TRANSCRIPTOMIC DATA SCIENCE USING BIOCONDUCTOR

DAVIS SR1, CAREY VJ2, CULHANE A3, GENTLEMAN R4, HANSEN KD5, HUBER W6, IRIZZARY RA3, LAWRENCE M7, WALDRON L8, and MORGAN M9

1Center for Cancer Research, National Cancer Institute

2Brigham and Women's Hospital, Harvard Medical School

3Dana-Farber Cancer Institute, Harvard School of Public Health

4Computational Biology, 23andMe

5Bloomberg School of Public Health, Johns Hopkins University

6European Molecular Biology Laboratory

7Genentech Research and Early Development

8CUNY School of Public Health at Hunter College

9Department of Biostatistics and Bioinformatics, Roswell Park Cancer Institute

High-throughput transcriptome profiling technologies are now commonplace and are available as commodity products. With the easy production of data has come an interest by an ever-expanding number of researchers asking increasingly complex and interesting questions. The value of these experiments and the level of the derived insights depend on the ability to organize, analyze, and interpret the data. Bioconductor is an open-source, open-development software project for the analysis and comprehension of high-throughput data in genomics and molecular biology. The project aims to enable interdisciplinary research, collaboration and rapid development of scientific software. Based on the statistical programming language R, Bioconductor comprises 1211 interoperable software packages contributed by a large, diverse community of scientists. Transcriptome analysis capabilities include primary data analysis such as normalization and quality control, secondary analysis such as differential expression and isoform discovery, and tertiary analysis such as pathway analysis. The breadth and depth of Bioconductor allows integration with orthogonal measurements such as functional screening, epigenomics, DNA sequencing, and clinical covariates. Functionality for advanced and interactive data visualization add further value to data. Bioconductor also has extensive capabilities for accessing and integrating with nearly all large publicly available datasets and repositories such as TCGA, GTex, NCBI GEO, the NCI Genomic Data Commons. Core values of reproducible research, interoperability, and open development have made Bioconductor a standard platform for genomic data science with over 15,000 downloads per month. I will review these capabilities in the context of cancer data science, transcriptome analysis, and data integration.