Mathematical Methods and Models in Biosciences
June 18-23, 2023, Pomorie, Bulgaria
https://biomath.math.bas.bg/biomath/index.php/bmcs

Identification and analysis of cell-specific expressed genetic variants from scRNA-seq data

Allen Kim[1], Kai Saito[2], Zhe Yu[1], Hovhannes Arestakesyan[1], Evgenia Unianova[1], Nathan Edwards[3], Anelia Horvath[1]

[1]McCormick Genomic and Proteomic Center,
School of Medicine and Health Sciences,
The George Washington University, Washington, DC 20037, USA
horvatha@email.gwu.edu

[2]Northeastern University, Boston, MA 02115, USA

[3]Department of Biochemistry and Molecular and Cellular Biology,
Georgetown University, Washington, DC 20057, USA

Low cellular frequency variants may indicate pre- or early-somatic clonality in cancer and normal tissues, or cell-specific RNA variance. Currently, most genetic variation is analyzed from bulk sequencing datasets, where low cellular frequency variants are difficult to distinguish from artifacts.

To address challenges posed by low frequency variation events, we have developed a computational framework for identification and analysis of Single Cell-specific Expressed Single Nucleotide Variants (sceSNVs) from single cell RNA-sequencing (scRNA-seq) data. Central for the framework is our new tool SCExecute, which enables the execution of various software designed for bulk sequencing data on barcode-stratified, extracted on-the-fly, single-cell alignments. Applying SCExecute in conjunction with tools for analysis of bulk sequencing data, we explored, for the first-time, expressed genetic variation at cell-level across 28 publicly available cancer and normal datasets, including prostate cancer, non-small cell lung carcinoma, cholangiocarcinoma, neuroblastoma, normal fetal adrenal and normal embryo. This analysis identified over 100,000 previously unreported expressed SNVs, including somatic mutations and RNA-originating variance, such as posttranscriptional modifications and locus-specific transcriptional infidelity. Our analysis shows that over 70% of these variants cannot be identified with the current bulk-based variant callers. Furthermore, approximately 10% of these novel variants show preferential expression in particular cell clusters and pseudo-time stages. Single-cell RNA e-QTL (scReQTL) analysis revealed that the expression of such sceSNVs correlates with increased expression of their harboring gene. Moreover, differential gene expression analysis between cells expressing these sceSNVs and the neighborhood cells expressing the reference allele, showed deregulation of functional gene-networks of the

SNV-harboring gene. Asymmetrically expressed sceSNVs across multiple samples are enriched in genes participating in DNA-repair, replication, and cell cycle pathways. We exemplify our analyses with a novel missense substitution – 6:26104128_G>T, expressed in a gene encoding one of the core histones (HIST1H4C^{V61F}). We demonstrate that HIST1H4C^{V61F} is correlated to high expression of HIST1H4C and deregulation of the HIST1H4C-related gene network, the observation being more pronounced in neurons, across multiple cancer samples.

Our findings suggest that there is an unappreciated repertoire of cell-level expressed nucleotide variation, possibly recurrent and common across samples, that participates in transcriptome function and dynamics in both cancer and normal cells. Their appearance and, for some, relationship to certain gene-sets and cell types, suggests novel mechanisms and function for the expressed genetic variation, including in cancer progression and cell fate.