2013) and classified them according to our classification scheme, resulting in a set of 3,987 pTSS and sTSS. We conducted a pairwise comparison of the positions to our data (4,261 pTSS and sTSS) based on a maximum allowed distance of 3 nt.

(vi) **Expression analysis and binning.** Expression values for predicted TSS classified as exclusively antisense or exclusively primary or secondary were calculated based on overlap counts for a 50-nt window downstream of the respective TSS position from which reads per kilobase per million mapped reads (RPKM) values were calculated (36). The TSS were grouped into six bins according to their RPKM values.

(vii) **Comparison of asRNAs detected in our and previous studies.** aTSS annotations were retrieved from the Materials and Methods sections in the supplemental material from published studies (13, 14, 16, 19) or were downloaded from RegulonDB (data set version 3.0 [21] and data set version 2.0 [20]). We excluded the study by Li et al. (15), which revealed >82,000 aTSS, as this number is very high compared to previous studies and our study and thus would bias the comparative analyses. We compared the aTSS from each data set, including our 6,379 predicted aTSS, to the aTSS of all other data sets in a pairwise manner, requiring either a precise match of the annotated positions or allowing a variation of 1, 2, 3, or 1 nt.

**Northern analysis.** RNA extraction for Northern analysis was performed using TRIzol reagent (Invitrogen). Northern analysis of 10 µg of total RNA was performed on denaturing 8% acrylamide–7 M urea gels as described previously (37), with minor changes for detection using riboprobes (for details and oligonucleotides used to create the riboprobes, see Materials and Methods in the supplemental material).

**RNA-seq data accession number.** Raw sequence reads were uploaded to the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo) under accession no. GSE55199.

**RESULTS**

**dRNA-seq reveals the primary transcriptome of *E. coli* MG1655.** To detect the transcripts expressed by *E. coli*, we collected two independent biological replicates (B1 and B2 samples) from MG1655 wild-type cells grown to the exponential phase (OD_{600} of ~0.4) or stationary phase (OD_{600} of ~2.0) in LB medium (samples LB 0.4 and LB 2.0, respectively) or grown to the exponential phase (OD_{600} of ~0.4) in M63 minimal glucose medium (sample M63 0.4) (Fig. 1; see also Table S3 in the supplemental material). For all six biological samples, total RNA was extracted and subjected to dRNA-seq library preparation for primary transcriptome analysis as described previously (3). Specifically, prior to cDNA library construction, half of each RNA sample was treated with 5′ terminator exonuclease (+ TEX samples), which degrades RNAs containing a 5′-monophosphate (5′-P), thereby enriching for primary transcripts containing 5′-triphosphates (5′-PPP). The other half of each sample was left untreated (−TEX samples) and thus contains both primary transcripts (5′-PPP) and processed RNAs.
(5’-P). Subsequently, the 5’-PPP ends in both samples were converted to 5’-P ends for cDNA library preparation.

The cDNA libraries of the first biological replicates (B1-L1) were sequenced on an Illumina Genome analyzer IIx (GA samples), while the second biological replicates (B2-L1) were sequenced on a HiSeq 2000 sequencer (HS1 samples). To examine variation between sequencing runs, the B2-L1 libraries were resequenced on a HiSeq 2000 sequencer (HS2 samples). To identify variation introduced during library preparation, technical replicates of the LB 2.0 libraries (B1-L2 and B2-L2 samples) were also generated and sequenced using the HiSeq 2000 (Fig. 1; see also Table S3 in the supplemental material).

Strand-specific sequencing resulted in a total number of ~1.8 to 3.6 million reads per sample for the GA set and ~5.3 to 9.8 million reads per sample for the HS sets after quality trimming (see Table S4 in the supplemental material). For all of the libraries, >70% of the reads could be mapped to the *E. coli* genome (NCBI accession no. NC_000913.2) indicating that the sequencing runs consisted of numerous high-quality reads. Read mapping analysis showed that for all three growth conditions, 65 to 80% of reads mapped to annotated regions of the genome while 2 to 6% mapped antisense to published annotations. The remainder of the reads mapped to unannotated intergenic regions, which also include UTRs (see Table S5 in the supplemental material). These data indicate the majority of transcripts correspond to the sense strand of genes; however, a small percentage of antisense transcription occurs, particularly opposite mRNAs.

**Correlation analysis reveals variation associated with library preparation and sequencing platform.** To assess the similarity between replicates, we calculated Spearman and Pearson correlation coefficients for nucleotide-wise expression values for both strands of all the −TEX and +TEX libraries (see Fig. S1 in the supplemental material). For each biological condition and both types of analysis, we noted the highest correlation among sequencing runs, the B2-L1 libraries were resequenced on the HiSeq 2000, likely due to differences in sequence coverage and cDNA library preparation protocols for the two platforms. Since the nucleotide-wise correlations are sensitive to slight fluctuations in cDNA read counts, we also assessed the correlation coefficients for gene-wise expression values, defined as the number of mapped reads within genes annotated by NCBI, among the −TEX and +TEX libraries. Overall the correlation increased but had a pattern similar to that seen for the nucleotide-wise comparisons.

Despite the high correlation between replicates and overall similar cDNA coverage patterns, a few regions showed variable expression or enrichment in the +TEX libraries across samples, likely due to the number of reads produced by the different sequencing instruments combined with differences in library preparation. However, as we had high correlation between replicates, similar read distributions across replicates, and agreement on the positions of transcript ends, we proceeded with automated genome-wide TSS annotation.

**The automated TSSpredator pipeline predicts previously unannotated TSS.** Several RNA-seq-based studies have reported genome-wide annotations of 5’ ends of *E. coli* genes, but most cannot distinguish between primary and processed transcripts, limiting the potential to identify these distinct types of transcripts. The TSSpredator pipeline recently developed to annotate TSS among multiple strains of *Campylobacter jejuni* (30). The TSSpredator prediction algorithm employs the RNA-seq data to determine the location of a TSS based on identifying positions with sharp increases in expression in the +TEX libraries relative to the −TEX libraries, which facilitates the differentiation between primary (5’-PPP) and processed (5’-P) transcripts (see Fig. S2A in the supplemental material) (3).

Using TSSpredator, TSS can be annotated in a comparative
manner among libraries through the integration of replicate information. If a strong enrichment is observed in one replicate, less stringent parameters can be applied to the same position in other replicates to ensure identification of TSS despite differences in library or sequencing preparations while still maintaining stringent criteria for detection. To perform such an analysis for our replicates of the three biological conditions (see Materials and Methods in the supplemental material), we adjusted the “matching replicates” parameter, which defines the minimum number of replicates in which a TSS must be detected for a particular biological condition. For the M63 0.4 and LB 0.4 conditions, where only three replicates were available, we required a TSS to be detected in at least two replicates, while for the LB 2.0 condition, we required detection in at least three of the five replicates. All other parameters were set to default values as established previously (30).

We predicted a total of 14,868 potential TSS mapping throughout the *E. coli* genome (see Data set S1 in the supplemental material). Of these, 6,297 were detected under all three conditions, 1,151 were detected only in cells growing exponentially in M63 minimal medium, 470 TSS were found in cells growing exponentially in LB, and 1,947 were found in stationary-phase cells growing in LB (Fig. 2A; see also Fig. S3A in the supplemental material for examples of TSS detected under only one condition). The higher number of TSS identified for the LB stationary-phase cells might be a result of changes in transcriptional programs required to survive in the stationary phase (38).

TSSpredator automatically assigns TSS to five different classes: primary TSS (pTSS; main transcription start of a gene or operon), secondary TSS (sTSS; alternative start with lower expression), internal TSS (iTSS; start within a gene), antisense TSS (asTSS; transcript start antisense to a gene ± 100 nt), and orphan TSS (oTSS; not associated with annotation) based on the location relative to existing gene annotation (see Fig. S2B in the supplemental material). A TSS can fall into more than one category, depending on its location relative to the surrounding gene annotations. For example, in the case of overlapping 5’ UTRs, a particular TSS can be both a pTSS and an asTSS. For downstream genes within operons, a pTSS can also be internal to the upstream genes. Among the 14,868 predicted TSS, we identified 2,672 pTSS (1,707 classified solely as pTSS), 1,589 sTSS (850 classified solely as sTSS), 5,574 iTSS (4,466 classified solely as iTSS), and 6,379 asTSS (5,495 classified solely as asTSS) (Fig. 2B).

To assess the coverage of our TSS predictions, we compared the number of TSS classified as pTSS only or pTSS and asTSS (2,057) and the number classified as pTSS and iTSS or pTSS, iTSS, and asTSS (615) with the number of genes classified as single-standing genes (1,526) or first genes within operons (851) in the Database of prOkaryotic OpeRons (DOOR) (35). In total, after excluding all TSS assigned to genes not annotated in DOOR (see Materials and Methods in the supplemental material), we used 2,441 of our TSS classified as pTSS. In agreement with the assumption that a TSS must precede genes annotated as single genes or first genes in DOOR, we detected a pTSS for ~78% of the single-standing or first genes in DOOR (1,847/2,377) (see *idrR* in Fig. S3B in the supplemental material). The ~22% of single or first genes of operons for which no pTSS was predicted by our data (530/2,377) (see *ybeT* in Fig. S3B) generally were missed due to low read coverage. For several of the genes without detected TSS, we found a processing site upstream, as indicated by an enrichment in the −100 nt region compared to the +100 nt region.

indicating that they could be cotranscribed with upstream genes (see *fbaA* in Fig. S3B).

Approximately 24% (594/2,441) of genes for which we detected a pTSS were not classified in DOOR as single or first genes in an operon. The majority of these TSS likely correspond to real promoters that are located internal to upstream genes within an operon defined by DOOR (see *thra* in Fig. S3B in the supplemental material). These TSS could drive transcription of unannotated alternative suboperons and thereby uncouple expression of the subset of genes from the longer operon. Some of these TSS are also found upstream of genes previously predicted to be in operons but are likely single genes (see *phCM* in Fig. S3B). Overall, these comparisons indicate that despite previous global transcriptome studies, the full complexity of the *E. coli* transcriptome is not yet known.

A comparison of our TSS predictions with TSS annotated in
RegulonDB (21), using a maximum distance of 3 nt, revealed that ~34% of our pTSS and sTSS overlap those annotated in RegulonDB (see Data set S1 in the supplemental material), while ~41% of the TSS from RegulonDB, classified as pTSS or sTSS, overlap our predictions. A TSS detected in our data but previously not annotated in RegulonDB is the pTSS for /espA (A) and /plsX (B) genes. The x axis depicts the genomic coordinates, while the y axis indicates the relative cDNA scores (normalized number of mapped cDNA reads). Red arrows indicate the previously unannotated TSS detected by our analysis. Promoter sequences for the new TSS, including the −10 and −35 sequences (boxed) and bases corresponding to TSS (red) are depicted below each plot.

**FIG 3** Examples of genes with newly detected pTSS. Screenshots showing the relative cDNA coverage plots for representative −TEX or +TEX libraries for the M63 0.4, LB 0.4, and 2.0 growth conditions across the genomic regions encompassing the /espA (A) and /plsX (B) genes. The x axis depicts the genomic coordinates, while the y axis indicates the relative cDNA scores (normalized number of mapped cDNA reads). Red arrows indicate the previously unannotated TSS detected by our analysis. Promoter sequences for the new TSS, including the −10 and −35 sequences (boxed) and bases corresponding to TSS (red) are depicted below each plot.

We identified 5,574 iTSS internal to annotated genes (Fig. 2B). It was recently reported that the majority of iTSS identified in the Gram-negative bacterium *Shewanella oneidensis* are present near the 5’ or 3’ ends of the genes (40). For a comparison, we examined the location of the 4,466 iTSS classified as iTSS only as one group and the 968 iTSS that are also annotated as pTSS or sTSS as a second group. Each annotated gene in which an iTSS was detected was divided into 10 equal sections, and the number of iTSS located in each section was counted for all genes. Those classified as iTSS only showed a broad distribution with similar numbers across the entire gene (see Fig. S4A in the supplemental material). In contrast, for the group of iTSS also classified as pTSS or sTSS, the majority (~86%) were located in the last 30% of the gene (see Fig. S4B). These 86% are likely TSS for downstream genes, driving alternative expression of suboperons (for an example, see *thrA* in Fig. S3B in the supplemental material) or the synthesis of small regulatory RNAs corresponding to the 3’ ends of mRNAs as was observed for the Micl RNA, whose promoter is within the *cutC* gene (41). Whether any of the iTSS in other categories result from spurious transcription or are generating functional alternative mRNAs or regulatory RNAs will require further characterization.

**pTSS and sTSS from divergently transcribed gene pairs could also serve as asRNA regulators.** In addition to the 5,495 TSS classified as asTSS only, we identified 350 pTSS and 386 sTSS that are also classified as asTSS. Examination of the regions encompassing these TSS revealed 212 divergently transcribed gene pairs with possible overlapping 5’ UTRs (see Data set S2A in the supplemental material), which could result in asRNA-mediated regulation of these genes (reviewed in reference 26) or could influence promoter occupancy (42). The set includes several gene pairs that encode proteins of opposing function, such as *entS* and *fepD*, encoding an enterobactin efflux system and a ferric enterobactin ABC transporter, respectively, and *pspF* and *pspA*, encoding the transcription factor *PspF* (phage shock protein F) and its antagonizing regulatory protein, *PspA* (see Fig. S5 in the supplemental material). Further characterization of these gene pairs will be required to determine if asRNA-mediated regulation occurs via the overlapping 5’ UTRs.

Some asTSS show high or differential levels of expression. Given that several asRNAs with characterized functions are expressed at high levels (reviewed in reference 43), we compared the relative expression levels for the 5,495 asTSS only (see Data set S3 in the supplemental material) to all pTSS only and sTSS only (see Data set S1 in the supplemental material) and TSS corresponding to known annotated asRNAs (see Table S6 in the supplemental material). We calculated reads per kilobase per million mapped reads (RPKM) values for all libraries utilizing a 50-nt window downstream of the predicted asTSS. The TSS were subsequently grouped into <10, 10 to 10^2, 10^2 to 10^3, 10^3 to 10^4, 10^4 to 10^5, and >10^5 bins according to their RPKM values. Using the highest expression value for a TSS among all conditions, we plotted the distribution of exclusively asTSS, exclusively pTSS or sTSS, and known annotated asRNAs for all conditions (Fig. 4A) and for each library individually (data not shown). For both the combined and individual sets, the distributions for pTSS or sTSS differ from the distributions for asTSS. Most pTSS or sTSS group within the 10^2 to 10^3 and 10^3 to 10^4 expression bins, while most asTSS are in the 10 to 10^2 and 10^2 to 10^3 expression bins (Fig. 4A). In contrast, most previously annotated asRNAs fall into the 10^3 to 10^4 and 10^4 to 10^5 bins. These distributions suggest that while a subset of asRNA candidates could be present at a high level, the majority of our predicted asRNA candidates might only exist as a few copies...
per cell or might be unstable transcripts that are rapidly degraded during RNA isolation and library preparation.

Since several functional asRNAs are expressed under specific conditions (44, 45), we also examined the distribution of the predicted asTSS across the different growth conditions (Fig. 4B). A total of 2,071 of the 5,495 asTSS were detected under all conditions. In general, candidate asRNAs in the >10^3 expression bin, showed a high signal for all growth conditions and library repli-

FIG 4 Comparison of asTSS. (A) Distribution of only asTSS, only pTSS or sTSS, and NCBI-annotated asRNAs in RPKM expression bins. The RPKM expression values were calculated based on cDNA read counts within 50-nt windows starting at the TSS. (B) Distribution of TSS classified exclusively as asTSS across the three biological conditions M63 0.4, LB 0.4, and LB 2.0. (C) Pairwise comparison of asTSS identified by our study and in previously published studies by Conway et al. (19), Dornenburg et al. (14), Raghavan et al. (16), Shinohara et al. (13), Mendoza-Vargas et al. (20), and Salgado et al. (21). The total numbers of annotated asTSS are shown on the main diagonal of the matrix. asTSS from the studies in the rows are compared to the studies in columns, and the number of TSS with exact matches is reported in the matrix entries. The background color depicts the percentage of overlapping asTSS relative to the total number of asTSS from the study in the particular row.
The majority of pTSS, iTSS, and asTSS are preceded by σ70 promoter elements. To detect potential differences between the promoters corresponding to the pTSS, iTSS, and asTSS, we compared the difference in expression for the pTSS only, iTSS only, and asTSS only detected in LB0.4 with those detected in M63 0.4 (see Fig. S6 in the supplemental material). Overall, there were proportionally more pTSS showing differential expression than iTSS and asTSS. This suggests that the pTSS generally are more highly regulated.

We also examined the sequences upstream of the 1,707 pTSS only, 4,466 iTSS only, and 5,495 asTSS using the MEME software (46). With a window of 0 to +1 relative to the TSS, the promoter motifs derived for the three classes of TSS overall were very similar (see Fig. S7 in the supplemental material). All had a potential −10 element resembling the TATAAT consensus for the housekeeping σ70 transcription factor (reviewed in reference 47). The enrichment for two T residues comprising a potential σ70 −35 element was significantly less than what was observed for the −10 element; however, both pTSS and iTSS logos showed some enrichment for a G at position −14, characteristic of an extended −10 sequence associated with σ70 promoters with weak −35 elements. A window of −50 to +5 relative to the TSS revealed that a subset of pTSS, iTSS, and asTSS show enrichment for a purine at +1 and a pyrimidine at −1, features of E. coli σ70 promoters reported previously (10). Overall, despite differences in the dRNA-seq signal, most of the pTSS, iTSS, and asTSS are likely transcribed by the σ70 holoenzyme.

Comparison of asTSS prediction with published data sets reveals limited overlap in candidate asRNAs. A number of transcriptome data sets have recently been published for E. coli with different extents of antisense transcription reported (13–16, 19–21). Given the discrepancy in numbers of annotated asRNAs, we were interested in the extent of overlap between our asRNA predictions and those of the other studies. For our cross-study comparison, we only included studies where detailed asRNA annotations were provided. We compared our asTSS only (see Data set S3 in the supplemental material) to the asRNA candidates reported by each group rather than to the primary data, given the differences in data generation, analysis, quality, and quantity of reads mapping to the E. coli genome (see Table S7 in the supplemental material). We first required the TSS positions between two studies to match precisely (Fig. 4C). This resulted in very limited overlap across the studies. The largest overlap occurred between our data set and that of Shinhara et al. (13), with 33% of their asRNAs overlapping our predictions. In some cases, increasing the window size within which an asTSS could match, to 1, 2, 3, or 10 nt, increased the overlap between studies (see Fig. S8 in the supplemental material). For example, with the 1-nt window, 79% (71/90) of the asRNAs detected by Raghavan et al. (16) corresponded to an asTSS in our data compared to ~12% (11/90) when an exact match was required. In other cases, the increase in window size did not make much difference. There was no overlap between the asRNAs predicted by Mendoza-Vargas et al. (20) compared to Raghavan et al. (16), Shinhara et al. (13), and Salgado et al. (21), regardless of the window size. The discrepancies between the asTSS reported likely result from combinations of differences in the quality of the sequencing reads, analysis pipelines, expression cutoffs, and definitions of what constitutes an asRNA.

We also compared our asTSS map to a recent study by Lybecker et al. examining the double-stranded transcriptome of E. coli (48). The premise of this study was that RNAs under asRNA-mediated control would be present in double-stranded RNA duplexes and thus should be identified by coimmunoprecipitation (co-IP) with a double-stranded RNA (dsRNA)-specific antibody followed by RNA-seq. We compared these reported IP dsRNAs to our asTSS set and considered them to match if an asTSS is found within the region 10 nucleotides upstream of an IP dsRNA 5′ end to 10 nucleotides upstream of the corresponding 3′ end on at least one strand (see Data set S2B in the supplemental material). We excluded the class of overlapping 3′ UTRs identified by Lybecker et al. from our analysis as they are not covered by our dRNA-seq, which sequences from the 5′ end of transcripts. This comparison yielded matching asTSS for 63% of the IP dsRNAs (193/308).

Candidate asRNAs are detected as distinct bands by Northern analysis. As independent verification of the predicted asRNAs, we selected a panel of 14 candidate asRNAs for Northern analysis (Fig. 5; see also Fig. S9 and Table S8 in the supplemental material). While we primarily selected candidates from the two highest-expression bins (see Data set S3 in the supplemental material), we also randomly selected a few candidates, which showed differences in expression among growth conditions or were not detected by others, from the third expression bin. We employed riboprobes covering the region of the dRNA-seq signal and importantly also probed total RNA from control strains where the region of mapped signal was deleted from the E. coli chromosome. In addition, we included total RNA isolated from strains defective for ribonucleases reported to be involved in asRNA processing and degradation; an rnc mutant lacking Rnase III, an endonuclease that cleaves double-stranded RNAs, and an rne-131 mutant with defective Rnase E, an essential endonuclease that associates with the RNA degradasome and cleaves single-stranded RNA. The C terminus of Rnase E is deleted in the rne-131 mutant, such that the enzyme can no longer associate with the degradasome, thus giving rise to reduced RNA turnover (49, 50).

We detected clear specific bands for RNA isolated from wild-type cells for six of the candidate asRNAs (as-gsiB, as-argR, as-ymfL, as-eutB, as-speA, and as-yliF) (Fig. 5; see also Fig. S9 in the supplemental material). Specific bands for five other candidates (as-gora, as-hoIE, as-serU, as-thwR, and as-ytfJ) were most evident in one or both of the Rnase mutant strains, while three candidates (as-yeal, as-gmr, and as-yggN) were only detected as smears. For 10 of the probes, we detected nonspecific bands present in all lanes serving as a loading control and emphasizing the importance of including samples from control deletion strains.

asRNAs show differential sensitivity to degradation by Rnase E and Rnase III. We were surprised to find that the Rnase mutants had varied impacts on the levels of our asRNA candidates. First, counter to expectations, the levels of some asRNAs, such as as-ymfL and as-speA were decreased in both Rnase mutant
strains (Fig. 5; see also Fig. S9 in the supplemental material). Possibly these asRNAs are destabilized by interactions with RNAs that are normally degraded by RNase III and RNase E, or alternatively, processing is required for stabilization of these transcripts (51).

The levels of three asRNAs (as-argR, as-qorA, and as-eutB) were elevated in the rne mutant, while the levels of four others (as-holE, as-serU, as-ytfJ, and as-thrW) were greatly elevated in the rnc mutant relative to the wild-type strain. Northern analysis carried out with RNA isolated from rnc mutants lacking the four chromosomal regions confirmed that the signal was specific (see Fig. S10 in the supplemental material; data not shown for as-thrW). Overall, these observations show that our detected asRNAs are substrates for different RNases and that regulation of asRNA levels by RNases may be more complex than previously thought.

DISCUSSION

In this study, we applied dRNA-seq and automated TSS prediction to the E. coli K-12 strain MG1655 grown under three different conditions to reveal >14,000 candidate TSS, of which >5,500 correspond to potential iTSS and >5,400 correspond to potential asRNAs. In contrast to previous E. coli transcriptome studies, dRNA-seq allowed us to globally map TSS since the approach specifically captures primary 5' ends and thus allows discrimination between processed and primary transcripts. Our global TSS
Automated TSS prediction has advantages and disadvantages. While the dRNA-seq analysis combined with automated TSS prediction used here provides a wealth of information, some reflection on the advantages and disadvantages is warranted. An automated approach for TSS annotation avoids potential bias introduced by manual annotation given that it follows defined rules and parameters. Automated annotation also facilitates rapid repetition of the analysis with different parameters or with additional data sets, a refinement that is impractical for manual annotation, especially for larger genomes or multiple strains or multiple conditions. However, choosing the right parameters for automated annotation, with the appropriate balance between sensitivity and specificity, can be difficult. Increasing the stringency for detection, for example, by filtering for those TSS whose step height is greater than 10 (see Data set S1 in the supplemental material) would reduce the number of TSS to ~4,400 (data not shown). Additionally, we analyzed the data for two of our LB 2.0 samples using another automated annotation program, TSSAR, with default parameters (52). This program predicted almost twice as many TSS as the TSSpredator program (data not shown) but is unable to integrate information from replicate samples. Therefore, for the TSS map presented in this study, we chose to use the parameters established on the basis of manual annotation of Helicobacter pylori dRNA-seq data and used for TSS annotation in Campylobacter jejuni (3, 30), which predicted TSS that were most consistent with manual annotation of selected regions of our E. coli data.

The automated TSSpredator program employed here led to the prediction of many more TSS candidates in E. coli than for other manually annotated data sets. To understand this difference between manual and automated annotations, we carried out automated TSS prediction using our E. coli parameters with the Salmonella dRNA-seq data sets from Kröger et al. (6). For the Salmonella dRNA-seq sets, we predicted ~22,000 potential candidate TSS, of which ~9,700 were found under all conditions (data not shown). These numbers are 4-fold higher than the TSS predicted by manual annotation. During manual TSS annotation, TSS corresponding to poorly expressed RNAs may not be annotated, resulting in underdetection of potential promoters transcribed at low levels. On the other hand, a higher discovery rate associated with automated TSS prediction may result in the false annotation of some promoters. It is also likely that our global map is still not saturated and that we have missed TSS that are not expressed under the limited growth conditions examined, as has been found for studies of Salmonella grown under a wide range of conditions (6).

Comparison of deep-sequencing data sets reveals sources of variation. When we compared our replicate deep-sequencing data sets, we found variation between different library preparations and sequencing platforms. The comparison of biological and technical replicates revealed that library preparation itself can lead to larger variation than found among biological replicates for which cDNA libraries were generated in parallel. ’t Hoen et al. similarly found that library preparation is a major source of variation for human samples (53).

Our comparisons of asRNAs predicted by different published RNA-seq data sets further highlighted discrepancies and led us to consider additional sources of variation. Differences in RNA isolation protocols might limit the ability to capture unstable transcripts or RNAs of certain sizes. For example, small RNA fractions are often lost in column-based purification methods, and rRNA depletion kits can lead to unintended removal of non-rRNA transcripts. The use of terminator exonuclease (TEX) treatment to enrich for primary transcripts may miss the TSS of RNAs that are monophosphorylated due to the pyrophosphate removal by the enzyme RppH (54). However, we identified TSS for the majority of validated RppH targets, suggesting this is not a significant limitation in our data set (data not shown). Other inherent properties of the RNA molecules also can be a source for bias as it has been reported that RNAs with high GC content are less readily amplified and that linker ligation is more efficient when certain nucleotides are at the 3′ and 5′ ends (55).

Additionally, differences in data analysis, including differences in read quality filtering, mapping protocols (using all or only uniquely mapped reads), and especially different methods and thresholds for assembling and annotating transcripts, can lead to significantly different results. Despite a rapid increase in data generation, the availability of standardized RNA-seq analysis pipelines is still limited (56), particularly for bacterial transcriptomes. Nevertheless, RNA-seq has been an invaluable resource and has revolutionized bacterial, archaeal, and eukaryotic transcriptome analyses. Hopefully, as the field of deep sequencing continues to mature, standards for sample preparation, depth of sequencing, number of replicates sequenced, and data analysis as well as simple platforms for shared data visualization can be developed that will facilitate the comparisons of data generated by different groups.

Independent documentation of asRNAs is advised before functional analysis. Our dRNA-seq approach revealed more than 5,400 asTSS. We do not know how many of these predicted asTSS correspond to spurious transcripts rather than functional RNAs, although some show differential expression under the growth conditions examined (see Data set S3 in the supplemental material). The above-mentioned RNA-seq study of Salmonella, which analyzed RNAs from 22 different growth conditions, reported <500 asRNAs (5, 6). These authors found that ~1.75% of their reads mapped antisense to annotations (5) which is similar to what we observed (2 to 4%) (see Table S5 in the supplemental material) and to what has been reported for another E. coli RNA-seq study (~2%) (16). Thus, the high number of asTSS we detect probably is not due to large differences in general transcriptome coverage but rather is due to differences in data analyses and annotation. Moreover, we specifically enriched for the 5′ ends of transcripts, which might be more stable than internal degradation fragments, and did not include fragmentation steps that could result in the lower numbers of sequenced 5′ ends of transcripts.

As our comparison among different E. coli studies showed, there is extensive variation in asRNA annotation. Nevertheless, we found that several asRNAs were detected in multiple RNA-seq studies (Fig. 4C; see also Fig. S8 and Data set S3 in the supplemental material). Given the laborious process of functional investigation, however, we propose that further validation of asRNAs with appropriate controls is critical for defining candidates for further study. We independently validated expression of 14 candidate as-
ACKNOWLEDGMENTS

We thank R. Reinhardt (Max Planck Genome Center, Cologne, Germany) for help with deep sequencing and vertis Biotechnologie AG for cDNA library preparation, E. P. Greenberg for work conducted by M.K.T. in his lab, and T. Conway for sharing data prior to publication. We also thank J. Hinton for discussions and J. Chen, S. Gottesman, and participants in the EMBO Lecture Course on the Biology of Bacterial Noncoding RNAs for comments on the manuscript.

Work in the Sharma laboratory is supported by the ZINF Young Investigator program of the Research Center for Infectious Diseases (ZINF) at the University of Würzburg, the Bavarian BiosysNet Program, and the Deutsche Forschungsgemeinschaft Project Sh580/1-1. Work in the Storz laboratory is supported by the Intramural Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development. For study stays in the Storz laboratory, S.K.E. was supported by a PROMOS travel scholarship (DAAD) of the University of Würzburg and C.M.S. was supported by a Boehringer Ingelheim Fonds travel allowance.

REFERENCES


