MOSim: bulk and single-cell multi-layer regulatory network simulator

Carolina Monzó[1,A], Carlos Martínez-Mira[2,A], Ángeles Arzalluz-Luque[1,3], Ana Conesa[1,4,B], Sonia Tarazona[3,B]

1. Genomics of Gene Expression Lab, Institute for Integrative Systems Biology, Spanish National Research Council (CSIC-UV), Paterna, 46980, Spain

2. Biobam Bioinformatics S.L., València, 46024, Spain

3. Applied Statistics, Operational Research and Quality Department, Universitat Politècnica de València, València, 46022, Spain

4. Microbiology and Cell Science Department, Institute for Food and Agricultural Research, University of Florida, Gainesville, 32603, United States

A. These authors contributed equally to this work.

B. Co-corresponding authors

Contributing authors: carolina.monzo@csic.es; cmarmir@gmail.com; angeles.arzalluz@gmail.com; ana.conesa@csic.es; sotacam@eio.upv.es

Abstract

As multi-omics sequencing technologies continue to advance, the need for simulation tools that can generate realistic and diverse bulk and single-cell multi-omics datasets becomes increasingly important. We present MOSim, an R package which enables the simulation of both bulk and single-cell multi-omics data. MOSim allows for the generation of bulk transcriptomics data (RNA-seq), the incorporation of additional regulatory omics layers (ATAC-seq, miRNA-seq, ChIP-seq, Methyl-seq and Transcription Factors) and the simulation

of single-cell transcriptomics data (scRNA-seq) with scATAC-seq as a regulatory layer. The tool supports various designs, including simulation of co-expression patterns of genes, simulation of replicates and differential expression between experimental conditions.

MOSim provides users with the ability to generate count matrices for each simulated omics data type, capturing the heterogeneity and complexity of bulk and single-cell multi-omics datasets. Moreover, MOSim identifies differentially expressed features within each omics layer and elucidates the active regulatory relationships between regulatory omics and gene expression data at the single-cell level.

By harnessing the capabilities of MOSim, researchers will be able to generate realistic and customizable bulk and single-cell multi-omics datasets, allowing them to benchmark and validate analytical methods specifically tailored for integrative analysis of diverse regulatory omics data.

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, a multitude of bioinformatics tools has emerged to assist researchers in processing multi-omics data, with a specific focus on unravelling multi-layer 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 the precise integration of these multiple omic layers.

Therefore, the importance of benchmarking, tuning, and validating multi-omics integration pipelines becomes evident.

Synthetic data, serving as a ground truth, emerges as an indispensable resource to define sets of true positive and negative features for rigorous benchmarking, tuning and validation of analytical methods. Despite the paramount role of synthetic data, there are few publicly available algorithms that can simulate several omic data types. To our knowledge, only three methods support comprehensive multi-omics simulation of gene expression regulation for bulk datasets. The first, InterSIM R package [3], generates datasets for DNA methylation, gene, protein expression and their relationships. Although the method allows customization of the number of biological replicates and proportion of differentially expressed features, it does not offer options for time series simulation and fails to easily 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, without the inclusion of customizable options for time points or replicates. The third tool, sismonr R package [5], simulates RNA-seq count data in conjunction with pre- and post- transcriptional regulatory networks, offering time-series simulation capabilities. However, this method lacks the flexibility to customise expression profiles and dynamics; also lacking the generation of omic quantification data other than gene expression. Considering the cell-type-specific nature of regulatory regions, it is surprising to note 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, neither provide essential customization options, such as the number of experimental groups, biological replicates, differentially expressed genes, accessible chromatin, and the reporting of

interaction between features. This examination of tools underscores the existing gaps and limitations in current multi-omics simulation tools.

Here, we present MOSim, a multi-layer regulatory network simulator for 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 differentially expressed features within each omics layer, precisely controlling the active regulatory relationships between regulatory omics and gene expression data. Moreover, MOSim empowers users to customise 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 benchmarking and validation of analytical methods specifically tailored for integration of multi-omics data and analysis of GRNs.

Results

Overview of MOSim's workflows

MOSim is a bulk and single cell simulation environment of multi-omic regulatory networks with precise control of regulator-gene relationships. In order 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 to each gene. 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 are ChIP-seq, miRNA-seq,

Methyl-seq, ATAC-seq and scATAC-seq. The algorithm also supports the modelling of Transcription Factor (TF) - target gene interactions from the bulk 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, number of replicates per experimental condition, degree of variability in the data, number of differentially expressed genes and number of regulators having an activator or repressor effect. MOSim returns the simulated count matrices for each expression and regulatory data type. Moreover, it generates a record of all parameters used to generate the data (constraints), indispensable to perform accurate testing of GRN inference bioinformatic tools (i.e. mean expression, variability, time profile, fold change etc.).

The workflow of bulk MOSim (Figure 1) consists of the following steps:

1. A negative binomial (NB) distribution is applied to generate the bulk RNA-seq count matrix, obtaining the mean and dispersion of each omic feature from the seed sample and the amount of variability 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) and flat, which is also the pattern for non-DEGs.
3. Expression profiles are simulated from the seed count values to recapitulate real data distributions. To model transitory profiles, the algorithm randomly chooses the time point where expression will be maximum or minimum. Both 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 more than two experimental groups, the first is used as reference and the fold-change is introduced in a random selection of the remaining groups.

4. Once gene expression values are generated for each condition, replicates are simulated from a NB distribution with mean equal to the count value for that condition and variance proportional to the mean.
5. Since all bulk MOSim data types are assumed to be measured by sequencing techniques, in all data types but Methyl-seq, the same NB pattern definition function is applied for the simulation of the remaining omics, but subjected to the constraints of the provided regulatory data and a randomly chosen direction of regulation. Regulators labelled as activators have the same profile as their associated gene, but the opposite if they have a repression effect.
6. For Methyl-seq, percentages are generated instead of counts based on the binomial distribution, following the strategy described in [8]. For simulating TF regulation, TF expression values are extracted from the simulated RNA-seq data.

The workflow of single-cell MOSim (Figure 1) consists of the following steps:

1. Following the co-expression strategy used by Acorde [9] , gene and peak expression values in the seed datasets are rearranged to build synthetic features following cross-cell type qualitative expression patterns, i.e. indicating low or high expression in a given cell type.
2. Peak expression values are rearranged to reproduce the regulatory relationship between scRNA-seq and scATAC-seq, where regulators labelled as activators have the same cross-cell type profile as their associated gene, but the opposite if they have a repression effect

3. Feature intensity, variability and library size of the rearranged seed scRNA-seq and scATAC-seq datasets are estimated using SPARSim and a reference dataset is simulated for each omic data type using a Gamma-Multivariate Hypergeometric model [10].
4. DEGs are randomly selected from the reference scRNA-seq. DEGs and their associated differentially expressed peaks between experimental groups are generated by introducing a fold-change in the experimental conditions, using the first as reference. Additionally, experimental group constraints are further modelled by introducing a variability following a normal distribution across features.
5. Feature intensity and library size of the experimental groups simulated count matrices for scRNA-seq and scATAC-seq datasets are estimated using SPARSim. Using the estimated parameters and a defined variability, biological replicates are simulated using the Gamma-Multivariate Hypergeometric model [10].

Besides the MOSim general wrapper functions to simulate bulk (mosim) or single-cell (scMosim) multi-omic datasets, other useful functions included in the package help users to modify seed data (omicData and scOmicData), change default omic parameters (omicSim) and recover simulation results and constraints (omicResults, omicSettings, scOmicResults and scOmicSettings).

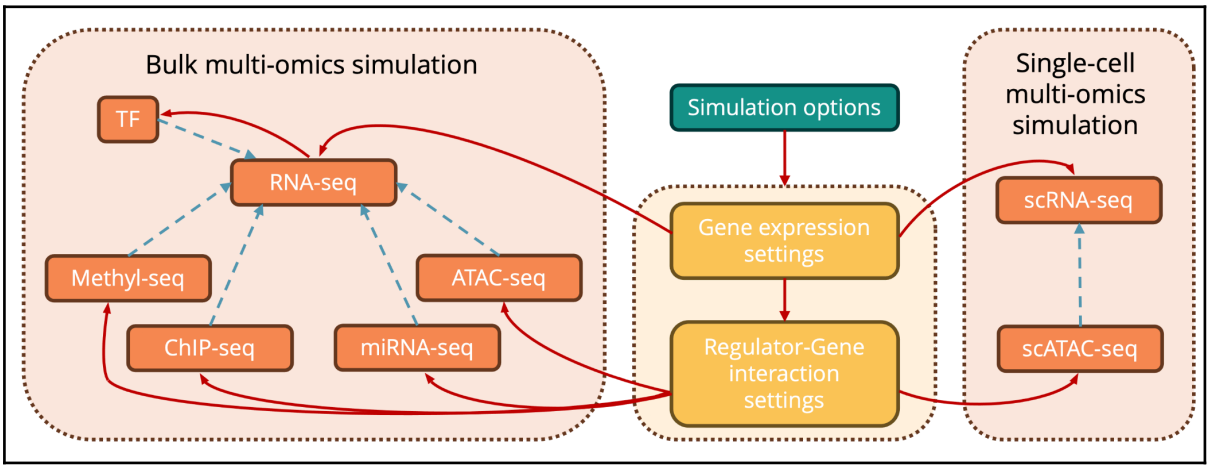

Figure 1: Schematic representation of the MOSim algorithms. Using user-defined simulation options, it defines gene expression settings and defines the regulator-gene interactions to simulate. Next it simulates RNA-seq for bulk multi-omics experiments or scRNA-seq for single-cell multi-omics experiments. Finally it simulates the count data for each regulatory layer fulfilling the regulator-gene requirements defined in the simulation options.

Validation of the bulk MOSim simulation approach

To illustrate bulk MOSim utilities, we simulated RNA-seq and ATAC-seq data with 5 time points, 2 experimental groups and 3 replicates using as seed data STATegra [11] samples included in the MOSim R package. We set the number of DEG to 15% and allowed modelling following the 5 temporal profiles described before. MOSim returns two types of output. The omicResults function retrieves a list containing the simulated data matrix for each omic with features in rows and observations in columns. The second results object, extracted with the omicSettings function, contains the MOSim-generated constraints 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 a DEG with 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 expressing the same temporal profiles as the regulated gene.

Table 1: MOSim-defined constraints for RNA-seq simulation example. 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.

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

Table 2: MOSim-defined constraints for ATAC-seq simulation example. 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.

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.

To check that the algorithm generates the expected profiles, we applied the K-means method to cluster simulated gene profiles. We applied a filtering step to remove features with an average expression per condition less than one count per million and compared the MOSim assigned profile with the average profile of the corresponding cluster and declared the gene correctly simulated if both profiles coincided (for example the gene had been assigned a constant induction profile and it clustered into a group with a mean profile presenting a constant increase of expression). We set k = 11 for K-means clustering, to allow one cluster per simulated pattern and time point of maximal or minimal expression. Figure 2A shows the results of K-means clustering for the simulated RNA-seq data in group 1, revealing that most genes in the cluster faithfully follow the mean cluster profile, as expected. Globally, less than 0.5% of the simulated profiles were assigned to an incorrect cluster.

We further evaluated the simulated data by Principal Component Analysis (PCA). We concluded that PCA (Figure 2B) indicates that the simulated data 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 have a high absolute correlation value when the regulator has a modelled effect on the gene, thus sharing the same type of profile for activation or having a totally opposite pattern (i.e. continuous induction vs continuous repression) for repression; if the interaction has no modelled effect then the profiles will have incomparable patterns (i.e. transitory vs continuous) therefore giving a presumably worse correlation value. Pearson correlations were calculated for each interaction and separately for each group. For interactions involving transitory profiles in both the regulator and the gene, a situation could occur in which the intermediate time point where the maximum- or minimum- is reached differs between regulator and gene, therefore affecting the correlation; in those cases instead of relying on conventional Pearson correlation, a method of lagged-correlation (cross-correlation of currency pairs) was used, limiting the sliding of time points to a maximum of 2 to prevent using too few values, and selecting the maximum correlation value from this process as the correct measure. In the ATAC-seq example (Figure 2C), it can be seen that 99.4% of interactions with a modelled activator or repression effect have a correlation value over 0.9, with 0% of interactions without modelled effect having values greater than that. For these "no effect" interactions it can be observed a distribution of correlation values over a wide range, from the expected low values to relatively high ones that can be traced to partial overlap between not comparable profiles, for instance a transitory induction profile in gene and continuous induction in the regulator sharing an increasing linear pattern at the same time points, which is an allowed and expected behaviour of the algorithm. In Figure 2D, we show the simulated temporal profiles in each experimental group

for two random pairs of gene-regulator. The first regulation (top plots) is an activation in both groups, while the second regulation (bottom plots) is a repression, again in both groups.

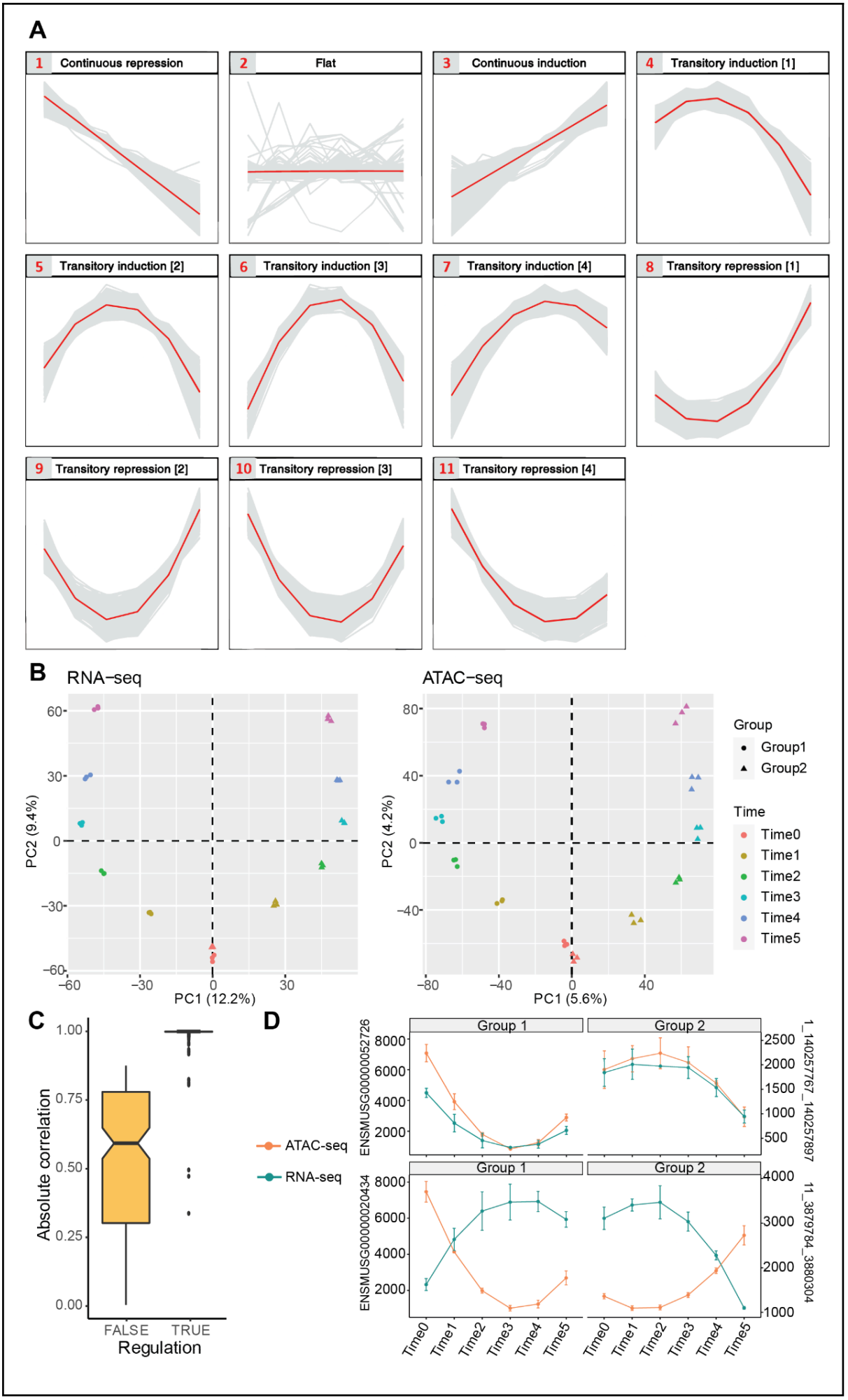
A
1 Continuous repression
2 Flat
3 Continuous induction
4 Transitory induction [1]
5 Transitory induction [2]
6 Transitory induction [3]
7 Transitory induction [4]
8 Transitory repression [1]
9 Transitory repression [2]
10 Transitory repression [3]
11 Transitory repression [4]
B
RNA−seq
ATAC−seq
PC1 (12.2%)
PC2 (9.4%)
PC1 (5.6%)
PC2 (4.2%)
Group
Group1
Group2
Time
Time0
Time1
Time2
Time3
Time4
Time5
C
Absolute correlation
FALSE
TRUE
Regulation
D
ATAC-seq
RNA-seq
Group 1
Group 2
ENSMUSG00000052726
ENSMUSG00000020434
1_140257767_140257897
11_3879784_3880304
Time0
Time1
Time2
Time3
Time4
Time5

Figure 2: A) Representation of K-means clusters for bulk RNA-seq in group 1. The red line represents the cluster mean profile. Each gene inside the cluster is represented by a gray line. The temporal simulated profiles associated to each cluster were: continuous repression (cluster 1), flat (cluster 2), continuous induction (cluster 3), transitory induction (from clusters 4 to 7 with the maximum in different intermediate time points) and transitory repression (from clusters 8 to 11 with the minimum in different intermediate time points). **B)** Exploratory analysis using Principal Component Analysis on low-count filtered data with logarithmic transformation. The first principal component separates the samples by experimental group, while the second one summarises the temporal profile. In X- and Y-axis, between brackets, the percentage of variability explained by the corresponding principal component. **C)** Boxplot of absolute Pearson 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 have an effect on gene expression. **D)** Two random examples of gene-regulator temporal profiles in each group. Left Y-axis shows gene expression values, while right Y-axis shows counts for ATAC-seq regions. Vertical bars at each time point show the standard deviation of the 3 simulated replicates.

Validation of the single-cell MOSim simulation approach

To illustrate the single-cell MOSim utilities, we simulated scRNA-seq and scATAC-seq data with four cell-types, three experimental groups and five replicates using as seed data the pbmcMultiome dataset available from SeuratData [12] and the gene-regulator association list available in the MOSim R package. We set the number of DEG to 30% upregulated and 20% downregulated in Group 2, and 10% upregulated and 40% downregulated in Group 3. We set a variance of 0.3 between replicates and 0.9 between experimental groups and allowed modelling of co-expression patterns across cell-types following eight random profiles. Lastly, we defined 20% activator and 10% repressor regulators in Group 1, 10% activators and 30% repressors in Group 2 and 10% activators and 20% repressors in Group 3.

Also in single-cell simulations, MOSim returns two types of results. The scOmicResults function retrieves a list containing the simulated data matrix for each omic, experimental group and biological replicate with features in rows and cells in columns.

The second results object, extracted with the scOmicSettings function, includes the MOSim-generated constraints associating gene and peak expression, as illustrated in Table 3. For instance, gene FAM138A is an upregulated DEG following across cell-type expression pattern 7 (Figure 3A). The chromatin accessible region chr2-200814169-200815379 is modelled as a significant activator of this gene, thereby expressing the same across cell-type profile as the regulated gene. On the other hand, the association of gene DHCR24 and chromatin accessible region chr3-144809419-144800982 is an example of a repressor effect of the regulator omic, where gene and peak follow opposite expression patterns (clusters 4 and 3, respectively) and the gene is downregulated when the regulator is upregulated (Table 3).

Table 3: MOSim-defined constraints for scRNA-seq and scATAC-seq for the simulation example. 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 expression profile across cell-types; G_DE: how the gene is differentially expressed; P_DE: how the peak is differentially expressed. 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 expression in Group 2 compared to Group 1.

Gene_ID	Peak_ID	RegEffect	G_cluster	P_cluster	G_DE	P_DE	G_FC	P_FC
FAM138A	2-200814169-200815379	Activator	7	7	Up	Up	70.2	87.6
TTC22	18-24394976-24395868	Activator	2	2	Up	Up	7.9	54.7
DHCR24	3-144809419-144809802	Repressor	4	3	Down	Up	0.13	66.5
GCC2-AS1	15-57324318-57324463	Repressor	8	6	Down	Up	0.13	4.6
ABCA4	NA	NE	7	NA	NE	NA	2.72	NA
NA	1-903617-907386	NE	NA	0	NA	NE	NA	1.69

To demonstrate the robustness of the single-cell MOSim framework's GRN simulation, we assessed it's ability to generate the expected across cell-type profiles. Employing the K-means method, we clustered the simulated feature profiles across cell-types and compared the simulated profile to the average expression of the MOSim-assigned cluster pattern. A feature was deemed correctly simulated if both profiles coincided. K-means clustering for the

simulated scRNA-seq and scATAC-seq unveiled the majority of features faithfully adhered to the mean cluster profile (Figure 3A).

We further evaluated the simulated data by examining the standard variance per average expression. Our findings revealed that the data dispersion recapitulated the distinct variance between replicates and experimental groups specified by the user (Figure 3B). Following individual sample variance assessment, Uniform Manifold Approximation and Projection (UMAP) was used as a dimensionality reduction framework to evaluate clustering of cells based on biological replicates, cell-types and experimental groups. UMAP exhibited robust clustering of the simulated data, reproducing a high-quality single-cell dataset where replicates, cell-types and experimental groups clustered cohesively (Figure 3C). Finally, Figure 3D, showcases the simulated across cell-type profiles for two pairs of gene-regulator associations, with one acting as an activator and the other as a repressor.

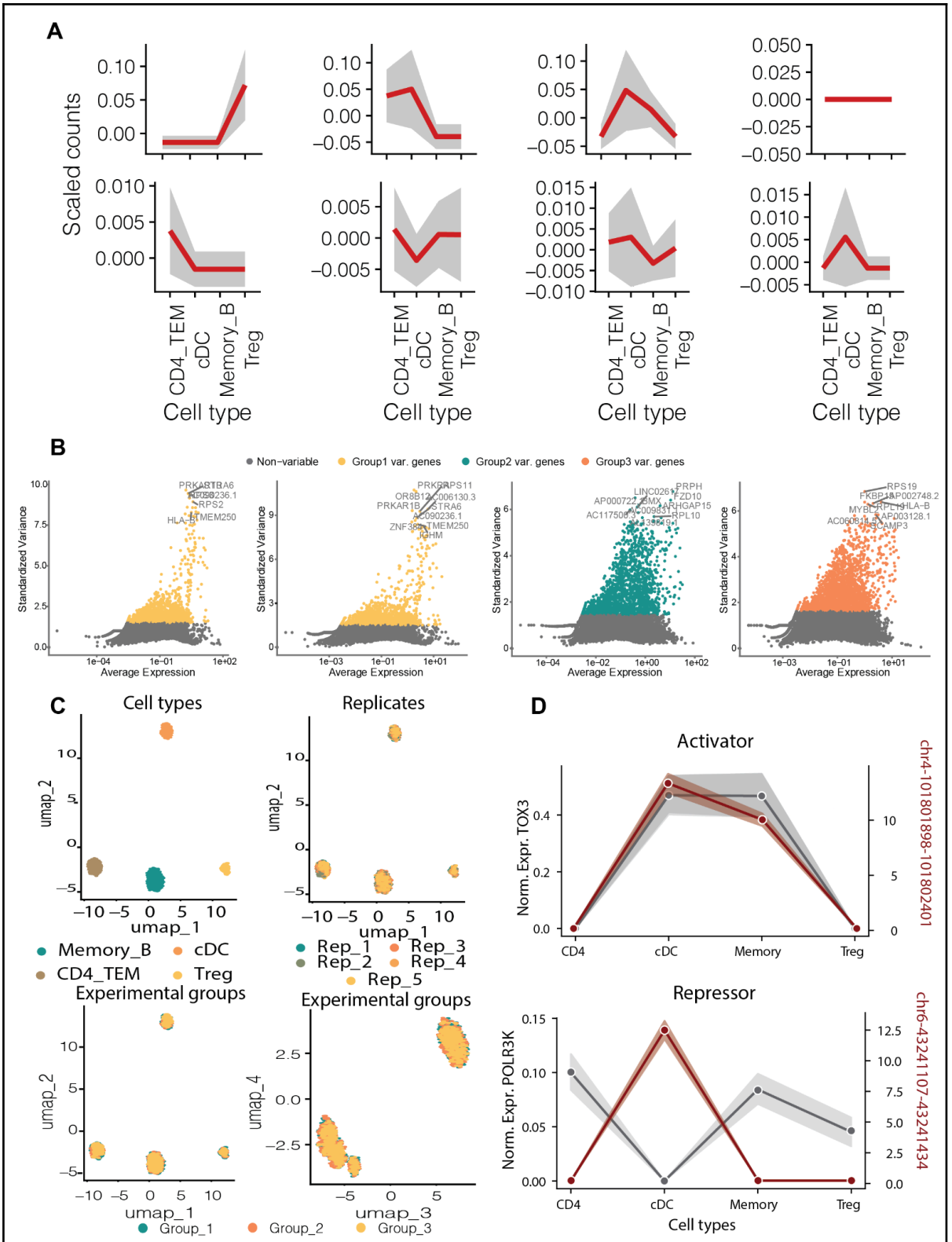

Figure 3: A) Representation of clustering patterns across cell-types. The red line represents the cluster mean profile. The shaded grey line represents standard deviation. **B)** Variance across gene expression in two biological replicates, and all 3 simulated experimental groups. Variable genes are coloured in yellow, green and orange. Black dots represent non-variable genes. **C)** UMAP clustering of cells coloured by biological replicate, cell-type and experimental group. **D)** Examples of two gene-regulator profiles across cell-types. Left Y axis shows gene expression values and right Y-axis shows counts for scATAC-regions.

Shaded areas correspond to 95% confidence intervals of feature expression between cells within each cell-type.

Discussion

Multi-omic assays, facilitated by massive sequencing technologies, have significantly improved our capacity to profile the regulatory mechanisms governing biological systems [1,2]. This progress has resulted in a better understanding of diseases and model organisms. However, benchmarking studies of bioinformatic tools designed for integrating multi-omics data, with the goal of unraveling multi-layered GRNs, have revealed significant discrepancies among various library preparation strategies and analysis methods[13]. This discrepancy underscores the intricate challenge of accurately identifying GRNs. As the adoption of multi-omic sequencing for studying regulatory mechanisms continues to expand, it becomes imperative to have access to tools capable of facilitating the precise assessment of GRN inference.

MOSim has been developed as an effective tool to provide a reliable framework for simulating bulk and single-cell multi-omics data in a controlled manner. Using a seed dataset and regulator-gene association matrix, MOSim generates simulated count matrices for both bulk and single-cell transcriptomics data, as well as count matrices for their associated regulatory omics. This approach employs the negative-binomial distribution for bulk data and leverages the well-established single-cell simulator SPARSim[10]. As a result, MOSim simulates count matrices that closely resemble real data, offering a more authentic representation compared to simulators that construct count matrices artificially without using a seed dataset as reference to infer distributions [14].

Moreover, MOSim operates at the level of count matrices rather than focusing on the read simulation task. This ensures a unified framework for simulating multi-omics data across different library preparation strategies (i.e. SmartSeq2, 10x Genomics), providing users with the flexibility to choose their preferred method and use it as seed dataset for MOSim.

MOSim enables for a fast and effortless generation of bulk and single-cell count data matrices for multiple omic data types, accommodating flexible experimental designs. Importantly, the algorithm is designed to simulate multiple regulatory relationships between gene expression and other molecular components in a manner consistent with a priori information, such as target mRNA-microRNA associations. The high flexibility in defining experimental designs, the number of DEGs, and active regulators makes the package a versatile tool for validating methods that aim to model complex multi-layered regulatory programs.

The MOSim framework has some limitations. It relies on a user inputted association matrix to establish the connection between regulators and transcripts. While this feature provides users with the flexibility to tailor MOSim for generating relationships resembling those intended for identification by their integration tool, there is a risk of users overfitting the MOSim simulation to meet the specific requirements of their tool. Furthermore, the current capabilities of single-cell simulation are limited to scRNA-seq and scATAC-seq, as these are the only two sequencing techniques that can be simultaneously executed on the same cell. Finally, both the bulk and single-cell modules are constrained to simulating data that could be generated by performing a sequencing experiment. Consequently, it may not fully capture the complexity of other molecular layers, such as proteomics or metabolomics, thereby limiting its capacity to simulate comprehensive multi-omics datasets.

In conclusion, the integration of multi-omics datasets for GRN identification remains a challenging task. We demonstrate that MOSim stands as an indispensable resource for benchmarking integration tools, addressing a significant gap in the multi-omics sequencing field.

Software 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/) and GitHub (https://github.com/ConesaLab/MOSim). The source code can also be found as supplemental Code. The software mainly uses R packages dplyr [15], purrr [15], Stats [16], Iranges [17], Seurat [18], SPARSim [10], and adapted scripts from Acorde [9] and WGBSSuite [8].

Data availability

Example bulk data available in MOSim was generated by STATegra project [11]. Example single-cell multiome data accessed by scMOSim is available via SeuratData R package [12].

Competing interests

The authors declare no competing interests.

Acknowledgements

This work was supported by the FP7 STATegra project (agreement no. 306000), the Spanish MINECO (BIO2012-40244), and the Spanish Bioinformatics Institute support

(PT17/0009/0015-ISCIII-SGEFI/ERDF). The authors thank Maider Aguerralde for rigorous testing and feedback on the software, and Arianna Febbo.

Authors' contributions

S.T., A.C. and C.M.M. conceptualised and designed the bulk MOSim approach. S.T., A.C., C.M. and A.A.L. conceptualised the scMOSim approach. C.M.M. developed and implemented bulk MOSim. C.M. developed and implemented scMOSim and contributed to implementing bulk MOSim. C.M.M, C.M. and S.T. performed the analysis and generated visualisations. A.A.L. and contributed to implementing scMOSim. 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.

1. Marku, M. & Pancaldi, V. From time-series transcriptomics to gene regulatory networks: A review on inference methods. *PLoS Comput. Biol.* **19**, e1011254 (2023).
2. Badia-I-Mompel, P. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* **24**, 739–754 (2023).
3. Chalise, P., Raghavan, R. & Fridley, B. L. InterSIM: Simulation tool for multiple integrative 'omic datasets'. *Comput. Methods Programs Biomed.* **128**, 69–74 (2016).
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* **8**, (2019).
5. Angelin-Bonnet, O., Biggs, P. J., Baldwin, S., Thomson, S. & Vignes, M. sismonr: simulation of in silico multi-omic networks with adjustable ploidy and post-transcriptional regulation in R. *Bioinformatics* **36**, 2938–2940 (2020).
6. Song, D. *et al.* scDesign3 generates realistic in silico data for multimodal single-cell and

spatial omics. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01772-1.

7. Li, H., Zhang, Z., Squires, M., Chen, X. & Zhang, X. scMultiSim: simulation of multi-modality single cell data guided by cell-cell interactions and gene regulatory networks. *bioRxiv* (2022) doi:10.1101/2022.10.15.512320.
8. Rackham, O. J. L., Dellaportas, P., Petretto, E. & Bottolo, L. WGBSSuite: simulating whole-genome bisulphite sequencing data and benchmarking differential DNA methylation analysis tools. *Bioinformatics* **31**, 2371–2373 (2015).
9. Arzalluz-Luque, A., Salguero, P., Tarazona, S. & Conesa, A. acorde unravels functionally interpretable networks of isoform co-usage from single cell data. *Nat. Commun.* **13**, 1828 (2022).
10. Baruzzo, G., Patuzzi, I. & Di Camillo, B. SPARSim single cell: a count data simulator for scRNA-seq data. *Bioinformatics* **36**, 1468–1475 (2020).
11. Gomez-Cabrero, D. *et al.* STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. *Sci Data* **6**, 256 (2019).
12. Satija Lab, L. pbmcMultiome.SeuratData: 10X Genomics PBMC Multiome Dataset. (2022).
13. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
14. Cao, Y., Yang, P. & Yang, J. Y. H. A benchmark study of simulation methods for single-cell RNA sequencing data. *Nat. Commun.* **12**, 6911 (2021).
15. Wickham, H. *et al.* Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
16. *R: A Language and Environment for Statistical Computing : Reference Index*. (R Foundation for Statistical Computing, 2010).
17. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
18. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).