Abstract:
Breast cancer is one of the most commonly diagnosed cancers. It is associated with DNA methylation, an epigenetic event with a methyl group added to a cytosine paired with a guanine, i.e., a CG site, in a human genome. The methylation levels of different genes in a genome are correlated in certain ways that affect gene functions. This correlation pattern is known as co-methylation. It is still not clear how different genes co-methylate in the whole genome of breast cancer samples. Previous studies are conducted using relatively small datasets (Illumina 27K data). In this study, we analyze much larger datasets (Illumina 450K data). We will show our correlation analysis results on overall co-methylation patterns, which are related to physical distance and sign (i.e., positive or negative correlation) in normal and tumor datasets. We will show specific sites whose co-methylation patterns change significantly from normal to tumor samples. We will report relationships of genes related to these sites. We will also show the super-connector sites/genes that have high correlations with many other genes and report pathway analysis results. Due to the large data size, our analysis is computationally challenging. For example, because of the memory/storage limit of R, we cannot even create a 450,000 by 450,000 matrix to store our correlation coefficients. Our ability of analyzing datasets of this size can provide researchers with a new and improved understanding of co-methylation patterns in breast cancer. Our new findings will furthermore allow researchers to help establish relationships and associations among different genes in the future.