# Global Analysis of Transcription Start Sites (TSSs) and Transcription Units (TUs) in Bacterial Genomes

Leonardo Collado-Torres[1], Alejandro Reyes[2], Gabriel Cuéllar[2], Víctor Moreno[2], Blanca Taboada[3], Leticia Vega[3], Verónica Jiménez[4], Alfredo Mendoza[2], Ricardo Grande[4], Leti Olvera[2], Maricela Olvera[2], Carlos Vargas[2], Katy Júarez[2], Julio Collado-Vides[5], Enrique Morett[2]

[1]IBT UNAM & Winter Genomics, [2]IBT UNAM, [3]CCADET UNAM, [4]UUMSD, UNAM, [5]CCG UNAM

## Introduction

With the surge of high throughput sequencing it is now possible to identify at the genomic scale the transcription start sites and transcription units in a bacterial genome. These analyses pose several challenges which involve filtering the data, visualizing global patterns and determining the high confidence results. To do so we have been developing R scripts (with plans to build a package) that will help us interpret the data.

## Objectives

1. Maximize the number of reads used while avoiding erroneous alignments.
2. Make several lanes comparable by removing background transcription noise.
3. Develop a visualization method that will enable us to identify global patterns for the transcription start sites (TSSs).
4. Develop a method that identifies the transcription units (TUs) using paired-end reads.
5. Combine the methods into an R package to guarantee the reproducibility of the work.

## Materials & Methods

TSSs Experimental methods

Using an Illumina Genome Analyzer IIx, we sequenced the 5' ends (36bp reads) of all transcripts from *E. coli* and *Geobacter sulfurreducens* using four experimental methods. All lanes were treated with DNase I, ribosomal RNA was removed with the RiboMinus Transcriptome Isolation Kit (Invitrogen, Carlsbad, CA), and after applying the method specific steps (see below) an adapter was ligated to the 5' end; all this guaranteed that the reads will not come from degradation products.

1. No special treament applied. (Tri+MonoPO4)
2. Enrichment for 5' monophosphate transcripts. (MonoPO4)
3. Degradation of all 5' monophosphate transcripts using a specific exonuclease. (TriPO4-Exo)
4. Ligation of an adapter only to the 5' triphosphate transcripts, which filters out monophosphate ends. (TriPO4-Ad)

TSSs Bioinformatics methods

Reads were mapped to the genome using Bowtie -v mode allowing maximum 3 mismatches. If a 36bp read did not align, we trimmed 1 base from the 3' end and re-used the same mapping parameters. We ended this iterative alignment if the reads were too short (15bp) or if there were less than 250 reads left to align. We used the start position of the read as the TSS position.

Using ShortRead and GenomicRanges, we loaded the data in R and filtered out reads mapping to ribosomal genes. After calculating the coverage per position, we assumed that positions with a frequency of 1 or 2 were the null distribution for a Poisson and used it to determine the lane-specific thresholds. We then standardized the frequencies using Z-scores.

In order to generate global perspectives of TSSs, we partitioned the genome into several regions and generated accumulated graphs of TSSs within them (see the TSSgram plots). The defined regions are: A) upstream of TUs, B) upstream within TUs, C) convergent and D) divergent non-coding regions, E) coding and F) antisense. We used a window size of 1bp and changed the frequencies into 0s or 1s before accumulating them. For regions of heterogeneous widths, we centered them to avoid misinterpretations.

TUs Experimental methods

Paired-end 50+48 bp (~90bp insert size) reads were sequenced using the Illumina GAIIx for *E. coli* grown in either minimal medium (MM) or LB. Samples were treated with DNase I and ribosomal RNA was removed with the kit specified above.

TUs Bioinformatics methods

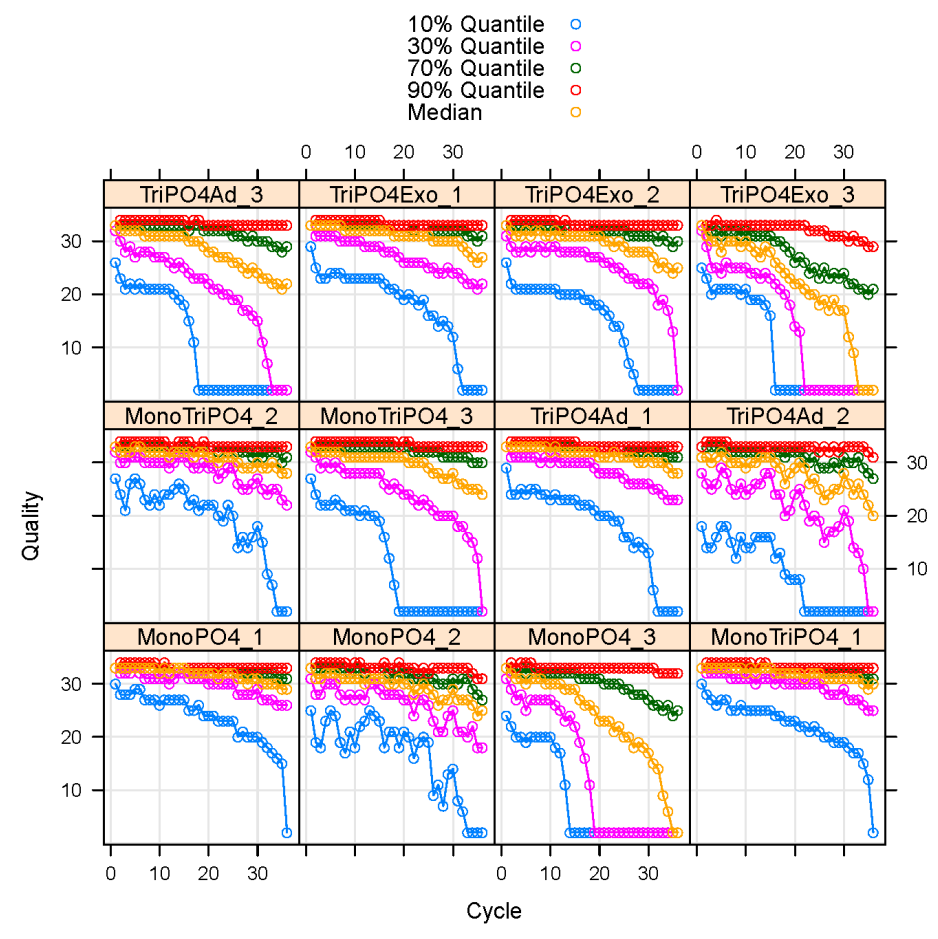
Reads were aligned to the genome using Bowtie -v mode allowing maximum 3 mismatches. All pairs of reads were joined and treated as a single longer one for the rest of the analysis.

*Gap method.* Multiple promoters and terminators can produce multiple TUs from a single operon. If two TUs overlap for a length smaller than the insert size, there will be a gap if you only take into account read start positions. Using the distribution of such gaps inside genes, at the start or end it should be possible to determine if any given pair of genes is part of the same TU. The challenge is that these start gaps can also be explained by low read coverage at a given region.

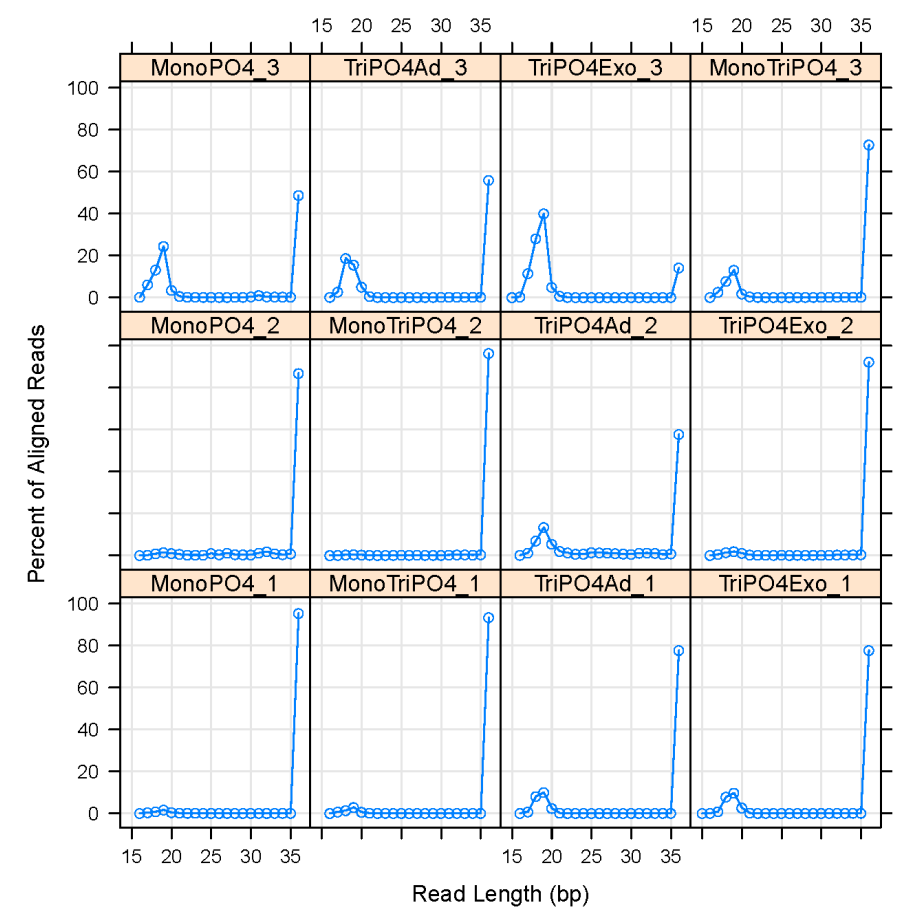
*Head minus tail method.* For every position of the genome we counted the number of reads starting (head) and subtracted the number of reads ending (tail). In theory, all TUs should be identifiable as a region limited by a strong positive peak (more heads) followed by a negative peak (more tails).

*Smooth moving average & differentiation method.* The coverage per position was smoothed by averaging 31bp windows. Regions are then delimited by coverage gaps and the coverage is transformed into relative frequency. Then, for each position we differentiated the values using the information from the 5 prior bases. In the final curve, if the region is composed by two TUs there are peaks partitioning the region.

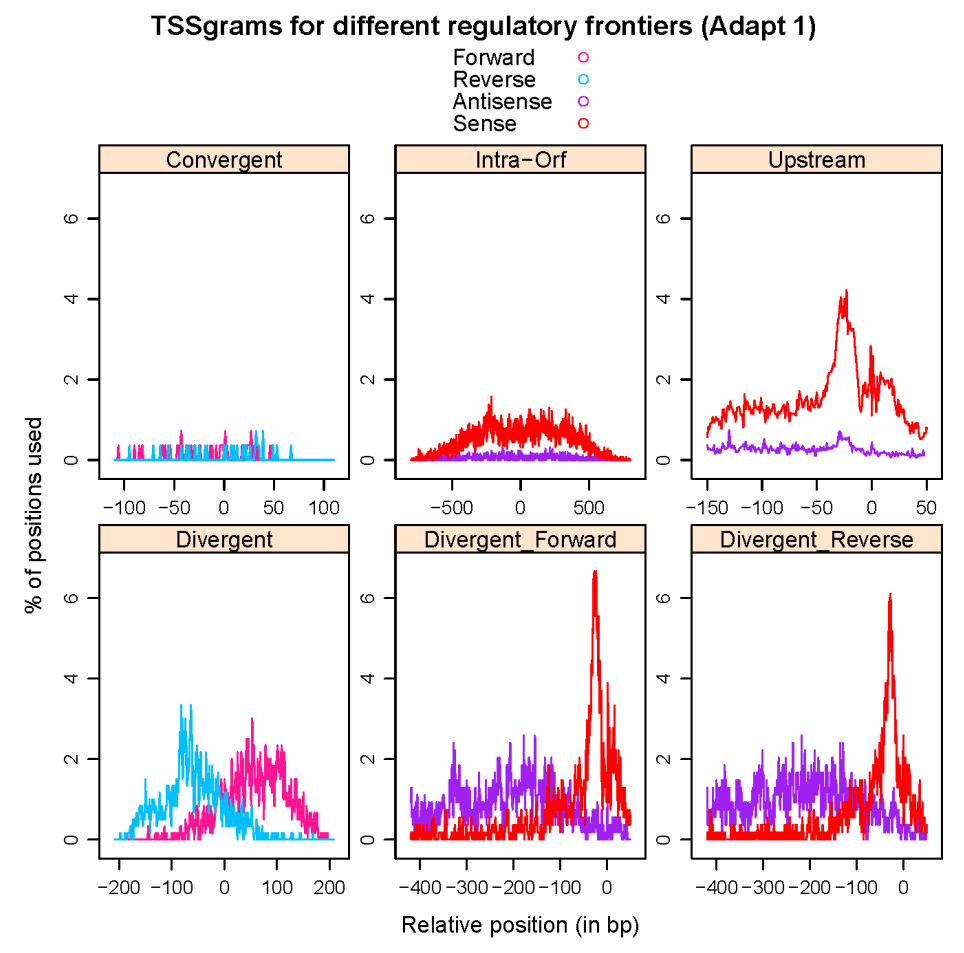
## Results



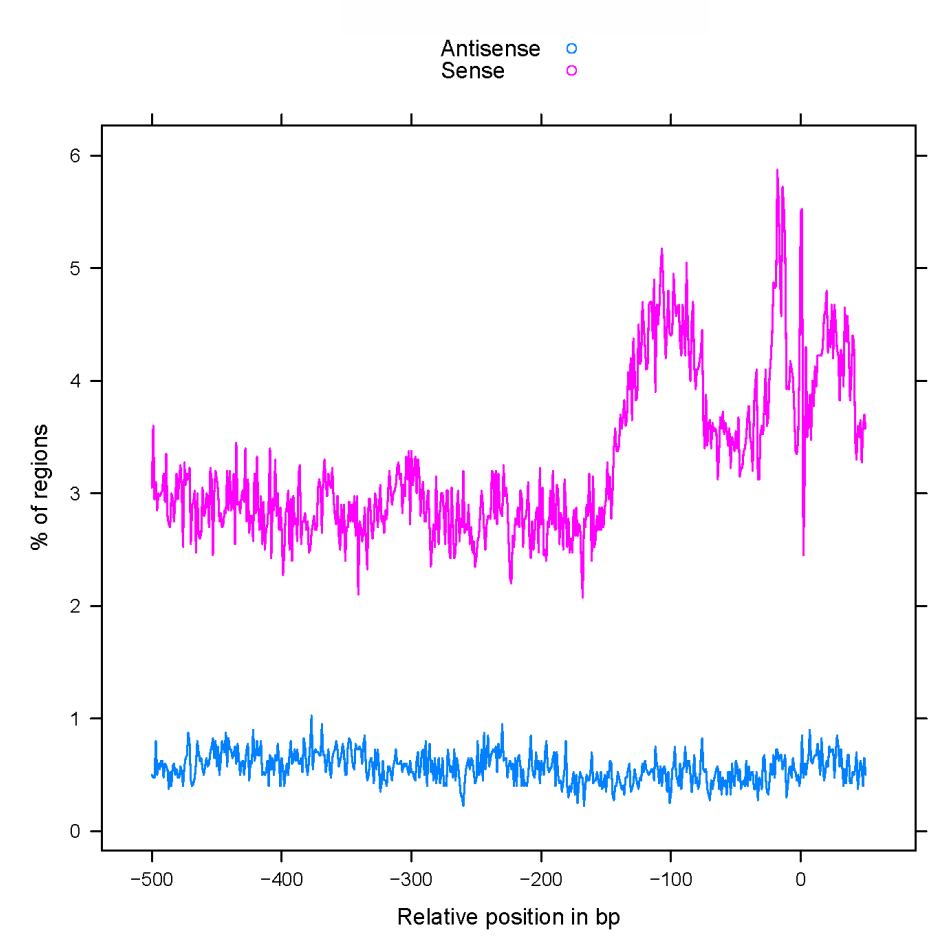
Quality per cycle for the *E. coli* TSS data



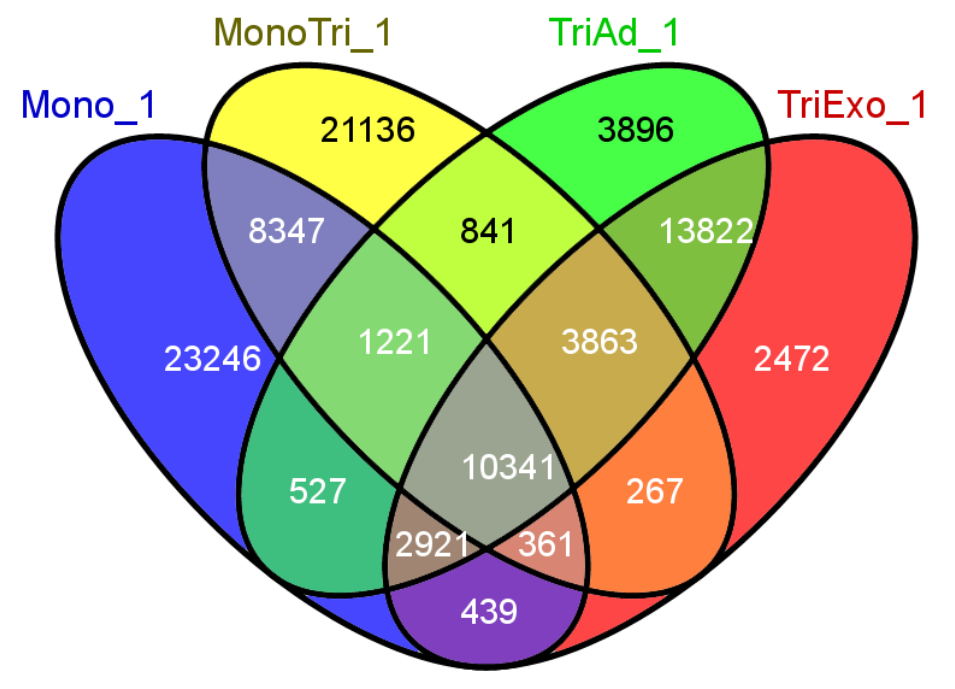
Iterative alignment results for *E. coli*



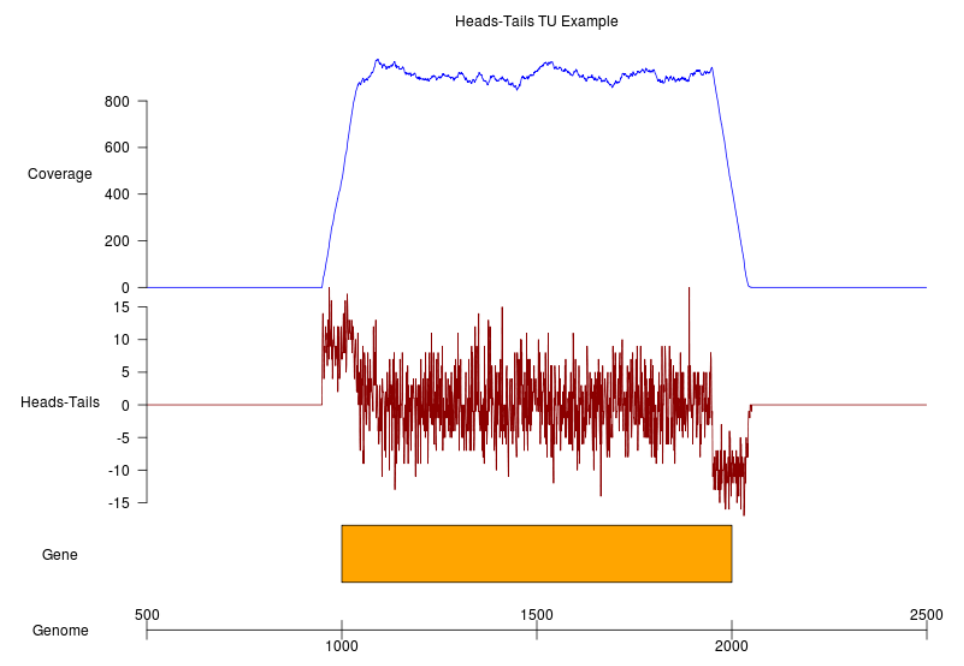
TSSgram plots for the *E. coli* lane TriPO4-Ad



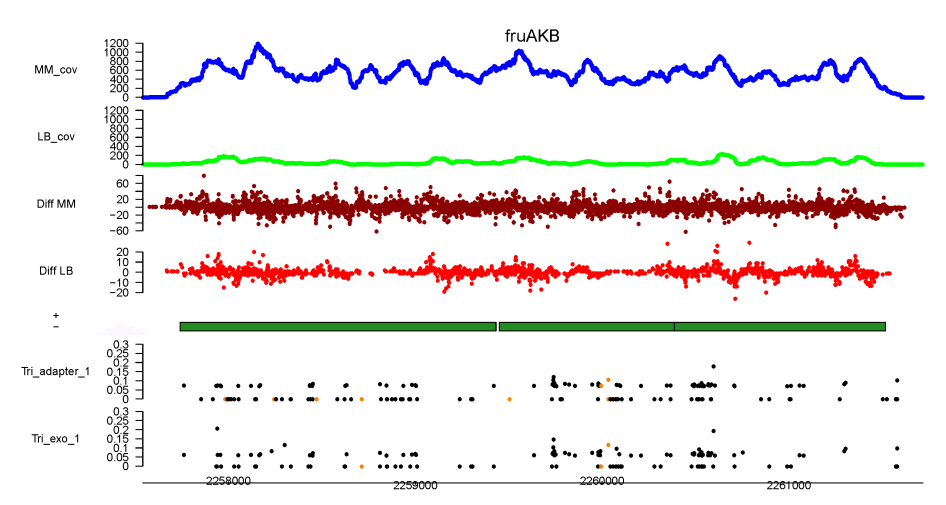
*G. sulf* TSS data for the TriPO4-Exo method



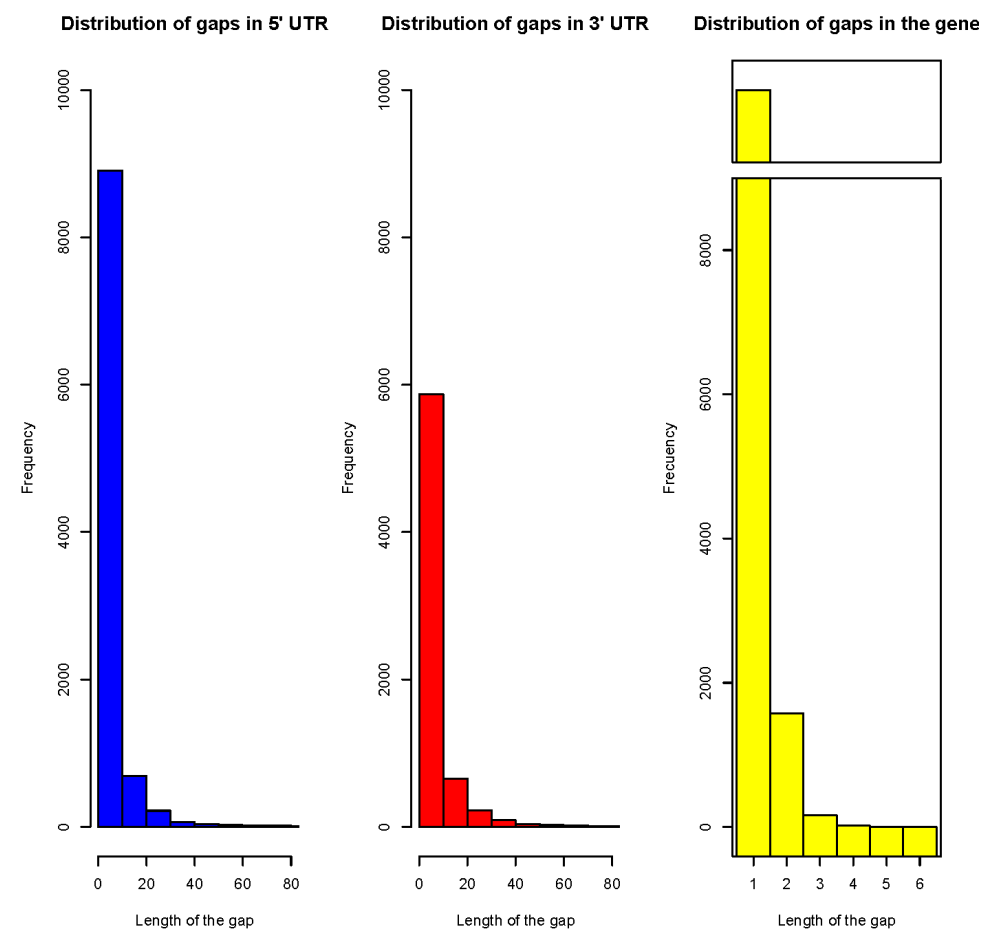
Positions overlap for *E. coli* TSS data



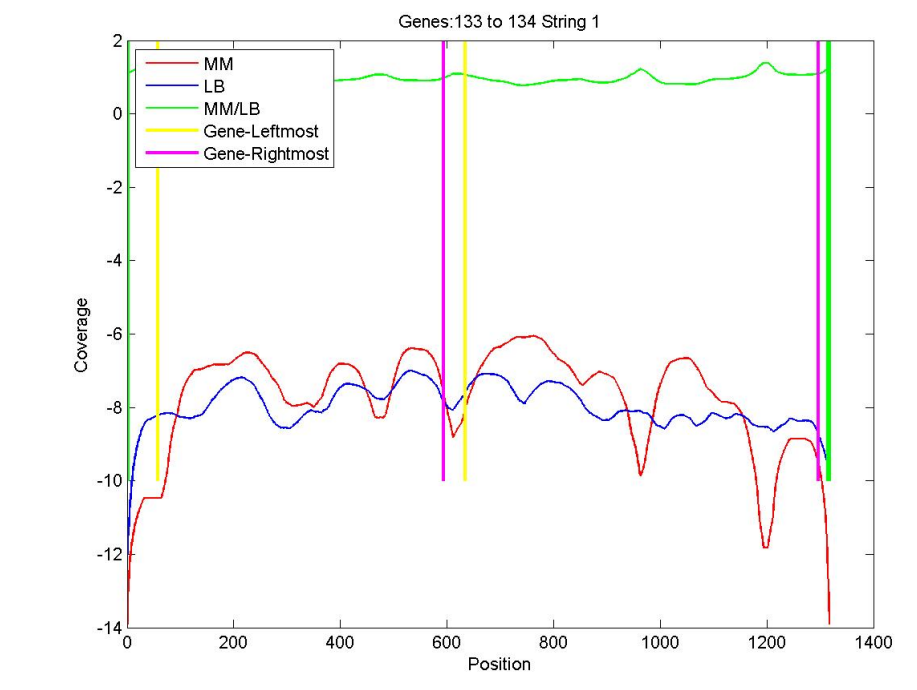
Head minus tail method theoretical example



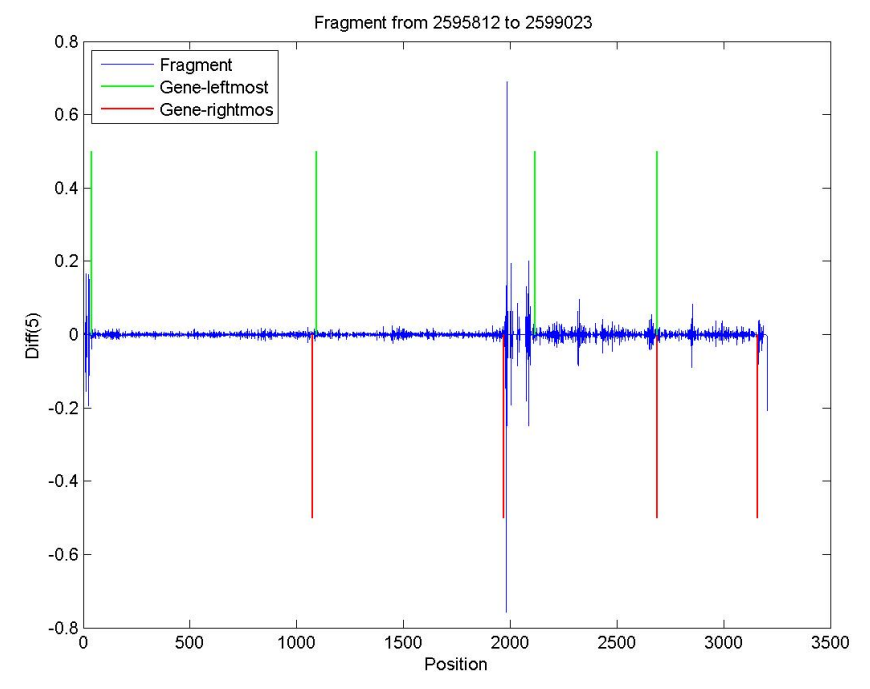
Visualizing TSSs and TUs data using GenomeGraphs



Distribution of start position gaps



TU smoothed coverage for a given region



TU differentiated curve for a given region

## Conclusions

**Transcription Start Sites (TSSs)**

- We were able to align nearly all of our reads with 2 peaks: the full length reads and those whose last 15bp were of low quality.
- The TSSgram provides a useful global view of TSSs at the genome level. Both in *E. coli* and *G. sulf*, TSSs were located within upstream regions of genes and TUs, whereas convergent regions have nearly no TSSs as expected. Basal levels of anti-sense transcription are present with a slight increase at the divergent regions. Interestingly, there is a second TSS peak barely inside genes on both organisms.
- For most TUs TSSs can be clustered, therefore the RNA polymerase is not as precise as expected.
- When comparing initiation sites in *E. coli* vs those reported in RegulonDB, there is no enrichment for known TSSs in the triphosphate vs. the monophosphate enriched lanes, thus the monophosphate method is not clearly enriched for degradation ends. An adequate interpretation requires further analysis of these datasets.

**Transcription Units (TUs)**

- Definitely, our methods work in regions when the sequencing coverage is appropriate. Low coverage regions produce false positive results with all methods.
- For our two growth conditions for *E. coli*, the coverage curve is non-uniform for any given gene which we did not expect. A key difference could be that in bacterial genomes it is not possible to retrieve mRNA using a poly A adapter. For both conditions, the curve has a similar form though they intersect each other several times for any given gene or transcription unit. While we cannot explain this phenomenon precisely, with the smooth moving average & differentiation method we can more easily identify TUs.

## References

1. Mendoza-Vargas, A. et al. Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in E. coli. PLoS ONE 4, e7526 (2009).
2. Work to be published *soon*.

## Acknowledgments

We would like to thank all the members from the Katy Juárez, Enrique Morett and Julio Collado-Vides groups involved in this project which we dub *Magallanes*. This project has been realized thanks to the enthusiasm, ideas and work done by Alejandro, Gabriel and Víctor who will soon finish their undergrad studies.

We acknowledge support from NIGMS-NH grant R01 GM071962-05 and from CONACYT México.