

refinebioc: A Bioconductor Package for Seamless Integration of refine.bio Transcriptome Data

Abstract

Genomic sequencing technologies generate vast amounts of data but often come with significant costs and resource demands. Refine.bio, a publicly available repository, offers uniformly processed and normalized transcriptome data from resources like Gene Expression Omnibus (GEO), Sequence Read Archive (SRA), and ArrayExpress. It provides over 60,000 harmonized transcriptome experiments, including RNA-Seq and microarray data, from more than 979 organisms and 1,200 platforms. Most of these data focus on human and mouse samples, with additional data spanning many species and conditions. To streamline the use of these data, we developed the refinebioc R/Bioconductor package.

The refinebioc package bridges the gap between the refine.bio repository and Bioconductor, a widely-used ecosystem for genomic data analysis. This integration simplifies data acquisition, analysis, and visualization, allowing users to leverage Bioconductor's extensive tools for downstream analyses like differential expression, clustering, and pathway analysis. Refinebioc enables users to search for experiments, retrieve metadata, and filter datasets by criteria such as organism, platform, or biological condition. Users can then download, extract, and load data directly into R as a SummarizedExperiment object, a Bioconductor-standard data container.

By utilizing the refine.bio Application Programming Interface (API) to automate data retrieval, refinebioc eliminates the need to perform manual searches on the refine.bio website. After downloading samples of interest, the package processes the data into a SummarizedExperiment, ensuring compatibility with Bioconductor's analysis methods. This design fosters computational reproducibility and transparency, making it easier for researchers to access relevant datasets from various organisms, platforms, and technologies.

The functionalities of refinebioc improve accessibility and enable data-driven insights. By applying these tools, researchers can focus more on data interpretation and discovery, rather than on the complexities of data retrieval and processing. Furthermore, as refine.bio continuously updates its dataset, refinebioc ensures users always have access to the latest data.

In conclusion, the refinebioc package empowers researchers to efficiently access and analyze data from the refine.bio repository. These tools bridge the gap between raw genomic data and powerful analysis methods, enhancing the effectiveness of genomic research while reducing barriers to data reuse. With over 2 million samples ready for analysis, these packages simplify the process of integrating large-scale transcriptome data into Bioconductor workflows, fostering better research outcomes and promoting reproducibility in bioinformatics.