Elucidating the Functional Roles of Antisense Transcripts with Computational Methods

Dongyan Lin[1], the FANTOM Consortium, Michiel de Hoon[2]

[1] Department of Physiology, University of Toronto, Canada [2] RIKEN Center for Integrative Medical Sciences (Laboratory for Applied Computational Genomics), Yokohama, Japan

Introduction

The human genome is pervasively transcribed into RNAs, but not all RNA transcripts will eventually be translated into protein (i.e. non-coding RNAs)[1]. Some of them regulate gene expression, which results in the drastically different products in the human body that come from an identical DNA sequence. Recently with regulatory RNA sequencing technologies, a novel family of RNAs called **CASPARs** (capped anti-sense promoter-associated regulatory RNAs) have been identified *(Fig. 1)*. As the name suggests, they have a 5' cap but lack a 3' poly-A tail, and their transcription start site (TSS) is near the promoters of coding genes on the opposite strand of DNA *(Fig. 2)*.

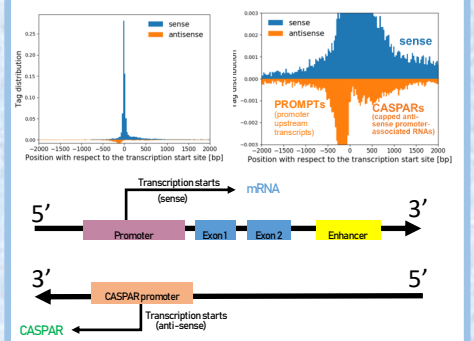

***Fig. 1** (top): The results of regulatory RNA sequencing*
***Fig. 2** (bottom): DNA double strand with exons of a protein-coding gene on one strand, and the TSS of CASPAR on the opposite strand.*

With almost 20 years of effort, **the FANTOM** (Functional Annotation of the Mammalian Genome) **Consortium** based in Japan has constructed an integrated database for human genome and their functionality, to a single-nucleotide resolution[2]. However, a large percentage of non-coding RNAs, including CASPARs, still have unknown biological functions. Therefore, with a focus on comparison to the sense gene promoters as well as the relationship to transcription factors, we aimed to elucidate the functional (especially regulatory) roles of CASPARs in human THP-1 leukemia cells using computational tools.

Methods

I. Transcription Factor Binding Site (TFBS) Distribution near CASPAR TSS

Given the genomic location (i.e. chromosome, start site, end site) of TFBS as well as the accurate 5' TSS of promoters and CASPARs , we used **BedTools**[3] to analyze the distribution of overlapping regions between TFBS and sense gene/CASPAR promoters, as well as the posterior probabilities of the TFBS near TSS. More specifically, we calculated the average posterior probability of the binding sites for each of the transcription factors, within ±500 bp of the TSS of sense gene/CASPAR, and compared the results *(Fig. 3)*.

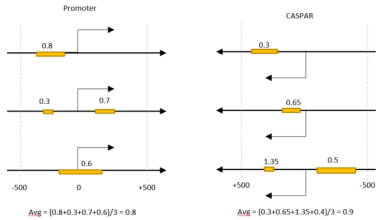

***Fig. 3**: Calculation for the average posterior probability of the binding sites of one transcription factor.*

II. Conservation near CASPAR TSS

With the phastCons conservation scores for multiple alignments of 45 other vertebrate genomes to the human genome (from the University of California, Santa Cruz), we plotted the conservation score distribution within ±500 bp of the TSS of either sense gene or CASPAR. Additionally, genome browsers such as **ZENBU** can be used to confirm that the computation for the average conservation score is correct, as it provides views of gene segments with additional information, including the conservation score at each nucleotide[4] *(Fig. 4)*. This way, we can check the average by looking at the scores base-by-base.

***Figure 4:** A screenshot of ZENBU which contains the human variation score at each nucleotide (top track) and FANTOM6 gene locations (bottom track).*

III. Gene Ontology (GO) Analysis

Gene ontology is the study that focuses on the functional annotation of genes. For this analysis, we selected human THP-1 genes that overlap with CASPAR TSS, and compared their GO functions to the matched background genes that have similar expression level but do not overlap with CASPAR TSS. To get the GO functions of each gene, we input their Ensembl IDs into **BioMart**, a data-mining tool that integrates different bioinformatics attributes of the same gene together.

Results

I. TFBS Distribution near CASPAR TSS

We found a positive correlation ($R^2 = 0.8475$) in the average posterior probabilities of TFBS between CASPAR promoters and sense gene promoters, with no TF strongly enriched or depleted at CASPAR promoters *(Fig. 5)*. This suggests that, similar to sense gene promoters, CASPAR promoters are also responsible for binding to TFs and regulating transcription and translation of the antisense genes.

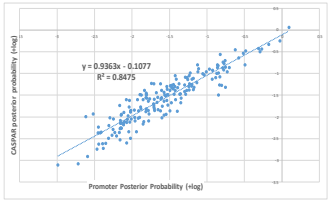

***Fig. 5**: Each dot represents one transcription factor, with their corresponding average number of binding sites for known sense gene promoters (on the x-axis) and for CASPAR promoters (on the y-axis).*

II. Conservation near CASPAR TSS

We compared human to 45 other vertebrate species and found higher conservation near sense gene TSS but not CASPAR TSS *(Fig. 6)*. This suggests that CASPAR transcripts could be human-specific.

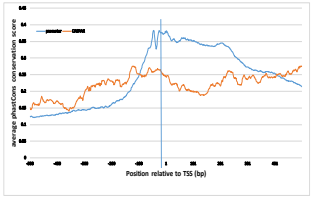

***Fig. 6**: The average phastCons score as a function of the distance to the TSS of sense gene or CASPAR.*

III. Gene Ontology (GO) Analysis

We found that, compared to the matched-background genes, CASPAR-associated genes statistically overrepresent many categories in biological processes, molecular functions and cellular component, particularly the category of "regulation of transcription" (odds ratio ~ 1.77, p ~ 2e-26). Since many genes that regulate transcription code for transcription factors, we further looked into details and found that genes that regulate transcription and code for transcription factors are more likely to overlap with CASPAR (odds ratio ~1.87, p ~ 6e-22). This suggests the important role of CASPARs in regulating gene expression, through regulating the expression levels of the sense genes that code for transcription factors.

Discussion

Through a series of analysis, many useful insights about the functional roles of CASPARs have slowly been revealed. CASPAR promoters seem to share some common attributes with gene promoters, such as the **posterior probability** of each TFBS near the TSS. The importance of transcription factors in assisting CASPARs to regulate gene expression was confirmed by **GO analysis**. Previously a lncRNA antisense to the mouse protein-coding gene Uchl1 has been shown to increase UCHL1 protein synthesis[5]. Combined with the fact that promoters regulate gene expression by interacting with transcription factors, evidence suggests that antisense transcripts indeed have a regulatory role on the transcription and translation of the sense gene they overlap through the interaction with transcription factors. Now the question becomes: Which genes do CASPARs regulate? How does this regulation dynamically change over time?

Future Directions

To find out which genes CASPARs regulate, we can **knock down** the CASPARs using antisense oligos and observe the genome-wide transcriptome response.

To elucidate the dynamics of CASPAR regulation, we can use the **CAGE (Cap Analysis of Gene Expression)** data at different time points from the FANTOM Consortium, and see if they correspond to the expression dynamics of the transcription factors that are coded by the CASPAR-overlapping sense genes.

Additionally, we can extend our experiments to other cell types besides THP-1 leukemia cells to see if the regulatory roles of CASPARs are cell-type specific.

Acknowledgements

- Laboratory for Applied Computational Genomics, RIKEN:
Dr. Michiel de Hoon Jessica Severin
Dr. Jordan Ramilowski Dr. Saumya Agrawal
- Department of Cell & System Biology, University of Toronto
Janet Mannone Prof. Leslie Buck

References

1. Hon, Chung-Chau, et al. "An atlas of human long non-coding RNAs with accurate 5' ends." *Nature* 543.7644 (2017): 199.
2. de Hoon, Michiel, Jay W. Shin, and Piero Carninci. "Paradigm shifts in genomics through the FANTOM projects." *Mammalian Genome* 26.9-10 (2015): 391-402.
3. Quinlan, Aaron R., and Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features." *Bioinformatics* 26.6 (2010): 841-842.
4. Severin, Jessica, et al. "Interactive visualization and analysis of large-scale sequencing datasets using ZENBU." *Nature Biotechnology* 32.3 (2014): 217-219 .
5. Carrieri, Claudia, et al. "Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat." *Nature* 491.7424 (2012): 454.