Màster Universitari
Anàlisi de Dades Òmiques / Omics Data Analysis
FACULTAT DE CIÈNCIES I TECNOLOGIA
UVIC | UVIC·UCC

Master of Science in Omics Data Analysis

Master Thesis

Bioinformatic tools for Big Data in Omic studies with application to genomic inversion calling and multi-omic data integration

by

Mª Dolors Pelegrí Sisó

Supervisor: [Juan Ramón González Ruíz, Bioinformatics Research Group in Epidemiology, ISGlobal]

Academic tutor: [M. Luz Calle Rosignana, Biosciences Department, University of Vic]

Biosciences Department

University of Vic – Central University of Catalonia

September 2020

doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date:
Manuscript Category

Application Note

Bioinformatic tools for Big Data in Omic studies with application to genomic inversion calling and multi-omic data integration

Dolors Pelegri-Siso [1,2,3], Juan R Gonzalez [2,3],

[1]Universitat de Vic - Universitat Central de Catalunya (UVic)
[2]ISGlobal, Centre for Research in Environmental Epidemiology (CREAL)
[3]Bioinformatics Research Group in Epidemiology (BRGE)

To whom correspondence should be addressed. E-mail: juanr.gonzalez@isglobal.org

Associate Editor: XXX

Received on XXX; revised on XXX; accepted on XXX

Abstract

Motivation: The diversity and huge omics data take biology and biomedicine research and application into a big data era. Most of the current statistical analyses required to analyze omic data are not designed to deal with big data. Principal component analyses and multivariate methods to integrate multi-omic data are one of those examples. Therefore, having efficient and scalable functions are required to exploit the large amount of omic data which is currently available.

Results: We developed a library called *BigDataStatMeth* which includes functions to perform basic matrix operations and linear algebra for big matrices using HDF5 and DelayedArray Bioconductor's infrastructure. We tested its performance by comparing the computational time with the one obtained with R base functions. Our results showed that our implementation outperforms existing functions and that the improvement increases when sample size is also increasing. This package can be the basis for implementing statistical methods required in omic data with large number of samples or features. As a proof-of-concept, we implemented PCA and Lasso regression within the same package and we also created another Bioconductor package, *mgcca*, which implements Generalized Canonical Correlation Analysis (GCCA) that is used in multi-omic data integration. We implemented an algorithm that allows the possibility of having missing individuals in one or more tables. The implemented methods have been used to analyze real omic data. We first used PCA to call genotype inversions of more than 400K individuals from UKBiobank. Then, data from TCGA was used to integrate multiple omic layers using GCCA.

Availability: Both packages are available at BRGE's GitHub repository: `https://github.com/isglobal-brge`

Contact:juanr.gonzalez@isglobal.org

Supplementary information: We have four supplementary material files. One of them (Supp_Mat.pdf) includes supplementary information about the methods used in this work as well as suplementary tables and figures corresponding to the benchmarking. The other three files correspond to one vignete describing *BigDataStatMeth* package another for *mgcca* and a last one having a real data example to integrate multi-omic data using GCCA.

1 Introduction

The diversity and huge omics data take biology and biomedicine research and application into a big data era. Encouraged by constant cost reduction, data-sharing initiatives, and availability of public data large amounts of genomic, transcriptomic, and other omics data have become available ready to be analyzed. Therefore, omics data analyses require scalable and computationally efficient algorithms. These include methods to analyze a single omic dataset as well as methods to get an integrated view from the different samples and omic tables with the aim of getting an accurate and comprehensive view about the diseases and different biological processes (Subramanian *et al.*, 2020).

Multivariate methods designed to analyze one or several high-dimensional datasets are widely used in omics data analyses. For instance, Principal Component Analysis (PCA) through Single Value Decomposition (SVD) has been used in genomics to address population stratification (Price *et al.*, 2006)(Price *et al.*, 2010a) or to genotype polymorphic inversions (Cáceres and González, 2015a). In transcriptomics and epigenomics SVD is used to estimate surrogate variables that is required to correct for bath effect (Leek *et al.*, 2012) or to estimate cell proportions (Houseman *et al.*, 2014)(Alquicira-Hernández *et al.*, 2018). On the other hand, Canonical Correlation Analysis (CCA) (Hotelling, 1936), Generalized Canonical Correlation (GCCA) (Kettenring, 1971), multi factorial analysis (MFA) (Abdi *et al.*, 2013), co-inertia analysis (CIA) (Culhane *et al.*, 2003) and Multi-Omics Factor Analysis (MOFA) (Argelaguet *et al.*, 2018) have been used to perform the multivariate analysis of multiple omic tables (see (Subramanian *et al.*, 2020) and (Csala A, 2019) for a review of these methods applied to multi-omic data integration).

Bioconductor includes several packages that are designed to performed most of these multivariate methods such as BiocSinular (Lun, 2020), PCAtools (Blighe and Aaron, 2020), SVA (Leek *et al.*, 2019) for surrogate variables and omicade4 (Meng *et al.*, 2014) which implements multiple co-inertia analysis. Most of these implementations, mainly those created to integrate multiple tables, were not designed to be deal with big data sets, and hence, they are not computationally efficient. Another limitation that current multivariate methods have to integrate multi-oimc data is how to handle missing data individuals. The presence of missing values in multi-omics data is inevitable since, in most cases, omic data are obtained in different time points and quality control can remove individuals form a single omic dataset. Most of the currently implemented methods works only with complete cases which is an underpowered approach (van de Velden and Takane, 2012). *missRows* is a Bioconductor library based on multi factorial analysis that addressed this issue (I and V, 2020) . However, the analysis of large datasets cannot be performed using this method. Actually, the authors recommend to filter out those features with less variability to reduce the dimensionality (Voillet *et al.*, 2016).

Most of the existing inefficient implementations are due to the fact that developers use base R functions to perform basic matrix operations or algebra. In order to overcome these difficulties and provide the user with efficient and scalable functions to implement any statistical method required to analyzed large omic dataset, we have developed BigDataStatMeth Bioconductor library that uses C++ language with Rcpp (Eddelbuettel and François, 2011) (Eddelbuettel and Balamuta, 2017) and RcppEigen (Bates and Eddelbuettel, 2013) from R-CRAN that provides an efficient tool for process and analyze omics data. BigDataStatMeth also works with HDF5 file format (Koranne, 2011) (Fischer *et al.*, 2019) and Delayed Arrays (Pagès *et al.*, 2020) directly from C++ using APIs developed from Bioconductor and other specific to C++. The implemented algorithms also use parallel methods that will make our functions scalable using OpenMP - OMP (Dagum and Menon, 1998) This library will allow us to implement a new package called mgcca that will allow us to analyze multi-omic data using GCCA including a method that is designed to analyze data with missing individuals (Velden and Bijmolt, 2006),(van de Velden and Takane, 2012).

In order to demonstrate the usability and the good performance of our proposed method, a benchmark analysis is performed to compare the behavior of our functions with those implemented in base R. Our libraries also include a vignette where the use of the functions is shown in a practical way as well as a brief theoretical explanation. The practical use is illustrated using data from two public databases: the UK Biobank (UKB) (Sudlow *et al.*, 2015) and The Cancer Genome Atlas (TCGA) (Tomczak *et al.*, 2015). The UKB data is used to perform inversion calling in about 500K samples by using PCA. The TCGA dataset is used to illustrate how to integrate different omic data when each table have information on different individuals.

2 Methods

2.1 Databases

The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov), is a project, to catalogue genetic mutations responsible for cancer, using genome sequencing and bioinformatics. TCGA began as a three-year pilot in 2006 with an investment from the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI). The TCGA pilot project proved that making the data freely available would enable researchers anywhere around the world to make and validate important discoveries. In our case, the TCGA data serves as an illustrative example of how to analyze big genomic datasets using our scalable algorithms. TCGA has one of the largest collections of multi-omics and clinical data sets for more than 33 different tumour types chosen because of their poor prognosis and availability of samples. The project contains molecular data from multiple types of assays including DNA and RNA sequencing, array-based expression and DNA methylation among others. Several pre-processed omic data tables are available for each tumour. Clinical data along with molecular tables can be used to decipher the role of different omics in cancer survival, prognosis; to find biomarkers of treatment response; or to determine individuals with different multi-omics profiles that can be used in personalized medicine.

TCGA data was downloaded using TCGAutils package (Ramos *et al.*, 2020). Gene annotations were made using the Bioconductor package biomaRt (Durinck *et al.*, 2009), (Durinck *et al.*, 2005) using the ensembl database (Yates *et al.*, 2020) and the hsapiens_gene_ensembl dataset. CpGs annotations were made using 450K Human Illumina methylation dataset.

UK Biobank

UK Biobank (https://www.ukbiobank.ac.uk) is an international health resource supported by the UK National Health Service (NHS). UK Biobank aims to improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses – including cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia.

The UK Biobank is a prospective cohort of 502,536 adults aged between 40 and 69 years (229,182 men and 273,474 women). At recruitment, participants provided electronic signed consent, answered questions on socio-demographic, lifestyle and health-related factors, and completed a range of physical measures. They also provided blood, urine and saliva samples, which were stored in such a way as to allow many different types of assay to be performed (for example, genetic, proteomic and metabonomic analyses). Once recruitment was fully underway, further enhancements were introduced to the assessment visit, including a range of eye measures, an electrocardiograph test, arterial stiffness and a hearing test. The baseline information has been, and will continue to be, extended in several ways, for example, repeat assessments are planned to be conducted in subsets of the cohort every few years. Here we are focusing on a total of 488,377 individuals with European ancestry to whom inversion calling will be performed using PCA (Cáceres and González, 2015a).

2.2 Library Implementation

In omics data, we used to deal with large datasets with thousands of variables and a small/moderate number of samples. Currently, this paradigm has become even more challenging since we also have information for thousands of individuals. The analysis of this data requires a great amount of computational resources and optimized algorithms. Bioconductor is software project for the analysis and comprehension of genomic data generated by wet lab experiments in molecular biology. It is based primarily on the statistical R programming language, but does contain contributions in other programming languages.

The challenges of dealing with big data sets in Bioconductor are those found in R. By default, R runs only on data that can fit into your computer's memory that is the biggest issue that researchers face when trying to use Big Data in R. Another big issue for doing Big Data work in R is that data transfer speeds are extremely slow relative to the time it takes to actually do data processing once the data has transferred. Finally, R is an interpreted programming language which means that it is not translated into machine language in a process prior to execution. R has a process that interprets the code in real time, this affects the efficiency at execution time and sometimes R code is not as fast as you would expect. Nevertheless, there are effective methods for working with big data in R that will allow the efficient and scalable implementation of omic data analyses. These include methods to: 1) program in low level language (C or C++), 2) work directly on disk and load in memory only the required data to be analyzed, and 3) implement parallel algorithms.

We have developed a Bioconductor package, *BigDataStathMeth*, based on previous methodologies that allows efficient and scalable computation of matrix operations and basic algebra required to implement statistical method in omic data analyses. First, in order to optimize the use of available resources for data processing, we have use Eigen (Guennebaud *et al.*, 2010) which is a C++ template library for linear algebra, matrices, vectors, numerical solvers, and related algorithms using RcppEigen package (Bates and Eddelbuettel, 2013). In conjunction with eigen we used LAPACK (Linear Algebra Package) which is a library for numerical linear algebra with low-level functions (Anderson *et al.*, 1999). Second, in order to optimize memory usage, Hierarchical Data Format (HDF) (Fortner, 1998) in its version 5 (HDF5) (Koranne, 2011) and Delayed Array data objects (Pagès *et al.*, 2020) were used. HDF5 and Delayed arrays allow the effective management of extremely large and complex omics data collections including genomic, RNA-seq, methylation, copy number, mutations or microRNA among others. It also allows to deal with other type of data and metadata associated with an assay like clinical or pathological data. The link between Delayed Arrays and C++ functions have been performed using *beachmat* library. Finally, in order to speed up computation, OpenMP (Dagum and Menon, 1998) have been used to implement parallel algorithms. Section 1 in Supplementary Material provides a general overview of these methodologies including some figures describing two methods to parallelize matrix multiplication and single value decomposition (SVD) (Figures S1 and S2).

2.3 Omics data Analysis - Statistical methods

Methods implemented in *BigDataStatMeth* can be used to program efficient and scalable statistical methods required in omics data analysis. Principal Component Analysis (PCA) is one of the widely used methods in several omic data analyses. PCA can serve, not only as a dimensional reduction technique, but also to visualize cluster of individuals that are created given a hidden structure. For instance, PCA has been used for a long time in population genetics studies to produce maps summarizing human genetic variation across geographic regions(Menozzi *et al.*, 1978). recently it can be used to explore the potential of disease identification in high dimensional blood microRNA data (SL *et al.*, 2020) or to cluster subjects depending on their genomic inversion status (Cáceres and González, 2015b). In *BigDataStatMeth* we have implemented an scalable and efficient function to perform PCA using our parallel implementation of SVD that allows, among other, to visualize cluster of individuals given their population of origin (Price *et al.*, 2010b) to address population stratification or to perform inversion calling (Cáceres and González, 2015b) in very large datasets such as UK Biobank.

The study of multi-omics data and its integration with clinical data has become an active line of research since it can provide useful insights into the cellular functions and help understanding the complex underlying biology (Canzler *et al.*, 2020). There are different multi-omics data integration methodologies, but these methodologies do not cover all possible cases (Tarazona *et al.*, 2020). Dimension reduction techniques have been proposed as a promising tool for the integrative analysis of multi-omics data (Meng *et al.*, 2016). These methods include Generalized Canonical Correlation (GCCA), Multi Factorial Analysis (MFA) or Multiple Co-Inertia Analysis (MCIA) among others unsupervised methods that are reviewed and evaluated in a recent paper (Pierre-Jean *et al.*, 2019). Multi-Omics Factor Analysis (MOFA) is a recent promising approach based on a factor analysis model that provides a general framework for the integration of multi-omics data sets (Argelaguet *et al.*, 2018). Section 2 in our Supplementary Material provides a global overview of these methods along with key references, R/Bioconductor packages and reviews.

These methodologies used to integrate multi-omic data are not properly implemented in R when dealing with big data. In order to overcome this difficulty, we have created another package, *mgcca* which implements GCCA using Delayed Arrays that allows the efficient access to multi-omic data. It considers as input *MultiAssayExperiment* objects which provides for the coordinated representation of, storage of, and operation on multiple diverse omic data in Bioconductor (Ramos *et al.*, 2017). Additionally, the presence of missing information for some individuals in multi-omics data is another important issue. It is normal to have individuals with no information in some tables since each omic is normally obtained independently and also because quality control may remove information from some individuals in a given table. To our knowledge, there is only one method to integrate multiple omic datasets using a MFA (MI-MFA) (Voillet *et al.*, 2016). This method generates multiple imputed datasets from a MFA model, then the yield results are combined in a single consensus solution but is extremely time consuming. Also, it is not properly implemented to deal with large datasets. The authors state that before using MI-MFA it is recommended to remove those variables with low variability in order to reduce the computational burden. In *mgcca* we have implemented a version of GCCA when having missing individuals that is a generalization of the Test Equating method available for PCA that only requires to perform matrix operations which are implemented in *BigDataStatMeth* (van de Velden and Takane, 2012).

3 Results

3.1 Bioconductor libraries

We developed a library called *BigDataStatMeth* which includes functions to perform basic matrix operations and linear algebra for big matrices using HDF5 and DelayedArray Bioconductor’s infrastructure. This package can be the basis for implementing statistical methods required in omic data with large number of samples or features. As a proof-of-concept, we implemented PCA and Lasso regression within the same package. We also created another Bioconductor package, *mgcca* which implements GCCA to be used in multi-omic data integration allowing the possibility of having missing individuals in one or more tables.

***BigDataStatMeth* library**

The methods described in previous section were used to implement *BigDataStatMeth* library that allow us to work in a efficient manner with omic data. *BigDataStatMeth* aims to be a practical, versatile and easy-to-use tool for researchers and omic data analysts. In omics data we can have different assays for the same set of samples with different omics data. For that reason, *BigDataStatMeth* allows the user to store different omic data in on same file. Additionally, *BigDataStatMeth* internally store all the results obtained with omics data in the same file as original data in an organized way. Figure 1 shows an example of how data is organized within *BigDataStatMeth* using a hdf5 file.

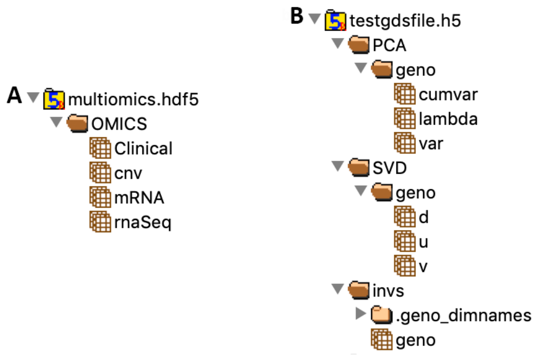
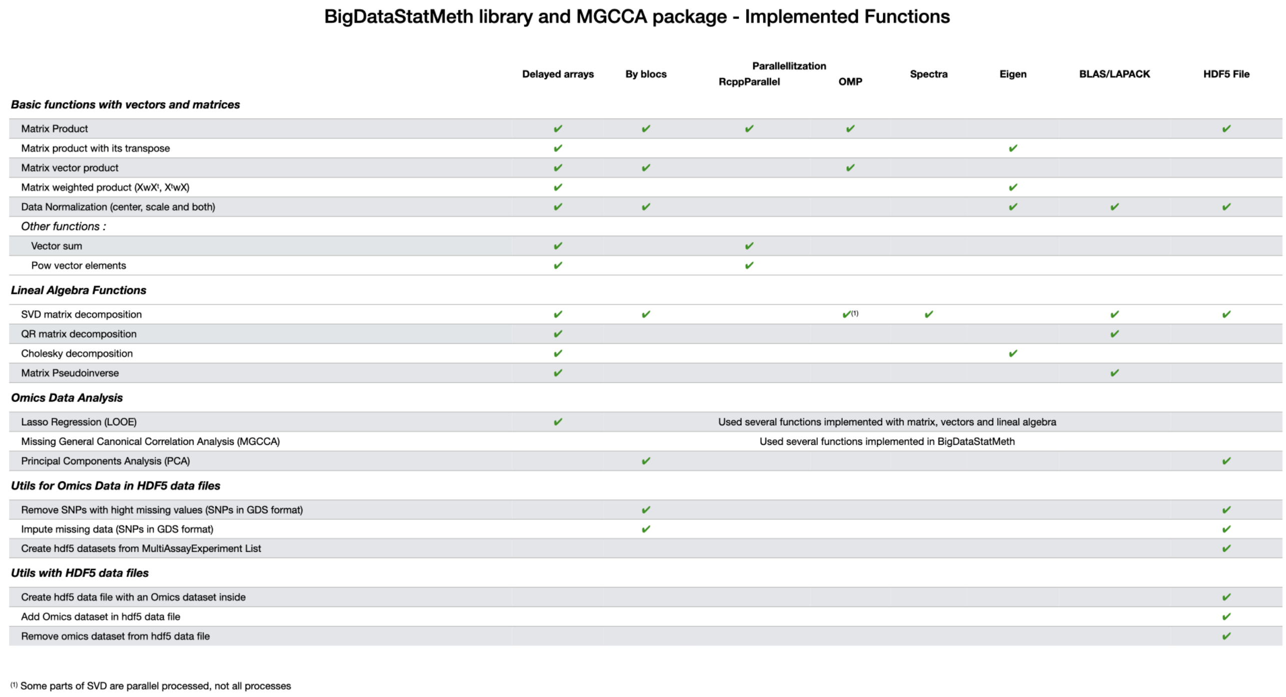


Fig. 1. Data structure inside hdf5 file. Figure A shows how data can be stored inside hdf5 file where all omics data can be under the same group. Figure B shows how the results from an omic datasets are stored using BigDataStatMeth. The results are saved in special groups in the same file as the original omic data for ease of use and allow for reuse and sharing. Here we have results stored in PCA and SVD folders and results for genomic inversions are store at invs folder joint with the original data (geno)

As stated in methods section, due to large amount of information associated with omic data, working with data blocks is a key issue. One of the biggest challenges in implementing *BigDataStatMeth* has been working with small blocks of data because: a) we have to take in to account at each moment the precise coordinates where we have to start to read and the exact block size to read; and b) not all operations can be performed in, since existing algorithms are highly complex. To the challenge of working in blocks we must add the complexity of working directly on files. It is possible when working with data in memory compute calculations directly with the complete dataset, but working with blocks of data from files only allows to have in memory the last read block. This requires to know the data we have on memory and which data we need to load or unload to



BigDataStatMeth library and MGCCA package - Implemented Functions

			Parallellitzation					
	Delayed arrays	By blocs	RcppParallel	OMP	Spectra	Eigen	BLAS/LAPACK	HDF5 File
Basic functions with vectors and matrices								
Matrix Product	✓	✓	✓	✓				✓
Matrix product with its transpose	✓					✓		
Matrix vector product	✓	✓		✓				
Matrix weighted product (XwX^t, X^twX)	✓					✓		
Data Normalization (center, scale and both)	✓	✓				✓	✓	✓
Other functions :								
Vector sum	✓		✓					
Pow vector elements	✓		✓					
Lineal Algebra Functions								
SVD matrix decomposition	✓	✓		✓(1)	✓		✓	✓
QR matrix decomposition	✓						✓	
Cholesky decomposition	✓					✓		
Matrix Pseudoinverse	✓						✓	
Omics Data Analysis								
Lasso Regression (LOOE)	✓		Used several functions implemented with matrix, vectors and lineal algebra					
Missing General Canonical Correlation Analysis (MGCCA)			Used several functions implemented in BigDataStatMeth					
Principal Components Analysis (PCA)		✓						✓
Utils for Omics Data in HDF5 data files								
Remove SNPs with hight missing values (SNPs in GDS format)		✓						✓
Impute missing data (SNPs in GDS format)		✓						✓
Create hdf5 datasets from MultiAssayExperiment List								✓
Utils with HDF5 data files								
Create hdf5 data file with an Omics dataset inside								✓
Add Omics dataset in hdf5 data file								✓
Remove omics dataset from hdf5 data file								✓

(1) Some parts of SVD are parallel processed, not all processes

Fig. 2. Summary table with the functions implemented in BigDataStatMeth and the methods and types of resources used to be implemented.

compute calculations. This is needed to avoid overloading memory and worsening the overall performance of the function.

In *BigDataStatMeth* we have developed different functions to work with omic data. The implemented functions can be classified into five groups 1) basic functions with vectors and matrices; 2) linear algebra functions 3) methods for omics data analyses; 4) pre-processing data analyses; and 5) utilities to work with HDF5 data files that allows basic omics data organization. Figure 2 provides a detailed summary with all the implemented functions and methodologies for each group.

We have creates a reproducible vignette illustrating how to use *BigDataStatMeth* that can be downloaded from (https://github.com/isglobal-brge/BigDataStatMeth/blob/master/vignettes/). The vignette document explains in detail how *BigDataStatMeth* operates. The vignette also contains detailed examples using real datasets as well as some benchmarking to compare its performance with other existing approaches.

mgcca library

BigDataStatMeth can be easily used to implement any statistical method that requires any of the basic functions described in Figure 2 when analyzing omic data. As a proof-of-concept, we have implemented functions to integrate multi-omic data using GCCA as well as an algorithm that allow integrating multiple omic tables having missing individuals (van de Velden and Takane, 2012). The library can be found in (https://github.com/isglobal-brge/mgcca). The package also includes a vignette having a complete pipeline to integrate transcriptomic and epigenomic data from TCGA using *MultiAssayExperiment* object which is the default method to handle with multi-omic data in Bioconductor.

3.2 Real data analyses

We have applied the implemented methods in *BigDataStatMeth* and *mgcca* to analyze real omics data. We first used PCA to call genotype inversions of more than 400K individuals from UKBiobank. Then, data from TCGA was used to integrate multiple omic layers using GCCA.

PCA with UKBiobank omics dataset

The PCA implementation in *BigDataStatMeth* was used to call polymorphic inversions in to well known inversions located at 8p23.1 and 17q21.31. We recall that a genomic inversion is a specific DNA interval that runs backward with respect to a reference genome and that it can be genotype from SNP data using bioinformatic tools based on PCA methodologies (Cáceres and González, 2015a)

Genomic data between coordinates 8,055,789 to 11,980,649 in chromosome 8 was obtained from UKB data. This data was downloaded in Genomic Data Structure format (GDS), and treated with gdsfmt library (Zheng *et al.*, 2012)(Zheng *et al.*, 2017). *BigDataStatMetht* has functions to transform GDS data to hdf5 datasets. Initially, data for 1,591 SNPs and 488,377 samples were available. We perform a quality control step by removing those individuals having more than 10% of missing data (n=1411) letting the total samples in 486966. The remaining missing data were imputed using the observed allele frequency at each SNP. After

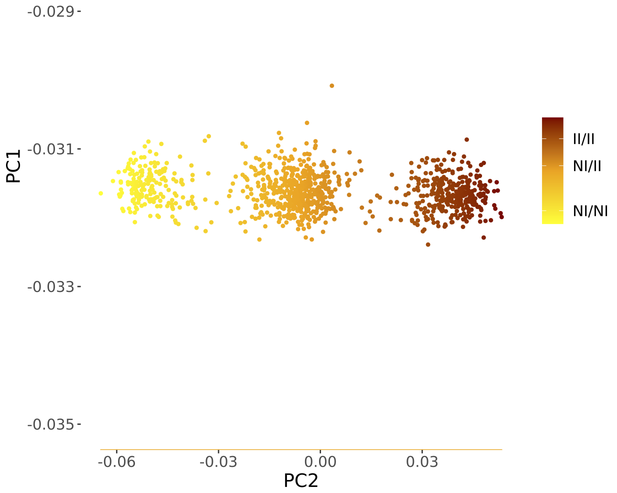


Fig. 3. Inversion calling at 8p23.1 from UK Biobank data

removing individuals with high percentage of missing values, we impute the rest of the missing values. Figure 3 shows the results of the calling procedure where a perfect clustering is observed.

The same procedure was applied for inversion at 17q21.31.In that case, we selected coordinates 43,661,775 to 44,372,665 from chromosome 17 that includes 462 SNPs. In this case, 15,147 samples were removed with due to the large number of missing information letting the total samples in 473230. Figure 4 depicts the results for this inversion.

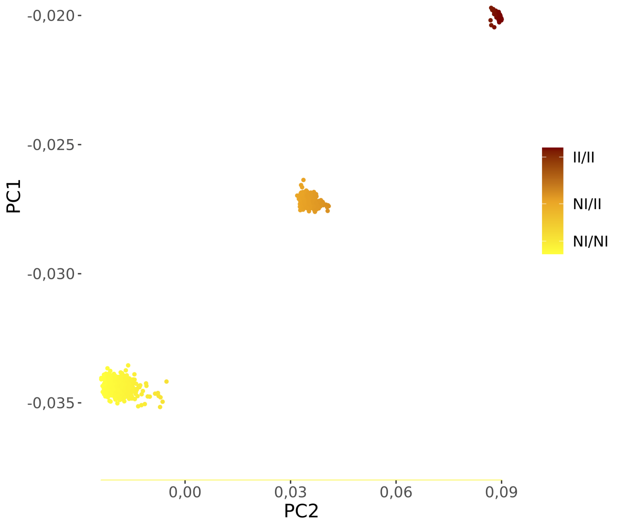


Fig. 4. Inversion calling at 17q21.31 from UK Biobank data

GCCA with TCGA data

GCCA method with missing individuals was used to analyzed data from TCGA. Our aim was to illustrate how to perform multi-omic data integration with GCCA. Bioconductor library curatedTCGAData (Ramos *et al.*, 2018) was used to download the data, this library provides available data from TCGA as a MultiAssayExperiment object. We illustrate how to perform multi-omic data integration at a whole genome level and then, in a region of interest given by a specific genomic coordinate range that the researcher may be interested in.

We downloaded data from Adrenocorticoical carcinoma (ACC) which is a rare endocrine malignancy. In particular, we analyzed RNA-seq (Normalized) and Methylation data. One of our supplementary materials also available at ((https://github.com/isglobal-brge/mgcca)) contains a complete description of how to get this data. The downloaded *MultiAssayExperiment* encapsulates gene expression (RNA-seq) data for 79 samples and 20,501 genes, while methylation data has data from 80 samples and 485,577 CpGs. We first imputed missing values from methylation data (gene expression did not have any). To this end, we use `impute` function implemented in *mgcca* package that uses a knn algorithm with 10 neighbours. This function is a wrapper of `impute` package form Bioconductor adapted to *MultiAssayExperiment* objects. GCCA analyses revealed a total of 3,906 genes and 6,335 CpGs associated with either first or second global axis with a false discovery rate (FDR) lower that 1% (Figure 5 A and B). In order to interpret the axes we can project the scores of the individuals and color them using any illustrate variable. Figure 5 C depicts the individuals given their vital status. We are aware that this analysis would deserve another type of method (i.e. survival analyses) but we are using it as an illustrative example. We can observe as individuals who died are located in the top-right part of the figure. Therefore, features associated with those axes will be important for survival status. The top five genes related to survival are Protein Coding genes. Main genes are shown in the top right part of the Figure 5 A. These protein coding genes are the Lysine Demethylase 4B (KDM4B), the Poly (ADP-Ribose) Polymerase 2 (PARP2), the Nicalin (NCLN), the Autophagy Related 4D Cysteine Peptidase (ATG4D) and the RNA Exonuclease 1 Homolog (REXO1). Three of the five most statistically significant CpGs are located near a Protein Coding gene. These CpGs are cg00161225, cg00330929 and cg00256231 that are near Purinergic Receptor P2X 1 (P2RX1), Complement C1q Like 1 (C1QL1) and TBC1 Domain Family Member 16 (TBC1D16) respectively. The fourth CpG is cg00362657 located near pseudokinase PEAK3/C19orf35 and the last is cg00164949 with an unknown near gene. Section 3 (tables S1 and S2) in our Supplementary Material provides an extended annotated list about top significant genes and CpGs obtained with mgcca associated with either first or second global axis. Figure 5 shows the main results of this analysis.

Adrenocortical tumors occur as sporadic tumors, as part of the multiple endocrine neoplasia type 1 (MEN1), syndrome or as part of other hereditary disorders (Heppner *et al.*, 1999), (Griniatsos *et al.*, 2011), (Wang

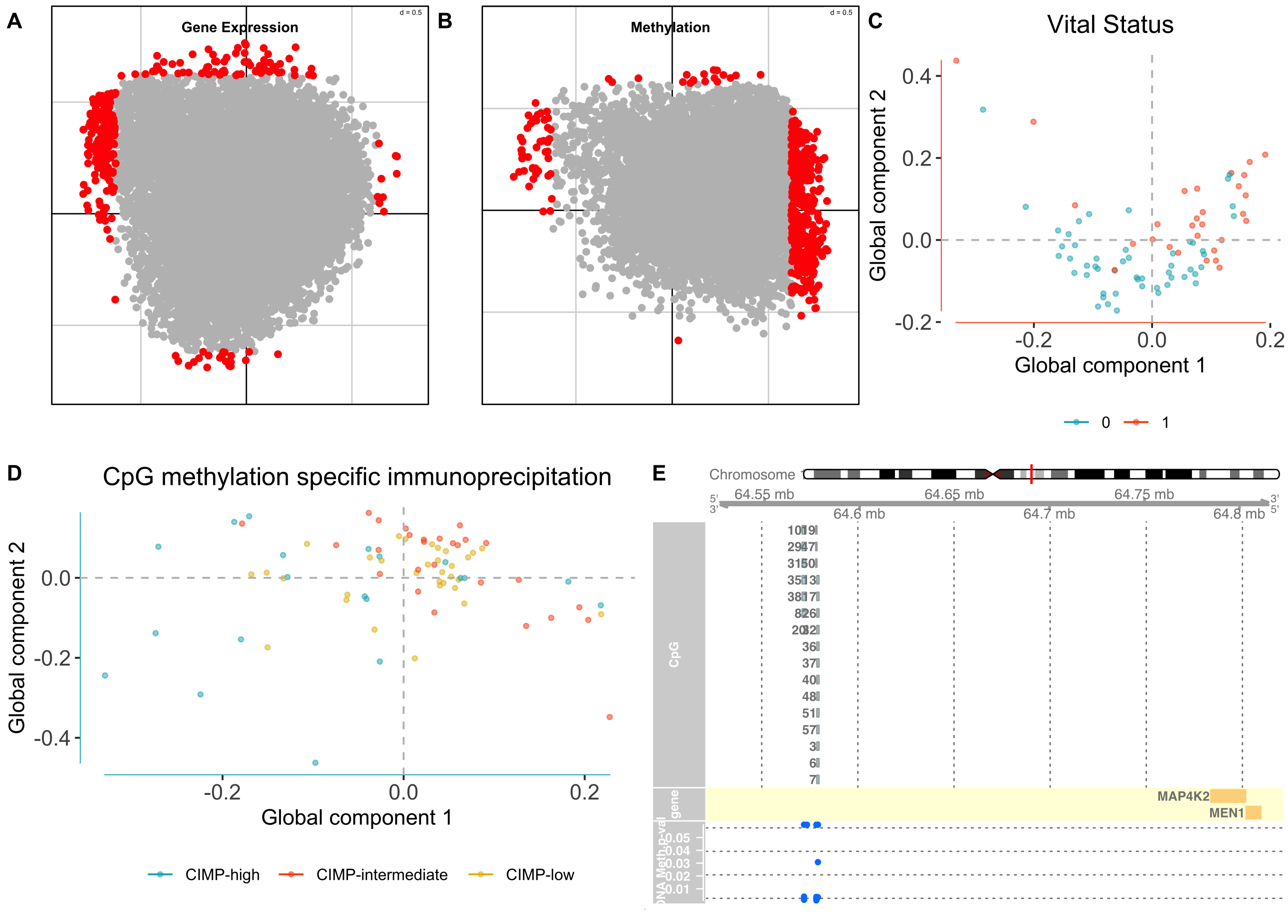


Fig. 5. GCCA with Adrenocortical carcinoma results obtained with implemented mgcca package. Figures A, B and C refer to the analysis with complete data, in Figure A, in red we observe the statistically significant genes at 10^{-9} significance level. In Figure B plot refers to significant CpGs detected 10^{-9} significance level. In Figure C we can observe how samples are distributed in two clusters attending to clinical variable vithal status. Figures D and E refer to the data analysis performed on the genomic coordinates related to the MEN1 gene. In figure D we show how samples are distributed taking in to account methylation level, high methylated samples are disperse distributed but major intermediate and low methylated samples are centered, in figure E we show significant genes detected near MEN1 and his location on chormosome 11 we also show the p-values obtained for CpGs and genes location (MEN1 and MAP4K2)

et al., 2019). Menin 1 gene (MEN1), is a tumor-suppressor gene located on chromosome 11q13 with genomic coordinates (11:64570986-64578766). Therefore, researchers may be interested in performing analyses in that region to find new biomarkers of ACC. We use the MEN1 gene coordinates with 2kb upstream and downstream for subsetting features in both RNA-seq and methylation. We have data for MEN1 and mitogen-activated protein kinase kinase kinase kinase 2 (MAP4K2), and 104 CpGs. The GCCA analysis in that region ended up with 19 statistically significant CpGs close to MEN1 genomic coordinates that are associated with the two first global components (Figure 5 D and E). Figure 5 D depicts the CpG methylation specific immunoprecipitation in each individual. We can observe as individuals with CpG island methylator phenotype-low (CIMP-low) are located in the top right part and slightly extended to the left in figure. Individuals with CpG island methylator phenotype-intermediate (CIMP-intermediate) are located in the top right part and slightly extended to the bottom right. Finally, individuals with CpG island methylator phenotype-high (CIMP-high) are located in the left part. The significant CpGs related to CIMP-low and CIMP-intermediate are those CpGs depict mainly in right part of Figure 5 C and those significant CpGs related to CMP-high are those CpGs depict on left part in the figure. Annotated list of genes and CpGs can be found in section 3 (tables S3 and S4) in our Supplementary Material.

3.3 Benchmarking

We tested the performance of some of the functions implemented in *BigDataStatMeth* with respect to those implemented using the basic functions or even more advanced in R. To perform the benchmark, we use the microbenchmark function (Mersmann, 2019), a program or routine to measure and test the performance of a single component or task, this function is implemented in microbenchmark package available in CRAN. The device used for the benchmark was an iMac with a quad-core i5 processor (I5-6500) at 3.2GHz, 24Gb 1867 MHz DDR3 of RAM and a fusion disk dive (hybrid drive that combines a hard disk drive with a NAND flash storage)

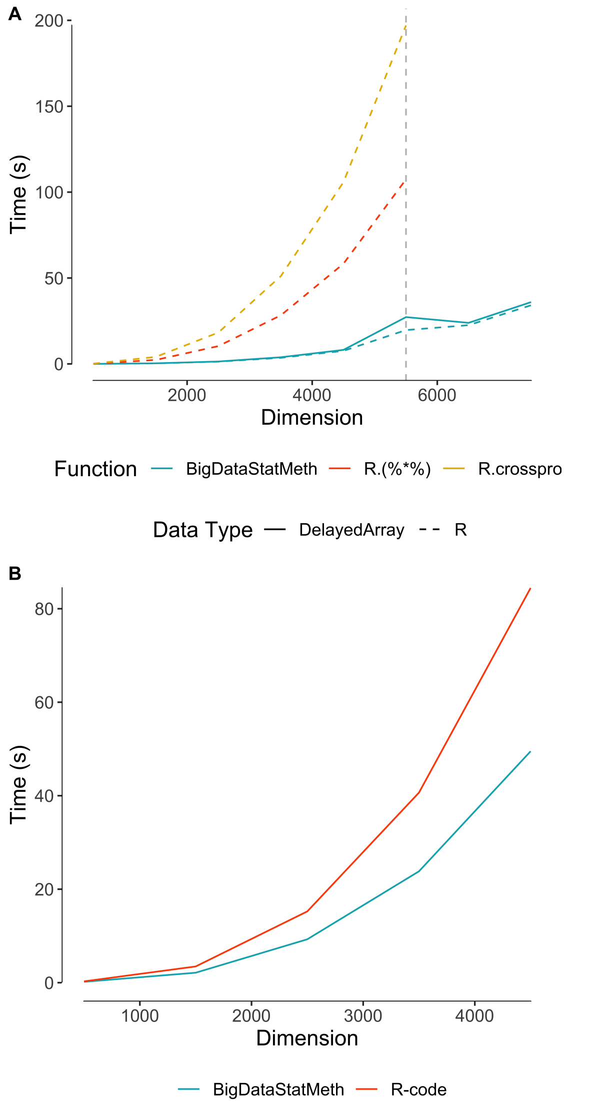


Fig. 6. Graphic performance for different functions implemented in BigDataStatMeth and R. Panel A compares matrix multiplication times using BigDataStatMeth's functions that uses C++ with basic R multiplication function and crossproduct function implemented in R. Panel C compares the computing time using mgcca function implemented in R and in BigDataStatMeth. The multiplications with R were carried out until reaching the data size of 4500x4500, the operations implemented in BigDataStatMeth continued to be applied until reaching a data size of 7500x7500. It was observed that the execution time with BigDataStatMeth was still much lower than that obtained with the R functions

Figure 6 compares the performance for some of the implemented method. More results can be found in **Benchmarking in Section 4 in our Supplementary Material**. In general, our implementations outperformed others available in R. We can also observe that the improvement increases when data dimension increases.

4 Conclusion

Omics technologies are bringing a revolution in transforming the medicine and the health care sector, especially with regard to personalized medicine. This work is only a very basic approximation of the tool that can be developed to support omics data analysis, and to advance in the field of personalized medicine where are needed tools capable of analyzing big data efficiently and accurately in a few seconds. With methods and technologies applied in BigDataStatMeth it has been seen that there are important improvements in terms of performance and system resource management that can help in personalized medicine to obtaining the results derived from omics data analysis with effectively and accurately results.

Considering that *BigDataStatMeth* is a scalable library, future work would go through creating more functionalities adapted to new requirements and methods in omic data analysis field that allow progress in biomedicine and the personalized medicine. Some of these features would be a) managing the missing data in different ways in order to prevent distorted results in omics analysis, b) creating more statistical methods able to analyze the distinct omics and multi-omics datasets from distinct perspectives and c) generating complex plots from omics analysis to help the scientific and medical community to understand the complex underlying biology.

Acknowledgements

I would like to thank the supervisor of this project, Juan Ramón González, for his great interest, patience, motivation and help during all these months and for giving me the opportunity to be a part of ISGlobal. And i would like to thank to my family and friends, specially to Jordi Amenós for all the unconditional support this year.

References

Abdi, H., Williams, L. J., and Valentin, D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary reviews: computational statistics*, **5**(2), 149–179.

Alquicira-Hernández, J., Nguyen, Q., and Powell, J. E. (2018). scpred: Single cell prediction using singular value decomposition and machine learning classification. *bioRxiv*, page 369538.

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., and Sorensen, D. (1999). *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*, **14**(6), e8124.

Bates, D. and Eddelbuettel, D. (2013). Fast and elegant numerical linear algebra using the RcppEigen package. *Journal of Statistical Software*, **52**(5), 1–24.

Blighe, K. and Aaron, L. (2020). *PCAtools: Everything Principal Components Analysis*.

Cáceres, A. and González, J. R. (2015a). Following the footprints of polymorphic inversions on snp data: from detection to association tests.

Nucleic Acids Res, **43**(8), e53.

Cáceres, A. and González, J. R. (2015b). Following the footprints of polymorphic inversions on snp data: from detection to association tests. *Nucleic acids research*, **43**(8), e53–e53.

Canzler, S., Schor, J., Busch, W., Schubert, K., Rolle-Kampczyk, U. E., Seitz, H., Kamp, H., von Bergen, M., Buesen, R., and Hackermüller, J. (2020). Prospects and challenges of multi-omics data integration in toxicology. *Archives of Toxicology*, pages 1–18.

Csala A, Z. A. (2019). *Multivariate Statistical Methods for High-Dimensional Multiset Omics Data Analysis.*, volume Chapter of *Computational Biology [Internet]*. Husi H, editor.

Culhane, A. C., Perriere, G., and Higgins, D. G. (2003). Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, **4**(1), 59.

Dagum, L. and Menon, R. (1998). Openmp: An industry-standard api for shared-memory programming. *IEEE Comput. Sci. Eng.*, **5**(1), 46–55.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols*, **4**, 1184–1191.

Eddelbuettel, D. and Balamuta, J. J. (2017). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *PeerJ Preprints*, **5**, e3188v1.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, **40**(8), 1–18.

Fischer, B., Pau, G., and Smith, M. (2019). *rhdf5: R Interface to HDF5*. R package version 2.30.1.

Fortner, B. (1998). Hdf: The hierarchical data format. *Dr Dobb's J Software Tools Prof Program*, **23**(5), 42.

Griniatsos, J. E., Dimitriou, N., Zilos, A., Sakellariou, S., Evangelou, K., Kamakari, S., Korkolopoulou, P., and Kaltsas, G. (2011). Bilateral adrenocortical carcinoma in a patient with multiple endocrine neoplasia type 1 (men1) and a novel mutation in the men1 gene. *World journal of surgical oncology*, **9**, 6–6.

Guennebaud, G., Jacob, B., *et al.* (2010). Eigen v3. http://eigen.tuxfamily.org.

Heppner, C., Reincke, M., Agarwal, S. K., Mora, P., Allolio, B., Burns, A. L., Spiegel, A. M., and Marx, S. J. (1999). Men1 gene analysis in sporadic adrenocortical neoplasms. *The Journal of Clinical Endocrinology & Metabolism*, **84**(1), 216–219.

Hotelling, H. (1936). RELATIONS BETWEEN TWO SETS OF VARIATES*. *Biometrika*, **28**(3-4), 321–377.

Houseman, E. A., Molitor, J., and Marsit, C. J. (2014). Reference-free cell mixture adjustments in analysis of dna methylation data. *Bioinformatics*, **30**(10), 1431–1439.

I, G. and V, V. (2020). *missRows: Handling Missing Individuals in Multi-Omics Data Integration*. Bioconductor, https://www.bioconductor.org/packages/release/bioc/html/missRows.html.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. **58**(3), 433–451.

Koranne, S. (2011). Hierarchical data format 5: Hdf5. In *Handbook of Open Source Tools*, pages 191–200. Springer.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**(6), 882–883.

Leek, J. T., Johnson, W. E., Parker, H. S., Fertig, E. J., Jaffe, A. E., Storey, J. D., Zhang, Y., and Torres, L. C. (2019). *sva: Surrogate Variable Analysis*. Bioconductor.

Lun, A. (2020). *BiocSingular: Singular Value Decomposition for Bioconductor Packages*.

Meng, C., Kuster, B., Culhane, A. C., and Gholami, A. M. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, **15**, 162.

Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., and Culhane, A. C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in bioinformatics*, **17**(4), 628–641.

Menozzi, P., Piazza, A., and Cavalli-Sforza, L. (1978). Synthetic maps of human gene frequencies in europeans. *Science*, **201**(4358), 786–792.

Mersmann, O. (2019). *microbenchmark: Accurate Timing Functions*. R package version 1.4-7.

Pagès, H., with contributions from Peter Hickey, and Lun, A. (2020). *DelayedArray: A unified framework for working transparently with on-disk and in-memory array-like datasets*. R package version 0.12.3.

Pierre-Jean, M., Deleuze, J.-F., Le Floch, E., and Mauger, F. (2019). Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. *Briefings in Bioinformatics*.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, **38**(8), 904–909.

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010a). New approaches to population stratification in genome-wide association studies. *Nat Rev Genet*, **11**(7), 459–463.

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010b). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, **11**(7), 459–463.

Ramos, M. *et al.* (2018). curatedtcga data: Curated data from the cancer genome atlas (tcga) as multiassayexperiment objects. *R package version*, **1**(5).

Ramos, M., Schiffer, L., Re, A., Azhar, R., Basunia, A., Rodriguez, C., Chan, T., Chapman, P., Davis, S. R., Gomez-Cabrero, D., *et al.* (2017). Software for the integration of multiomics experiments in bioconductor.

Cancer research, **77**(21), e39–e42.

Ramos, M., Schiffer, L., and Waldron, L. (2020). *TCGAutils: TCGA utility functions for data management*. R package version 1.6.2.

SL, S., SG, W., DS, P., and HL, H. (2020). Principal component analysis of blood microrna datasets facilitates diagnosis of diverse diseases. *PLoS ONE*, (15).

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights*, **14**, 1177932219899051.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. (2015). Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*.

Tarazona, S., Balzano-Nogueira, L., Gómez-Cabrero, D., Schmidt, A., Imhof, A., Hankemeier, T., Tegnér, J., Westerhuis, J. A., and Conesa, A. (2020). Harmonization of quality metrics and power calculation in multi-omic studies. *Nature communications*, **11**(1), 1–13.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*, **19**(1A), A68–77.

van de Velden, M. and Takane, Y. (2012). Generalized canonical correlation analysis with missing values. *Computational Statistics*, **27**(3), 551–571.

Velden, M. v. d. and Bijmolt, T. H. A. (2006). Generalized canonical correlation analysis of matrices with missing rows: a simulation study. *Psychometrika*, **71**(2), 323–331.

Voillet, V., Besse, P., Liaubet, L., San Cristobal, M., and González, I. (2016). Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinformatics*, **17**(1), 402.

Wang, W., Han, R., Ye, L., Xie, J., Tao, B., Sun, F., Zhuo, R., Chen, X., Deng, X., Ye, C., *et al.* (2019). Adrenocortical carcinoma in patients with men1: a kindred report and review of the literature. *Endocrine connections*, **8**(3), 230–238.

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., *et al.* (2020). Ensembl 2020. *Nucleic acids research*, **48**(D1), D682–D688.

Zheng, X., Levine, D., Shen, J., Gogarten, S., Laurie, C., and Weir, B. (2012). A high-performance computing toolset for relatedness and principal component analysis of snp data. *Bioinformatics*, **28**(24), 3326–3328.

Zheng, X., Gogarten, S., Lawrence, M., Stilp, A., Conomos, M., Weir, B., Laurie, C., and Levine, D. (2017). Seqarray – a storage-efficient high-performance data format for wgs variant calls. *Bioinformatics*.