Indrani Datta, MS, Biostatistician, 3138746229, idatta1@hfhs.org

Co-Expression of Long non-coding RNAs with Epigenetically regulated genes in TCGA Glioma subtypes

Indrani Datta[1,2,3], Laila M. Poisson[1,2,3]

Center for Bioinformatics[1], Public Health Sciences[2], Hermelin Brain Tumor Center[3], Henry Ford Health System, Detroit, Michigan

Background- In recent years RNA-seq deep sequencing technology has emerged as a revolutionary tool to precisely measure transcriptome profiling in eukaryotic genomes. Beyond protein coding RNAs, long non-coding RNAs (lncRNAs) have become recognized as a gene regulators as well as prognostic markers in cancer. In this study, we initiated an *in-silico* analysis of co-expression of lncRNAs with epigenetically regulated genes (EReg) in TCGA Glioblastoma multiform (GBM) and Lower Grade Glioma (LGG) RNA-seq data.

Method- Open-source RNA-seq data set which is manufactured with Illumina HiSeq platform from TCGA GBM and LGG cohort were integrated to capture highly correlated bio-molecules, in our case, lncRNAs and ERegs. A set of 12382 differentially regulated lncRNAs transcripts were identified across various cancers including GBM & LGG samples (372) were derived from Chinnaiyan *et.al* identified by the Tuxedo suite (i.e, Tophat, Cufflink) which perform many aspects of complete RNA-seq analysis in *ab initio* assembly mode. A set of 809 EReg transcripts were obtained from Cecceralli *et al*. which categorizes 7 distinct glioma subtypes in IDHmutant (codal=69,G-CIMP-high=104,G-CIMP-low=8) and IDHwildtype (Classic-like=54,LGm6-GBM=12,Mesenchymal-like=69,PA-like=15) by unsupervised clustering of Illumina methylation 27k and 450k array probes. The expression estimates of these EReg transcripts were generated by Mapsplice/RSEM workflow constructed by broad Institute, were downloaded from GDAC. Expression estimates of lncRNAs were in FPKM and ERegs were in estimated transcript fraction, as these two measures were generated by two different algorithms/workflow, so they were made compatible by converting to transcripts per million (TPM). Following data processing and QC on this integrated data, 315 samples and 12991 (lncRNA=12195 and EReg=796 transcripts) molecules were carried forward for analysis with Weighted Correlation network analysis. Following detection of networks which consists of lncRNAs and ERegs, association of these co-expression networks to glioma subtypes were analyzed with Anova. EReg genes from significantly associated modules were further analyzed by Ingenuity's IPA to delineate biological association as majority of lncRNAs have unknown functions, so "guilt by association" mechanism was used for retrieving functional relevance to these lncRNAs by EReg genes.

Results: There were 27 lncRNA-EReg gene modules were detected. Among these, 2 modules were significantly associated 2 glioma subtypes (IDHWt = PA-Like and IDHWt = LGm6-GBM) at pvalue < 0.05. After multiple testing corrections, both of these modules remain as significant at FDR level < 0.05. EReg genes which were extracted from module associated with LGm6-GBM are working together in cell-To-cell Signaling and Interaction, cellular Growth and Proliferation while EReg genes associated with PA-Like glioma subtype were working together in cell cycle, cellular development, cellular growth and proliferation biological functions. So it can be assumed that lncRNA transcripts which were co-expressed with ERegs transcripts from above mentioned modules will participate in these cellular functions.

Conclusion- This study demonstrates the application of existing bioinformatics algorithms to analyze open source RNA-seq data to capture gene-lncRNA association in respect to glioma subtypes.