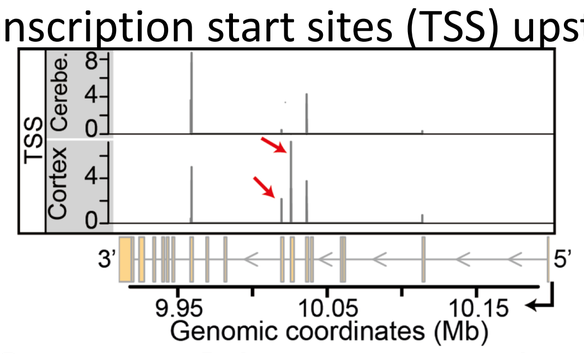
Alternative Promoter Usage using Transcription Start Sites

Mojca Stampar, Computer Science (Biomedical Informatics) (MS)
Mentor: Heewook Lee, Assistant Professor
Ira A. Fulton Schools of Engineering

Introduction

- With relatively small genomes, species have evolved mechanisms for diversifying their transcriptome
- One of less explored mechanisms is alternative promoter usage , which generates different transcripts by selecting different transcription start sites (TSS) upstream of a gene



- Different TSSs usage of the same gene lead to different transcripts
- Alternative promoter usage has the potential to regulate processes like alternative splicing, tissue specificity, regional specificity and subcellular specificity of gene expression and gene activation during development
- Recent approach to studying the alternative promoter usage is to cluster multiple TSSs into a region with the same biological function defined as transcription start region (TSR) to get coarser landscape of TSS

Objective

To study and improve methods for analyzing differential TSS usage, specifically by studying normalization and TSS clustering methods available to current users.

Methodology

Types of assays to probe TSS:

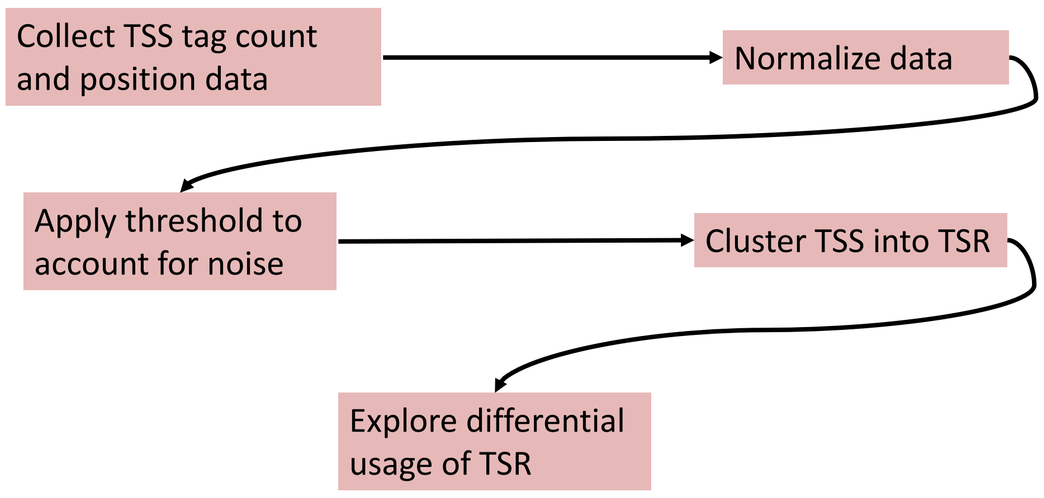
- CAGE
- RAMPAGE
- STRIPE-seq

Normalization methods used for accounting for different sizes of libraries:

- MOR (median of ratios)
- TMM (trimmed median of M values)
- TPM (tags per million)
- Power normalization

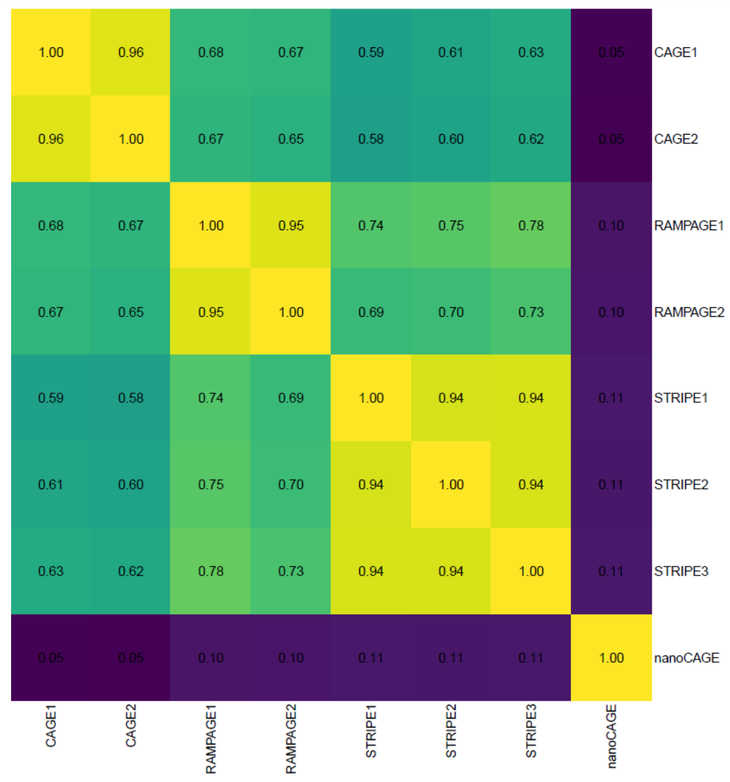
Clustering methods used for grouping TSS into clusters representing TSR:

- Distance based clustering (greedy algorithm)
- Density based parametric clustering (paraclu)

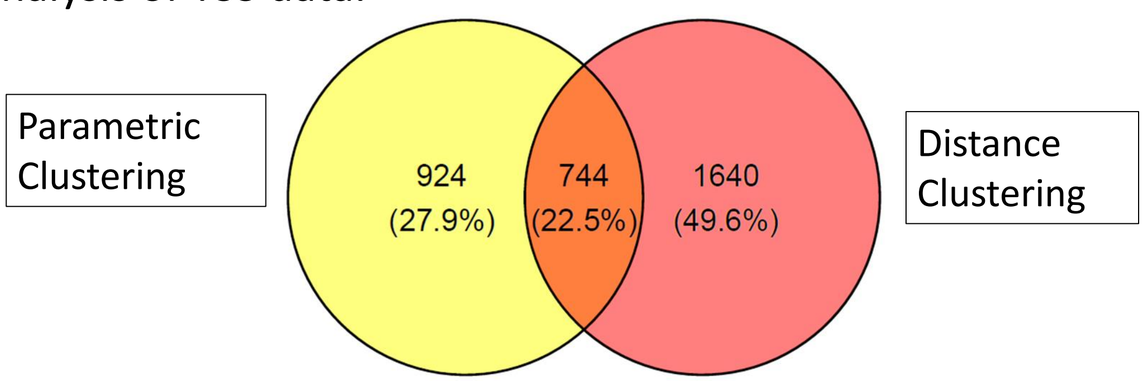


Work done

When comparing different types of assays after normalization of tag numbers we observed that different assays, when used on same sample produce similar TSS abundance data , with exception of nanocage.



When applying different normalization techniques and filtering data based on a chosen threshold of number of tags we obtain the same subset of data. This indicates that the choice of normalization method does should not affect downstream analysis of TSS data.



Preliminary comparison of clusters obtained by distance based clustering and density based parametric clustering gives vastly different results. This could present a big problem, since these techniques are being used to make biological claims based on TSR usage.

Future Work

- Studying the origin of discrepancies in clusters produced by different clustering methods
- Implementing a more accurate and reliable clustering method based on Kernel Density Estimation algorithm
- Apply and modify previously described statistical techniques for quantifying gene expression based on RNA-seq data in order to analyze and discover tissue specific TSRs

References

- Policastro, R. A., Raborn, R. T., Brendel, V. P. & Zentner, G. E. *Simple and efficient profiling of transcription initiation and transcript levels with STRIPE-seq*. Genome Res 30, 910–923 (2020).
- Raborn, R.T., Sridharan, Krishnakumar,Volker P. *Brendel TSRchitect. (Bioconductor, 2017).* doi:10.18129/B9.BIOC.TSRCHITECT.
- Rejes, A and Huber, W. *Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues.* Nucleic Acid Res 46, 582-592 (2018)

MORE
Masters Opportunity for Research in Engineering
ASU Ira A. Fulton Schools of Engineering
Arizona State University