Focus on Computer Resources

Cancer Research

# Software for the Integration of Multiomics Experiments in Bioconductor

Marcel Ramos[1,2,3], Lucas Schiffer[1,2], Angela Re[4], Rimsha Azhar[1,2], Azfar Basunia[5], Carmen Rodriguez[1,2], Tiffany Chan[1,2], Phil Chapman[6], Sean R. Davis[7], David Gomez-Cabrero[8], Aedin C. Culhane[5,9], Benjamin Haibe-Kains[10,11,12,13], Kasper D. Hansen[14,15], Hanish Kodali[1,2], Marie S. Louis[1,2], Arvind S. Mer[10], Markus Riester[16], Martin Morgan[3], Vince Carey[5,17], and Levi Waldron[1,2]

## Abstract

Multiomics experiments are increasingly commonplace in biomedical research and add layers of complexity to experimental design, data integration, and analysis. R and Bioconductor provide a generic framework for statistical analysis and visualization, as well as specialized data classes for a variety of high-throughput data types, but methods are lacking for integrative analysis of multiomics experiments. The MultiAssayExperiment software package, implemented in R and leveraging Bioconductor software and design principles, provides for the coordinated representation of, storage of, and operation on multiple diverse genomics data. We provide the unrestricted multiple `omics data for each cancer tissue in The Cancer Genome Atlas as ready-to-analyze MultiAssayExperiment objects and demonstrate in these and other datasets how the software simplifies data representation, statistical analysis, and visualization. The MultiAssayExperiment Bioconductor package reduces major obstacles to efficient, scalable, and reproducible statistical analysis of multiomics data and enhances data science applications of multiple omics datasets. *Cancer Res; 77(21); e39–42. ©2017 AACR.*



## Introduction

Multiassay experiments collect multiple, complementary data types for a set of specimens. Bioconductor (1) provides classes to ensure coherence between a single assay and patient data during data analysis, such as eSet and SummarizedExperiment (2). However, novel challenges arise in data representation, management, and analysis of multiassay experiments (3) that cannot be addressed by these or other single-assay data architectures. These include (i) coordination of different assays on, for example, genes, miRNAs, or genomic ranges; (ii) coordination of missing or replicated assays; (iii) sample identifiers that differ between assays; (iv) reshaping data to fit the variety of existing statistical and visualization packages; (v) doing the above in a concise and reproducible way that is amenable to new assay types and data classes.

The need for a unified data model for multiomics experiments has been recognized in other projects, such as MultiDataSet (4) and CNAMet (5). Our developments are motivated by an interest in bridging effective single-assay application program interface (API) elements, including endomorphic feature and sample subset operations, to multiomic contexts of arbitrary complexity and volume (Supplementary Table S1). A main concern in our work is to allow data analysts and developers to simplify the management of both traditional in-memory assay stores for smaller datasets, and out-of-memory assay stores for very large data in such formats as HDF5 (6), tabix-indexed variant call format (VCF; ref. 7), or Google BigTable (8).

MultiAssayExperiment provides data structures and methods for representing, manipulating, and integrating multiassay genomic experiments. It integrates an open-ended set of R and Bioconductor single-assay data classes, while abstracting the complexity of back-end data objects and providing a sufficient set of data manipulation, extraction, and reshaping methods to interface with most R/Bioconductor data analysis and visualization tools. We demonstrate its use by representing unrestricted data from The Cancer Genome Atlas as a single

[1]Graduate School of Public Health & Health Policy, City University of New York, New York, New York. [2]Institute for Implementation Science in Population Health, City University of New York, New York, New York. [3]Roswell Park Cancer Institute, University of Buffalo, Buffalo, New York. [4]Centre for Sustainable Future Technologies, Istituto Italiano di Tecnologia, Corso Trento, Torino, Italy. [5]Harvard TH Chan School of Public Health, Boston, Massachusetts. [6]Computational Biology Support Team, Cancer Research UK Manchester Institute, The University of Manchester, Manchester, United Kingdom. [7]Center for Cancer Research, NCI, NIH, Bethesda, Maryland. [8]Mucosal and Salivary Biology Division, King's College London Dental Institute, London, United Kingdom. [9]Dana-Farber Cancer Institute, Boston, Massachusetts. [10]Princess Margaret Cancer Center, University Health Network, Toronto, Ontario, Canada. [11]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [12]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [13]Ontario Institute of Cancer Research, Toronto, Ontario, Canada. [14]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland. [15]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland. [16]Novartis Institutes for BioMedical Research, Cambridge, Massachusetts. [17]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts.

**Note:** Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

**Corresponding Author:** Levi Waldron, CUNY School of Public Health, 55 W 125th Street, 6th Floor, New York, NY 10027. Phone: 646-364-9616; Fax: 646-219-6444; E-mail: levi.waldron@sph.cuny.edu

**doi:** 10.1158/0008-5472.CAN-17-0344

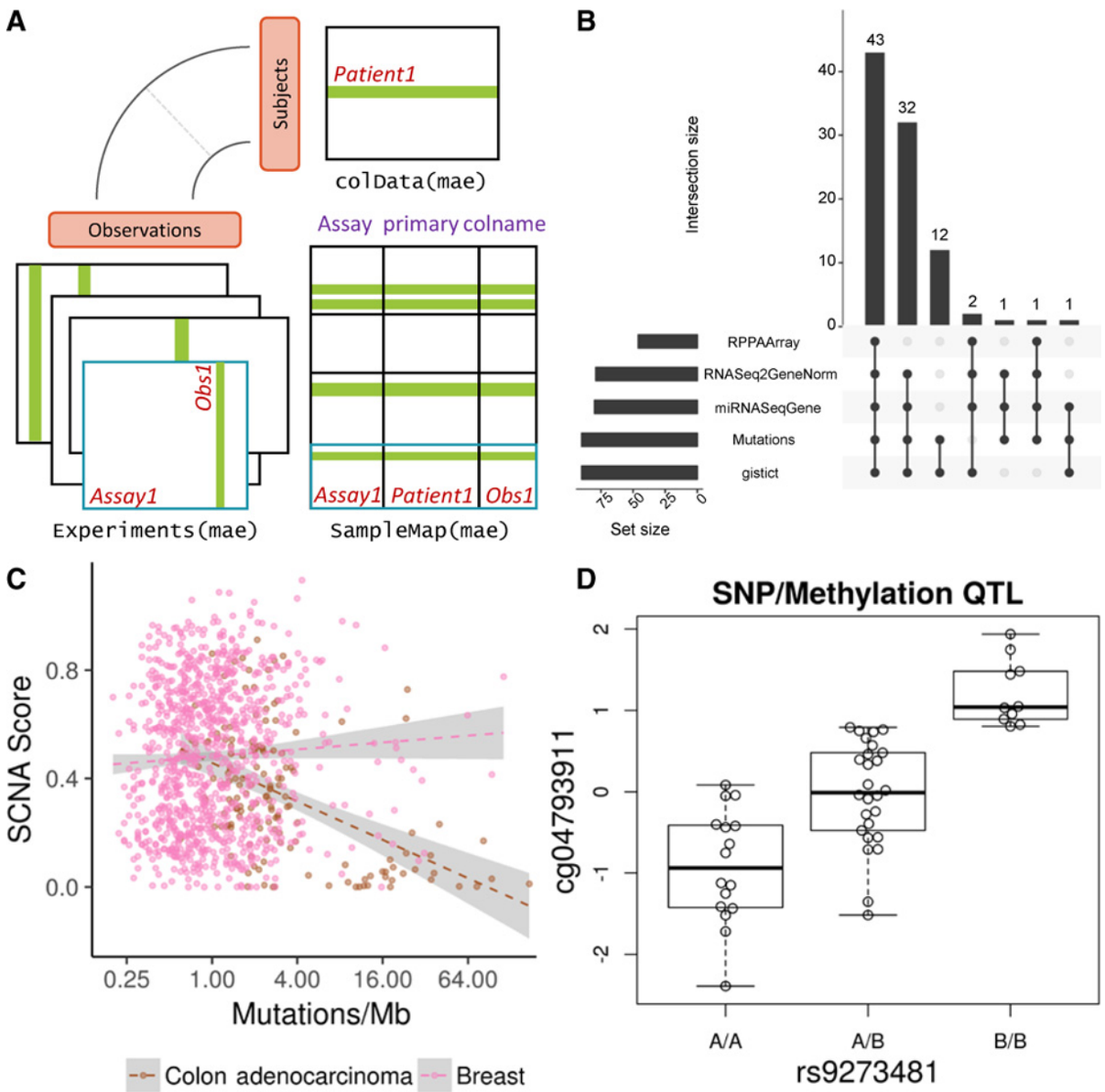


**Figure 1.**

MultiAssayExperiment design and applications. **A,** The MultiAssayExperiment object schematic shows the design of the infrastructure class, detailed in Supplementary Table S3. The colData provides data about the patients, cell lines, or other biological units, with one row per unit and one column per variable. The experiments are a list of assay datasets of arbitrary class, with one column per observation. The sampleMap links a single table of patient data (colData) to a list of experiments via a simple but powerful table of experiment:patient edges (relationships) that can be created automatically in simple cases or in a spreadsheet if assay-specific sample identifiers are used. sampleMap relates each column (observation) in the assays (experiments) to exactly one row (biological unit) in colData; however, one row of colData may map to zero, one, or more columns per assay, allowing for missing and replicate assays. Green stripes indicate a mapping of one subject to multiple observations across experiments. **B,** The UpSetR (9) graphic represents a complex Venn diagram of assay availability for patients in a MultiAssayExperiment. This reduced adrenocortical carcinoma object is provided as an example dataset in the MultiAssayExperiment package. The barplot on the left shows sample size of each experiment; links to its right indicate combinations of one to four experiments, with top bars showing the number of patients having exactly those data types. **C,** Extent of copy number alteration versus somatic mutation burden. Cancer types with high levels of aneuploidy often show a positive correlation of mutation load and chromosomal instability (10), perhaps due to a higher tolerance of deleterious mutations, as shown here in pink for breast cancer. Tumors with a hypermutator phenotype rarely display extensive chromosomal instability, resulting in a negative correlation of mutation load and chromosomal instability in cancer types where hypermutation is common (shown in brown for colon adenocarcinoma). **D,** Methylation quantitative trait locus identified from an on-disk representation of VCF files of the 1000 Genomes Project integrated with 450K methylation array data as a MultiAssayExperiment.

MultiAssayExperiment object per cancer type and demonstrating greatly simplified multiassay analyses with these and other public multiomics datasets.

## Materials and Methods

MultiAssayExperiment (https://bioconductor.org/packages/MultiAssayExperiment) introduces a Bioconductor object-oriented S4 class, defining a general data structure for representing multiomics experiments. This data class has three key components: (i) colData, a "primary" dataset containing patient or cell line–level characteristics, such as pathology and histology; (ii) ExperimentList, a list of results from complementary experiments; and (iii) sampleMap, a map that relates these elements (Fig. 1A). ExperimentList data elements may be of any data class that has standard methods for basic subsetting (single square bracket "[") and dimension names and sizes ["dimnames()" and "dim()"]. Key methods available for manipulating the MultiAssayExperiment data class include:

(i) A constructor function and associated validity checks that simplifies creating MultiAssayExperiment objects while allowing for flexibility in representing complex experiments.
(ii) Subsetting operations allowing data selection by genomic identifiers or ranges, clinical/pathologic variables, available complete data (subsets that include no missing values), and by specific assays.
(iii) Robust and intuitive extraction and replacement operations for components of the MultiAssayExperiment.

The MultiAssayExperiment API is based wherever possible on SummarizedExperiment while supporting heterogeneous multiomics experiments. MultiAssayExperiment design, constructor, subsetting, extraction, and helper methods, as well as methods and code for the examples demonstrated here, are detailed in the Supplementary Methods.

## Results

The MultiAssayExperiment class and methods (Table 1) provide a flexible framework for integrating and analyzing complementary assays on an overlapping set of samples. It integrates any data class that supports basic subsetting and dimension names, so that many data classes are supported by default without additional accommodations. The MultiAssayExperiment class (Fig. 1A) ensures correct alignment of assays and patients, provides coordinated subsetting of samples and features while maintaining correct alignment, and enables simple integration of data types to formats amenable to analysis by existing tools. Basic usage is outlined in

**Table 1.** Summary of the MultiAssayExperiment API

| Category and function | Description | Returned class |
|---|---|---|
| Constructors | | |
| MultiAssayExperiment | Create a MultiAssayExperiment object | MultiAssayExperiment |
| ExperimentList | Create an ExperimentList from a List or list | ExperimentList |
| Accessors | | |
| colData | Get or set data that describe the samples | DataFrame |
| experiments | Get or set the list of experimental data objects as original classes | ExperimentList |
| assays | Get the list of experimental data numeric matrices | SimpleList |
| assay | Get the first experimental data numeric matrix | Matrix, matrix-like |
| sampleMap | Get or set the map relating observations to subjects | DataFrame |
| metadata | Get or set additional data descriptions | List |
| rownames | Get row names for all experiments | CharacterList |
| colnames | Get column names for all experiments | CharacterList |
| Subsetting | | |
| mae[i, j, k] | Get rows, columns, and/or experiments | MultiAssayExperiment |
| mae[i, ,] | GRanges, character, integer, logical, List, list | MultiAssayExperiment |
| mae[,j,] | Character, integer, logical, List, list | MultiAssayExperiment |
| mae[,,k] | Character, integer, logical | MultiAssayExperiment |
| mae[[i]] | Get or set object of arbitrary class from experiments | (Varies) |
| mae[[i]] | Character, integer, logical | |
| mae$column | Get or set colData column | Vector (varies) |
| Management | | |
| complete.cases | Identify subjects with complete data in all experiments | Vector (logical) |
| duplicated | Identify subjects with replicate observations per experiment | List of LogicalLists |
| mergeReplicates | Merge replicate observations within each experiment | MultiAssayExperiment |
| intersectRows | Return features that are present for all experiments | MultiAssayExperiment |
| intersectColumns | Return subjects with data available for all experiments | MultiAssayExperiment |
| prepMultiAssay | Troubleshoot common problems when constructing main class | List |
| Reshaping | | |
| longFormat | Return a long and tidy DataFrame with optional colData columns | DataFrame |
| wideFormat | Create a wide DataFrame, one row per subject | DataFrame |
| Combining | | |
| c | Concatenate an experiment | MultiAssayExperiment |

NOTE: Assay refers to a procedure for measuring the biochemical or immunologic activity of a sample, e.g., RNA-seq, segmented copy number, and somatic mutation calls would be considered three different assays. Experiment refers to the application of an assay to a set of samples. In general, it is assumed that each experiment uses a different assay type, although an assay type may of course be repeated in different experiments. mae refers to a MultiAssayExperiment object. Subject refers to patient, cell line, or other biological unit. Observation refers to results of an assay, e.g., gene expression, somatic mutations, etc. Features refer to measurements returned by the assays, labeled by row names or genomic ranges.

Supplementary Video S1 (https://www.youtube.com/watch?v=w6HWAHaDpyk&feature=youtu.be) and in the QuickStart-MultiAssay vignette accompanying the package.

We coordinated over 300 assays from over 11,000 patients of 33 different cancer types from The Cancer Genome Atlas as one MultiAssayExperiment per cancer type (Supplementary Table S2). These data objects link each assay to their patient of origin, allowing more straightforward selection of cases with complete data for assays of interest, and integration of data across assays and between assays and clinical data. We demonstrate applications of MultiAssayExperiment for visualizing the overlap in assays performed for adrenocortical carcinoma patients (Fig. 1B), confirming recently reported correlations between somatic mutation and copy number burden in colorectal cancer and breast cancer (Fig. 1C), identifying an SNP/methylation quantitative trait locus using remotely stored tabix-indexed VCF files for the 1000 genomes project (Fig. 1D), multiassay gene set analysis for ovarian cancer (Supplementary Figs. S1 and S2), and calculating correlations between copy number, gene expression, and protein expression in the NCI-60 cell lines (Supplementary Fig. S3). Demonstrative code chunks and fully reproducible scripts are given to demonstrate the simple and powerful flexibility provided by MultiAssayExperiment.

## Discussion

MultiAssayExperiment enables coordinated management and extraction of complex multiassay experiments and clinical data, with the same ease of user-level coding as for a single experiment. Its extensible design supports any assay data class meeting basic requirements, including out-of-memory representations for very large datasets. We have confirmed "out-of-the-box" compatibility with on-disk data representations, including the DelayedMatrix class via an HDF5 backend (6), and the VcfStack class based on the GenomicFiles infrastructure. Future work will focus on higher level visualization, integration, and analysis tools using MultiAssayExperiment as a building block. This project will receive long-term support as a necessary element of multiassay data representation and analysis in Bioconductor.

## Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

## Authors' Contributions

**Conception and design:** M. Ramos, L. Schiffer, P. Chapman, D. Gomez-Cabrero, K.D. Hansen, M. Morgan, V. Carey, L. Waldron

**Development of methodology:** M. Ramos, L. Schiffer, T. Chan, P. Chapman, K.D. Hansen, M. Morgan, V. Carey, L. Waldron

**Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.):** M. Ramos, S.R. Davis, H. Kodali, V. Carey

**Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis):** A. Re, R. Azhar, A. Basunia, P. Chapman, S.R. Davis, A.C. Culhane, B. Haibe-Kains, A.S. Mer, M. Riester, V. Carey, L. Waldron

**Writing, review, and/or revision of the manuscript:** M. Ramos, L. Schiffer, A. Re, P. Chapman, S.R. Davis, D. Gomez-Cabrero, A.C. Culhane, B. Haibe-Kains, H. Kodali, A.S. Mer, M. Riester, V. Carey, L. Waldron

**Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases):** M. Ramos, A. Basunia, C. Rodriguez, T. Chan, H. Kodali, M.S. Louis

**Study supervision:** V. Carey, L. Waldron

## Acknowledgments

This work was also supported by the CUNY High Performance Computing Center, which is operated by the College of Staten Island and funded, in part, by grants from the City of New York, State of New York, CUNY Research Foundation, and National Science Foundation grants CNS-0958379, CNS-0855217, and ACI 1126113.

## Grant Support

The authors' work was funded by the NCI of the NIH (U24CA180996 to M. Morgan).

Received February 9, 2017; revised May 30, 2017; accepted July 27, 2017; published online November 1, 2017.

## References

1. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods 2015;12:115–21.
2. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol 2013;9:e1003118.
3. Kannan L, Ramos M, Re A, El-Hachem N, Safikhani Z, Gendoo DMA, et al. Public data and open source tools for multi-assay genomic investigation of disease. Brief Bioinform 2016;17:603–15.
4. Hernandez-Ferrer C, Ruiz-Arenas C, Beltran-Gomila A, González JR. MultiDataSet: an R package for encapsulating multiple data sets with application to omic data integration. BMC Bioinformatics 2017;18:36.
5. Louhimo R, Hautaniemi S. CNAmet: an R package for integrating copy number, methylation and expression data. Bioinformatics 2011;27:887–8.
6. Folk M, Heber G, Koziol Q, Pourmal E, Robinson D. An overview of the HDF5 technology suite and its applications. Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases; 2011 Mar 25; Uppsala, Sweden. New York, NY: ACM; 2011. p. 36–47.
7. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. Bioinformatics 2014;30:2076–8.
8. Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, et al. Bigtable: a distributed storage system for structured data. ACM Trans Comput Syst 2008;26:4.
9. Conway JR, Lex A, Gehlenborg N. UpSetR: An R package for the visualization of intersecting sets and their properties. Bioinformatics; 2017 Jun 22. [Epub ahead of print].
10. Davoli T, Uno H, Wooten EC, Elledge SJ. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science 2017;355:pii:eaaf8399.