Normalizing cancer RNA-seq data for library size, tumor purity and batch effects

A seminar by Ramyar Molania

Walter and Eliza Hall Institute of Medical Research - Australia

Friday 14 Oct 2022 | 2:30 p.m.
Room Benvenuti and live Zoom
Department of Statistical Sciences

Accurate identification and effective removal of unwanted variation is essential to derive meaningful biological results from RNA sequencing (RNA-seq) data, especially when the data come from large and complex studies. Using RNA-seq data from The Cancer Genome Atlas (TCGA), we examined several sources of unwanted variation and demonstrate here how these can signifi- cantly compromise various downstream analyses, including cancer subtype identification, association between gene expression and survival outcomes and gene co-expression analysis. We propose a strategy, called pseudo-replicates of pseudo-samples (PRPS), for deploying our recently developed normalization method, called removing unwanted variation III (RUV-III), to remove the variation caused by library size, tumor purity and batch effects in TCGA RNA-seq data. We illustrate the value of our approach by comparing it to the standard TCGA normalizations on several TCGA RNA-seq datasets. RUV-III with PRPS can be used to integrate and normalize other large transcriptomic datasets coming from multiple laboratories or platforms.