1:

Next Generation Sequencing (NGS) alignment, variant calling, annotation

Next Generation Sequencing (NGS) RNASeq

2:

Differential Expression

3:

Bioinformatics data mining

|  |
| --- |
| **3:**  **Bioinformatics data mining** |
| integrative data analysis  BMC bioinformatics  "data mining" "datasets re-use"  Integrative bioinformatics  "data model"  -----------------------------------------------------------------------------------------------------------------------------------  **Data mining**  uses statistical methods to search for patterns in existing data. This method generally returns many patterns, of which some are spurious and some are significant, but all of the patterns the program finds must be evaluated individually. Currently, some research is focused on incorporating existing data mining techniques with novel pattern analysis methods that reduce the need to spend time going over each pattern found by the initial program, but instead, return a few results with a high likelihood of relevance.[10] One drawback of this approach is that it does not integrate multiple databases, which means that comparisons across databases are not possible. The major advantage to this approach is that it allows for the generation of new hypotheses to test.  The challenge is not only to extract meaningful information from this data, but to gain knowledge, to discover previously unknown insight, look for patterns, and to make sense of the data [20], [21]. Many different approaches, including statistical and graph theoretical methods, data mining, and machine learning methods, have been applied in the past - however with partly unsatisfactory success [22, 23] especially in terms of performance [24].  -----------------------------------------------------------------------------------------------------------------------------------  "Mining and Analysing Spatio-Temporal Patterns of Gene Expression in An Integrative database Framework."  Journal of Integrative Bioinformatics (2010)  -----------------------------------------------------------------------------------------------------------------------------------  **An Integrated Encyclopedia of DNA Elements in the Human Genome**  **The ENCODE Project Consortium**  PMC 2013 Mar 6  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3439153/>  -----------------------------------------------------------------------------------------------------------------------------------  **Big Data Bioinformatics**  PMCID: PMC5604462  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5604462/>  Unsupervised Methods Discover New Patterns  Unsupervised methods are used with the overarching goal to uncover the underlying organizational structure of the data.  For example, "What patterns exist in gene expression of cancers?”  These algorithms tend to discover dominant recurrent features in the data.  Principle Components Analysis (PCA) is an unsupervised approach to identify hidden features in the data that provide the most signal.  The first principle component is the feature that explains most of the variability in the data.  When we perform PCA analysis on a dataset combined from two large studies of breast cancer (Cancer Genome Atlas, 2012)  Such study, platform, or batch effects can confound unsupervised analyses,  so these algorithms are often best applied within a single homogenous dataset.  Random Forest (RF)  Support Vector Regression (SVR)  multivariable linear regression model (MLR)  Machine Learning supervised unsupervised  supervised learning model    labeled unlabeled data  **European Molecular Biology Laboratory / European Bioinformatics Institute (EBI)** Hinxton, UK |
| **MultiDataSet: an R package for encapsulating multiple data sets with application to omic data integration**  BMC Bioinformatics volume 18, (2017)  11 Citations  <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1>  Results To cover this need, we have developed **MultiDataSet**, a new R class based on Bioconductor standards, designed to encapsulate multiple data sets. MultiDataSet deals with the usual difficulties of managing multiple and non-complete data sets while offering a simple and general way of subsetting features and selecting samples. We illustrate the use of MultiDataSet in three common situations: 1) performing integration analysis with third party packages; 2) creating new methods and functions for omic data integration; 3) encapsulating new unimplemented data from any biological experiment. Conclusions **MultiDataSet is a suitable class for data integration under R and Bioconductor framework.**  A standard infrastructure has been created to represent biological data comprising, amongst others, two basic R classes to load experiment’s information: ***eSet*** and ***SummarizedExperiment***. The main objective of these classes is that biological data and phenotypic descriptions are well coordinated. In particular, methods such as subsetting are easily applied simultaneously to experiment and phenotypic data. While *eSet* is designed for microarray data, *SummarizedExperiment* is for next generation sequencing data.  Major public projects have performed experiments to a group of individuals generating different types of datasets [[5](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR5)]. For instance, the Cancer Genome Atlas (TCGA) [[6](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR6), [7](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR7)], is the largest resource available for multi-assay cancer genomics data; the 1000 Genome Project [[8](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR8), [9](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR9)] aims to provide a comprehensive resource for human genetic variants and gene-expression across populations and; the International Cancer Genome Consortium (ICGC) [[10](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR10), [11](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR11)] coordinates 55 research projects to characterize the genome, transcriptome and epigenome of multiple tumors. In addition, large repositories collect data of several smaller projects allowing unified storage and stimulating data sharing. Gene Expression Omnibus (GEO) [[12](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR12)–[14](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR14)] is the primary database where data from multi-assay experiments is shared publicly. Other reference databases are dbSNP [[15](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR15), [16](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR16)], a deposit for short genetic variations and Database of Genomic Variants archive (DGVa), for longer structural variants [[17](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR17), [18](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR18)].  All of these data resources are accessible through standard Bioconductor classes (*eSet* and *SummarizedExperiment*). There are numerous packages used to retrieve and transform data from public repositories. For instance, GEOquery [[19](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR19)] that obtains GEO data as an *eSet* object. Such packages aim to facilitate downstream analyses for Bioconductor’s packages. However, Bioconductor lacks a standard class to efficiently manage different datasets obtained from the same individuals.  Several R/Bioconductor packages have implemented methods to integrate and visualize biological data: *PMA* [[20](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR20)–[22](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR22)], *mixOmics* [[23](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR23)–[25](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR25)], *made4* [[26](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR26), [27](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR27)], *RGCCA* [[28](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR28), [29](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR29)], *omicade4* [[30](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR30), [31](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR31)], *CNAmet* [[32](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR32), [33](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR33)], *RTopper* [[34](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR34), [35](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR35)], *iClusterPlus* [[36](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR36), [37](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR37)] and *STATegRa* [[38](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR38)] among others. Each of these packages implements a different strategy to face the integration analysis. They typically use their own data structure, which is usually a list of matrices. The use of such structure makes it difficult to perform usual operations such as subsetting data across data sets and selecting samples (e.g. complete cases are usually required in all integration analysis). The specificity of the data structures to each method further hinders the user’s disposition to perform different integration analyses on one study. Therefore, a standard structure to manage the different datasets of the same individuals will promote the use of current and future integration methods, allowing the implementation of general methods for management and processing.  In this article, we present *MultiDataSet*, a new R class based on Bioconductor standards developed to encapsulate multiple datasets. *MultiDataSet* deals with the usual difficulties of managing multiple and non-complete datasets while offering a simple way of subsetting features and selecting individuals. We describe the internal structure of *MultiDataSet* and illustrate its use in three examples (Additional files [1](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#MOESM1), [2](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#MOESM2) and [3](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#MOESM3)) which cover three common situations in integration analyses. **Design and implementation** ***MultiDataSet* is a S4 class of R implemented under Bioconductor guidelines** [[39](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#ref-CR39)]. Its structure is an extension of the abstract *eSet* class. *MultiDataSet* is therefore a data-storage class that comprises datasets of different omic data (assay data), feature data and phenotypic data. Despite its general form, *MultiDataSet* maintains the specific characteristics of the datasets (e.g. it preserves matrices of calls and probabilities of a ***SnpSet***). Internal structure of MultiDataSet ***MultiDataSet* comprises five fields that are R standard lists.** Their names match other Bioconductor classes: ***assayData*** that contains the measurement values; ***phenoData*** that stores the description of the samples; ***featureData*** and ***rowRanges*** that have the description of the features; and ***return\_method*** that allows recovering the original dataset. Relation between fields is shown in Fig. [1](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1455-1#Fig1). In each dataset, samples are shared between *assayData* and *phenoData*, and features between ***assayData***, ***featureData*** and ***rowRanges***. We have programmed a function to recover the original datasets. The class is designed such that the different data is coordinated. A particular feature of *MultiDataSet* is the storing of datasets from different experiments that may not share the full set of samples between them. |