MOSim: Multi-Omics Simulation in R

Carlos Martínez-Mira[1], Ana Conesa[2,3]* and Sonia Tarazona[1,4]

1 Genomics of Gene Expression Laboratory, Centro de Investigación Príncipe Felipe, Valencia, Spain
2 Microbiology and Cell Science Department, institute for Food and Agricultural Research, University of Florida, USA
3 Genetics Institute, University of Florida, Gainesville, FL, USA
4 Applied Statistics, Operational Research and Quality Department, Polytechnic University of Valencia, Valencia, Spain

* aconesa@ufl.edu

Abstract

Motivation: As new integrative methodologies are being developed to analyse multi-omic experiments, validation strategies are required for benchmarking. *In silico* approaches such as simulated data are popular as they are fast and cheap. However, few tools are available for creating synthetic multi-omic data sets.
Results: `MOSim` is a new R package for easily simulating multi-omic experiments consisting of gene expression data, other regulatory omics and the regulatory relationships between them. `MOSim` supports different experimental designs including time series data.
Availability: The package is freely available under the GPL-3 license from the Bitbucket repository (https://bitbucket.org/ConesaLab/mosim/).
Contact: cmartinez@cipf.es
Supplementary information: Supplementary material is available at *bioRxiv* online.

1 Introduction

Advances in massive sequencing technologies are favoring the proliferation of experiments applying several omics assays on the same biological system. Consequently, there is an increasing need of bioinformatics tools to help scientists in the processing of multi-omics data, including the validation of novel integration methodologies and the tuning of multi-omics analysis pipelines. A common strategy for the validation of analysis methods is the utilization of synthetic data where researchers define which features (e.g. genes) change across conditions and how these features are regulated by other features (e.g. microRNAs).

Several simulation algorithms exist for specific data types: compcodeR, polyester, FluxSimulator, MetaSim, dwgsim, ART, to cite a few [1]. However, there is a lack of tools for multi-omics simulation, given the complexity of the data structure. Some publicly available algorithms can simulate several omic data types as well as interactions among features [2], but allow for very limited experimental designs and do not offer flexible and user-friendly ways of modifying regulatory relationships.

In this work, we present `MOSim`, an R algorithm to simulate multi-omics data sets. `MOSim` generates count data for different sequencing assays with flexible choices for experimental designs. More importantly, the tool also simulates regulatory programs that link gene expression with other omic features (CpG sites, transcription factors, miRNAs, etc.) by defining the values of gene regulators as a function of their regulatory effect on gene expression (activation or repression). `MOSim` is a useful tool to test the performance of integrative methodologies, benchmark analysis pipelines before experimental data are available and generate examples for teaching purposes or user manuals.

2 Methods

In order to create a synthetic multi-omic dataset, `MOSim` requires as input a list of omics to simulate, one seed data file for each of them, information on *a priori* or potential regulatory features of each gene, and several configuration parameters (experimental design, dispersion, number of features, number of differentially expressed genes (DEGs), etc). Supported omic data types are RNA-seq, ATAC-seq (or DNase-seq), ChIP-seq, miRNA-seq and Methyl-seq. In case transcription factor (TF) regulation is also modeled, this feature type must be indicated and the corresponding association file included. The package contains a seed dataset obtained from the STATegra project (GEO accession numbers GSE75395, GSE38169 and GSE42462) with these data types and associations. Users must provide experimental design information by indicating the number of experimental groups, time-points if applicable, and the number of replicates per experimental condition. A detailed description of the algorithm can be found in the Supplementary material.

The simulation starts by creating the gene expression dataset. DEGs are randomly selected from the seed RNA-seq sample. For a time course design, DEGs are labeled with one of the following patterns in each experimental group: continuous induction (increasing linear pattern), continuous repression (decreasing linear pattern), transitory induction (quadratic pattern with a intermediate maximum), transitory repression (quadratic pattern with a intermediate minimum) and flat, which is also the pattern for non-DEGs (Table 1). Expression profiles are simulated from the seed count values to recapitulate real data distributions. DEGs with flat profiles or at case-control designs are modeled by introducing a fold-change in one of the experimental conditions. Once gene expression values are generated for each condition, replicates are simulated from a negative binomial (NB) distribution with mean equal to the count value for

that condition and variance proportional to the mean.

ID	DE	Group1	Group2
ENSMUSG00000097082	TRUE	transitory.induct	transitory.induct
ENSMUSG00000020205	TRUE	transitory.induct	continuous.induct
ENSMUSG00000055493	TRUE	transitory.induct	continuous.repress
ENSMUSG00000087802	FALSE	flat	flat
ENSMUSG00000017204	TRUE	transitory.induct	continuous.repress
ENSMUSG00000017221	TRUE	transitory.induct	continuous.induct

Table 1: RNA-seq settings for a simulation example. *ID*: gene identifier; *DE*: indicates if the gene is differentially expressed (TRUE) or not (FALSE); *Group1*: temporal profile of the gene in experimental group 1; *Group2*: temporal profile of the gene in experimental group 2.

The simulation of the remaining omics uses the same pattern definition function subjected to the constrains of the provided regulatory data and a randomly chosen direction of regulation. Regulators labeled as activators will have the same profile as their associated gene, but the opposite if they have a repression effect (see an example in Table 2). For Methyl-seq, percentages are generated instead of counts based on the binomial distribution, following the strategy described in [3]; while for simulating TFs regulation, the expression values are extracted from the simulated RNA-seq data. Users may indicate the percentage of active regulators and the algorithm verifies that the regulatory network is consistent with the input association data.

ID	Gene	Effect.Group1	Effect.Group2	Group1	Group2
10_111588324_111588448	ENSMUSG00000097082	activator	activator	transitory.induc	transitory.induc
10_111588324_111588448	ENSMUSG00000020205	activator	NA	transitory.induc	transitory.induc
10_11358301_11358431	ENSMUSG00000055493	activator	activator	transitory.induc	continuous.repress
10_11358301_11358431	ENSMUSG00000087802	NA	NA	transitory.induc	continuous.repress
11_98682094_98682786	ENSMUSG00000017204	repressor	activator	transitory.repress	continuous.repress
11_98682094_98682786	ENSMUSG00000017221	repressor	repressor	transitory.repress	continuous.repress

Table 2: ATAC-seq settings for the simulation example in Table 1. *ID*: genomic coordinates of ATAC-seq region (chromosome, and start and end positions for chromatin-accessible regions); *Gene*: regulated gene; *Effect.Group1*: regulatory effect of the ATAC-seq region on gene expression in experimental group 1; *Effect.Group2*: regulatory effect of the ATAC-seq region on gene expression in experimental group 2; *Group1*: temporal profile of the ATAC-seq region in experimental group 1; *Group2*: temporal profile of the ATAC-seq region in experimental group 2.

Besides the `MOSim` general wrapper function to simulate a multi-omic dataset (`mosim`), other useful functions included in the package help users to modify seed data (`omicData`) or default omic parameters (`omicSim`) and to recover simulation results as explained in the next section.

3 Results

To illustrate `MOSim` utilities, we simulated RNA-seq and ATAC-seq data with 5 time points, 2 experimental groups, 3 replicates and STATegra samples as seed data. `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 object, extracted with the `omicSettings` function, contains the settings used to generate each omic data type and the modeled relationships between gene expression and the rest of omics, as illustrated in Tables 1 and 2. For instance, gene ENSMUSG00000055493 is a DEG with transitory induction in condition 1 and continuous repression in condition 2. The chromatin-accessible region 10_11358301_11358431 is modeled as a significant activator of this gene in both conditions, thereby expressing the same temporal profiles as the regulated gene.

4 Discussion

The new `MOSim` R package allows for a fast and effortless generation of count data matrices for multiple omic data types with flexible experimental designs. More importantly, the algorithm has been designed to simulate multiple regulatory relationships between gene expression and other molecular components in a way consistent with *a priori* information, such as target mRNA-microRNA associations. High flexibility in the definition of experimental designs, number of DEGs and active regulators makes the package a versatile tool to validate methods that aim to model complex multi-layered regulatory programs.

Funding

This work has been funded 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).

References

1. Zhao M, Liu D, Qu H. Systematic review of next-generation sequencing simulators: computational tools, features and perspectives. Briefings in Functional Genomics. 2017;16(3):121–128. doi:10.1093/bfgp/elw012.

2. Chalise P, Raghavan R, Fridley BL. InterSIM: Simulation tool for multiple integrative 'omic datasets'. Comput Methods Programs Biomed. 2016;128:69–74.

3. Rackham OJL, Dellaportas P, Petretto E, Bottolo L. WGBSSuite: simulating whole-genome bisulphite sequencing data and benchmarking differ-

ential DNA methylation analysis tools. Bioinformatics. 2015;31(14):2371–2373. doi:10.1093/bioinformatics/btv114.