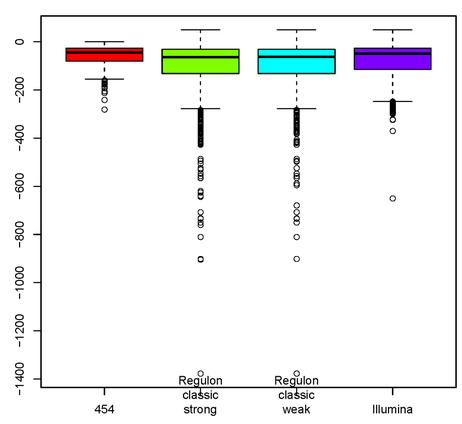
Global Analysis of Transcription Start Sites and Transcription Units in Bacterial Genomes

Collado-Torres L.[1, 5], Reyes-Quiroz A.[2], Cuellar-Partida G.[2], Moreno-Mayar V.[2], Vargas-Chávez C.[1, 5], Taboada B.[3], Vega-Alvarado L.[3], Jímenez-Jacinto V.[4], Mendoza-Vargas A.[5], Grande R.[4], Olvera L.[5], Olvera M.[5], Júarez K.[5], Collado-Vides J.[6], Morett E[5].

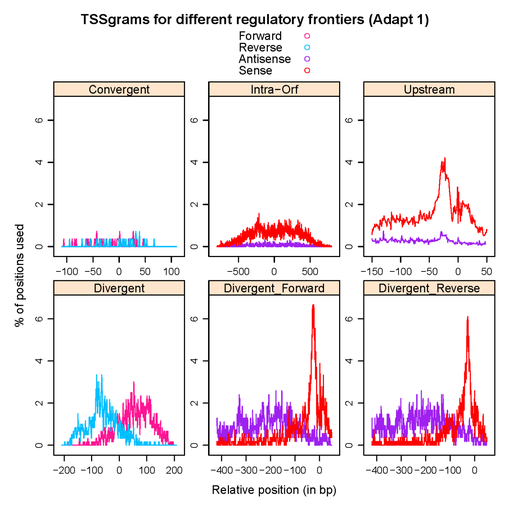
[1]Winter Genomics, [2]LCG-UNAM, [3]CCADET-UNAM, [4]UUSMD-UNAM, [5]IBT-UNAM, [6]CCG-UNAM

Summary

With high throughput sequencing it is possible to identify at the genomic scale the transcription start sites (**TSSs**) and transcription units (**TUs**) in bacterial genomes. Due to the biological and data complexity, these analyses are challenging and require the development of custom algorithms. Critical steps involve 1) maximizing the number of reads that can be used without introducing false alignments, 2) removing biological noise: random transcription and degradation products, 3) identifying the TSSs, 4) visualizing global TSSs patterns, and 5) identifying TUs. Condensing the analyses tools into a Bioconductor package will guarantee the reproducibility of the work. The methods have been developed with data from *Escherichia coli* and *Geobacter sulfurreducens*.

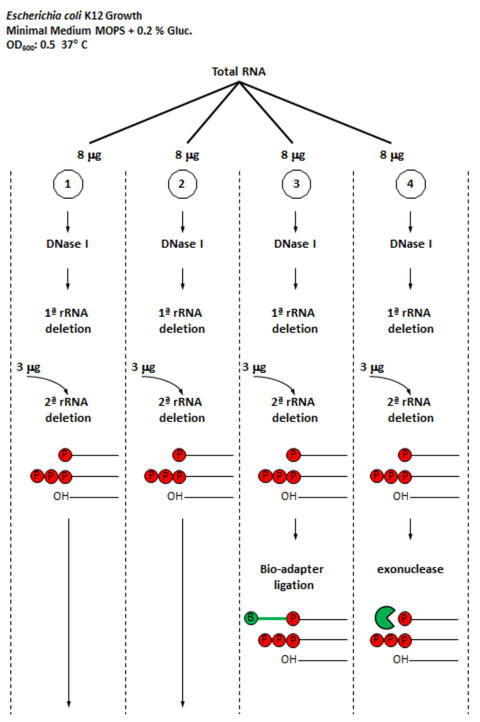


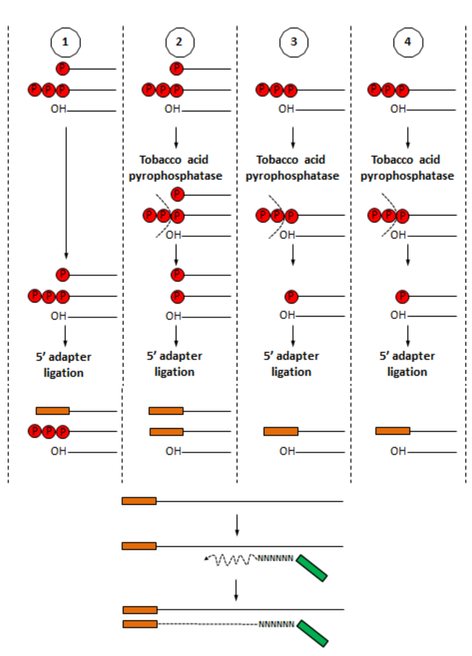
Distribution of the distance to gene start



TSSgram sample

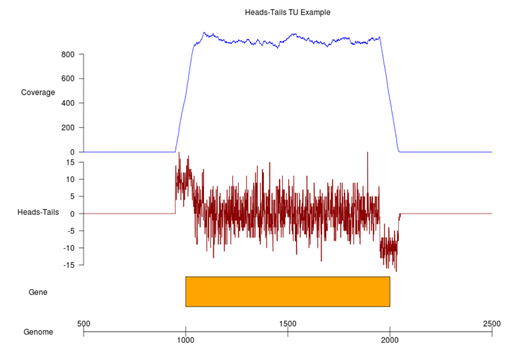
Methods



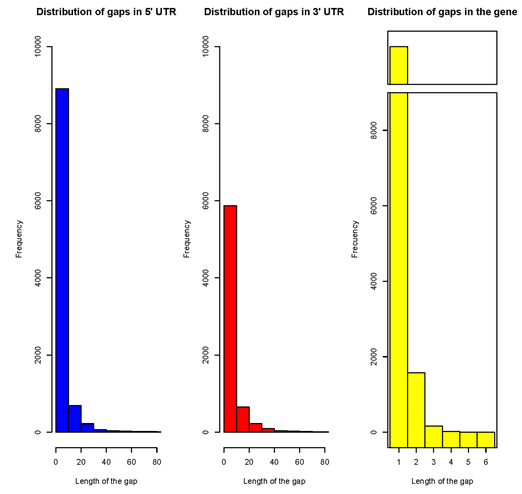


TSSs experimental methods

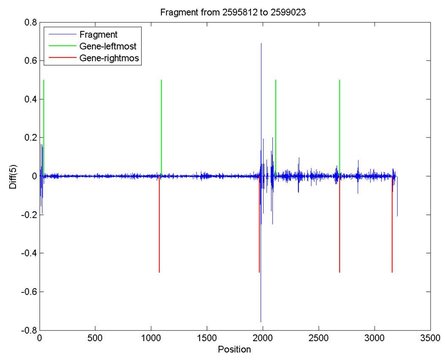
Transcription Units



Heads minus tail method

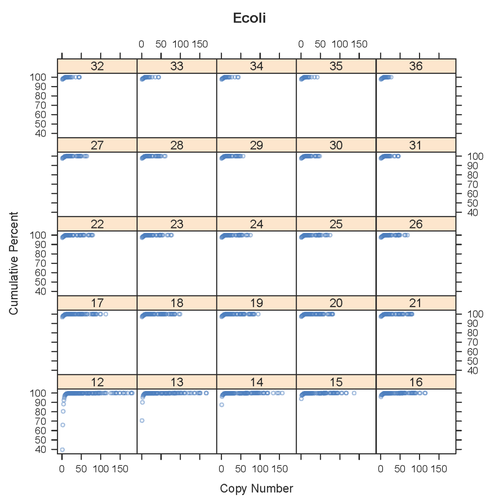


Left position gaps method

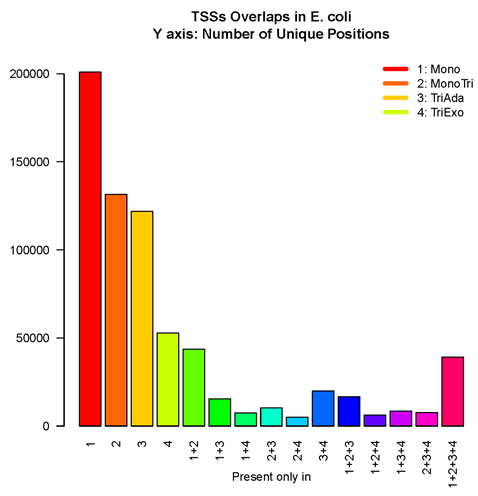


Differentiation method

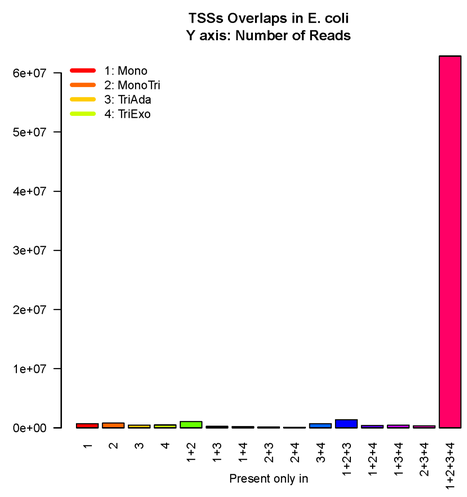
Transcription Start Sites



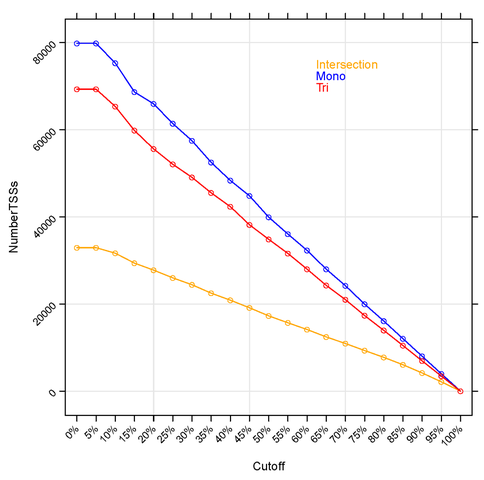
Choosing the minimum read length



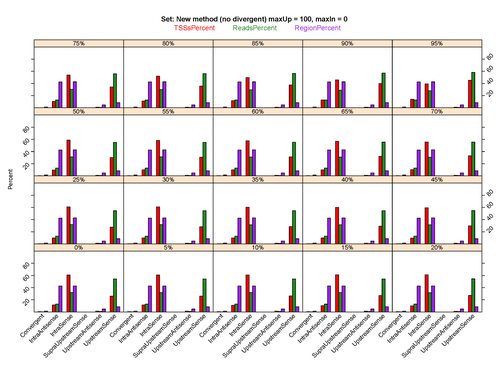
Number of putative TSSs positions



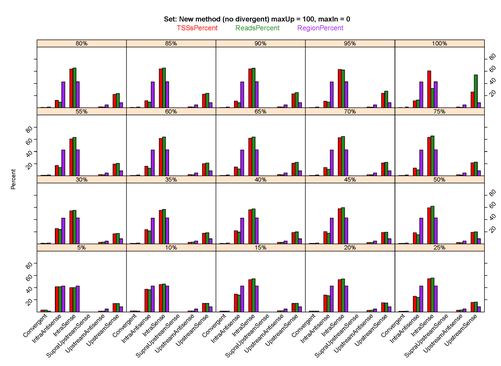
Read coverage (frequency) of the putative TSSs



Number of TSSs related to the stringency



TSSs in genomic regions be removing low frequency TSSs



TSSs in genomic regions be removing high frequency TSSs

Conclusions

These preliminary analyses will lead to the improvement of the accuracy of promoter prediction, operon structure and regulatory networks and support a new understanding from a genome perspective of the complex regulatory network that governs transcription and regulation in bacterial genomes such as *E. coli* and *G. sulfurreducens*.

References

Mendoza-Vargas, A. *et al.* Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in *E. coli*. PLoS ONE (2009).

Gama-Castro, S. *et al.* RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). Nucleic Acids Res (2010).

Collado-Torres, L. *et al.* manuscript *in preparation*.

BacterialTranscription Bioconductor package *in preparation*. More on the EMBL 2010 Bioconductor Developer Meeting.

We acknowledge support from NIGMS-NH grant R01 GM071962-05 and from CONACYT México (83686 G.I.).