ABSTRACT

Identification of cancer patient subgroups using high throughput genomic data is of critical importance to clinicians and scientists because it can offer opportunities for more personalized treatment and overlapping treatments of cancers. Despite tremendous efforts, this problem remains challenging because of low reproducibility and instability of identified cancer subgroups and molecular features. In order to address this challenge, we developed Integrative Genomics Robust iDentification of cancer subgroups (InGRiD), a statistical approach that integrates information from biological pathway databases with high-throughput genomic data to improve the robustness for identification and interpretation of molecularly-defined subgroups of cancer patients. We applied InGRiD to the gene expression data of high-grade serous ovarian cancer from The Cancer Genome Atlas and the Australian Ovarian Cancer Study. The results indicate clear benefits of the pathway-level approaches over the gene-level approaches. In addition, using the proposed InGRiD framework, we also investigate and address the issue of gene sharing among pathways, which often occurs in practice, to further facilitate biological interpretation of key molecular features associated with cancer progression. The R package “InGRiD” implementing the proposed approach is currently available in our research group GitHub webpage (https://dongjunchung.github.io/INGRID/).