3-4-2021

STATegra: Multi-Omics Data Integration - A Conceptual Scheme With a Bioinformatics Pipeline

Nuria Planell
Universidad Pública de Navarra (UPNA)

Vincenzo Lagani
Ilia State University

Patricia Sebastian-Leon
Instituto Valenciano de Infertilidad – Reproductive Medicine Associates

Frans van der Kloet
University of Amsterdam

Ewoud Ewing
Karolinska University Hospital

See next page for additional authors

frontiers in Genetics

TECHNOLOGY AND CODE
published: 04 March 2021
doi: 10.3389/fgene.2021.620453

STATegra: Multi-Omics Data Integration – A Conceptual Scheme With a Bioinformatics Pipeline

*Nuria Planell[1], Vincenzo Lagani[2,3], Patricia Sebastian-Leon[4], Frans van der Kloet[5], Ewoud Ewing[6], Nestoras Karathanasis[7,8], Arantxa Urdangarin[1], Imanol Arozarena[9], Maja Jagodic[6], Ioannis Tsamardinos[3,10], Sonia Tarazona[11], Ana Conesa[12,13], Jesper Tegner[14,15,16] and David Gomez-Cabrero[1,14,15,17]**

[1]Translational Bioinformatics Unit, Navarrabiomed, Complejo Hospitalario de Navarra (CHN), Universidad Pública de Navarra (UPNA), IdiSNA, Pamplona, Spain, [2]Institute of Chemical Biology, Ilia State University, Tbilisi, Georgia, [3]Gnosis Data Analysis P.C., Heraklion, Greece, [4]Department of Genomic and Systems Reproductive Medicine, IVI-RMA (Instituto Valenciano de Infertilidad – Reproductive Medicine Associates) IVI Foundation, Valencia, Spain, [5]Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, Netherlands, [6]Department of Clinical Neuroscience, Karolinska Institutet, Center for Molecular Medicine, Karolinska University Hospital, Stockholm, Sweden, [7]Institute of Computer Science, Foundation for Research and Technology-Hellas, Heraklion, Greece, [8]Computational Medicine Center, Thomas Jefferson University, Philadelphia, PA, United States, [9]Cancer Signalling Unit, Navarrabiomed, Complejo Hospitalario de Navarra (CHN), Universidad Pública de Navarra (UPNA), Health Research Institute of Navarre (IdiSNA), Pamplona, Spain, [10]Computer Science Department, University of Crete, Heraklion, Greece, [11]Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, València, Spain, [12]Microbiology and Cell Science, Institute for Food and Agricultural Sciences, University of Florida, Gainesville, FL, United States, [13]Genetics Institute, University of Florida, Gainesville, FL, United States, [14]Biological and Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, [15]Unit of Computational Medicine, Department of Medicine, Center for Molecular Medicine, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden, [16]Science for Life Laboratory, Solna, Sweden, [17]Mucosal & Salivary Biology DivisionKing's College London Dental Institute, London, United Kingdom*

OPEN ACCESS

Edited by:
Liqing Tian,
St. Jude Children's Research Hospital, United States

Reviewed by:
Darong Yang,
University of Tennessee Health Science Center (UTHSC), United States
Yi Zhang,
Dana-Farber Cancer Institute, United States

***Correspondence:**
David Gomez-Cabrero
david.gomez.cabrero@navarra.es

Specialty section:
This article was submitted to Genomic Medicine, a section of the journal Frontiers in Genetics

Received: *22 October 2020*
Accepted: *20 January 2021*
Published: *04 March 2021*

Citation:
Planell N, Lagani V, Sebastian-Leon P, van der Kloet F, Ewing E, Karathanasis N, Urdangarin A, Arozarena I, Jagodic M, Tsamardinos I, Tarazona S, Conesa A, Tegner J and Gomez-Cabrero D (2021) STATegra: Multi-Omics Data Integration – A Conceptual Scheme With a Bioinformatics Pipeline. Front. Genet. 12:620453. doi: 10.3389/fgene.2021.620453

Technologies for profiling samples using different omics platforms have been at the forefront since the human genome project. Large-scale multi-omics data hold the promise of deciphering different regulatory layers. Yet, while there is a myriad of bioinformatics tools, each multi-omics analysis appears to start from scratch with an arbitrary decision over which tools to use and how to combine them. Therefore, it is an unmet need to conceptualize how to integrate such data and implement and validate pipelines in different cases. We have designed a conceptual framework (STATegra), aiming it to be as generic as possible for multi-omics analysis, combining available multi-omic anlaysis tools (machine learning component analysis, non-parametric data combination, and a multi-omics exploratory analysis) in a step-wise manner. While in several studies, we have previously combined those integrative tools, here, we provide a systematic description of the STATegra framework and its validation using two The Cancer Genome Atlas (TCGA) case studies. For both, the Glioblastoma and the Skin Cutaneous Melanoma (SKCM) cases, we demonstrate an enhanced capacity of the framework (and beyond the individual tools) to identify features and pathways compared to single-omics analysis. Such an integrative multi-omics analysis framework for identifying features and components facilitates the discovery of new biology. Finally, we provide several options for applying the STATegra

framework when parametric assumptions are fulfilled and for the case when not all the samples are profiled for all omics. The STATegra framework is built using several tools, which are being integrated step-by-step as OpenSource in the STATeg**R**a Bioconductor package.[1]

Keywords: multi-omic analyses, data-integration, next-generation sequencing, component analysis, non-parametric combination, GeneSetCluster

[1]https://bioconductor.org/packages/release/bioc/html/STATegra.html

INTRODUCTION

Computational and experimental developments have enabled the profiling of multiple layers of cell regulation: genome, transcriptome, epigenome, chromatin conformation or metabolome, among many globally known "omics" (Ramos et al., 2017; Gomez-Cabrero et al., 2019). The development of such technologies was driven by the understanding that a single-omic does not provide enough information to allow dissecting biological mechanisms (Joyce and Palsson, 2006; Gomez-Cabrero et al., 2014). For instance, while specific DNA variations have been linked with multiple diseases, the associated mechanisms are not fully understood (Gilad et al., 2008; James et al., 2018). As a result, multi-omics data-sets are increasingly applied across biological domains such as cancer biology (Gerstung et al., 2015; Tomczak et al., 2015; Iorio et al., 2016; Mertins et al., 2016; de Anda-Jáuregui and Hernández-Lemus, 2020). Furthermore, single-cell multi-omics analysis (Macaulay et al., 2017; Colomé-Tatché and Theis, 2018; Chen et al., 2019; Welch et al., 2019) has just become a reality.

However, from the necessity of multi-omics profiling came the need for multi-omics analysis tools. Thus, integrative approaches are expected to generate significantly more comprehensive insights into the biological systems under study (SuS). A myriad of such tools in the literature may be categorized and classified differently (possibly in complex ways; Gomez-Cabrero et al., 2014; Hofmann-Apitius et al., 2015; Kannan et al., 2016; Meng et al., 2016; Rohart et al., 2017; Argelaguet et al., 2018; Stein-O'Brien et al., 2018). While each of the tools is a valuable resource for any multi-omics research, combining them into a *conceptually unified framework* is key. Equally important is the fact that each framework must be as generic as possible. Thus, we introduce the STATegra framework, in which we integrate three multi-omics based approaches *into a single pipeline*: (a) Component Analysis (CA) to understand the coordination among omics data-types (Måge et al., 2019); (b) Non-Parametric Combination (NPC) analysis to leverage on paired designs to increase statistical power (Karathanasis et al., 2016); and (c) an integrative exploratory analysis (Ewing et al., 2020). Furthermore, this framework may be extended by including additional tools such as network analysis (Barabási et al., 2011; Yugi et al., 2016). We incorporated most of these tools into the STATeg**R**a Bioconductor package to facilitate their use.[2] The package is continuously being updated and developed. Furthermore, as described in the framework, additional tools are planned to be incorporated into the Bioconductor package, e.g., the pESCA (Song et al., 2020) for multi-omics CA and the GeneSetCluster (Ewing et al., 2020) for multi-omics exploratory analysis.

To demonstrate the added value of the STATegra framework as a whole, we applied it to two data-sets from The Cancer Genome Atlas (TCGA): the glioblastoma data-set (Turcan et al., 2012) and the melanoma data-set (Akbani et al., 2015). We also explored (i) the use of samples for which only a subset of omics profiles is available and (ii) the use of parametric vs. non-parametric analysis.

[2]https://bioconductor.org/packages/release/bioc/html/STATegra.html

MATERIALS AND METHODS

Additional information is included in **Supplementary Material**, and an html-R Markdown document is provided for each data-set in **Supplementary Material**; each document provides a comprehensive overview of the code used to enhance their reproducibility.

Downloading and Preprocessing Data

We selected the Glioblastoma Multiforme (GBM) and the SKCM data-sets from TCGA. The level 3 publicly available data for gene expression (gene expression calls), miRNA (miRNA expression calls), and DNA methylation (beta values per CpG, DNAm) were obtained per sample through the NCI's Genomic Data Commons (GDC) portal (Tomczak et al., 2015). The associated metadata for each project was also obtained. Additionally, for the SKCM data-set, curated metadata generated in a previous TCGA study was also used (Akbani et al., 2015).

Glioblastoma multiforme: three data types were downloaded: array-based expression (mRNA) – Affymetrix Human Genome HT U133A, array-based expression (miRNA) – Agilent Microarray, and array-based DNA Methylation (DNAm) – Illumina Human Methylation 450 K. The number of available samples differed depending on the omic: mRNA, miRNA, and DNAm profiles are available for 523, 518, and 95 samples, respectively (**Supplementary Table 1**; **Supplementary Figure 2A**).

Skin Cutaneous Melanoma: three data types were downloaded: RNA-seq-based expression (mRNA) – Illumina HiSeq 2000, miRNA-Seq-based expression (miRNA) – Illumina HiSeq 2000, and array-based DNA Methylation (DNAm) – Illumina Human Methylation 450 K. The data from these three omics are available for all the individuals (n = 425); however, divergences between

the initial date of diagnosis (driving the metadata information) and the TCGA specimen date were identified (**Supplementary Figure 1**). Consequently, we decided to include only those cases for which specimens were obtained within a 1-year window from diagnosis (n = 104).

Supplementary Table 1 describes the characteristics of the two data-sets and the pre-processing steps applied before starting the integrative workflow of multi-omics data. We conducted an exhaustive exploration for each data type assessing the need for data normalization and/or filtering (**Supplementary Material**). Metadata is available for GBM and SKCM (summarized in **Supplementary Table 2** and described in **Supplementary Tables 3, 4** for a detailed description of the variables). In general, the data provided by TCGA contains information on demographic features (age, gender, race, and ethnicity), tumor characteristics (age at diagnosis, the primary site of the disease, stage of the neoplasm, prior glioma, ulceration in melanoma, Karnofsky score for GBM, and Breslow thickness for SKCM), survival outcome (vital status, days to death, days to the last follow-up), and technical processes (batch number, tissue source site – TSS, i.e., centers which collect samples and clinical metadata).

At the end of the preprocessing, numerous matrices, i.e., one matrix per every omics data-type (mRNA, miRNA, and DNAm), plus one additional matrix containing the metadata of the samples, compose each data-set (GBM, SKCM). Omics data-type matrices are arranged placing measurements (a.k.a. features) on rows and samples in columns, while metadata matrices include samples as rows and metadata information (e.g., age, gender, etc.) in columns.

Component Analysis for Two Data-Types (omicsPCA)

To perform joint exploration of data, the two data-types must fulfill the following criteria: (i) each feature must be scaled and (ii) only samples that are common to the two data types can be analyzed. Each feature was mean-centered and then normalized to the unit sum of squares (Frobenius normalization). Due to sample availability, component analysis for two data-type matrices was restricted for each analysis for common samples (**Supplementary Figure 2A**).

Once input data were ready, the two main omicsPCA steps were applied: model selection and subspace recovery. For model selection, we aimed to identify the correct model, which means the exact number of common (shared) components and the number of distinctive components per data-type. We investigated the following methodologies: JIVE (Lock et al., 2013; the jive R package), PCA-GCA (Gu and Van Deun, 2019; RegularizedSCA R package), and pESCA (Song et al., 2020; RpESCA and Rspectra R packages; **Supplementary Table 5** and **Supplementary html-R Markdown document**).

Finally, the association between metadata and the shared/individual components obtained was assessed using the Kruskal-Wallis test, Spearman's correlation, or the Cox regression model, depending if the variable of interest was categorical, numerical or time-to-event, respectively.

All analyses were conducted in R (R Core Team, 2017).

Non-Parametric Combination for Two Data-Types (omicsNPC)

Non-Parametric Combination techniques allow combining statistical evidence (p-values) across data-types to obtain a more precise characterization of the changes associated with the outcome of interest (Karathanasis et al., 2016).

The above-described approach allows to integrating data matrices defined on overlapping sets of samples. Taking advantages of this possibility, we explored the NPC following two strategies: analyzing only common samples or analyzing all available samples (including non-overlapping ones, when applicable).

Importantly, NPC methods require linking the features across data-types. To that end, the relation between mRNA and miRNA and mRNA and methylation were obtained using the *SpidermiR* R package (Cava et al., 2017) and RGmatch (Furió-Tarí et al., 2016), respectively.

In the case of mRNA and miRNA mapping, different versions of annotation were found; we combined the following two: the *miRNAmeConverter* (Haunsberger et al., 2017) and *anamiR* (Wang et al., 2019) R packages.

Finally, the NPC may be run using the *omicsNPC* function from the STATeg**R**a package using the two data-types, the mapping file (i.e., mRNA – miRNA), and the variables to include in the model (see R-code below) as inputs.

In our analyses, the outcomes of interest were survival for the GBM data-set and the primary site of tumor for the SKCM data-set. Additionally, age was included as a co-variable in all the models. Depending on the nature of the outcome of interest the analysis performed during NPC differs. In the case of GBM, the association between each molecular quantity and the time-to-event was assessed through a Cox Regression model (Cox, 1972). Since age is by itself a relevant factor (**Supplementary Figure 5**), it was treated as a time-varying factor by specifying a time-transform function (Therneau et al., 2020). In SKCM associations between each molecular quantity and the primary site of the tumor were assessed through a differential expression analysis using Limma (Robinson, 2009; highlighted lines from the R-code).

-------CODE-------

```
# Detailed version of the code is provided as Supplementary Material (RMarkDown)
#NPC input
mRNA_data #mRNA expression data matrix
miRNA_data #miRNA expression data matrix
mapping_gene #mapping of mRNA to genes
mapping_mirna #mapping of miRNA to genes

#1 – Generate the mapping between mRNA and miRNA; a data frame describing how to map measurements across data-sets
dataMappingExprMirna <- combiningMappings(mappings=list(expr = mapping_gene, mirna = mapping_mirna), retainAll=TRUE, reference = 'Gene')

#2 – Specify data type.
# The type of analysis to be performed is defined here.
```

```
# For GBM, as the output of interest is the survival outcome, we must define a coxph function that considers the age as a co-variable. This defined function is called "ttCoxphContinuous"
dataTypesExprMirna <- list(ttCoxphContinuous,ttCoxphContinuous)

#For SKCM, as our output of interest is the differential expression between primary site of tumor, it is only necessary to define that our data-types are continuous.
dataTypesExprMirna <- c(expr = 'continuous', mirna = 'continuous').

#3 - Preparing the data-sets as an ExpressionSet object (outcome variable refers to our variable of interest, in that case, "survival" for GBM data-set and "primary site of tumor" for SKCM data-set).
mRNA <- createOmicsExpressionSet(Data = mRNA_data, pData = metadata[,c("age","outcome")]0029

miRNA <- createOmicsExpressionSet(Data = miRNA_data, pData = metadata[,c("age""outcome")])
dataInputExprMirna <- list(expr = mRNA, mirna = miRNA)

#4 - Setting methods to combine p-values
combMethods <- c("Fisher","Liptak","Tippett")

# Setting number of permutations
numPerms <- 1000

# Setting number of cores
numCores <- 4

# Setting omicsNPC to print out the steps that it performs.
verbose <- TRUE

#Run the omicsNPC
omicsNPC_output <- omicsNPC(dataInput = dataInputExprMirna,
dataMapping = dataMappingExprMirna,
dataTypes = dataTypesExprMirna,
combMethods = combMethods,
numPerms = numPerms,
numCores = numCores,
verbose = verbose)
------------------------
```

GeneSetClustering

Significant genes from omicsNPC in the different approaches (Adj.value of $p < 0.05$ or Fisher p-value <0.05 in NPC) were uploaded to the Ingenuity Pathway Analysis (IPA; Krämer et al., 2014) database (Qiagen), and core expression analysis was performed to identify affected canonical pathways and functional annotations. Right-tailed Fisher's exact test was used to calculate a p-value. Canonical pathways/functional annotations were clustered together using *GeneSetCluster* (Ewing et al., 2020). Briefly, the gene-sets were grouped into clusters by calculating the similarity of pathways/annotations of the gene content using the relative risk (RR) of each e-set appearing with each other. Only significant gene-sets (values of $p < 0.05$) with a minimum of three genes were selected for functional exploration. RR scores were clustered into groups using k-means with the optimal number of genes determined using gap statistics.

RESULTS

We designed the STATegra framework as a four-step analysis (**Figure 1**). In the first step, each data-type was analyzed separately using state-of-the-art tools for each omic. Next, in a second step, we explored the shared variability between the different data-types using unsupervised techniques such as *Joint and Individual Variation Explained* (JIVE; Lock et al., 2013), implemented in OmicsPCA. This analysis provided qualitative and quantitative insights into how much the different data-types (e.g., different omics) and their features were "*coordinated*." Moreover, the analysis provided useful information for targeting specific omics combinations (Gomez-Cabrero et al., 2019). In the third step, for those combinations of omics characterized as *coordinated*, NPC analysis allowed increasing the statistical power to identify significant features as we have recently demonstrated (Ewing et al., 2019; Fernandes et al., 2019). For that purpose, we used the NPC within the omicsNPC function (Karathanasis et al., 2016). In the final step, clustering tools (e.g., OmicsClustering) and gene-set enrichment analysis summarizing tools (such as GeneSetCluster, Ewing et al., 2020) allowed an integrated approach.

Selected Case Studies

We selected two case studies: GBM and SKCM. GBM is the first cancer studied by TCGA (McLendon et al., 2008; Brennan et al., 2013). The TCGA GBM data-set consists of primary tumor samples from roughly 600 cases. The data-set contains gene expression, miRNA, and DNA methylation microarrays. Several findings have been reported on these data, including a molecular classification of glioblastoma based on gene expression profiles (classical, proneural, neural, and mesenchymal; Verhaak et al., 2010). The TCGA Consortium published the landscape of SKCM in 2015 (Akbani et al., 2015). The TCGA SKCM data-set consists of melanoma samples from patients diagnosed with either primary or metastatic cutaneous melanoma or metastatic melanoma of unknown primary from ~400 cases. The data-set contains genotype information, gene expression, and methylation microarrays. Based on these data, several findings have been reported, including the genomic identification of four mutant subtypes (BRAF hotspot, NF1 mutant, RAS hotspot, and triple wild-type) and a molecular classification based on gene expression profiles (immune, keratin, and MITF-low related profiles) associated with survival time. In general, patients from both studies were Caucasian with a median age of 58–59 years and a higher proportion of males (~60%). The mortality rate in GBM was high (78%) with a median life expectancy of around 1 year. For SKCM, 42% of patients died during follow-up and median life expectancy was of 1 year and 3 months (**Supplementary Tables 1, 2**).

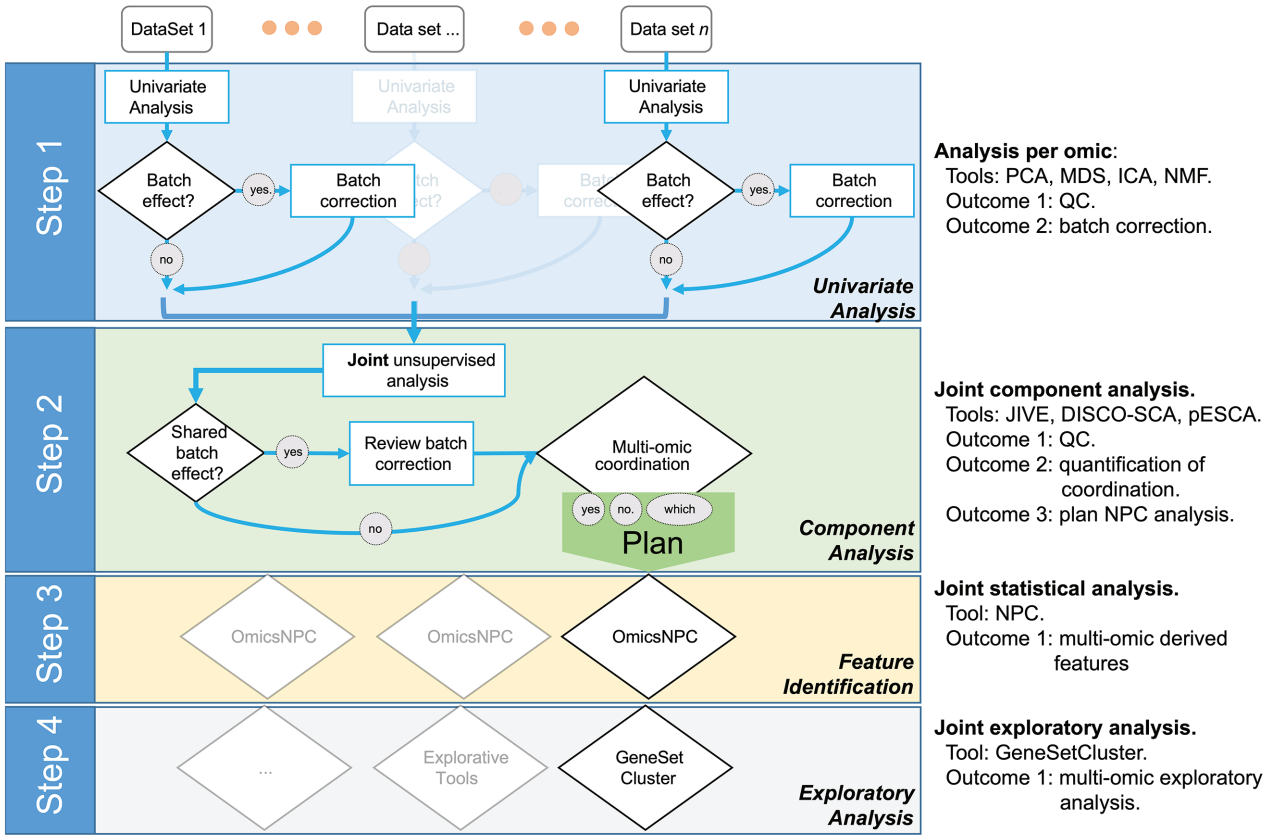

FIGURE 1 | Workflow diagram of the multi-omics analysis framework.

Step 1: Independent Data-Type Exploration and Characterization

Once the data is pre-processed, we recommend conducting quality controls for each individual data-type as the first step in the STATegra framework. In our example we made use of principal component analysis (PCA) as an unsupervised exploratory analysis. However, other matrix-factorization techniques may be used, e.g., Independent Component Analysis (ICA; Lee and Batzoglou, 2003) or Non-negative Matrix Factorization (NMF; Lee and Seung, 1999). It is important to emphasize the relevance of setting up a proper study design to avoid possible batch-effects not to be confounded with the biological effects under study: a component analysis will not overcome a wrong design.

In the GBM data-set case, the two first PCA components showed a limited amount of variability explained for all omics (**Supplementary Figure 2B**), suggesting a large per sample variability. As expected from the original TCGA publication (Verhaak et al., 2010), we found a significant association between the previously defined "gene expression subtypes" (Verhaak et al., 2010) and the first PCs of mRNA (Bonferroni adjusted value of $p < 0.001$; refer to **Supplementary Material**). Interestingly, such association was also found for miRNA and DNAm (**Supplementary Figure 2C**; adjusted value of $p < 0.005$). Moreover, we identified several clinical variables associated with at least one of the first three main components of omics data (refer to **Supplementary Material**; Bonferroni adjusted value of $p < 0.05$): survival outcome (mRNA, miRNA, DNAm) and TSS (mRNA).

In the case of the SKCM data-set, the two first PCA components showed a limited amount of variability explained for all omics (**Supplementary Figure 3A**). We identified several clinical variables associated with at least one of the first three main components of omics data (refer to **Supplementary Material**; Bonferroni adjusted value of $p < 0.05$): primary site of disease (mRNA, miRNA), neoplasm (mRNA), and pathological stage of the disease (mRNA, miRNA).

It is worth noting that some of the clinical variables were associated with at least one of the first three components in the individual data-type exploration *for more than one omics* data type. Such results apply to both GBM and SKCM data-sets. Consequently, we hypothesize that several omics are coordinated and their analytical integration would bring more statistical power and synergistic insights. In Step 2, we investigated such assumptions.

Step 2: Joint Exploration and Characterization

As previously shown, several clinical variables were associated with more than one omics data-type in both selected data-sets. Such observations may indicate that some (if not all) those

omics profiles are coordinated (or at least some of their features are). Therefore, the next step in the STATegra framework was to investigate and quantify a potential coordination.

Thus, instead of looking at the PCA-derived components of mRNA and miRNA separately, we investigated the existence of components (or factors) shared by both omics (Gomez-Cabrero et al., 2019). Intuitively, while in PCA we projected using the main components per omic (refer to **Supplementary Figures 2B, 3A** as examples), we next aimed to identify projections where the components are informative for more than one data-type simultaneously (refer to **Figures 2A,C**). In summary, when analyzing the variability of data-types A and B, we aimed to identify components associated to both A and B (shared components), components associated only to A, and components associated only to B (distinctive components).

Multi-data-set component analysis methodologies have three key steps: (a) model selection, (b) subspace recovery, and (c) estimation of robustness. In (a) model selection, we aimed to identify the correct model, which means the exact number of common (*shared*) components and the number of *distinctive* components per data-type. The determination of model selection, although fundamental, remains an open question (van der Kloet et al., 2016; Måge et al., 2019); hence, no final function has yet been included in the STATegRa package. However, we explored several methods [JIVE (Lock et al., 2013), PCA-GCA (Smilde et al., 2017), and pESCA (Song et al., 2020)]. Both, *common* and *distinctive* components obtained for each method are summarized in **Supplementary Table 5**. In our experience, the selected method depends on the nature of the data [as shown in (Måge et al., 2019)]. We do however recommend the use of several methodologies to establish more robust insights. While identifying the best model is an open challenge, we considered – based on the estimates – using the results from pESCA (Song et al., 2020), specifically pESCA (1%). Once the number of *shared* and *distinctive* components was determined, the subspace recovery (identification of loads and scores for the components) should be conducted using the same methodology used to identify space. Finally, to address robustness estimation we refer to the method in Måge et al. (2019).

In the current data-sets we were prioritizing a gene-centric analysis for both data-sets (GBM and SKCM); therefore, we posed two scenarios; the joint analysis of mRNA and miRNA, and the joint analysis of mRNA and methylation. We acknowledge that there are tools in development for integrating more than two omics; see for instance (Srivastava et al., 2013) and its application in Gomez-Cabrero et al. (2019).

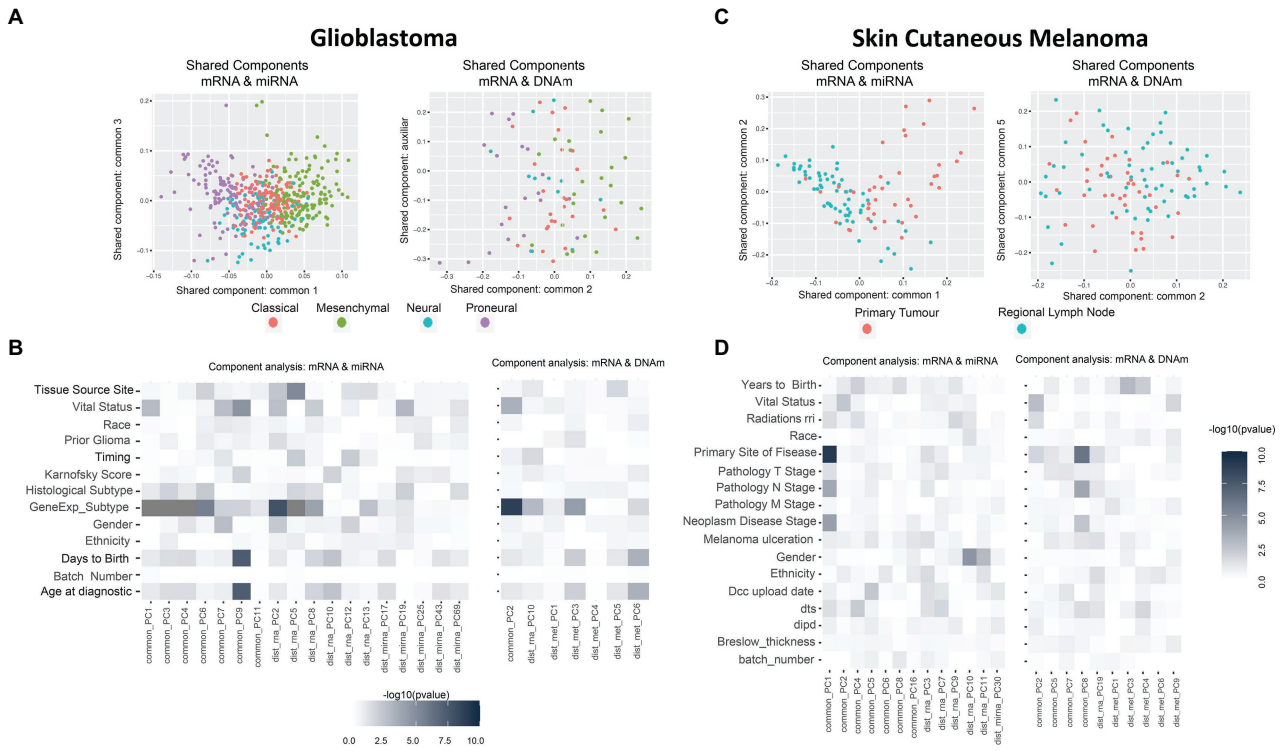

FIGURE 2 | Multi-omics component analysis. **(A,C)** Component-based representation of Glioblastoma Multiforme (GBM) and Skin Cutaneous Melanoma (SKCM) joint exploration; mRNA + miRNA (left) and mRNA + methylation (right). First and second common components (or auxiliary if only one common component is found) are shown. Samples are colored based on gene expression subtype for GBM and primary site of disease for SKCM. **(B,D)** Heatmap representation of –log10 (*p*-values) of the statistical test between metadata and common and distinctive components of GBM and SKCM joint analysis; mRNA + miRNA (left) and mRNA + methylation (right). Color ranges from white to black, understood as *p*-values with no significance to significant *p*-values. Based on the nature of the variables, *p*-values were obtained by association, correlation, or using a survival test. *Radiations rri*, *dts*, and *dipd* denote, respectively, "*Radiations radiation regimen indication*," "*Days to submitted specimen dx*," and "*Date of initial pathologic diagnosis*." See **Supplementary Tables 3, 4** for a detailed description of the variables.

GBM data-set: we identified seven shared components between mRNA and miRNA and one between mRNA and DNAm (refer to **Supplementary Table 5**). **Figure 2A** shows two PC score plots; the association between components (shared and distinctive) and clinical variables is shown in **Figure 2B**. After investigating all pairs of "share components vs. factors," we observed that at least one shared component was significantly associated (Bonferroni adjusted value of $p < 0.05$) with: "gene expression subtype" derived from (Verhaak et al., 2010; mRNA-miRNA, mRNA-DNAm), survival outcome (mRNA-miRNA), and age (mRNA-miRNA; **Figure 2B**). No significant relationship was seen between gene expression subtype and survival outcome (**Supplementary Figure 4**, value of $p = 0.06$), although a relationship between age and survival outcome was observed (adjusted value of $p < 0.05$). Based on these results, we hypothesized a coordination between the mRNA and miRNA profiles, and such coordination is associated with survival. Consequently, we also considered that integrating both data types will contribute to increasing the knowledge regarding GBM survival. We identified a limited global coordination when considering the mRNA and DNAm profiles.

SKCM data-set: seven shared components were identified between mRNA and miRNA profiles, and four common components between mRNA and DNAm profiles (refer to **Supplementary Table 5**). **Figure 2C** shows two PC score plots, and the association between components (shared and distinctive) and clinical variables is shown in **Figure 2D**. At least one component identified is significantly associated with the primary site of the disease for both mRNA-miRNA and mRNA-DNAm pairs and the disease stage for the mRNA-miRNA pair (refer to **Supplementary Material**; Bonferroni adjusted value of $p < 0.05$). Based on these results, we concluded that mRNA, miRNA and DNAm are globally coordinated, and this is mainly associated with the primary site of the disease. Therefore, the integration of the three data-types may contribute to increase the knowledge on SKCM primary site.

Importantly, based on the *complexity of the data*, the joint exploration may allow data-type specific related batch effects (identified in Step 1) from batch effects associated with sample collection (which will be associated to all omics). Interestingly, more than two omics (*blocks*) can be analyzed to identify shared components (Srivastava et al., 2013; Argelaguet et al., 2018; Song et al., 2020).

The next challenge, Step 3, was to leverage the coordination identified among omics to gain statistical power to identify the relevant features that explain the SuS.

Step 3: Integrative Differential Analysis, omicsNPC

In Step 3 we used NPC to increase the statistical power for the analysis of the SuS (Pesarin and Salmaso, 2010). Briefly, NPC non-parametrically combines *p*-values from associated features, such as a miRNA and one of its target genes measured on overlapping sets of samples. We used the omicsNPC (Karathanasis et al., 2016) included in the STATegRa package, specifically tailored for the characteristics of omics data.

The main advantages of the NPC include: (a) high statistical power with minimal assumptions; (b) wide applicability on different study designs; (c) it allows integrating data modalities with different encodings, ranges, and data distributions; and (d) it models the correlation structures present in the data producing unbiased/calibrated *p*-values, an interpretable metric (Pesarin and Salmaso, 2010).

OmicsNPC first analyses each data-type separately through a permutation-based scheme. Currently, omicsNPC uses the package limma or survival (coxph) for computing statistics and *p*-values; however, the user may also customize the functions (refer to "Materials and Methods"). The resulting permuted-based *p*-values may be combined using Tippett's (aimed to identify findings supported *by at least one omics modality*), Liptak's (*by most omics modalities*), or Fisher's (intermediate behavior between Tippett and Liptak) combination function. Following the original NPC, omicsNPC (Karathanasis et al., 2016) makes minimal assumptions: as permutation is employed throughout the process, no parametric form is assumed for the null distribution of the statistical tests, and the main requirement is that samples are freely exchangeable under the null-hypothesis. This frees the researcher from the need of defining and modeling between dataset dependencies. Most importantly, it provides global *p*-values for assessing the overall association of related features across different data modalities with the specified outcome (Pesarin and Salmaso, 2010).

GBM analysis: we aimed to investigate GBM survival through its relationship with omic features corrected for age, based on the association identified in **Supplementary Figure 5**. We only used samples profiled for all data-types ($n = 515$ and $n = 83$ for the mRNA-miRNA and mRNA-DNAm pairs, respectively). **Table 1** (*Overlapping samples* column) presents the NPC outputs. When the NPC is applied on "mRNA and miRNA," the integration allowed identifying 23 new genes and four new miRNAs. For "mRNA and DNAm," the integration allowed identification of 106 new genes and 150 new CpG sites.

TABLE 1 | Non-parametric combination analysis results of two-omics data from the GBM and SKCM projects.

	GBM		SKCM
mRNA + miRNA	**Overlapping samples**	**Whole dataset**	
mRNA dimension	7.814 × 515	7.814 × 523	9,491 × 104
mRNA significant	**1**	**4**	**216**
miRNA dimension	325 × 515	323 × 518	239 × 104
miRNA significant	**1**	**1**	**6**
mRNA-miRNA total pairs	24,665	24,665	20,225
NPC_Fisher significant pairs	27	50	114
New mRNA from NPC	**23**	**43**	**48**
New miRNA from NPC	**4**	**7**	**14**
mRNA + DNAm			
mRNA dimension	9,620 × 83	9,620 × 523	9,564 × 104
mRNA significant	**2**	**7**	**277**
Methylation dimension	57,645 × 83	57,645 × 95	55,729 × 104
Methylation significant	**1**	**0**	**12**
mRNA-methylation total pairs	57,645	57,645	55,729
NPC_Fisher significant pairs	150	332	432
New mRNA from NPC	**106**	**174**	**116**
New methylation sites from NPC	**150**	**332**	**428**

Significance was considered for a False Discovery Rate <0.05. Bold values highlight the number of significant features.

SKCM analysis: we explored the omics characterization associated to the primary site of the disease. When the NPC was applied on "mRNA and miRNA," the integration allowed identifying 48 new genes and 14 new miRNAs. For "mRNA and DNAm," the integration allowed identifying 116 new genes and 428 new CpG sites. This increase of the statistical power was expected based on the results from the joint exploration (**Figure 2**).

Alternatives to Step 3

Including samples available for a sub-set of data-types: when doing the NPC analysis, we considered samples available for both omics. However, in the case of GBM we discarded a large number of samples. In (Karathanasis et al., 2016; Ewing et al., 2019), we modified the NPC permutation protocol to include the discarded samples. We observed that the use of all samples allowed us to identify a larger number of novel features ("mRNA and miRNA" identified 43 new genes instead of 23; for complete results refer to **Table 1**, Column *Whole data-set*).

Parametric version: The NPC requires a large number of permutations, which is time consuming. To address this, the STATeRra package includes a parametric combination methodology (Benjamini and Heller, 2008; Karathanasis et al., 2016). This parametric approach is a faster alternative to NPC, which we suggest to use in preliminary explorations. In our analyses, the parametric approach generated a larger number of significant results in comparison to the non-parametric counterpart (**Supplementary Table 6**), which may be explained by unaccounted inter-data-sets correlations that inflate the significance of the p-values.

Step 4: Exploratory Analysis and Determination of the Framework's Added Value

The STATegra framework provided novel genes, miRNAs, and CpG sites for the two selected cases in comparison to unimodal analyses. We investigated if such novel elements could also provide new insights at gene-set level. For this, we made use of the GeneSetCluster (Ewing et al., 2020), a tool that summarizes gene-set analysis (GSA) results derived from multiple analyses. It allows identifying core-results by clustering gene-sets and posterior exploration; furthermore, it analyzes the integration of more than one gene-set (which could be derived from more than one omic) simultaneously. When investigating SKCM, we compared three GSA: (Ramos et al., 2017) using genes derived from mRNA single-omic analysis, (Gomez-Cabrero et al., 2019) using genes derived from mRNA-miRNA NPC analysis, and (Gomez-Cabrero et al., 2014) genes in (Gomez-Cabrero et al., 2019) not identified in (Ramos et al., 2017). We observed that the set of genes in (Ramos et al., 2017) identified several relevant canonical pathways, which are also identified in (Gomez-Cabrero et al., 2019) and (Gomez-Cabrero et al., 2014); but, especially, (Gomez-Cabrero et al., 2014, 2019)

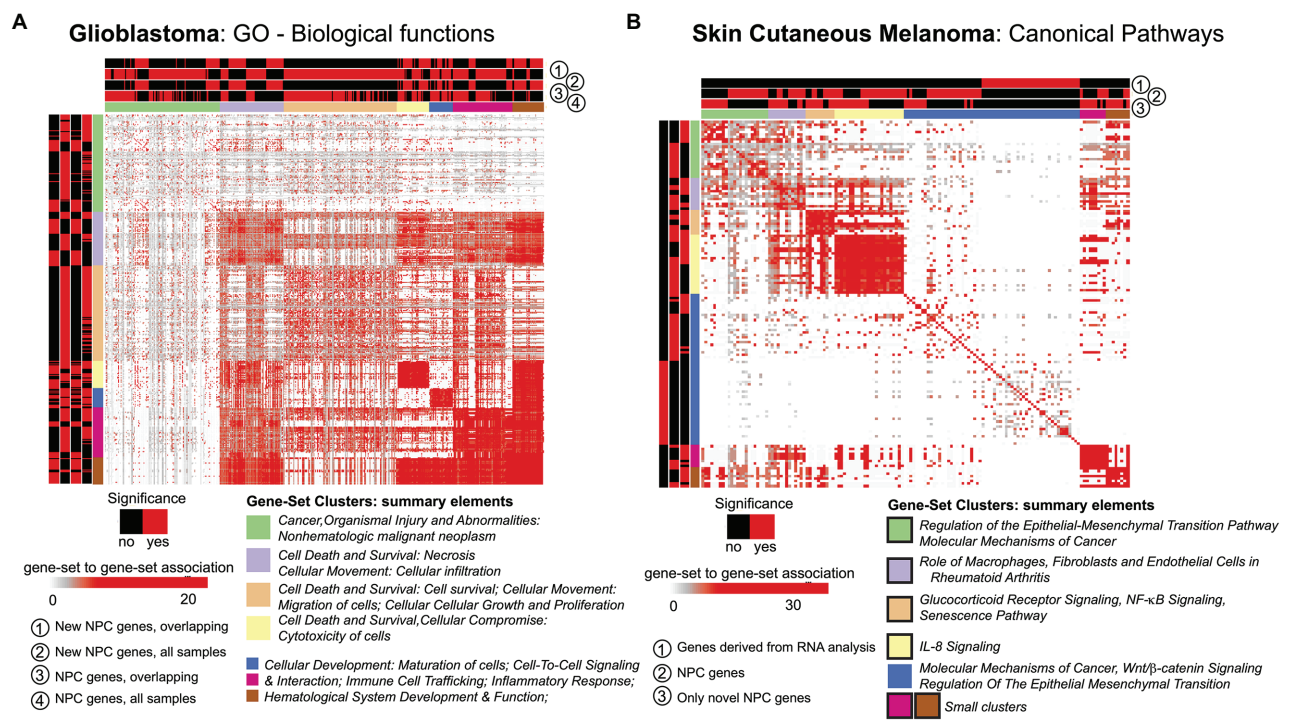

FIGURE 3 | GeneSetCluster analysis. Each heatmap depicts the gene-set to gene-set RR distance (Ewing et al., 2020). In each case several gene-set analyses have been conducted. Red (yes)/black (no) shows which gene-sets have been identified for each gene-set analysis. **(A)** GBM and **(B)** SKCM. In SKCM, clusters presented as black lines are those associated to discoveries through multi-omics integration.

GSA identified many additional relevant pathways as shown in **Figure 3B** for Canonical Pathways analysis (see box strokes on clusters). In the case of GBM, four GSAs were conducted with the following pair combinations: (a) "*considering only samples with all omics available* (OVERLAP)" or "*considering all samples* (ALL)," and (b) "*considering all identified genes*" or "*considering genes only identified by NPC*." We observed major differences in the summarized gene-sets between OVERLAP vs. ALL; see for instance **Figure 3A**, when analyzing "Gene Ontology – Biological Functions" (Blake et al., 2015). The use of GeneSetCluster allowed us to demonstrate the added value of the STATegra framework. Furthermore, it is also a tool for multi-omics GSA integrative analysis that we consider as part of the STATegra framework. We plan to integrate such tools continuously to the STATeg**R**a package.

DISCUSSION

There are many bioinformatics integrative tools (Gomez-Cabrero et al., 2014; Yugi et al., 2016; Hasin et al., 2017; Argelaguet et al., 2018; Shafi et al., 2019). However, when carrying out multi-omics analysis, as a rule, researchers use custom pipelines that combine some of the available tools. While every multi-omics data combination is different, we believe that a general framework is key to gain knowledge for an "*optimized*" integrated research analysis in the future. We here present the STATegra framework, a multi-omics integrative pipeline, the result of integrative analyses done over the last decade (Karathanasis et al., 2016; Carlström et al., 2019; Ewing et al., 2019, 2020; Fernandes et al., 2019). In the two chosen case studies used to evaluate the STATegra framework, GBM and SKCM, we show through a consecutive four-step process (**Figure 1**), how single omics integration generates additional information. Step 2, Component Analysis, quantifies the coordination of the different data-types, a key phase to identify where omic-combination can be leveraged, and Step 3 -Non-Parametric Combination is used to gain statistical power. In both case studies, we detect a greater number of genes as shown in **Table 1**. Interestingly and following the gene expression vs. DNA Methylation relation, in the case of the statistically significant pair of features identified in the mRNA-DNAm analysis, were showing a bimodal – but mostly negative – distribution of the correlation between gene expression and DNA methylation (see **Supplementary Figure 6**). Step 4 examines the added value of the biological-insights of the features identified by the integration process.

In GBM we examine the association of the omics profiles with survival. In comparison to single-omic analysis, the STATegra framework identifies additional genes already known to be associated with GBM such as CAST, ATF5, GANAB [glycoprotein associated with GBM cancer stem cells (Dai et al., 2011)], ICAM [overexpressed in bevacizumab-resistant GBM (Piao et al., 2017)], CORO1A [upregulated in GBM (Berezovsky et al., 2014)], LYN [*in vitro* association of enhanced survival of GBM cells (Liu et al., 2013)], MET (proto-oncogene) and STAT5 [enhances GBM cells migration, survival (Roos et al., 2018), and proliferation (Feng and Cao, 2014)], among others. Most have been previously associated with cancer and particularly to glioblastoma. We also compare the identified miRNAs with existing miRNA-derived survival signatures (Srinivasan et al., 2011); only miR222 is identified in the single-omic analysis, while three additional miRNAs (miR31, miR221, and miR200b) are identified by STATegra.

With the analysis of GSA, STATegra identifies new gene-sets, e.g., the TREM1 signaling pathway, previously associated with GBM (Kluckova et al., 2020). In SKCM we investigated the omics association with the primary site of disease. In addition to the newly identified genes (refer to **Table 1**), the major STATegra-associated novel insights are derived from GSA analysis as shown in **Figure 3B**, particularly regarding the identification of the IL8 signaling, which is known to be relevant in SKCM (Shoshan et al., 2016; Tobin et al., 2019).

Importantly, the new results are not derived only because of the application of the tools, but also because the application of their combination as a framework (see also **Figure 1**). For instance, the outcome of the Component Analysis provides insights into which clinical variables to investigate or which combination of omics to prioritize in the next steps. Furthermore, as shown, the outcome of the NPC (identification of features by a single-omic or by paired-multi-omic-features) can be leveraged in the GeneSetCluster tool to identify pathways derived from single-omic or coordinated among omics as shown in **Figure 3**. Adding new tools to the framework or modifying the existing ones should aim to generate greater synergies between the selected tools.

It is important to point out that we are not comparing our analysis against the original publications: GBM (McLendon et al., 2008; Brennan et al., 2013) and SKCM (Akbani et al., 2015). The idea is to compare a generic framework with single-omic approaches. Moreover, since the questions and data-sets used are different from those in the original TCGA publications, a back-to-back comparison is not justified.

The results generated by STATegra show the *added value* of a general integrative framework. Still, we acknowledge that, similarly to Operations Research there is "*no-free-lunch*" (Wolpert and Macready, 1997). Generic frameworks provide an initial approximation to any integrative analysis. Once completed, they may be further customized – and therefore further optimized – to account for the characteristics of the data and considered SuS. Still, the STATegra framework's value is its solid integration starting point, and - after being applied in many projects – generic rules can be extracted to allow an easier and faster customization.

Frameworks as the one we present here or complementary ones aimed to supervised learning (Rohart et al., 2017) are becoming increasingly necessary due to the amount of growing multi-omics data, particularly in the context of single-cell multi-omics (Colomé-Tatché and Theis, 2018). Further developments are required in multi-omics visualization (González et al., 2012), simulated data (Martínez-Mira et al., 2018), or further exploitation of Component Analysis as shown in (Stein-O'Brien et al., 2018), among others. Thus, we consider that the STATegra framework is the starting point that will be further developed over time. The next immediate steps are the inclusion of pESCA (Song et al., 2020) for multi-omic component analysis

and GeneSetCluster (Ewing et al., 2020) for multi-omic exploratory analysis within the STATegRa Bioconductor package.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: TCGA Data Portal.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by TCGA. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DG-C, JT, AC, and ST designed the global pipeline. NP implemented the global pipeline and conducted the analysis. NP and DG-C wrote the manuscript. VL, PS-L, FK, EE, NK, and AU implemented specific parts of the analysis and provided supervision. All the authors reviewed the manuscript and were part of the review of the results of the analysis. All authors contributed to the article and approved the submitted version.

FUNDING

This work has been funded by the European Union Seventh Framework Programme (FP7/2007–2013) under the grant agreement 306000-STATegra.

ACKNOWLEDGMENTS

We thank all members of the STATegra consortium for their contributions to this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.620453/full#supplementary-material

REFERENCES

Akbani, R., Akdemir, K. C., Aksoy, B. A., Albert, M., Ally, A., Amin, S. B., et al. (2015). Genomic classification of cutaneous melanoma. *Cell* 161, 1681–1696. doi: 10.1016/j.cell.2015.05.044

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., et al. (2018). Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* 14:e8124. doi: 10.15252/msb.20178124

Barabási, A. -L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68. doi: 10.1038/nrg2918

Benjamini, Y., and Heller, R. (2008). Screening for partial conjunction hypotheses. *Biometrics* 64, 1215–1222. doi: 10.1111/j.1541-0420.2007.00984.x

Berezovsky, A. D., Poisson, L. M., Cherba, D., Webb, C. P., Transou, A. D., Lemke, N. W., et al. (2014). Sox2 promotes malignancy in glioblastoma by regulating plasticity and astrocytic differentiation. *Neoplasia* 16, 193–206. doi: 10.1016/j.neo.2014.03.006

Blake, J. A., Christie, K. R., Dolan, M. E., Drabkin, H. J., Hill, D. P., Ni, L., et al. (2015). Gene ontology consortium: going forward. *Nucleic Acids Res.* 43, D1049–D1056. doi: 10.1093/nar/gku1179

Brennan, C. W., Verhaak, R. G. W., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., et al. (2013). The somatic genomic landscape of glioblastoma. *Cell* 155, 462–477. doi: 10.1016/j.cell.2013.09.034

Carlström, K. E., Ewing, E., Granqvist, M., Gyllenberg, A., Aeinehband, S., Enoksson, S. L., et al. (2019). Therapeutic efficacy of dimethyl fumarate in relapsing-remitting multiple sclerosis associates with ROS pathway in monocytes. *Nat. Commun.* 10:3081. doi: 10.1038/s41467-019-11139-3

Cava, C., Colaprico, A., Bertoli, G., Graudenzi, A., Silva, T., Olsen, C., et al. (2017, 2017). SpidermiR: an R/bioconductor package for integrative analysis with miRNA data. *Int. J. Mol. Sci.* 18:274. doi: 10.3390/ijms18020274

Chen, S., Lake, B. B., and Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* 37, 1452–1457. doi: 10.1038/s41587-019-0290-0

Colomé-Tatché, M., and Theis, F. J. (2018). Statistical single cell multi-omics integration. *Curr. Opin. Syst. Biol.* 7, 54–59. doi: 10.1016/j.coisb.2018.01.003

Cox, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B* 34, 187–202. doi: 10.1111/j.2517-6161.1972.tb00899.x

Dai, L., Liu, Y., He, J., Flack, C. G., Talsma, C. E., Crowley, J. G., et al. (2011). Differential profiling studies of N-linked glycoproteins in glioblastoma cancer stem cells upon treatment with γ-secretase inhibitor. *Proteomics* 11, 4021–4028. doi: 10.1002/pmic.201100014

de Anda-Jáuregui, G., and Hernández-Lemus, E. (2020). Computational oncology in the multi-omics era: state of the art. *Front. Oncol.* 10:423. doi: 10.3389/fonc.2020.00423

Ewing, E., Kular, L., Fernandes, S. J., Karathanasis, N., Lagani, V., Ruhrmann, S., et al. (2019). Combining evidence from four immune cell types identifies DNA methylation patterns that implicate functionally distinct pathways during multiple sclerosis progression. *EBioMedicine* 43, 411–423. doi: 10.1016/j.ebiom.2019.04.042

Ewing, E., Planell-Picola, N., Jagodic, M., and Gomez-Cabrero, D. (2020). GeneSetCluster: a tool for summarizing and integrating gene-set analysis results. *BMC Bioinformatics* 21:443. doi: 10.1186/s12859-020-03784-z

Feng, C., and Cao, S. (2014). Activation of STAT5 contributes to proliferation in U87 human glioblastoma multiforme cells. *Mol. Med. Rep.* 10, 203–310. doi: 10.3892/mmr.2014.2223

Fernandes, S. J., Morikawa, H., Ewing, E., Ruhrmann, S., Joshi, R. N., Lagani, V., et al. (2019). Non-parametric combination analysis of multiple data types enables detection of novel regulatory mechanisms in T cells of multiple sclerosis patients. *Sci. Rep.* 9:11996. doi: 10.1038/s41598-019-48493-7

Furió-Tarí, P., Conesa, A., and Tarazona, S. (2016). RGmatch: matching genomic regions to proximal genes in omics data integration. *BMC Bioinformatics.* 17:427. doi: 10.1186/s12859-016-1293-1

Gerstung, M., Pellagatti, A., Malcovati, L., Giagounidis, A., Della, P. M. G., Jädersten, M., et al. (2015). Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat. Commun.* 6, 5901. doi: 10.1038/ncomms6901

Gilad, Y., Rifkin, S. A., and Pritchard, J. K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* 24, 408–415. doi: 10.1016/j.tig.2008.06.001

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., et al. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* 8:11. doi: 10.1186/1752-0509-8-S2-I1

Gomez-Cabrero, D., Tarazona, S., Ferreirós-Vidal, I., Ramirez, R. N., Company, C., Schmidt, A., et al. (2019). STATegra, a comprehensive multi-omics dataset

of B-cell differentiation in mouse. *Sci. Data* 6:256. doi: 10.1038/s41597-019-0202-7

González, I., Cao, K.-A. L., Davis, M. J., and Déjean, S. (2012). Visualising associations between paired "omics" data sets. *BioData Min.* 5:19. doi: 10.1186/1756-0381-5-19

Gu, Z., and Van Deun, K. (2019). RegularizedSCA: regularized simultaneous component analysis of multiblock data in R. *Behav. Res. Methods* 51, 2268–2289. doi: 10.3758/s13428-018-1163-z

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18:83. doi: 10.1186/s13059-017-1215-1

Haunsberger, S. J., NMC, C., and JHM, P. (2017). miRNAmeConverter: an R/bioconductor package for translating mature miRNA names to different miRBase versions. *Bioinformatics* 33, 592–593. doi: 10.1093/bioinformatics/btw660

Hofmann-Apitius, M., Ball, G., Gebel, S., Bagewadi, S., De Bono, B., Schneider, R., et al. (2015). Bioinformatics mining and modeling methods for the identification of disease mechanisms in neurodegenerative disorders. *Int. J. Mol. Sci.* 16, 29179–29206. doi: 10.3390/ijms161226148

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., et al. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell* 166, 740–754. doi: 10.1016/j.cell.2016.06.017

James, T., Lindén, M., Morikawa, H., Fernandes, S. J., Ruhrmann, S., Huss, M., et al. (2018). Impact of genetic risk loci for multiple sclerosis on expression of proximal genes in patients. *Hum. Mol. Genet.* 27, 912–928. doi: 10.1093/hmg/ddy001

Joyce, A. R., and Palsson, B. Ø. (2006). The model organism as a system: integrating "omics" data sets. *Nat. Rev. Mol. Cell Biol.* 7, 198–210. doi: 10.1038/nrm1857

Kannan, L., Ramos, M., Re, A., El-Hachem, N., Safikhani, Z., Gendoo, D. M. A., et al. (2016). Public data and open source tools for multi-assay genomic investigation of disease. *Brief. Bioinform.* 17, 603–615. doi: 10.1093/bib/bbv080

Karathanasis, N., Tsamardinos, I., and Lagani, V. (2016). OmicsNPC: applying the non-parametric combination methodology to the integrative analysis of heterogeneous omics data. *PLoS One* 11:e0165545. doi: 10.1371/journal.pone.0165545

Kluckova, K., Kozak, J., Szaboova, K., Rychly, B., Svajdler, M., Suchankova, M., et al. (2020). TREM-1 and TREM-2 expression on blood monocytes could help predict survival in high-grade glioma patients. *Mediat. Inflamm.* 2020, 1–13. doi: 10.1155/2020/1798147

Krämer, A., Green, J., Pollard, J., and Tugendreich, S. (2014). Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* 30, 523–530. doi: 10.1093/bioinformatics/btt703

Lee, S. -I., and Batzoglou, S. (2003). Application of independent component analysis to microarrays. *Genome Biol.* 4:R76. doi: 10.1186/gb-2003-4-11-r76

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791. doi: 10.1038/44565

Liu, W. M., Huang, P., Kar, N., Burgett, M., Muller-Greven, G., Nowacki, A. S., et al. (2013). Lyn facilitates glioblastoma cell survival under conditions of nutrient deprivation by promoting autophagy. *PLoS One* 8:e70804. doi: 10.1371/journal.pone.0070804

Lock, E. F., Hoadley, K. a., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (Jive) for integrated analysis of multiple data types. *Ann. Appl. Stat.* 7, 523–542. doi: 10.1214/12-AOAS597

Macaulay, I. C., Ponting, C. P., and Voet, T. (2017). Single-cell multiomics: multiple measurements from single cells. *Trends Genet.* 33, 155–168. doi: 10.1016/j.tig.2016.12.003

Måge, I., Smilde, A. K., and van der Kloet, F. M. (2019). Performance of methods that separate common and distinct variation in multiple data blocks. *J. Chemom.* 33:e3085. doi: 10.1002/cem.3085

Martínez-Mira, C., Conesa, A., and Tarazona, S. (2018). MOSim: Multi-Omics Simulation in R. bioRxiv [Preprint]. Placeholder Text.

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, M. G., et al (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 455, 1061–1068, doi: 10.1038/nature07385.

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. *Brief. Bioinform.* 17, 628–641. doi: 10.1093/bib/bbv108

Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534, 55–62. doi: 10.1038/nature18003

Pesarin, F., and Salmaso, L. (2010). Finite-sample consistency of combination-based permutation tests with application to repeated measures designs. *J. Nonparametr. Stat.* 22, 669–684. doi: 10.1080/10485250902807407

Piao, Y., Henry, V., Tiao, N., Park, S. Y., Martinez-Ledesma, J., Dong, J. W., et al. (2017). Targeting intercellular adhesion molecule-1 prolongs survival in mice bearing bevacizumab-resistant glioblastoma. *Oncotarget* 8, 96970–96983. doi: 10.18632/oncotarget.18859

R Core Team (2017). R: A language and environment for statistical computing. Viena.

Ramos, M., Schiffer, L., Re, A., Azhar, R., Basunia, A., Rodriguez, C., et al. (2017). Software for the integration of multiomics experiments in bioconductor. *Cancer Res.* 77, e39–e42. doi: 10.1158/0008-5472.CAN-17-0344

Robinson, M. D. (2009). Linear models and Limma (August).

Rohart, F., Gautier, B., Singh, A., and mixOmics, L. C. K. -A. (2017). An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* 13:e1005752. doi: 10.1371/journal.pcbi.1005752

Roos, A., Dhruv, H. D., Peng, S., Inge, L. J., Tuncali, S., Pineda, M., et al. (2018). EGFRvIII–Stat5 signaling enhances glioblastoma cell migration and survival. *Mol. Cancer Res.* 16, 1185–1195. doi: 10.1158/1541-7786.MCR-18-0125

Shafi, A., Nguyen, T., Peyvandipour, A., Nguyen, H., and Draghici, S. A. (2019). Multi-cohort and multi-omics meta-analysis framework to identify network-based gene signatures. *Front. Genet.* 10:159. doi: 10.3389/fgene.2019.00159

Shoshan, E., Braeuer, R. R., Kamiya, T., Mobley, A. K., Huang, L., Vasquez, M. E., et al. (2016). NFAT1 directly regulates IL8 and MMP3 to promote melanoma tumor growth and metastasis. *Cancer Res.* 76, 3145–3155. doi: 10.1158/0008-5472.CAN-15-2511

Smilde, A. K., Måge, I., Næs, T., Hankemeier, T., Lips, M. A., Kiers, H. A. L., et al. (2017). Common and distinct components in data fusion. *J. Chemom.* 31:e2900. doi: 10.1002/cem.2900

Song, Y., Westerhuis, J. A., and Smilde, A. K. (2020). Separating common (global and local) and distinct variation in multiple mixed types data sets. *J. Chemom.* 34:e3197. doi: 10.1002/cem.3197

Srinivasan, S., Patric, I. R. P., and Somasundaram, K. A. (2011). Ten-microRNA expression signature predicts survival in glioblastoma. *PLoS One* 6:e17438. doi: 10.1371/journal.pone.0017438

Srivastava, V., Obudulu, O., Bygdell, J., Löfstedt, T., Rydén, P., Nilsson, R., et al. (2013). OnPLS integration of transcriptomic, proteomic and metabolomic data shows multi-level oxidative stress responses in the cambium of transgenic hipI- superoxide dismutase Populus plants. *BMC Genomics* 14:893. doi: 10.1186/1471-2164-14-893

Stein-O'Brien, G. L., Arora, R., Culhane, A. C., Favorov, A. V., Garmire, L. X., Greene, C. S., et al. (2018). Enter the matrix: factorization uncovers knowledge from omics. *Trends Genet.* 34, 790–805. doi: 10.1016/j.tig.2018.07.003

Therneau, T., Crowson, C., and Atkinson, E. (2020). Using time dependent covariates and time dependent coefficients in the cox model. R survival package vignette.

Tobin, R. P., Jordan, K. R., Kapoor, P., Spongberg, E., Davis, D., Vorwald, V. M., et al. (2019). IL-6 and IL-8 are linked with myeloid-derived suppressor cell accumulation and correlate with poor clinical outcomes in melanoma patients. *Front. Oncol.* 9:1223. doi: 10.3389/fonc.2019.01223

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–A77. doi: 10.5114/wo.2014.47136

Turcan, S., Rohle, D., Goenka, A., Walsh, L. A., Fang, F., Yilmaz, E., et al. (2012). IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* 483, 479–483. doi: 10.1038/nature10866

van der Kloet, F. M., Sebastián-León, P., Conesa, A., Smilde, A. K., and Westerhuis, J. A. (2016). Separating common from distinctive variation. *BMC Bioinformatics.* 17:195. doi: 10.1186/s12859-016-1037-2

Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities

in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi: 10.1016/j.ccr.2009.12.020

Wang, T. -T., Lee, C. -Y., Lai, L. -C., Tsai, M. -H., Lu, T. -P., and Chuang, E. Y. (2019). anamiR: integrated analysis of MicroRNA and gene expression profiling. *BMC Bioinformatics*. 20:239. doi: 10.1186/s12859-019-2870-x

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. Z. (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 177, 1873–1887.e17. doi: 10.1016/j.cell.2019.05.006

Wolpert, D. H., and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* 1, 67–82.

Yugi, K., Kubota, H., Hatano, A., and Kuroda, S. (2016). Trans-omics: how to reconstruct biochemical networks across multiple 'Omic' layers. *Trends Biotechnol.* 34, 276–290. doi: 10.1016/j.tibtech.2015.12.013

Conflict of Interest: VL and IT were employed by Gnosis Data Analysis P.C., Greece.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.