Briefings in Bioinformatics, 2025, **26(2)**, bbaf110

https://doi.org/10.1093/bib/bbaf110

Problem Solving Protocol

MOSim: bulk and single-cell multilayer regulatory network simulator

Carolina Monzó[1], Maider Aguerralde-Martin[2], Carlos Martínez-Mira[3], Ángeles Arzalluz-Luque[12], Ana Conesa[1,*], Sonia Tarazona[2,*]

[1]Genomics of Gene Expression Lab, Institute for Integrative Systems Biology, Spanish National Research Council (CSIC-UV), C/ Catedràtic Agustín Escardino Benlloch, Paterna 46980, Spain

[2]Applied Statistics, Operational Research and Quality Department, Universitat Politècnica de València, Camí de Vera s/n, València 46022, Spain

[3]Biobam Bioinformatics S.L., Marina de Valencia Base 5, BioHub, C/ de la Travesía, s/n, Sector Puerto 14 E, València 46024, Spain

[*]Corresponding authors. Ana Conesa, Genomics of Gene Expression Lab, Institute for Integrative Systems Biology, Spanish National Research Council (CSIC-UV), Paterna 46980, Spain. E-mail: ana.conesa@csic.es; Sonia Tarazona, Applied Statistics, Operational Research and Quality Department, Universitat Politècnica de València, València 46022, Spain. E-mail: sotacam@eio.upv.es

Abstract

As multi-omics sequencing technologies advance, the need for simulation tools capable of generating realistic and diverse (bulk and single-cell) multi-omics datasets for method testing and benchmarking becomes increasingly important. We present MOSim, an R package that simulates both bulk (via mosim function) and single-cell (via sc_mosim function) multi-omics data. The mosim function generates bulk transcriptomics data (RNA-seq) and additional regulatory omics layers (ATAC-seq, miRNA-seq, ChIP-seq, Methyl-seq, and transcription factors), while sc_mosim simulates single-cell transcriptomics data (scRNA-seq) with scATAC-seq and transcription factors as regulatory layers. The tool supports various experimental designs, including simulation of gene co-expression patterns, biological replicates, and differential expression between conditions. MOSim enables users to generate quantification matrices for each simulated omics data type, capturing the heterogeneity and complexity of bulk and single-cell multi-omics datasets. Furthermore, MOSim provides differentially abundant features within each omics layer and elucidates the active regulatory relationships between regulatory omics and gene expression data at both bulk and single-cell levels. By leveraging MOSim, researchers will be able to generate realistic and customizable bulk and single-cell multi-omics datasets to benchmark and validate analytical methods specifically designed for the integrative analysis of diverse regulatory omics data.

Keywords: multi-omic simulator; bulk; single cell; transcriptomics

Introduction

Rapid advancements in massive sequencing technologies have significantly facilitated the widespread adoption of multi-omic assays, enabling a comprehensive exploration of the regulatory mechanisms governing biological systems. Consequently, numerous bioinformatics tools have emerged to assist researchers in processing multi-omics data, with a specific focus on unravelling multilayer gene regulatory networks (GRNs) [1, 2]. These GRNs serve as interpretable computational models, providing insights into the intricate regulation of gene expression through interconnected networks. Notably, GRNs encompass diverse regulatory components, including transcription factors (TF), chromatin accessibility, long non-coding RNAs, micro-RNAs, and methylation, among others [2]. Despite the experimental capacity to generate both bulk and single-cell multi-omic sequencing datasets, a significant challenge in GRN studies lies in precisely integrating these multiple omic layers. Therefore, the importance of benchmarking, tuning, and validating multi-omics integration pipelines becomes evident.

Synthetic data, serving as ground truth, provide an indispensable resource for defining true positive and negative feature sets, enabling rigorous benchmarking, tuning, and validation of analytical methods. Despite the paramount role of synthetic data, there are few publicly available algorithms capable of simulating multiple omic data types. To our knowledge, only three methods support comprehensive multi-omics simulation of gene expression regulation for bulk datasets. The first, the InterSIM R package [3], generates datasets for DNA methylation, gene expression, protein abundance, and their relationships. Although the method allows for customization of the number of biological replicates and the proportion of differentially expressed features, it lacks options for time-series simulation and fails to report the interaction among features. The second tool, OmicsSIMLA C++ [4], can simulate genomics, transcriptomics, methylation, and proteomics data. Nevertheless, it restricts the generation of count data matrices to the transcriptomics module and does not include customizable options for time points or replicates. The third tool, the sismonr R package [5], simulates RNA-seq count data in conjunction with pre- and post-transcriptional regulatory networks, offering time-series simulation capabilities. Nonetheless, this method lacks the flexibility to customize expression profiles and dynamics, and the only omic quantification data it generates is gene expression.

Received: November 5, 2024. **Revised:** February 13, 2025. **Accepted:** February 21, 2025

Given the cell type–specific nature of regulatory regions, it is surprising that only two methods currently support multi-omics simulation for single-cell datasets. The statistical simulator scDesign3 [6] encompasses scRNA-seq, scATAC-seq, CITE-seq, and methylation. Meanwhile, scMultiSim [7] can simulate scRNA-seq and scATAC-seq datasets. While both methods accurately simulate datasets closely resembling real data, none of them provide essential customization options, such as the number of experimental groups, biological replicates, differentially expressed genes, accessible chromatin, and reporting of interaction between features. Importantly, none of these tools is designed to simulate gene regulatory relationships across omics features, which underscores the existing gaps and limitations in current multi-omics simulation tools. GRouNdGAN [8] partially addresses this limitation by modelling GRNs with genes and TFs with single-cell resolution. However, it does not support other omic modalities, multiple experimental conditions, or multiple samples, further underscoring the need for more comprehensive simulation tools.

Here, we present MOSim, a multilayer regulatory network simulator for both bulk (RNA-seq, ATAC-seq, miRNA-seq, ChIP-seq, and Methyl-seq) and single-cell datasets (scRNA-seq and scATAC-seq), implemented as an R Bioconductor package. In a nutshell, MOSim generates quantification data for each omics layer, precisely controlling active regulatory relationships between regulatory omics and gene expression data for differentially expressed genes. Moreover, MOSim empowers users to customize data generation, enabling the inclusion of experimental groups, biological replicates, time series, and diverse cell types. By harnessing the capabilities of MOSim, bioinformatic tool developers will be able to generate realistic and customizable bulk and single-cell multi-omics datasets, facilitating the benchmarking and validation of analytical methods tailored explicitly for integrating multi-omics data and inference of multilayer GRNs.

Results

Overview of MOSim’s workflows

MOSim is a bulk and single-cell simulation environment designed for generating multi-omic regulatory networks with precise control over regulator–gene relationships. To create a synthetic ground-truth multi-omic dataset, MOSim requires as input the list of omic data types to be simulated, a single sample of seed count data for each of them, and an association file for each regulatory omic type, indicating the *a priori* or potential regulatory features associated with each gene (Fig. 1A). While MOSim provides users with example multi-omics datasets to use as seed count data for simulation, the algorithm may also be fed with the user’s count dataset of choice, regardless of organism, disease, or platform of origin. Besides simulation of RNA-seq or scRNA-seq data depending on the type of study (i.e. bulk or single cell), currently supported omic regulatory data types include ChIP-seq, miRNA-seq, Methyl-seq, ATAC-seq, and scATAC-seq. The algorithm also supports modelling transcription factor (TF)–target gene interactions from both bulk and single-cell RNA-seq data.

Users can define various configuration parameters related to the experimental design, such as the number of experimental groups, time points or cell types (if applicable), replicates per experimental condition, data dispersion, number of differentially expressed genes, and number of regulators with activator or repressor effects. MOSim outputs simulated count matrices for each expression and regulatory data type. Moreover, it generates a record of all parameters used in data creation (MOSim simulated settings), which is indispensable for accurately testing GRN inference bioinformatic tools (i.e. mean expression, dispersion, time profile, fold-change etc.) (Fig. 1A).

The MOSim package includes two main functions: mosim, for bulk dataset simulation, and sc_mosim, for single-cell dataset simulation.

The simulation results include three distinct outputs: (1) the simulated omic count data, represented as a matrix for each omic modality. Each matrix contains the same number of omic features as provided in the seed data and the number of samples, groups, cells, etc., specified by the user; (2) for gene expression, a table listing the genes simulated as differentially expressed, along with their temporal profile (only for bulk; see Table 1); and (3) a table for each omic modality detailing the regulatory relationships provided by the user, including the simulated activator or repression regulations (see Tables 2 and 3).

In addition to the primary MOSim functions for simulating bulk (mosim) or single-cell (sc_mosim) multi-omic datasets, the package provides several other useful functions. These include functions for modifying seed data (omicData and sc_omicData), adjusting default omic parameters (omicSim), and retrieving simulation results and settings (omicResults, omicSettings, sc_omicResults, and sc_omicSettings).

Bulk multi-omic GRNs: the mosim functionality

The mosim workflow (Fig. 1B) consists of the following steps:

1) As the more extended assumption for RNA-seq data, a negative binomial (NB) distribution is applied to generate a bulk RNA-seq count matrix, obtaining the mean and dispersion from the seed sample and the amount of variability across replicates set by the user.
2) Differentially expressed genes (DEGs) are randomly selected from the seed RNA-seq sample. DEGs are labelled with one of the following time-course patterns in each experimental group: continuous induction (increasing linear pattern), continuous repression (decreasing linear pattern), transitory induction (quadratic pattern with an intermediate maximum), transitory repression (quadratic pattern with an intermediate minimum), or flat, which is also the default pattern for non-DEGs.
3) Expression profiles are simulated based on the seed count values to closely reflect real-data distributions. For transitory profiles, the algorithm randomly selects the time point at which the expression reaches its maximum or minimum and simulates a quadratic pattern. For continuous profiles, the algorithm randomly defines both the expression value at the first time point and the slope of change over time, simulating a linear pattern. These patterns vary depending on the coefficient values of the simulation function, particularly as the number of time points increases, and thus, although there are four theoretical temporal profiles, the simulated profiles encompass a wider range of patterns. DEGs with flat profiles or DEGs in a two-group design with no time points are modelled by introducing a fold-change in one of the experimental conditions. For designs with more than two experimental groups, the first serves as the reference and the fold-change is applied to a random selection of the remaining groups.
4) After generating gene expression values for each condition, replicates are simulated using a NB distribution.
5) All bulk MOSim data types (except Methyl-seq) are assumed to follow a NB distribution. Therefore, the NB is also used to simulate replicates for the remaining omics, but subjected to

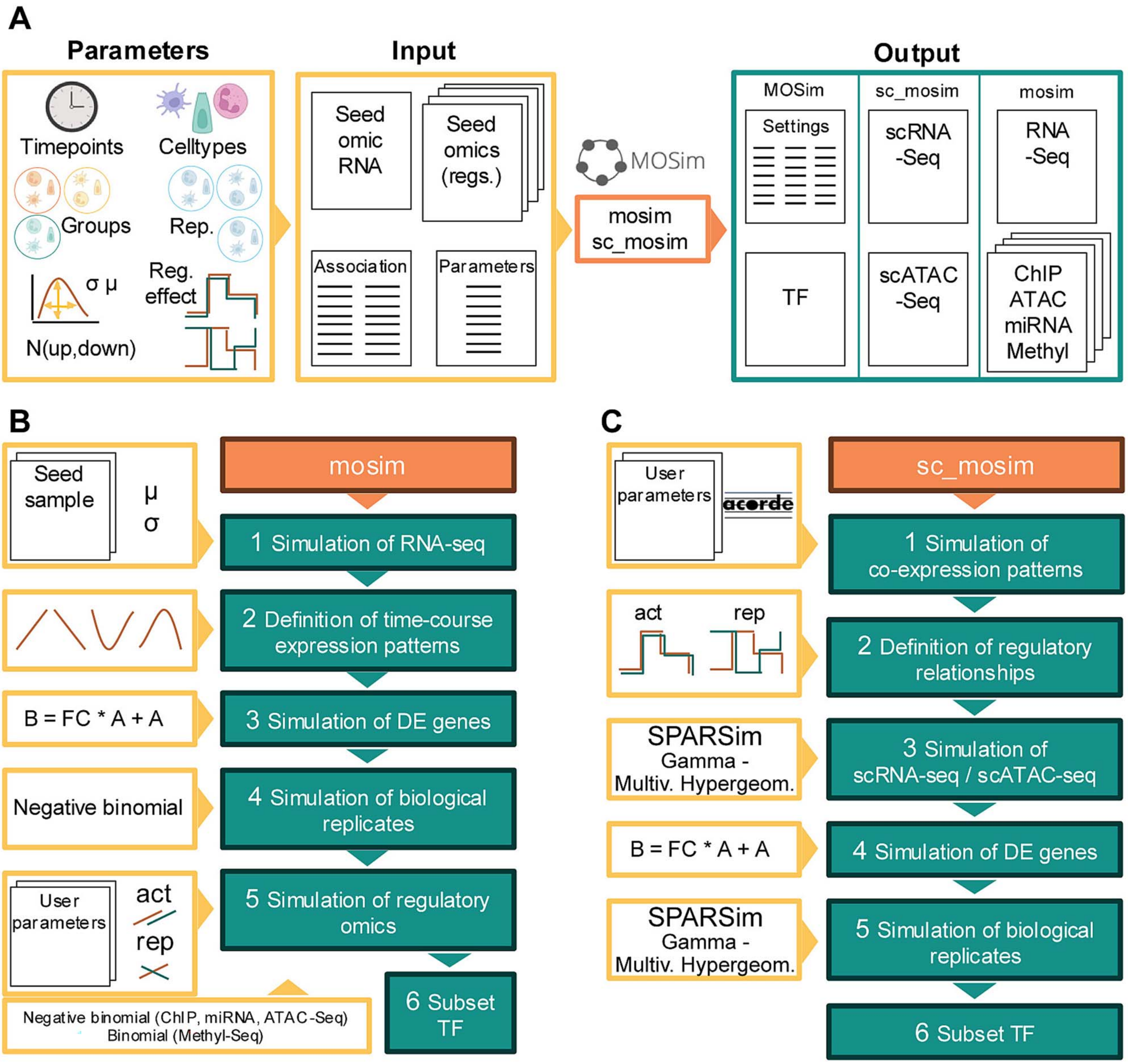

Figure 1. Schematic representation of the MOSim algorithms. (A) MOSim's simulation functionalities. (B) Flowchart of the mosim pipeline to simulate bulk multi-omics datasets. (C) Flowchart of the sc_mosim pipeline to simulate single-cell multi-omics datasets.

Table 1. MOSim-defined settings for RNA-seq simulation example

ID	DE	Gr1	Gr2	Tmax.Gr1	Tmax.Gr2
ENSMUSG00000097082	TRUE	Tran.Ind.	Tran.Ind.	1.872	1.311
ENSMUSG00000020205	TRUE	Tran.Ind.	Cont.Ind.	2.114	NA
ENSMUSG00000055493	TRUE	Tran.Ind.	Cont.Rep.	3.062	NA
ENSMUSG00000087802	FALSE	Flat	Flat	NA	NA
ENSMUSG00000017204	TRUE	Tran.Ind.	Cont.Rep.	2.610	NA
ENSMUSG00000017221	TRUE	Tran.Ind.	Cont.Ind.	1.359	NA
ENSMUSG00000052726	TRUE	Tran.Ind.	Tran.Ind.	3.178	1.626

ID: gene identifier; DE: whether the gene is differentially expressed (TRUE) or not (FALSE); GrX: type of gene temporal profile in experimental group X; Tmax.GrX: for transitory profiles, time point where the minimum or maximum is reached in the corresponding group X; Tran.Ind.: for transitory induction profiles; Cont.Rep.: for continuous repression profiles.

the simulated settings of the provided regulatory data and a randomly chosen direction of regulation. Regulators labelled as activators adopt the same profile as their associated genes, while repressors follow the opposite pattern.

6) For Methyl-seq, proportions are generated instead of counts based on the binomial distribution, following the strategy described in [9]. TF expression values are extracted from the simulated RNA-seq data to simulate TF regulation.

Table 2. MOSim-defined settings for ATAC-seq simulation example

ID	Gene	Effect.Gr1	Effect.Gr2	Gr1	Gr2
10_111588324_111588448	ENSMUSG00000097082	Activator	Activator	Trans.Ind.	Trans.Ind.
10_111588324_111588448	ENSMUSG00000020205	Activator	NA	Trans.Ind.	Trans.Ind.
10_11358301_11358431	ENSMUSG00000055493	Activator	Activator	Trans.Ind.	Cont.Rep.
10_11358301_11358431	ENSMUSG00000087802	NA	NA	Trans.Ind.	Cont.Rep.
11_98682094_98682786	ENSMUSG00000017204	Repressor	Activator	Trans.Rep.	Cont.Rep.
11_98682094_98682786	ENSMUSG00000017221	Repressor	Repressor	Trans.Rep.	Cont.Rep.
1_140257767_140257897	ENSMUSG00000052726	Activator	Activator	Trans.Rep.	Trans.Ind.

ID: genomic coordinates of ATAC-seq region (chromosome, and start and end positions for chromatin-accessible regions); Gene: regulated gene; Effect.GrX: regulatory effect of the ATAC-seq region on gene expression in experimental group X; GrX: temporal profile of the ATAC-seq region in experimental group X; Trans.Ind.: for transitory induction profiles; Trans.Rep.: for transitory repression profiles; Cont.Rep.: for continuous repression profiles.

Table 3. MOSim-defined settings for scRNA-seq and scATAC-seq for the simulation example

Gene_ID	Peak_ID	RegEffect	G_cluster	P_cluster	G_DE	P_DE	G_FC	P_FC
PRXL2B	chr19-46542333-46543301	Activator	1	1	Up	Up	67.443	67.443
SPSB1	chr22-39902952-39911753	Activator	7	7	Up	Up	73.932	73.932
PTPN22	chr12-31742761-31743451	Activator	5	5	Up	Up	45.054	45.054
PLEKHG5	chr2-132267871-132268833	Repressor	2	5	Down	Up	0.135	57.516
RBP7	chr3-101753518-101753798	Repressor	2	5	Down	Up	0.159	88.201
FRRS1	chr2-88765163-88766080	Repressor	2	5	Down	Up	0.203	74.137

Gene_ID: gene identifier; peak_ID: peak identifier; RegEffect: regulatory effect of the scATAC-seq region on gene expression in experimental group 2; G_cluster: gene expression profile across cell types; P_cluster: peak accessibility profile across cell types; G_DE: how the gene is differentially expressed; P_DE: how the peak is differentially accessible; G_FC: fold-change applied to induce differential gene expression in group 2 compared to group 1; P_FC: fold-change applied to induce differential peak accessibility in group 2 compared to group 1.

A detailed explanation of the bulk mosim algorithm implementation is provided in Supplementary File 1.

Single-cell multi-omic GRNs: the sc_mosim functionality

The workflow of sc_mosim (Fig. 1C) consists of the following steps:

1) Following the approach used by the acorde R package for defining isoform profiles across cell types in single-cell RNA-seq [10], gene expression and peak accessibility values in the seed datasets are reorganized to build synthetic features following cross-cell type patterns, i.e. indicating low or high expression in a given cell type.
2) Peak accessibility values are rearranged to reflect the regulatory relationship between scRNA-seq and scATAC-seq. Regulators labelled as activators share the same cross-cell type profile as their associated gene, while repressors have the opposite pattern.
3) Feature intensity, variability (variance of normalized counts across cells of the same cell type), and library size of the rearranged seed scRNA-seq and scATAC-seq datasets are estimated using SPARSim. A reference dataset is then simulated for each omic data type using a Gamma-Multivariate Hypergeometric model [11].
4) DEGs are randomly selected from the reference scRNA-seq. DEGs and their associated differentially accessible peaks between experimental groups are generated by introducing a fold-change in the experimental conditions, using the first condition as the reference. Additionally, random noise is added to the quantification values to introduce realistic variability between experimental groups and across features.
5) Feature intensity and library size of the simulated scRNA-seq and scATAC-seq count matrices for each experimental group are estimated using SPARSim. Biological replicates are then simulated using the Gamma-Multivariate Hypergeometric model [11], with the estimated parameters and a small random variability. TF expression values are extracted from the simulated scRNA-seq data to simulate TF regulation.

A detailed explanation of the single-cell sc_mosim algorithm implementation is provided in Supplementary File 1.

Validation of the bulk (mosim) simulation approach

To demonstrate mosim's capabilities for bulk sequencing data, we simulated RNA-seq and ATAC-seq data with five time points, two experimental groups, and three replicates, using the STATegra [12] samples included in the MOSim R package as seed data. We set the number of DEGs to 15% and modelled the five temporal profiles previously described. MOSim returns two types of output. The omicResults function returns a list containing the simulated data matrix for each omic, with features in rows and observations in columns. The second results object, accessible via the omicSettings function, includes the mosim-generated settings for the simulated relationships between gene expression and the rest of omics, as illustrated in Tables 1 and 2 containing simulation settings for RNA-seq and ATAC-seq, respectively. For instance, gene ENSMUSG00000052726 is identified as a DEG, displaying transitory repression in condition 1 and transitory induction in condition 2. The chromatin-accessible region 1_140257767_140257897 is simulated as a significant activator of this gene in both conditions, thereby following the same temporal profiles as the regulated gene.

We applied the *k*-means method to cluster simulated gene profiles, aiming to verify that the algorithm generates the expected profiles. Features with an average expression per condition of <1 count per million were filtered out. We compared the MOSim assigned profile with the average profile of the corresponding cluster and classified a gene as correctly simulated if both profiles coincided (e.g. if a gene was assigned a constant induction profile and clustered with a group exhibiting a continuous increase in expression). The optimal number of clusters was found to be $k = 7$ for k-means clustering, which resulted in one cluster per simulated pattern and time point of maximal or minimal expression. Figure 2A displays the k-means clustering results for the simulated RNA-seq data in group 1, revealing that most genes in the cluster faithfully follow the mean cluster profile, as expected. Overall, <0.5% of the simulated profiles were assigned to an incorrect cluster.

We further evaluated the simulated data using principal component analysis (PCA). The PCA score plot (Fig. 2B) indicates that the simulated data effectively recapitulated a quality time-course dataset, where replicates were clustered together and consecutive time points were proximate.

Following the validation of individual omic data, relationships between gene expression and regulatory omics were evaluated by measuring correlations. An interaction between a regulator and a gene is expected to yield a high absolute correlation value when the regulator exerts a modelled effect on the gene, sharing the same profile type for activation or exhibiting an opposite pattern (i.e. continuous induction versus continuous repression) for repression. When no effect is modelled between the gene and regulator, the profiles will exhibit uncorrelated patterns (e.g. transitory versus continuous). Pearson's correlations were calculated for each interaction and separately for each group. Interactions involving transitory profiles in both the regulator and the gene may include a delayed response, where the signal maxima—or minima—occur at different time points, with Pearson's correlation failing to capture these regulatory relationships. To address these scenarios, we also computed a lagged correlation, limiting the sliding of time points to a maximum of two to control for false positives and selecting the maximum value from Pearson and lagged correlations as the correct measure. In the ATAC-seq example (Fig. 2C), 99.4% of interactions with a modelled activator or repressor effect displayed a correlation value >0.9, while 0% of interactions without a modelled effect reached this threshold. Correlation values varied widely for these 'no effect' interactions, ranging from the expected low values to relatively high ones. The latter can often be attributed to partial overlap between noncomparable profiles, such as a transient induction profile in the gene alongside a continuous induction profile in the regulator, both sharing an increasing linear trend over the same time points. This pattern aligns with the algorithm's intended and expected behaviour. Figure 2D presents simulated temporal profiles for each experimental group, showcasing two randomly selected gene–regulator pairs. In the first pair (top plots), the regulation is activation in both groups, while in the second pair (bottom plots), the regulation is repression, also consistent across both groups.

Validation of the single-cell (sc_mosim) simulation approach

To demonstrate the utilities of sc_mosim for single-cell sequencing data, we simulated scRNA-seq and scATAC-seq data with six cell types, two experimental groups, and three replicates. We used the pbmcMultiome dataset available from SeuratData [13] as seed data and the gene–regulator association list provided in the MOSim R package. We set the number of DEGs to 30% upregulated and 20% downregulated. Variances were set to 0.1 between replicates and 0.3 between experimental groups, and we allowed for the modelling of co-expression patterns across cell types, following seven random profiles. Finally, we defined 20% activator and 10% repressor regulators in group 1, and 10% activators and 20% repressors in group 2.

In single-cell simulations, MOSim generates two main types of output. The sc_omicResults function retrieves a list containing the simulated data matrices for each omic, experimental group, and biological replicate, with features in rows and cells in columns. The second results object, extracted with the sc_omicSettings function, includes the MOSim-generated settings that associate genes and peaks (Table 3), and specify TFs with their target genes, along with the type of regulatory relationship between them. For example, gene PTPN22 is identified as an upregulated DEG that follows the across-cell-type expression pattern 5 (Fig. 3A). The chromatin-accessible region chr12-31742761-31743451 is modelled as a significant activator of this gene, following the same across-cell-type profile as the regulated gene. Conversely, the association between the gene RBP7 and chromatin-accessible region chr3-101753518-101753798 exemplifies a repressor effect of the regulator omic, where gene and peak follow opposite patterns (clusters 2 and 5, respectively), with the gene downregulated when the regulator is upregulated (Table 3).

To demonstrate the robustness of the single-cell MOSim framework for GRN simulation, we assessed its capacity to generate the expected across-cell-type expression profiles. Single-cell data are typically characterized by a high abundance of zeros and many cells belonging to the same cell type, leading to increased noise and outliers. Given the robustness of Spearman's correlation distance and k-medoids clustering techniques in noisy scenarios, we used them to extract and cluster the simulated feature profiles across cell types. The cluster average profiles were then compared to the sc_mosim simulated profiles after excluding genes with flat expression profiles. A feature was deemed correctly simulated if both profiles matched. To achieve this, we set the optimal number of clusters to $k = 10$ for k-medoids clustering, which resulted in one or two clusters per simulated co-expression pattern, minus flat expression. Clustering of the simulated scRNA-seq and scATAC-seq revealed that most features closely adhered to the mean cluster profiles as expected (Fig. 3A), with only 3.3% of simulated profiles assigned to an incorrect cluster.

We further assessed whether cells from the same cell types, experimental groups, and biological replicates clustered according to the defined simulation settings using PCA for dimensionality reduction (Fig. 3B). PCA results showed robust clustering of the simulated data, capturing a high-quality single-cell dataset where PC1 separated cells by experimental group, while PCs 1 to 4 represented the cohesive clustering of cell types (Fig. 3B). Additionally, while the majority of data variability was due to differences specified between groups, small variability between biological replicates was also observable (Fig. 3B).

To evaluate whether simulated regulatory relationships presented stronger correlations than non-regulatory peak–gene associations, Kendall's correlations between gene and peak profiles were computed within each simulated experimental group. A strong absolute correlation is expected for pairs when a regulatory effect was modelled, reflecting similar activation or opposite repression profiles. In contrast, non-regulatory peak–gene interactions typically display lower and more variable correlation values due to differences in absolute terms. As shown in Fig. 3C, 79.5% of interactions with modelled activator or repressor effects had absolute correlation values >0.7, while 'no effect' interactions displayed a broader range, centred at 0.32 absolute correlation. This range is likely due to partial overlaps, such as shared

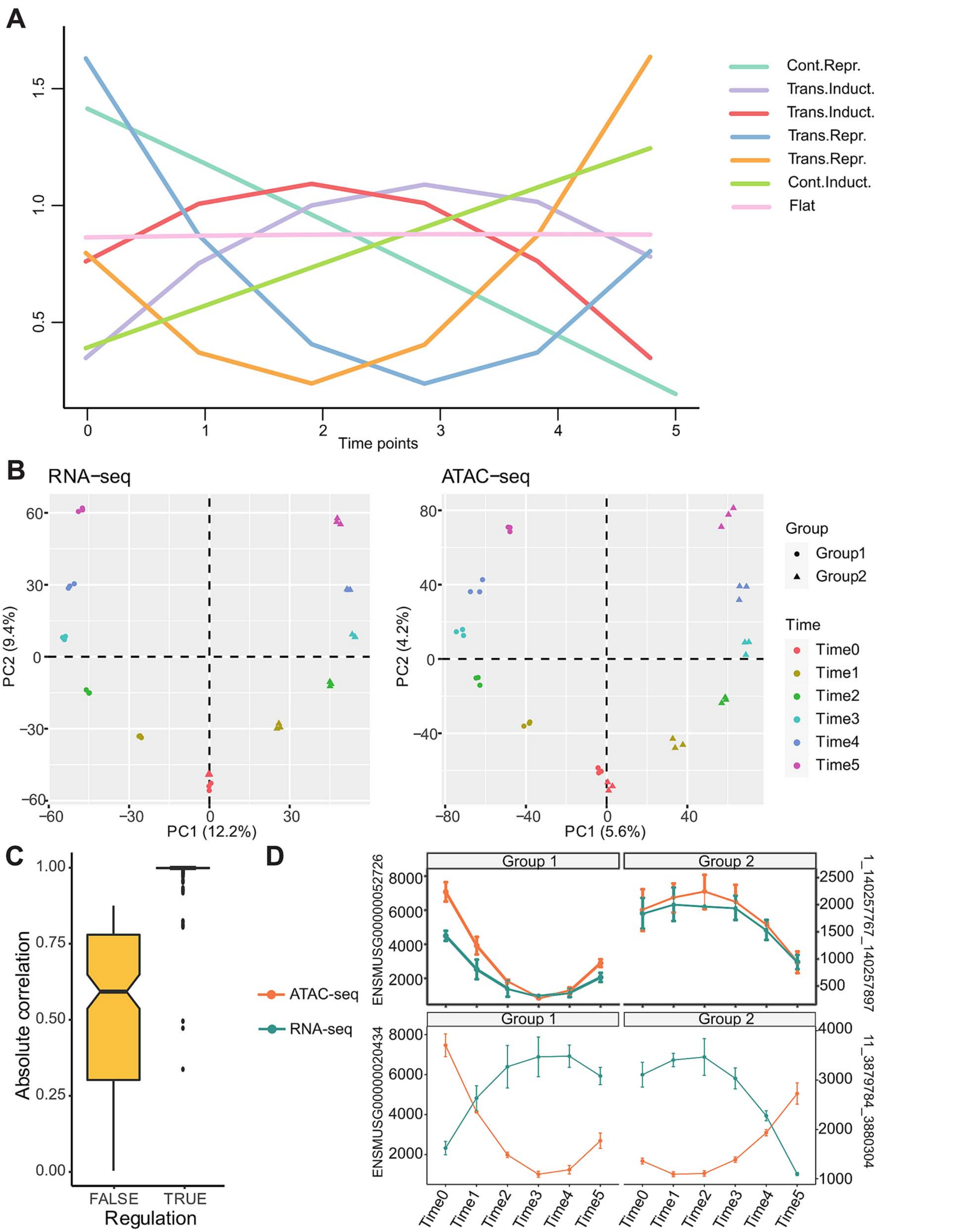

Figure 2. (A) Representation of *k*-means clusters for bulk RNA-seq in group 1. The lines represent the cluster mean profiles. The temporal simulated profiles associated with each cluster are indicated in the figure legend: One cluster corresponds to continuous repression, one to continuous induction, and one to a flat profile. Additionally, there are two clusters showing a maximum peak at different intermediate time points for transient induction and two clusters showing a minimum peak at different time points for transient repression. (B) Exploratory analysis using principal component analysis on low-count filtered data with logarithmic transformation. The first principal component separates the samples by the experimental group, while the second summarizes the temporal profile. X- and Y-axis labels indicate the percentage of variability explained by the corresponding principal component. (C) Boxplot of absolute Pearson's correlation values from interactions of ATAC-seq regulators with genes in group 1. Regulation is TRUE when the regulator has been simulated to activate or repress gene expression. Regulation is FALSE for interactions where the regulator has not been simulated to affect gene expression. (D) Two random examples of gene–regulator temporal profiles in each group. The left Y-axis shows gene expression values, while the right Y-axis shows counts for ATAC-seq regions. Vertical bars at each time point show the SD of the 3 simulated replicates.

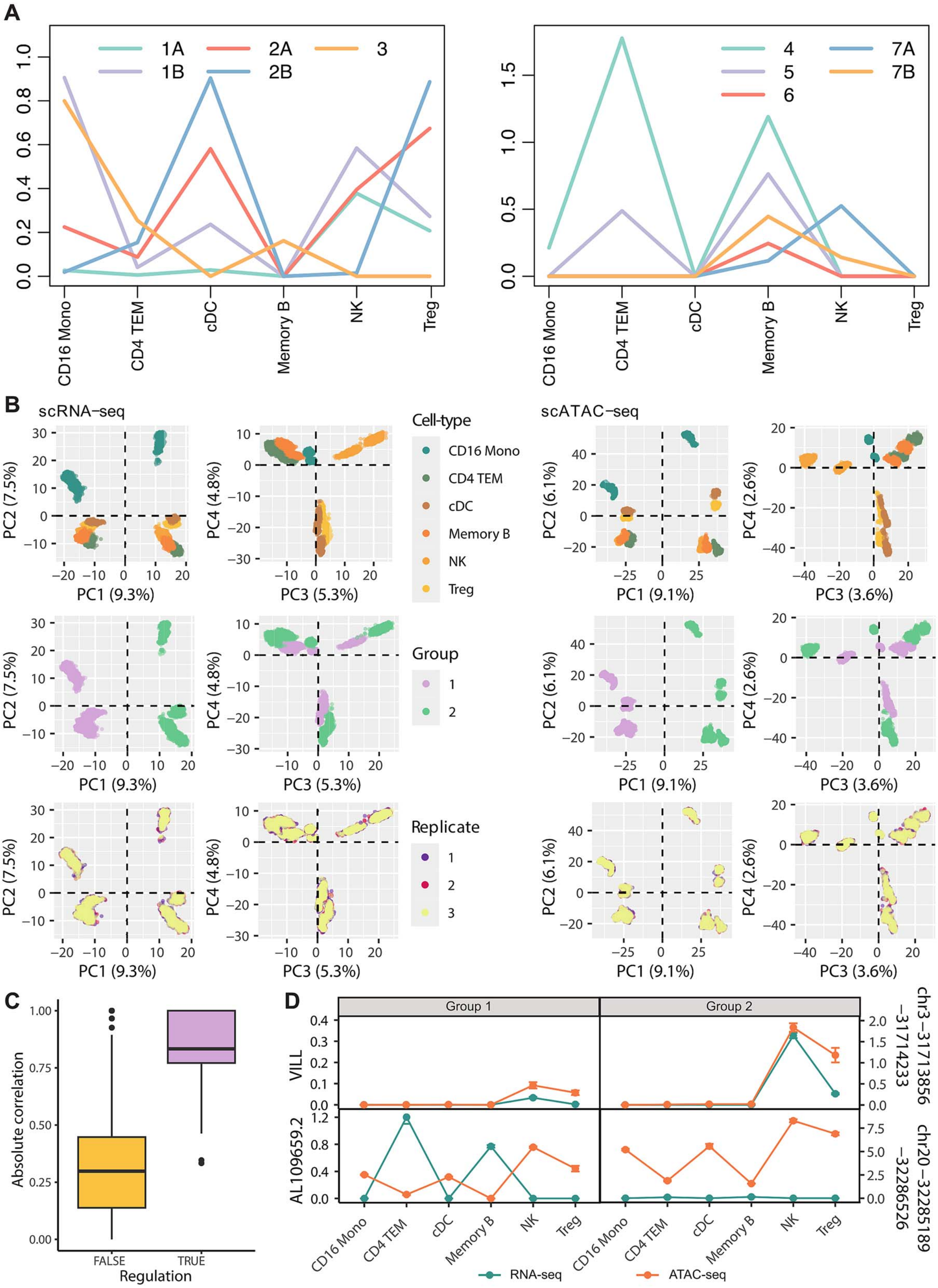

Figure 3. (A) Representation of clustering patterns for single-cell RNA-seq across cell types for group 1, split into two plots to improve visualization and cluster differentiation. The lines represent the cluster mean profiles. (B) Exploratory analysis using principal component analysis to visualize clustering of cells per cell type, experimental group, and replicate. X- and Y- axis labels indicate the percentage of variability explained by the corresponding principal component. (C) Boxplot of absolute Kendall correlation values from interactions of scRNA-seq genes in group 1 with scATAC-seq regulators. Regulation is TRUE when the regulator has been simulated to activate or repress gene expression. Regulation is FALSE for interactions where the regulator has not been simulated to affect gene expression. (D) Two examples of gene–regulator single-cell simulated profiles in each group. The left Y-axis shows gene expression values, while the right Y-axis shows counts for scATAC-seq regions. Vertical bars at each time point show the SEM of the cells for the 3 simulated replicates.

trends between cell types, which are expected outcomes of the simulation.

Finally, Fig. 3D illustrates simulated feature profiles across cell types for two pairs of gene–regulator associations, one with an activator effect and the other with a repressor effect. The first regulation (top plots) represents activation in both groups, whereas the second regulation (bottom plots) represents repression across both groups.

Simulation of multilayered gene regulatory networks

Finally, we illustrate how MOSim effectively simulates multilayered GRNs. Simulating GRNs is challenging due to the complex many-to-many relationships among some regulators and their target genes. For example, a TF or microRNA might regulate multiple target genes with varying regulatory relationships, while the same gene could be influenced by multiple factors. A multimodal GRN simulation algorithm must therefore produce a consistent dataset with expression patterns reflecting these different regulatory patterns. In MOSim, users can specify a desired percentage of active regulatory relationships, and the algorithm adjusts regulatory pairs and profiles to achieve this level of regulation across layers (Figs 2 and 3).

To demonstrate MOSim's capabilities in modelling multilayered regulatory interactions, we used the STATegra dataset [12] to simulate RNA-seq, miRNA-seq, and TF data. The simulation was performed with a sequencing depth of 30 million reads, two experimental groups, three replicates per group, and six time points, forming a detailed experimental design. Additionally, we specified that 5% of genes be differentially expressed, and 40% of miRNA-seq over the total number of regulators should be repressor effects.

Given the complexity of visualizing the simulated GRN, we selected the first 100 differentially expressed genes and plotted their corresponding GRNs for each experimental group (Fig. 4A and B). To illustrate the profiles of features in these simulated subnetworks and the efficiency of MOSim in creating consistent expression patterns across different layers, we generated heatmaps for each experimental group (Fig. 4C). To facilitate visualization and interpretation, we calculated the mean expression across replicates for each time point and experimental group, scaling the expression values across modalities, since each omic layer may have different value ranges. Figure 4C demonstrates MOSim's capacity to simulate distinct feature profiles across layers, accurately reflecting both activator and repressor regulatory effects.

This example demonstrates that MOSim can generate consistent, complex modules with both positive and negative regulatory relationships, spanning multiple layers and including one-to-many and many-to-many interactions—providing a unique capability to simulate the complexity of gene regulation.

Application of MOSim for benchmarking a GRN inference tool

To demonstrate one potential application of MOSim simulations, we used MOSim-generated data to test MORE (Multi-Omics Regulation), a tool designed to infer GRNs from bulk multi-omics data [14]. Specifically, we simulated RNA-seq, miRNA-seq, and TF data with MOSim using the STATegra dataset [12]. The simulation was configured with a sequencing depth of 30 million reads, two experimental groups, 20 time points per group, and one replicate per time point. Additionally, we set the percentage of differentially expressed genes to 50% and the percentage of significant regulations to 60%.

Prior to applying MORE, the RNA-seq count matrix was preprocessed. Low-count genes were filtered out with the NOISeq R package [15], using a threshold of 1 count per million. Count data were normalized with the weighted trimmed mean of M-values normalization in the NOISeq package and voom-transformed [16]. Differential expression analysis between groups 1 and 2 was performed with the limma R package [17], yielding 10 593 DEGs (adjusted p-value < 0.05). These DEGs were set as the target omic features required by MORE. The miRNA-seq and TF data were used as the regulatory omics. For GRN inference, we applied the MORE PLS1 option with auto-scaling and Jack-Knife resampling for the selection of significant regulators.

MORE fitted 5573 models, one for each gene with potential regulators. The MOSim simulation provided a total of 370 566 potential regulatory interactions (gene–regulator pairs), 47% of which were simulated as significant in at least one of the groups (174 051 in group 1 and 174 067 in group 2). These significant regulations served as the ground truth, or positive instances, to evaluate MORE's performance. At a significance level of 0.05, MORE identified 233 598 significant regulations in group 1 and 240 474 in group 2 that were compared to the positive instances. The analysis yielded similar error metrics for both groups, with a slightly better performance observed in miRNA-seq compared to TFs. Overall, MORE achieved a sensitivity of 85.5% and an F1-score of 62.9%. These results demonstrate MORE's ability to detect significant regulatory interactions, while also indicating areas where the tool could be improved or where hyperparameter tuning might enhance its performance.

This example highlights how MOSim can serve as a reliable ground-truth framework for evaluating the performance of GRN inference tools during their development.

Benchmarking scMOSim's scRNA-Seq simulations using a deep learning algorithm

To further demonstrate other applications of MOSim simulations, we tested it using a variational autoencoder (VAE)–based tool. VAEs are capable of learning meaningful latent representations of single-cell data. Unlike standard autoencoders, VAEs impose a probabilistic structure on the latent space, enabling more robust feature extraction and better generalization across datasets. This makes VAEs particularly useful for clustering, dimensionality reduction, and transcription factor perturbation analysis [18]. Examples of VAE models for single-cell data include scGen [19], VEGA [20], siVAE [21], scVAE [22], scDHA [23], scVI [24], manatee [25], and ScInfoVAE [26].

We tested scMOSim-generated single-cell RNA-Seq data using the VAE-based tool, single-cell Decomposition using Hierarchical Autoencoder (scDHA) [23]. scDHA first removes noise using a non-negative kernel autoencoder and then projects the data into a low-dimensional space using a stacked Bayesian autoencoder. Finally, it applies iterative perturbations to reduce overfitting and create a more generalized representation.

We used one replicate from a single experimental group of scRNA-Seq data simulated with scMOSim to evaluate cell clustering with scDHA. The clustering identified five of six simulated cell types, with one cluster combining cDC and Treg cells (Table 4). The Adjusted Rand Index score was 0.949, showing high agreement between predicted and true labels. These results demonstrate scMOSim's, and its underlying algorithm SPARSim's [11], ability to reliably simulate single-cell RNA-Seq ground-truth datasets with

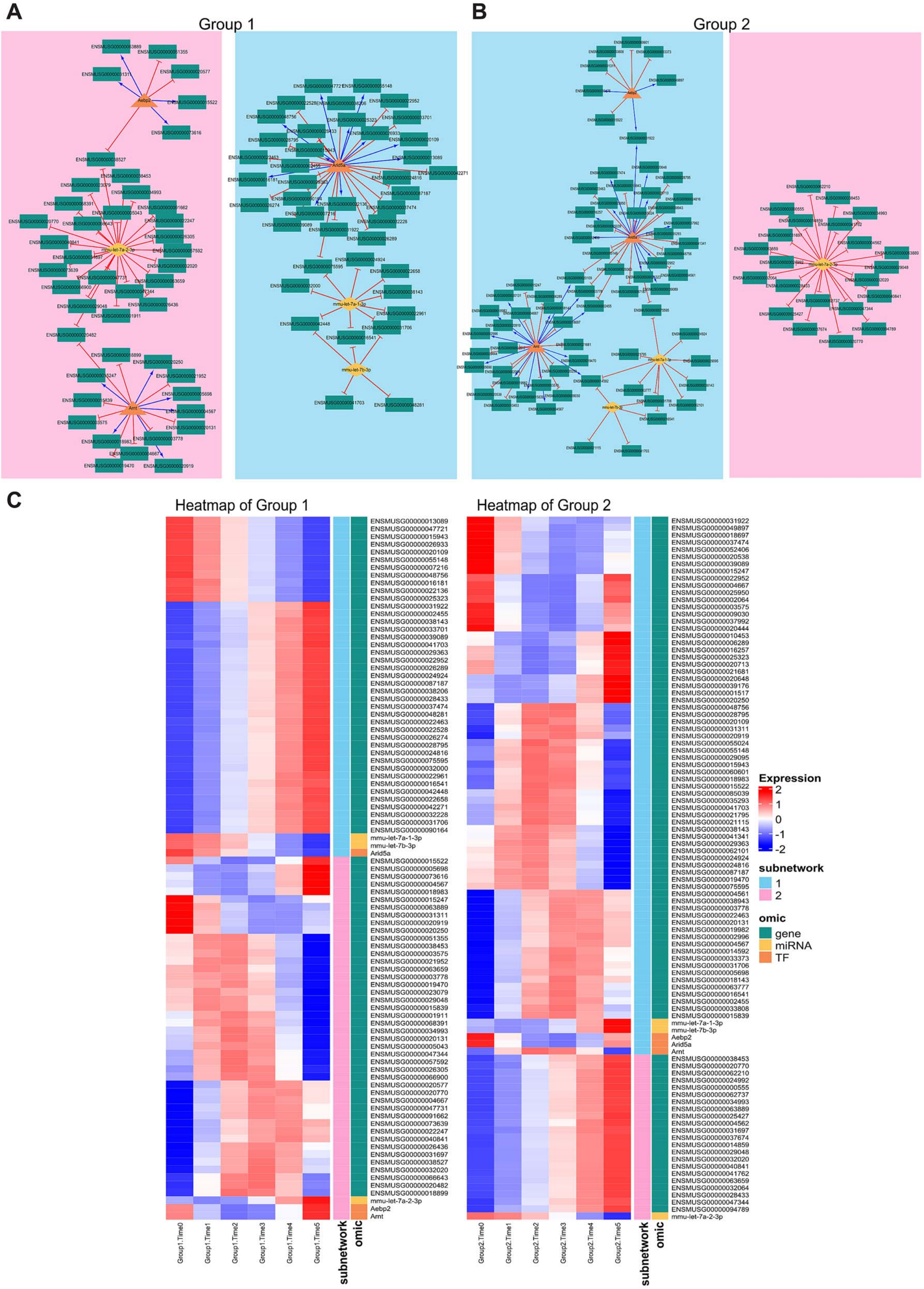

Figure 4. Representation of multilayer regulatory networks simulated by MOSim. Genes are represented in green, transcription factors are in orange, and miRNAs are in yellow. Blue arrows represent activator regulations, while red arrows represent repressor regulations. (A) Gene regulatory network for group 1. (B) Gene regulatory network for group 2. (C) Heatmaps for the expression profiles of the genes, miRNAs, and transcription factors in gene regulatory networks of groups 1 and 2. The right Y-axis shows the omic data type and the subnetwork they belong to (which refers to the connected subnetworks observed in (A) and (B) framed in pink and blue rectangles).

Table 4. Number of cells per cell cluster identified using scDHA, compared with ground-truth cell-type groups simulated using scMOSim

		scDHA predicted clusters				
		1	2	3	4	5
scMOSim simulated cell types	CD16 mono	514	0	0	0	0
	CD4 TEM	0	0	0	298	0
	cDC	0	0	198	0	0
	Memory B	0	0	0	1	370
	NK	0	468	0	0	0
	Treg	0	0	162	0	0

different cell types sufficiently distinguishable as to be identified by a VAE algorithm such as scDHA.

Discussion

Multi-omic assays, facilitated by massively parallel sequencing technologies, have greatly enhanced our ability to profile regulatory mechanisms in biological systems [1, 2], leading to a deeper understanding of diseases and model organisms. However, benchmarking studies of bioinformatic tools designed to elucidate multilayered GRNs by integrating multi-omics datasets have exposed notable discrepancies in library preparation strategies and analysis methods [27]. These discrepancies underscore the complex challenge of accurately identifying GRNs. As multi-omic sequencing continues to gain traction in the study of regulatory mechanisms, there is a pressing need for tools that support rigorous GRN inference assessment.

MOSim was developed to provide a robust framework for simulating bulk and single-cell multi-omics data in a controlled setting. Using a seed dataset and a regulator–gene association matrix, MOSim generates realistic simulated count matrices for both bulk and single-cell transcriptomics data, as well as for associated regulatory omics. For bulk data, the simulation is based on the negative binomial distribution, while for single-cell data, it leverages the well-established simulator SPARSim [11]. By using a seed dataset as a reference to infer distributions, MOSim generates count matrices that closely mirror real omic data, offering a more authentic representation than simulators that artificially construct count matrices without a real-data foundation [28].

Additionally, MOSim operates at the count matrix level rather than simulating read data, providing a unified framework for generating multi-omics data across different library preparation methods (e.g. SmartSeq2, 10x Genomics). This allows users to select a preferred method as the seed dataset for MOSim, adding flexibility to the simulation process.

MOSim enables a fast and effortless generation of bulk and single-cell count data matrices for multiple omic types, supporting flexible experimental designs. Importantly, the algorithm can simulate complex regulatory relationships between gene expression and other molecular components, guided by prior knowledge, such as target mRNA–microRNA associations. This flexibility in defining experimental designs, DEGs, and active regulators makes MOSim a versatile tool for a variety of different applications, including (1) validating methods aimed at modelling complex, multilayered regulatory networks, (2) benchmarking multi-omics data integration pipelines, (3) benchmarking GRN inference tools [2], (4) evaluating differential expression and accessibility analysis tools [24], (5) testing single-cell data clustering methods (Supplementary File 1) [24], (6) evaluating multi-omics visualization tools, and (7) testing methods for time-series analysis in RNA-seq data [29], among others. Several tools have already been tested using MOSim simulations, including DEGRE [30], scAI [31], JISAE [32], GR-NIC [33], and scLRTD [34], highlighting that MOSim's ability to specify an association matrix for linking regulators with transcripts further allows users to tailor MOSim outputs to align with the intended integration goals of their analysis tools.

The MOSim framework has some limitations. Currently, single-cell simulation is restricted to scRNA-seq and scATAC-seq, as these are presently the only two commercially available sequencing techniques that can be simultaneously performed on the same cell. As additional single-cell omics techniques become widely available, extending MOSim to other data types will be straightforward based on its bulk framework. At this point, MOSim is not prepared to simulate GRN with spatial resolution, which could be inferred from spatial multi-omics data. While these datasets are not yet widespread, they might be in the near future. We envision that the flexible MOSim simulation framework could incorporate the spatial information either as covariates of the regulatory model or by modelling cell-to-cell communication signals as an additional regulatory layer. These possibilities are to be explored in future work. Finally, both bulk and single-cell modules are designed to simulate gene regulatory relationships based on sequencing data, limiting applicability to other omics layers like proteomics and metabolomics, which may influence gene regulation in more complex or uncertain ways. Future work will also explore extending MOSim to simulate interactions between gene expression, the proteome, and the metabolome.

Conclusion

The integration of multi-omics datasets for GRN identification remains a challenging task. We demonstrate that MOSim serves as an essential resource for benchmarking integration tools, filling a critical gap in the multi-omics sequencing field.

Methods

The MOSim algorithms are introduced in the results section and extended in Supplementary File 1. The algorithms are implemented in R and mainly use R packages dplyr [35], purrr [35], Stats [36], Iranges [37], Seurat [38], SPARSim [11], and adapted scripts from Acorde [10] and WGBSSuite [9].

MOSim algorithm assessment

The performance of the MOSim bulk simulation was tested with mouse multi-omics data from the STATegra project [12], while single-cell simulation performance was evaluated using the human pbmc.multiome 10x Genomics dataset from the SeuratData R package [13].

For the bulk data, k-means clustering [39] was applied to the simulated feature profiles to assess the correct simulation of temporal expression patterns. For single-cell data, the simulated count matrix was aggregated to obtain the average count per cell type. Spearman's distance (1 – Spearman's correlation [40]) and partition around medoids (k-medoids [41]) clustering were then used to cluster gene expression profiles across cell types. In both cases, the optimal number of clusters was obtained by combining the maximization of Silhouette's coefficient and minimizing the intra-cluster variability.

In both bulk and single-cell simulations, a log transformation ($\log(x + 1)$) [42] was applied to the data. PCA was used to confirm that clustering aligned with the simulation settings. Finally, to validate gene–regulator relationships, Pearson's correlation was computed for bulk data and Kendall's T_b correlation for single-cell data [43]. These correlations were compared with 20 000 random feature pairs with no simulated regulatory effects.

Key Points

- MOSim is capable of generating synthetic datasets for a broad spectrum of omics types, supporting bulk RNA-seq, ChIP-seq, ATAC-seq, miRNA-seq, Methyl-seq, and transcription factor data, as well as single-cell omics, including scRNA-seq, scATAC-seq, and transcription factors.
- MOSim enables the robust simulation of complex, many-to-many regulatory relationships across molecular layers, faithfully capturing intricate regulatory patterns.
- Offering extensive options for customization, MOSim's flexible experimental design and parameterization empowers users to simulate count matrices and multilayer regulatory networks, tailoring simulations to diverse experimental scenarios and omics types.

Acknowledgements

The authors thank Arianna Febbo for preliminary analyses. The development of the tool and test simulations were performed on the high-performance computing cluster Garnatxa at the Institute for Integrative Systems Biology (I2SysBio). I2SysBio is a mixed research centre formed by the University of Valencia (UV) and Spanish National Research Council (CSIC).

Author contributions

S.T., A.C., and C.M.M. conceptualized and designed the mosim approach. S.T., A.C., C.M., and A.A.L. conceptualized the sc_mosim approach. C.M.M. developed and implemented mosim. C.M. developed and implemented sc_mosim and contributed to implementing mosim. C.M.M., C.M., M.A., and S.T. performed the analysis and generated visualizations. A.A.L. contributed to implementing sc_mosim. S.T. and A.C. envisioned the study and supervised the work. C.M., C.M.M., A.C., and S.T. drafted the manuscript. All authors read and approved the final manuscript.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported by the Scientific Foundation of the Spanish Association Against Cancer through the project PERME224336TARA and by Instituto de Salud Carlos III through the project AC22/00058 (co-funded by the European Union as part of the Next Generation EU programme and the Recovery and Resilience Mechanism, MRR), both projects under the frame of ERA PerMed (ERAPERMED2022-141—OVA-PDM). This work also received funding from the FP7 STATegra project (agreement no. 306000), the Spanish MINECO (BIO2012-40244), the Spanish MICIN (PID2020-119537RB-100), the Spanish Bioinformatics Institute support (PT17/0009/0015-ISCIII-SGEFI/ERDF), the Scientific Foundation of the Spanish Association Against Cancer (PERME224336TARA), under the frame of ERA PerMed (ERAPERMED2022-141—OVA-PDM), and the European Union's programme Horizon Europe under a Marie Skłodowska-Curie grant (101149931).

Data availability

The package is released under the GNU Public License to the community as a package named MOSim, for Multi-Omics Simulator, at Bioconductor (https://bioconductor.org/packages/MOSim/). Bulk example data in MOSim were generated by the STATegra project [12]. Single-cell example data are available in the pbmc.multiome dataset in the SeuratData R package [13]. Code to reproduce the figures in the manuscript is available on GitHub (https://github.com/BiostatOmics/MOSim_plots).

References

1. Marku M, Pancaldi V. From time-series transcriptomics to gene regulatory networks: a review on inference methods. *PLoS Comput Biol* 2023;**19**:e1011254. https://doi.org/10.1371/journal.pcbi.1011254.
2. Badia-I-Mompel P, Wessels L, Müller-Dott S. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet* 2023;**24**:739–54. https://doi.org/10.1038/s41576-023-00618-5.
3. Chalise P, Raghavan R, Fridley BL. InterSIM: simulation tool for multiple integrative 'omic datasets'. *Comput Methods Programs Biomed* 2016;**128**:69–74. https://doi.org/10.1016/j.cmpb.2016.02.011.
4. Chung R-H, Kang C-Y. A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. *Gigascience* 2019;**8**:8. https://doi.org/10.1093/gigascience/giz045.
5. Angelin-Bonnet O, Biggs PJ, Baldwin S. *et al.* Sismonr: simulation of in silico multi-omic networks with adjustable ploidy and post-transcriptional regulation in R. *Bioinformatics* 2020;**36**:2938–40. https://doi.org/10.1093/bioinformatics/btaa002.
6. Song D, Wang Q, Yan G. *et al.* scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nat Biotechnol* 2023;**42**:247–52. https://doi.org/10.1038/s41587-023-01772-1.
7. Li H, Zhang Z, Squires M. *et al.* scMultiSim: simulation of multi-modality single cell data guided by cell-cell interactions and gene regulatory networks. *bioRxiv* 2022, 10.15.512320. https://doi.org/10.1101/2022.11.23.517678.
8. Zinati Y, Takiddeen A, Emad A. GRouNdGAN: GRN-guided simulation of single-cell RNA-seq data using causal generative

adversarial networks. *Nat Commun* 2024;**15**:4055. https://doi.org/10.1038/s41467-024-48516-6.

9. Rackham OJL, Dellaportas P, Petretto E. *et al.* WGBSSuite: simulating whole-genome bisulphite sequencing data and benchmarking differential DNA methylation analysis tools. *Bioinformatics* 2015;**31**:2371–3. https://doi.org/10.1093/bioinformatics/btv114.
10. Arzalluz-Luque A, Salguero P, Tarazona S. *et al.* Acorde unravels functionally interpretable networks of isoform co-usage from single cell data. *Nat Commun* 2022;**13**:1828. https://doi.org/10.1038/s41467-022-29497-w.
11. Baruzzo G, Patuzzi I, Di Camillo B. SPARSim single cell: a count data simulator for scRNA-seq data. *Bioinformatics* 2020;**36**:1468–75. https://doi.org/10.1093/bioinformatics/btz752.
12. Gomez-Cabrero D, Tarazona S, Ferreirós-Vidal I. *et al.* STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. *Sci Data* 2019;**6**:256. https://doi.org/10.1038/s41597-019-0202-7.
13. Satija LL. *pbmcMultiome.SeuratData: 10X Genomics PBMC Multiome Dataset*. 2022.
14. Aguerralde-Martin M, Clemente-Císcar M, Conesa A. *et al.* MORE interpretable multi-omic regulatory networks to characterize phenotypes. *bioRxiv* 2024, 01.25.577162; https://doi.org/10.1101/2024.01.25.577162.
15. Tarazona S, Furió-Tarí P, Turrà D. *et al.* Data quality aware analysis of differential expression in RNA-seq with NOISeq R/bioc package. *Nucleic Acids Res* 2015;**43**:e140. https://doi.org/10.1093/nar/gkv711.
16. Law CW, Chen Y, Shi W. *et al.* Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;**15**:R29. https://doi.org/10.1186/gb-2014-15-2-r29.
17. Ritchie ME, Phipson B, Wu D. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47. https://doi.org/10.1093/nar/gkv007.
18. Brendel M, Su C, Bai Z. *et al.* Application of deep learning on single-cell RNA sequencing data analysis: a review. *Genomics Proteomics Bioinformatics* 2022;**20**:814–35. https://doi.org/10.1016/j.gpb.2022.11.011.
19. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. *Nat Methods* 2019;**16**:715–21. https://doi.org/10.1038/s41592-019-0494-8.
20. Seninge L, Anastopoulos I, Ding H. *et al.* VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nat Commun* 2021;**12**:5684. https://doi.org/10.1038/s41467-021-26017-0.
21. Choi Y, Li R, Quon G. siVAE: Interpretable deep generative models for single-cell transcriptomes. *Genome Biol* 2023;**24**:29. https://doi.org/10.1186/s13059-023-02850-y.
22. Grønbech CH, Vording MF, Timshel PN. *et al.* scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics* 2020;**36**:4415–22. https://doi.org/10.1093/bioinformatics/btaa293.
23. Tran D, Nguyen H, Tran B. *et al.* Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nat Commun* 2021;**12**:1029. https://doi.org/10.1038/s41467-021-21312-2.
24. Svensson V, Gayoso A, Yosef N. *et al.* Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* 2020;**36**:3418–21. https://doi.org/10.1093/bioinformatics/btaa169.
25. Yang Y, Seninge L, Wang Z. *et al.* The manatee variational autoencoder model for predicting gene expression alterations caused by transcription factor perturbations. *Sci Rep* 2024;**14**:11794. https://doi.org/10.1038/s41598-024-62620-z.
26. Pan W, Long F, Pan J. ScInfoVAE: interpretable dimensional reduction of single cell transcription data with variational autoencoders and extended mutual information regularization. *BioData Min* 2023;**16**:17. https://doi.org/10.1186/s13040-023-00333-1.
27. Luecken MD, Büttner M, Chaichoompu K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* 2022;**19**:41–50. https://doi.org/10.1038/s41592-021-01336-8.
28. Cao Y, Yang P, Yang JYH. A benchmark study of simulation methods for single-cell RNA sequencing data. *Nat Commun* 2021;**12**:6911. https://doi.org/10.1038/s41467-021-27130-w.
29. Äijö T, Butty V, Chen Z. *et al.* Methods for time series analysis of RNA-seq data with application to human Th17 cell differentiation. *Bioinformatics* 2014;**30**:i113–20. https://doi.org/10.1093/bioinformatics/btu274.
30. Terra Machado D, Bernardes Brustolini OJ, Côrtes Martins Y. *et al.* Inference of differentially expressed genes using generalized linear mixed models in a pairwise fashion. *PeerJ* 2023;**11**:e15145. https://doi.org/10.7717/peerj.15145.
31. Jin S, Zhang L, Nie Q. scAI: An unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol* 2020;**21**:25. https://doi.org/10.1186/s13059-020-1932-8.
32. Wang C, O'Connell MJ. Autoencoders with shared and specific embeddings for multi-omics data integration. *bioRxiv* 2024, 08.14.607979. https://doi.org/10.1101/2024.12.24.630168.
33. Zhang S, Kong W. An improved multiomics data clustering algorithm based on graph regularization constraints. *Third International Conference on Biomedical and Intelligent Systems (IC-BIS 2024)* 2024;**13208**:436–41.
34. Ni Z, Zheng X, Zheng X. *et al.* scLRTD: a novel low rank tensor decomposition method for imputing missing values in single-cell multi-omics sequencing data. *IEEE/ACM Trans Comput Biol Bioinform* 2022;**19**:1144–53. https://doi.org/10.1109/TCBB.2020.3025804.
35. Wickham H, Averick M, Bryan J. *et al.* Welcome to the tidyverse. *J Open Source Softw* 2019;**4**:1686. https://doi.org/10.21105/joss.01686.
36. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2021.
37. Lawrence M, Huber W, Pagès H. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* 2013;**9**:e1003118. https://doi.org/10.1371/journal.pcbi.1003118.
38. Hao Y, Hao S, Andersen-Nissen E. *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.
39. Hartigan JA, Wong MA. Algorithm AS 136: a K-means clustering algorithm. *J R Stat Soc Ser C Appl Stat* 1979;**28**:100.
40. Zar JH. *Spearman Rank Correlation*. Encyclopedia of biostatistics, 2005, 7.
41. Reynolds AP, Richards G, de la Iglesia B. *et al.* Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *J Math Model Algorithms* 2006;**5**:475–504. https://doi.org/10.1007/s10852-005-9022-1.
42. Choudhary S, Satija R. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biol* 2022;**23**:27. https://doi.org/10.1186/s13059-021-02584-9.
43. Kendall MG. The treatment of ties in ranking problems. *Biometrika* 1945;**33**:239–51. https://doi.org/10.1093/biomet/33.3.239.

Briefings in Bioinformatics, 2025, **26(2)**, bbaf110
https://doi.org/10.1093/bib/bbaf110
Problem Solving Protocol