Reordering Analysis Operations to Improve Accuracy of Genomic Copy Number Analysis of Tumors

In genomic copy number analysis of tumors, a series of complex data analysis operations are performed to address a set of statistical problems. Most published methods perform data analysis operations in the following conventional order: (1) normalize raw data of tumors and controls, (2) compare normalized data of tumors to that of controls, (3) identify genomic loci that define change-points in the distribution of the comparison statistic, and (4) infer copy number status of segments defined by those loci. Careful consideration of analysis results indicates that this conventional ordering does not generate and utilize biological and statistical information in a manner to optimally reduce the bias and variance of the final result. A systematic characterization of how the order of operations impacts the availability, quality, and utilization of information indicates that more accurate results can be obtained by using the following unconventional order of analysis operations: (1) compare unnormalized data of tumors and controls, (2) iteratively normalize comparison statistics (e.g., differences) and identify change-point loci, (3) identify two-copy genomic segments, and (4) infer the copy number status of genomic segments. New methods that use this novel and non-conventional order of analysis operations substantially improve accuracy of copy number analysis of SNP microarray and whole-genome next-generation sequencing data as confirmed by comparison of analysis results with cytogenetic copy number data.