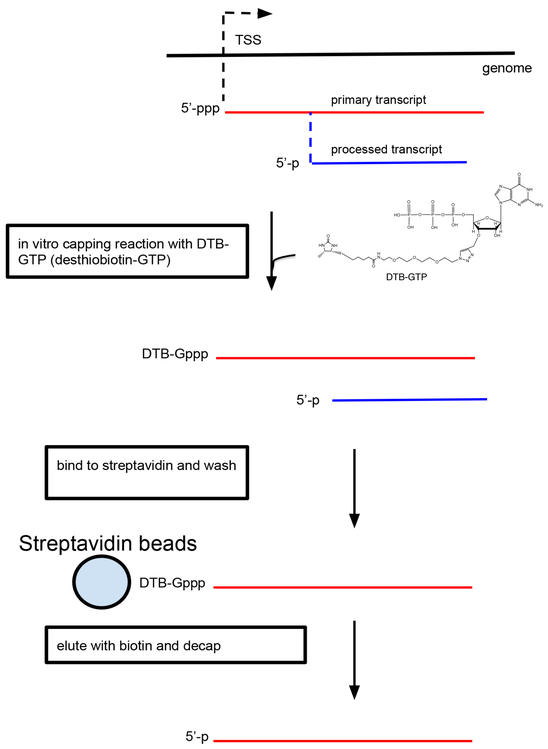
Capturing primary RNA transcripts, a novel strategy for analyzing transcriptomes

Laurence Ettwiller, John Buswell, Erbay Yigit, and Ira Schildkraut
New England Biolabs, Ipswich, MA

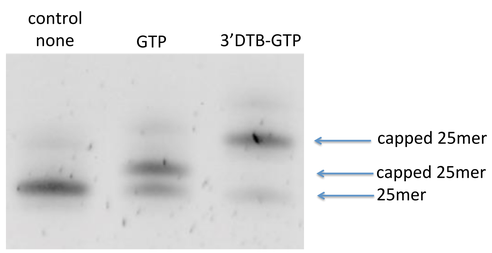
INTRODUCTION

The initiating nucleotide that is found at the 5' end of primary transcripts has a distinctive triphosphorylated end which distinguishes these transcripts from all other RNA species. Here we developed a novel method which specifically captures, for the first time, primary RNA transcripts by enzymatically modifying the 5' RNA triphosphate with a selectable tag. Rather than removing processed and degraded RNA, primary transcripts are directly and specifically modified at their 5' end. To achieve this modification, the RNA is enzymatically capped with a modified guanosine triphosphate that contains a 3' desthiobiotin moiety. The reaction requires the 5' end of the RNA to be triphosphate (or diphosphate), capping only primary transcripts. Desthiobiotinylated RNA is selectively bound to streptavidin and the remaining RNA is washed away. The bound fraction is eluted and decapped leaving the RNA with ligatable 5' monophosphate ends. Coupled with direct ligation-based library preparation and sequencing, this technique directly defines TSS of prokaryotes at single base resolution genome-wide. When applied to *E. coli* an unprecedented number of putative TSS are obtained. In addition this method is species agnostic and avoids uninformative sequence such as rRNA. Furthermore Cappable-Seq when applied to eukaryotes determines the TSS of RNA polymerase I and III primary transcripts.

Cappable-Seq

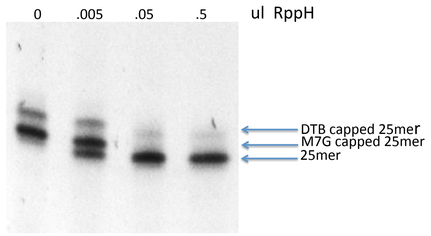


Capping with DTB-GTP



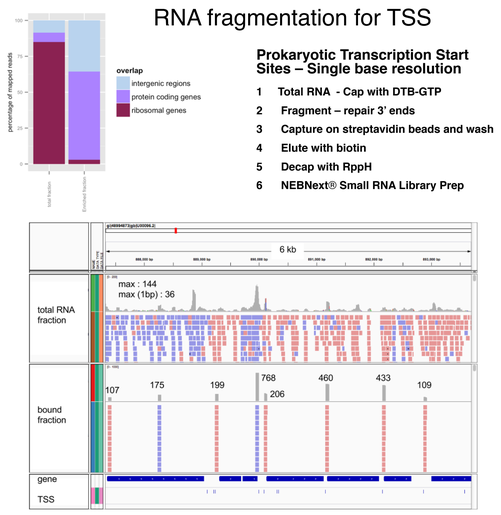
A 15% TBE Urea polyacrylamide gel of a 25mer T7 triphosphorylated transcript incubated with VCE in the absence of nucleotide (none) or the presence of 0.5 mM GTP or 0.5 mM 3'DTB-GTP.

Decapping with RppH



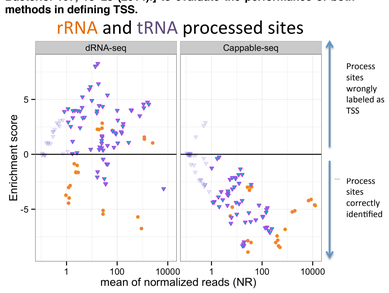
A mixture of the 25-mer transcript capped by VCE with 3' DTB-GTP or GTP. The transcripts were incubated with varying amounts of RppH for 30 minutes at 37 degrees in 1X Thermopol Buffer.

Transcription Start Sites – *E. coli*

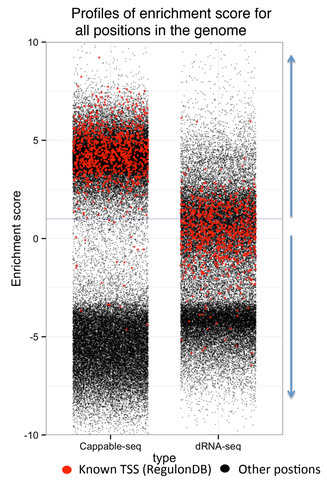


Total TSS	Intragenic TSS	Sense TSS	Anti-sense TSS	Ribo and tRNA TSS
16974	10816	7816	3000	34

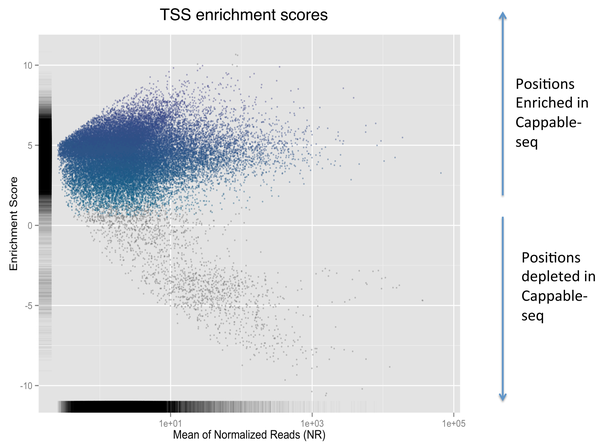
Below we compare Cappable-seq to published dRNA seq dataset of *E .coli* MG1655 grown in minimal media [Thomason, M. K. et al., Bacteriol 197, 18–28 (2014).] to evaluate the performance of both methods in defining TSS.



For dRNA-seq : Only a small subset of the annotated processed 5' ends are depleted after Xrn1 treatment.
For Cappable-seq : All of the rRNAs and all but one of the 82 tRNAs processed sites are strongly depleted with Cappable-Seq.
More process sites are correctly identified in Cappable-seq than in dRNA-seq

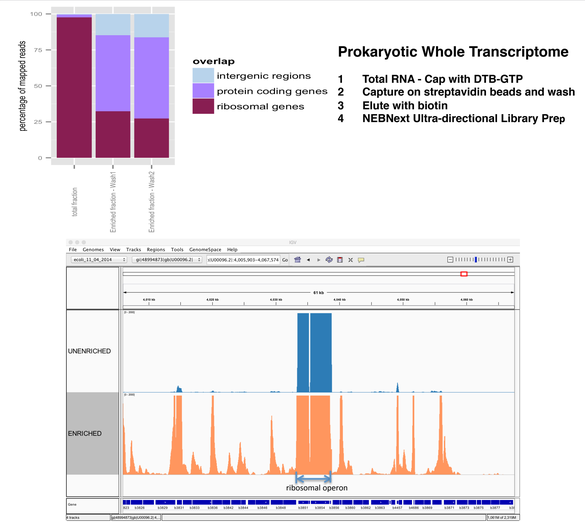


Enrichment score for all positions in the genome passing read threshold in either the assay or control library for both cappable-seq and dRNA-Seq experiments. Negative scores are depleted regions and positive scores are enriched regions in CAPPABLE-seq or dRNA-Seq compared to control. Red points are known TSS from Regulon DB.



Enrichment score function of the mean of relative read score for the 36077 putative TSSs found in *e.coli* grown on minimal media. In blue are TSS that are enriched in cappable-seq library compared to the unenriched control library. Grey are TSSs that are depleted in Cappable-seq and are discarded. Filtering of the TSS is based on the enrichment of Cappable-seq compare to the control library. This filtering removes 1738 spurious TSS mainly located in ribosomal loci.

Whole Transcriptome - *E. coli*



RNA-Seq data plotted on IGV browser where a 60 kb region of E coli chromosome is displayed. **Unenriched** refers to mapped reads from a library that has undergone DTB capping and decapping without the streptavidin enrichment step. **Enriched** refers to a Cappable-Seq library that has undergone streptavidin enrichment. As can be seen a significantly higher number informative non-ribosomal reads are obtained in the enriched library.

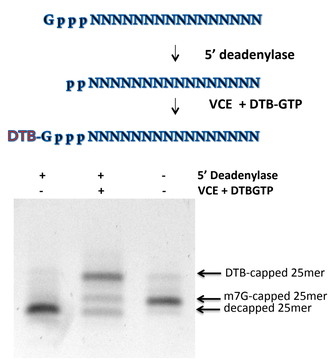
Applicable to Eukaryotes

RNA polymerase I + III transcripts are 5' triphosphate
RNA polymerase II transcripts are m7G capped

Cappable-Seq does enrich human pol I and III transcripts

ReCappable-Seq for m7G-capped mRNA - pol II transcripts

- ✓ m7G-capped mRNA can be decapped with yeast hnt3 (5'deadenylase) which leaves a 5' diphosphate RNA
- ✓ 5' diphosphate RNA is recapped with DTB-GTP



Decapping and Recapping: A 15% TBE Urea polyacrylamide gel of a 25mer T7 7mG capped transcript first incubated with 5' deadenylase and subsequently incubated with VCE and DTB-GTP.

CONCLUSION

Cappable-Seq is a novel method which enables direct modification and identification of the triphosphorylated RNA characteristic of primary transcripts. Cappable-seq determines TSS at one base resolution genome-wide in *E. coli* by pairing Cappable-Seq with direct 5' ligation of sequencing adaptors to the RNA. Despite being a very different approach, the results are consistent with the established methodologies for determining TSS in prokaryotes. Indeed, a large fraction (59 %) of the TSS found in *E. coli* by Cappable-Seq are coincident with known TSS, others are novel. Cappable-Seq discriminates the 5' triphosphate end characteristic of initiating 5' triphosphorylated nucleotide incorporated by the RNA polymerases from the processed 5' monophosphate RNAs.

As triphosphorylated transcription start sites are believed to be universal in prokaryotes, Cappable-Seq should perform well when applied to that large group of organisms and provide for the first time a solution for TSS determination in complex microbiome population. TSS derived from a microbiome can represent a unique signature derived from the 5' end of transcripts. Thus, Cappable-Seq can be used to derive sets of quantitative markers from which association to diseases or other states can be made. By providing a signature profile of the microbiome functional state, this technology should greatly facilitate metagenome-wide association studies.

Cappable-Seq can also be used for whole transcriptome analysis by avoiding the fragmentation and decapping steps. Unfragmented primary RNA can be enriched, effectively removing rRNA. Such depletion is ideally suited for microbiome studies as it should universally remove rRNA and most contaminating eukaryotic host RNA leaving prokaryotic transcripts intact.

Applying Cappable-Seq to eukaryotic RNA would reveal the triphosphorylated transcriptome such as the Pol I and Pol III transcripts and identify the TSS of such transcripts. While eukaryotic pol II mRNA differs from prokaryotic mRNA by virtue of the G cap, ReCappable-Seq could be utilized on the identification of the TSS of eukaryotic mRNA by first treating the RNA with a decapping enzyme such as 5'deadenylase that only removes the m7G monophosphate and leaves a 5'-diphosphate at the terminus of the mRNA. This can then be recapped with desthiobiotin GTP and then be treated in an analagous way as prokarytotic RNA to reveal the TSS of eukaryotic mRNAs.

- Applicable to Prokaryotes and Eukaryotes
- Highly Specific for TSS
- Single base resolution TSS
- Depletes ribosomal RNA