Mathematical Methods and Models in Biosciences
June 15–20, 2025, Sofia, Bulgaria
https://biomath.math.bas.bg/biomath/index.php/bmcs

ML models for functional SNV prioritization using allele-specific expression from single-cell RNA-seq

Vania Ballesteros Prieto[1], Sunisha Harish[1], Siera Martinez[1], Tushar Sharma[1], Luke Johnson[1], Hovhannes Arestakesyan[1], Jewel Dias[1], Joseph Goldfrank[2], Anelia Horvath[1]

[1]McCormick Genomics and Proteomics Center, Department of Biochemistry and Molecular Medicine, School of Medicine and Health Sciences, The George Washington University, Washington, DC 20037, USA
horvatha@email.gwu.edu

[2]Department of Computer Science, School of Engineering & Applied Science, Georgetown University, Washington, DC 20057, USA

This study presents an AI-driven framework for identifying and prioritizing functional single-nucleotide variants (SNVs) in cancer using allele-specific expression (ASE) patterns derived from single-cell RNA sequencing (scRNA-seq) data. Traditional approaches typically examine SNVs in isolation, neglecting potential interactions between co-occurring variants. We addressed this gap by analyzing both individual and combinatorial SNV effects on gene expression and cellular phenotypes within their native genetic contexts.

Two complementary strategies were employed: (1) machine learning (ML) models to detect SNVs exhibiting ASE signatures similar to known pathogenic variants, and (2) a method for identifying SNV combinations with synergistic transcriptomic effects. Preliminary analyses revealed strong correlations between predicted SNV functionality and ASE in scRNA-seq, supporting allele expression as a biologically meaningful indicator of variant impact.

The computational pipeline integrates Bayesian networks, classical ensemble models, neural networks, and hierarchical modeling to prioritize SNVs and interactions with high biological relevance. The framework incorporates standard scRNA-seq tools and custom modules for SNV-centric analyses, ensuring scalability and robustness across large, heterogeneous datasets.

Interpretability was prioritized through differential expression analysis, functional enrichment, and regulatory network reconstruction. Overrepresented pathways and master regulators linked to variant configurations were identified, including variant effects stratified by cell type and RNA-editing status.

This approach not only established a scalable strategy for pinpointing functional and combinatorial SNVs in cancer but also provided insights into tumor biology, candidate biomarkers, and potential predictors of therapeutic response. The methodological flexibility supports broader applicability across other complex diseases, extending the utility of ASE-informed variant analysis beyond oncology.