Bioinformatics, 38(3), 2022, 866–868
doi: 10.1093/bioinformatics/btab678
Advance Access Publication Date: 29 September 2021
Applications Note

Systems biology

Multiomix: a cloud-based platform to infer cancer genomic and epigenomic events associated with gene expression modulation

Genaro Camele[1], Sebastian Menazzi[2], Hernán Chanfreau[2], Agustin Marraco[2], Waldo Hasperué[1], Matias D. Butti[2,3,*] and Martin C. Abba [3,*]

[1]Instituto de Investigación en Informática (LIDI), Facultad de Informática, Universidad Nacional de La Plata, La Plata B1900, Argentina, [2]Centro de Altos Estudios en Tecnología Informática (CAETI), Facultad de Tecnología Informática, Universidad Abierta Interamericana, Caba C1270AAH, Argentina and [3]Centro de Investigaciones Inmunológicas Básicas y Aplicadas (CINIBA), Facultad de Ciencias Médicas, Universidad Nacional de La Plata, La Plata B1900, Argentina

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on June 7, 2021; revised on September 10, 2021; editorial decision on September 20, 2021; accepted on September 23, 2021

Abstract

Motivation: Large-scale cancer genome projects have generated genomic, transcriptomic, epigenomic and clinicopathological data from thousands of samples in almost every human tumor site. Although most omics data and their associated resources are publicly available, its full integration and interpretation to dissect the sources of gene expression modulation require specialized knowledge and software.

Results: We present Multiomix, an interactive cloud-based platform that allows biologists to identify genetic and epigenetic events associated with the transcriptional modulation of cancer-related genes through the analysis of multiomics data available on public functional genomic databases or user-uploaded datasets. Multiomix consists of an integrated set of functions, pipelines and a graphical user interface that allows retrieval, aggregation, analysis and visualization of different omics data sources. After the user provides the data to be analyzed, Multiomix identifies all significant correlations between mRNAs and non-mRNA genomics features (e.g. miRNA, DNA methylation and CNV) across the genome, the predicted sequence-based interactions (e.g. miRNA–mRNA) and their associated prognostic values.

Availability and implementation: Multiomix is available at https://www.multiomix.org. The source code is freely available at https://github.com/omics-datascience/multiomix.

Contact: matias.butti@gmail.com or mcabba@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genomic and epigenomic changes lead to transcriptomic alterations affecting key cellular processes such as cell proliferation and differentiation, which are crucial for the development of neoplastic diseases. Copy-number alterations (CNAs) are important genomic events that can, through gene dosage variations, promote tumor progression via alteration of gene expression levels (Bhattacharya *et al.*, 2020). Epigenomic alterations are important mechanisms regulating gene expression, and its role in cancer has been extensively studied. These epigenetic mechanisms include DNA methylation, chromatin remodeling, non-coding RNAs (ncRNAs) and transcription factors (Kagohara *et al.*, 2018). Gene expression silencing due to promoter hypermethylation or miRNA-mediated mRNA degradation of tumor suppressors is frequently detected at early cancer developmental stages and contributes to aberrant activation of oncogenic pathways (Abba *et al.*, 2015). All these genomics and epigenomics alterations ultimately result in aberrant gene expression of cancer driver genes that may also affect patients' survival. Several large-scale cancer genome projects such as The Cancer Genome Atlas (TCGA), the Tumor Alterations Relevant for Genomics-driven Therapy (TARGET) and Genomics Evidence Neoplasia Information Exchange (GENIE) have generated genomic data among thousands of samples in almost all human tumor sites (Jensen *et al.*, 2017). Although a plethora of web applications and software focused on data retrieval, visualization and analyses have been developed, the full integration and interpretation of these multilayer data require specialized knowledge and dedication (Das *et al.*, 2020). Here, we present Multiomix, an interactive Python/Rust cloud-based platform that allows biologists to explore the transcriptional effects of DNA methylation, miRNA expression

and CNAs profiles from public cancer multi-omics repositories, which can lead to novel biological insights and biomarker discovery.

2 Implementation and main functions

Multiomix is a Python and Rust web application developed with Django framework and implemented in an asynchronous server using Daphne protocol and Redis WebSocket server. This architecture allows users to queue Multiomix pipeline executions and then continue working on the platform while they are being processed. When the execution is ready, results will be posted for user analysis. The backend consists of algorithms developed in Rust programming language due to the high computational and parallelization requirements. The correlation analyses were numerically integrated using GNU Scientific Library (GSL) and the Rust bindings provided by RGSL. In addition, the SciPy library is used to compute descriptive statistics and distribution tests. PostgreSQL (v.13) is employed to manage all the required annotation data for mapping gene/miRNA identifiers, methylation probes, chromosome positions and related information. Multiomix also provides a MongoDB (Multiomix DB) containing preloaded datasets that were programmatically retrieved from cBioPortal (https://www.cbioportal.org/) (Cerami *et al.*, 2012). The graphical user interface (GUI) of Multiomix is provided by React using Typescript and the ApexCharts and Data-rich UI libraries for interactive data visualization (density plots, correlation plots, boxplots and survival plots). Comprehensive documentation, data examples, pipeline tutorials and the instructions for setting up a local instance are available and hosted on https://www.multiomix.org. Users can register or use the following credentials to try it *'user: demo/password: demo'*.

Overall, Multiomix consists of an integrated set of functions for multi-omics data retrieval and aggregation from local sources or publicly available databases, and the statistical setup of three downstream pipelines (Supplementary Fig. S1). These functionalities are combined with exploratory and visualization capabilities of the obtained results in the context of clinical and follow-up data (phenodata). Using the datasets dashboard, users can upload their own data to Multiomix DB. Omics and clinical data obtained from relevant resources such as UCSC-Xena browser (http://xena.ucsc.edu/) (Goldman *et al.*, 2020) can be directly uploaded to Multiomix DB. Moreover, several cBioPortal datasets from TCGA projects are available for public use at Multiomix DB. Using the Analysis dashboard, users select the required datasets (e.g. mRNA and miRNA profiles), filtering criteria (e.g. minimum standard deviation, correlation threshold) and setup of the correlation (e.g. Pearson, Spearman, Kendall) and *P*-value adjustment (e.g. Benjamini-Hochberg, Benjamini-Yekutieli, Bonferroni) methods that will be applied in downstream pipeline described below.

(i) miRNA–mRNA pipeline: After the user provides the preprocessed data files to be analyzed—mRNA and mature or pre/pri-miRNA profiles—Multiomix evaluates all paired miRNA-mRNAs expression levels and identifies significant correlations after careful statistical adjustment. For each significant result, it shows target predicted sequence-based interactions through the mirDIP. miRDIP comprises almost 152 million human miRNA–target predictions (Tokar *et al.*, 2018). In addition, the identified miRNAs are enriched with data retrieved from mirBase (Kozomara et al., 2019), SM2miR (Liu *et al.*, 2013), the Human microRNA Disease Database (Huang *et al.*, 2019) and miRTarBase (Hsu *et al.*, 2011). All these miRNA databases are integrated in Modulector, an own API that simplifies and standardizes the access to miRNA information (https://github.com/omics-datascience/modulector).

(ii) DNA methylation–mRNA pipeline: After the user provides the required data—mRNA and DNA methylation (Beta- or M-values of CpG sites)—Multiomix identifies all significant correlations between the methylation levels of all CpG regions that belong to a specific gene and their gene expression levels across the genome. Multiomix supports automatic mapping of CpG probes to gene for 27K and 450K Illumina-based arrays.

(iii) CNA–mRNA pipeline: After the user provides the two preprocessed files—mRNA and CNA in log2FC, GISTIC2 or GISTIC2 thresholded profiles—Multiomix identifies all significant correlations between the CNA and mRNA expression levels for each gene across the genome. Subsequently, correlated features are aggregated by chromosomal coordinates to facilitate the detection of chromosome regions of potential focal amplification.

Moreover, all described pipelines also perform survival analysis (Kaplan-Meier curve and Log-rank test computation) of the modulated transcripts to estimate the biological impact of the genetic or epigenetic events detected. Finally, the obtained results can be downloaded (.tsv file) and/or visually explored in the context of phenodata (e.g. sample type, tumor stage, intrinsic subtypes, etc.). To illustrate the use of Multiomix, we performed an integrative analysis of mRNA and miRNA profiles with clinical and follow-up data retrieved from the TCGA breast cancer project (Supplementary Fig. S2). Multiomix miRNA-pipeline allowed us to identify a group of novel miRNAs involved in the epigenetic modulation of coding RNAs associated with breast cancer progression.

3 Discussion and future works

Currently, there exists a diverse number of software platforms, web applications and cloud-based services that mainly provide general-purpose bioinformatic toolkits and/or workflows for processing and analysis of functional genomics data (e.g. QC/trimming, Exome-Seq, ChIP-seq, RNA-seq, microbiome analysis) such as VisRseq (Younesy *et al.*, 2015), Galaxy (Afgan *et al.*, 2018) and Terra.Bio (Van der Auwera and O'Connor, 2020) among others. In contrast, a limited number of R-based packages were developed to facilitate the integration and analysis of public and preprocessed genomics data such as TCGAbiolinks (Colaprico *et al.*, 2016), TCGA-Assembler 2 (Wan *et al.*, 2016), TCGA2stat (Wei *et al.*, 2018) among others. Multiomix is a dedicated cancer biomarker discovery platform that facilitates the integration and mining process of public and user-uploaded oncogenomic data by providing a friendly GUI to non-expert users. Furthermore, Multiomix offers a solid cloud platform that can be locally implemented. It is shared with the open-source community to leverage the development of new functionalities. In this sense, the Multiomix database will be expanded over time, providing integration with other databases and new pipelines will be implemented for other ncRNA gene expression modulators (e.g. lncRNA–miRNA–mRNA). It will also integrate models based on feature selection algorithms to help users identify potential prognostic/predictive cancer biomarkers.

Funding

This work was supported by the National Agency of Scientific and Technological Promotion PICT-2018-01403, Universidad Abierta Interamericana (UAI) and Universidad Nacional de La Plata (UNLP).

Conflict of Interest: none declared.

References

Abba,M.C. *et al.* (2015) A molecular portrait of high-grade ductal carcinoma in situ. *Cancer Res.*, **75**, 3980–3990.

Afgan,E. *et al.* (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, **46**, W537–W544.

Bhattacharya,A. *et al.* (2020) Transcriptional effects of copy number alterations in a large set of human cancers. *Nat. Commun.*, **11**, 1–2.

Cerami,E. *et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, **2**, 401–404.

Colaprico,A. *et al.* (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, **44**, e71.

Das,T. *et al.* (2020) Integration of online omics-data resources for cancer research. *Front. Genet.*, **11**, 578345.

Goldman,M.J. *et al.* (2020) Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.*, **38**, 675–678.
Hsu,S.D. *et al.* (2011) miRTarBase: a database curates experimentally validated microRNA–target interactions. *Nucleic Acids Res.*, **39**, D163–D169.
Huang,Z. *et al.* (2019) HMDD v3. 0: a database for experimentally supported human microRNA–disease associations. *Nucleic Acids Res.*, **47**, D1013–D1017.
Jensen,M.A. *et al.* (2017) The NCI Genomic Data Commons as an engine for precision medicine. *Blood*, **130**, 453–459.
Kagohara,L.T. *et al.* (2018) Epigenetic regulation of gene expression in cancer: techniques, resources and analysis. *Brief. Funct. Genomics*, **17**, 49–63.
Kozomara,A. *et al.* (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
Liu,X. *et al.* (2013) SM2miR: a database of the experimentally validated small molecules' effects on microRNA expression. *Bioinformatics*, **29**, 409–411.
Tokar,T. *et al.* (2018) mirDIP 4.1—integrative database of human microRNA target predictions. *Nucleic Acids Res.*, **46**, D360–D370.
Van der Auwera,G.A. and O'Connor,B.D. (2020) *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media, Sebastopol, CA, USA.
Wan,Y.W. *et al.* (2016) TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. *Bioinformatics*, **32**, 952–954.
Wei,L. *et al.* (2018) TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics*, **34**, 1615–1617.
Younesy,H. *et al.* (2015) VisRseq: R-based visual framework for analysis of sequencing data. *BMC Bioinformatics*, **16**, S2– 4.