Harmonizing Omics Data across Platforms for Post-GWAS of Lung Adenocarcinoma

陳孜圩
東海大學智慧計算暨應用數學系

Abstract

A recent genome-wide association study (GWAS) of lung adenocarcinoma (LUAD) in East Asian populations reported 28 independent susceptibility variants across 25 loci and, by integrating a Taiwanese lung tissue gene expression reference dataset, identified two genes whose genetically predicted expression levels are associated with LUAD risk through transcriptome-wide association studies (TWAS). While promising, these findings highlight the need to increase statistical power by expanding reference datasets and to incorporate additional regulatory information through the integration of DNA methylation data for methylome-wide association studies (MWAS). In this talk, we present a systematic strategy for cross-platform omics data harmonization to address these needs. By utilizing inverse-rank transformation, we harmonized gene expression data (across microarray and RNA-seq platforms) and DNA methylation data (across Illumina 450K, EPIC, and EPICv2 arrays) to construct expanded reference datasets. We then applied these datasets to TWAS and MWAS, identifying several LUAD-associated genes and CpGs. Notably, 23 of the 28 known susceptibility variants are located near these identified genes or CpGs, with some representing putative novel loci. These results demonstrate that cross-platform harmonization significantly enhances the discovery power of post-GWAS analyses, substantially advancing our understanding of the etiology of LUAD.