exSEEK: A bioinformatics tool for extra-cellular RNA biomarker discovery

Xupeng CHEN

Advisor: Prof. Zhi Lu

MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China

Abstract

Extracellular RNAs (exRNAs) in body fluid provide a large repository of biomarker candidates. Deep sequencing makes it possible to monitor exRNAs in a comprehensive way. Because of the unique properties of exRNA sequencing data, identification of potential biomarkers for clinical usage remains challenging. exSEEK, a bioinformatics tool for identification of biomarkers associated with certain diseases, which is suitable for analyzing of both small and long RNA sequencing data generated by difference library construction methods, was developed to overcome such challenges.

exRNAs are highly fragmented. In this work, we showed that some of these fragments have a recurring pattern. Compared to full length transcripts, the recurring fragments, or "domain features", perform better in predicting cancers, and has higher concordance with the result of PCR based assays. exSEEK can be customized to select both full length features and domain features for cancer prediction. The reliability of exSEEK was further confirmed by publicly available datasets and experimental validation.

We developed exSEEK, a tool for computational analysis of exRNA sequencing datasets and biomarker discovery. For small RNA-seq data, exSEEK assigns reads to multiple RNA types sequentially in a user-specified order. A unique feature of exSEEK is that regions with significantly higher read coverage than background are detected to generate "domain" features. Abundance of exRNA domains and miRNAs are combined to create a count matrix. We applied various combinations of normalization and batch removal methods to the count matrix to correct data heterogeneity and batch effects and suggest that correction is needed to spurious associations between features and the true biological signal. We also developed a machine learning framework to robustly select most important features that distinguish cancer from normal samples. We performed integrative analysis of three datasets: cell-free small RNA, exosomal small RNA and exosomal long RNA and evaluated the classification performance of HCC, CRC, PRAD and PAAD.

Keywords: liquid biopsy, biomarker, feature selection, machine learning