MORE interpretable multi-omic regulatory networks to characterize phenotypes

Maider Aguerralde-Martin, Mónica Clemente-Císcar,

Luis Lopez-Cárcel, Ana Conesa, Sonia Tarazona

January 2024

1 Abstract

The emergence of omics has revolutionized how we study and understand biological systems, enabling breakthrough discoveries with profound implications for medicine, agriculture, biotechnology, and more. However, with the help of advanced computational tools and artificial intelligence, meaningful patterns and relationships can now be uncovered in omics data, offering a unique opportunity to gain a deeper understanding and contribute to new insights into the complex regulatory mechanisms of biological systems. In this context, we have developed *MORE* (Multi-Omics REgulation), a tool designed to identify relevant regulations of the gene expression for the biological system under study and subsequently construct the regulatory networks for the considered experimental conditions. The presented method not only allows the incorporation of prior biological information into network construction but also can infer relationships de novo in the absence of such information. Moreover, we effectively addressed multicollinearity issues inherent in such data types, ensuring precise and reliable inference of regulatory networks when performing GLM models. In our comparison to KiMONo, our tool exhibited superior evaluation metrics, including F1-score, R2, and computational efficiency. Finally, applying our tool to a real ovarian cancer dataset yielded intriguing and biologically meaningful results. Our developed methodology represents a versatile and powerful multi-omic regulatory network inference approach, demonstrating good performance and applicability to real-world biological datasets. It is freely available at https://github.com/ConesaLab/MORE.git.

2 Introduction

In recent years, omics have become a powerful approach to unraveling complex biological systems. From genomics to proteomics, metabolomics to transcriptomics, and beyond, omics technologies have revolutionized how we study and

understand biological systems. The vast amount of information these technologies generate poses significant data analysis and interpretation challenges. However, with the help of advanced computational tools and artificial intelligence, meaningful patterns and relationships can now be uncovered in omics data, enabling breakthrough discoveries with profound implications for medicine, agriculture, biotechnology and more.

An illustrative example of the power of omics in understanding biological systems is the study of disease. Common diseases often result from changes in gene regulation, and the same genetic variants can contribute to different outcomes depending on environmental and genetic factors [Hasin et al., 2017]. By analyzing gene expression profiles, researchers aim to identify genes with closely aligned expression patterns, where the gene expression levels either ascend or descend concurrently under specific biological conditions [Randhawa and Pathania, 2020].

Although omic analysis has already revolutionized biological research, analyzing one data type may limit the identification of causal changes that drive biological phenomena, whereas combining different omics can provide a more comprehensive understanding of underlying regulatory mechanisms. Thus, multi-omics analysis could provide researchers with a more holistic and interdependent view of biological processes, offering a unique opportunity to gain a deeper understanding and contribute to new insights into the complex regulatory mechanisms of biological systems. However, as many researchers have already pointed out [Tarazona et al., 2021; Nguyen and Wang, 2020], multi-omic data analysis still faces many challenges due to the heterogeneity of the data across omic technologies, the missing values they present, and the difficulty of interpreting the multi-layered models.

Nevertheless, it is important to pursue solutions to these questions, as, for instance, the discovery of effective therapies hinges on a thorough comprehension of the molecular changes underlying diseases and their intricate interactions. Gene co-expression network analysis has proven effective in unveiling functional co-expressed gene modules associated with complex human diseases [Langfelder and Horvath, 2008]. In this context, a key question to be addressed is whether we can identify the regulatory mechanisms governing the behavior of specific biological components under varying conditions. Therefore, it would be necessary to understand the controlling mechanisms associated with changes in gene expression. The inference of regulatory networks constitutes a fundamental strategy to achieve this objective. Within this framework, we developed *MORE* (Multi-Omics REgultaion), a tool designed to identify potential regulators for the biological system under study and subsequently construct the regulatory networks for the considered experimental conditions.

In the gene regulatory network inference field, different approaches have been developed to analyze gene co-expression, and some examples are ARACNE

[Margolin et al., 2006], RegNetB [Alvarez and Woolf, 2011], TRENA [Pearl et al., 2017], DINA [Gambardella et al., 2013] or the one presented by W & H [Werhli and Husmeier, 2007]. Their main objective is to reconstruct and analyze transcriptional regulatory networks or identify gene sets whose co-regulation is condition-specific. While this approach can elucidate a wealth of information about the biological mechanisms underlying disease, as mentioned above, a multi-omics approach can enhance the understanding of these mechanisms compared to a single-omics approach.

At the multi-omics level, DCGRN [Kim et al., 2014] and BMNPGRN [Yuan et al., 2019] tools make an approach for the integration of DNA methylation and Copy Number Variation (CNV) data to delineate the regulation of the gene expression. Both tools employ a LASSO-based approach to discern the regulators influencing gene expression patterns. The second is an extension of the first one, which allows multiple methylation regulators to regulate a single gene. Theoretically, it should be extensible to other omics but the computational cost of doing so could be elevated. DINGO [Ha et al., 2015] and its extension, iDINGO [Class et al., 2017], employ a multi-omics approach for differential network analysis. DINGO, even if defined as a multi-omic approach, treats each omic separately and takes the results of each canal separately, potentially leading to the loss of information regarding causal relationships between omics. In the extension, iDINGO adopts a more integrated approach by creating chain graph models. These tools are valuable as they use both directed and undirected graphs in constructing regulatory networks, providing insights into the direction of regulations. However, due to potential computational limitations, the developers do not recommend using this model with datasets of more than 300 genes. There are several other tools that using a multi-omics approach for multi-omics network analysis are limited to use them on specific omics, we present several of them. IMNA [Chen et al., 2020] incorporates GWAS signals, gene-expression signatures, chromatin and protein interactions for the identification of potential key genes in regulatory networks. MoNET [Li et al., 2021] integrates the relationship between single nucleotide polymorphisms (SNPs), genes/proteins and metabolites for multi-omic network analysis. TFmiR [Hamed et al., 2015], a freely available web server, performs an integrative analysis of combinatorial regulatory interactions between transcription factors (TFs), microRNAs (miRNAs) and target genes that are involved in disease pathogenesis. SAMNetWeb [Gosline et al., 2014] analyzes gene expression datasets with upstream data, identifying condition-specific networks and pathways, and visualizes interactions through functional enrichment analysis. However, this software can only analyze two different types of data, i.e., mRNA and global proteomics. IntOmics [Pačínková and Popovici, 2023] uses Markov Chain Monte Carlo sampling to analyze gene expression, CNV, DNA methylation and biological prior knowledge to infer regulatory networks. It also estimates the empirical biological knowledge from the available experimental data to complement the missing biological prior knowledge. RACER [Li et al., 2014], uses CNV, DNA methylation, transcription factor and microRNA explanatory variables to model each

mRNA expression gene. It applies a two-stage regression, it first fits sample-specific regulatory activities by TFs and miRNAs, then used as inputs to infer TF/miRNA-gene interactions. However, we found some models that are not restricted to specific omics, so that can be more comparable to our method. CANTARE [Siebert et al., 2021], fits first pairwise regression models across all pairs of omics to construct the resulting network. Then it employs logistic regression to predict the outcome using the highly connected subnetworks and other non-omic variables. PARADIGM [Vaske et al., 2010] models genes as interconnected variables, providing insights into gene expression and activity via curated pathways. However, it lacks the capacity to identify specific regulators influencing individual genes from omics data. KIMONo [Ogris et al., 2021] optimises sparse group LASSO penalization in the multivariate regression to model gene expression of each gene separately by the multi-omics measurements combined with prior information from biological sources. Finally, these fitted models are aggregated in the final heterogeneous multi-omics network. The MINT version of mixOmics [Rohart et al., 2017], one of the most popular multi-omics approaches used currently for multi-omics data integration, creates a Partial Least Squares (PLS) model. They implemented several approaches, such as PLS, block PLS, sparse PLS (sPLS) and sparse block PLS models. Using a kind of Lasso approach for the variable selection in sparse approaches. In the context of dimensionality reduction approaches, we found the unsupervised factor analysis MOFA [Argelaguet et al., 2018], which derives a low-dimensional representation of multi-omics data in terms of a few latent factors. These factors are used to learn about the sources of variation and to identify cellular states or disease subgroups. Results can only be visualized.

In this context, we introduce *MORE*, which, to the best of our knowledge, stands as one of the first software tools capable of using prior biological knowledge to construct a specific gene regulatory network for each of the phenotypes under study (e.g. control/disease). *MORE* achieves it by modeling the gene expression through the incorporation of potential regulators to each gene using either Generalized Linear Models(GLM) with ElasticNet or Iterative Sparse Group Lasso (ISGL) penalizations or PLS models. In case the user does not have any prior biological knowledge it also enables to identify de novo regulations of the gene expression. In Figure 1, we introduce a graphical abstract so that the reader has a previous image of *MORE*'s construction.

3 Materials and Methods

3.1 Materials

3.1.1 Simulated data

In order to compare the performance of *MORE* first with its various variable selection approaches and then with other already available tools, it was necessary to have a database in which we had a complete understanding of the actual

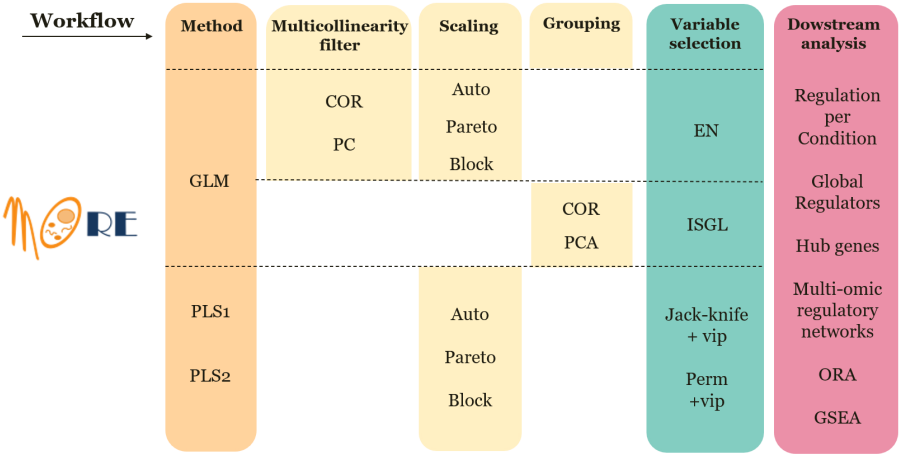


Figure 1: MORE overview

regulators of gene expression. Obtaining such information from a real database can be complicated, almost impossible. For this reason, we decided to create simulated databases. After comparing and investigating different tools for simulating multi-omics data, we finally chose *MOSim* [Monzó et al., 2024] since it gave us more flexibility to simulate data from different types of omics and data types. It also gave us more information about the associations between regulators and gene expression than other tools and more information that could be useful for calculating the predictive power of our tool.

To have a complete view of *MORE*'s performance, instead of simulating data only under one specific condition, we created the databases under several conditions to analyze whether *MORE*'s performance improved depending on the conditions of the databases. Under these premises, twelve scenarios were simulated using *MOSim*. Nevertheless, in all scenarios, two groups (cases and controls) and five types of omics data were simulated, including RNA-seq, ChIP-seq, miRNA-seq, Methyl-seq, TF, and ATAC-seq (still denoted as DNase-seq in the simulator). The sequencing depth of the data was set to 30 million reads, and the default number of genes, 38.293, was used for all simulated datasets. We simulated different numbers of observations per group: 5, 10, 20, 30 and 50.

The twelve scenarios were obtained from the combination of the following settings:

- Three different seeds: 123, 321, and 231.
- Two different differential expression scenarios ranging from 3% to 35% of genes exhibiting differential expression.
- Two different percentages of significant regulators: 40, 60. In all cases,

except for miRNA, the same percentage of activator and repressor regulators were simulated. In miRNA's case, all simulated significant regulators were activators following the biological sense.

Since the main goal of *MORE* is to compare multi-omics regulatory networks across experimental conditions, analyzing those genes that do not show differential expression across conditions makes little sense. For this reason, it is strongly recommended to fit *MORE* only to deferentially expressed (DE) genes.

3.1.2 TCGA Ovarian Cancer data

We used a high-grade serous ovarian adenocarcinoma (HGSOC) dataset from The Cancer Genome Atlas Ovarian Cancer Cohort [Cancer Genome Atlas Research Network, 2012] to exemplify the usage of MORE. The collection consisted of the following omic modalities and with the following number of features for each of them (after appropriate filtering described below): gene expression measured by RNA-seq (11.700 genes), Copy Number Variation (24.776 CNV), DNA Methylation (21.876 CpG sites), microRNA expression (756 miRNAs), and Transcription Factors extracted from the RNA-seq data (947 TFs). The first three omics were obtained from the UCSC Xena data portal [UCSC, 2020], miRNA data was extracted from the Broad Institute FileBrowse data portal, and their identifiers were updated to the latest available version of miRNA names using miRCarta identifier conversion.

In order to use biological prior knowledge when creating MORE networks, both the list of TFs and their target genes were obtained from the TF Link database [Liska et al., 2022]; CNVs were given at gene level; methylation probes were associated with genes according to information provided by Illumina; and miRNA target genes were obtained from the mirWalk database [Sticht et al., 2018].

The phenotypes to be compared with *MORE* were the four different cancer subtypes defined in [Cancer Genome Atlas Research Network, 2012]: differentiated, immunoreactive, mesenchymal, and proliferative, also obtained from the UCSC data portal. Finally, only the 292 patients with information for all the omics and the subtypes were considered.

The downloaded gene expression data were already normalized (RSEM values) and log-transformed. The NOISeq R package [Tarazona et al., 2015] was applied to filter out low-count genes. We generated MORE networks for genes presenting changes in expression for at least one cancer subtype. Such genes were identified by a differential expression analysis performed with the limma R package [Ritchie et al., 2015] (adjusted p-value < 0.01).

3.2 Methods

3.2.1 MORE general framework

Let $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_G]$ denote the response data matrix (e.g. gene expression or protein expression) of size N observations (rows) and G variables (e.g. genes or proteins) (columns). In advance, we will talk about gene expression as the response variable for simplicity, but it works the same with other such as protein expression. Let now denote $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_K]$ as the data matrix of N observations and K variables, which contains data of the expression of the regulators of the gene expression which come from different omics. Assuming that biological prior knowledge is provided, it would be obvious to say that not all regulators from all omics regulate the expression of all the genes, so it is clear that data would be partitioned in G different studies $(X^{(1)}, \mathbf{y}_1), ..., (X^{(G)}, \mathbf{y}_G)$, where $X^{(g)} = [\mathbf{x}_1^{(g)}, ..., \mathbf{x}_K^{(g)}]$ is a partition of $\mathbf{X}$ data matrix with data from the regulators that potentially regulate the gene expression of g gene, $\mathbf{y}_g$. However, in case there is no prior biological knowledge, the G different studies would reduce to a unique one $(\mathbf{X}, \mathbf{Y})$, where all regulators contained in $\mathbf{X}$ potentially regulate all genes in $\mathbf{Y}$.

In case biological prior knowledge is provided and G models want to be computed, *MORE* can use GLM (with any of the penalties) and PLS1 to model the gene expression and induce regulatory networks. However, we want to make clear that the main objective of *MORE* is not the prediction of the gene expression but the relation between the experimental conditions and the regulators in their influence on changes in the gene expression. In future work, we plan to incorporate variants into *MORE*. This enhancement aims to enable gene expression prediction, even the prediction of experimental groups based on their multi-omic signature.

Additionally, it is important to note that, regardless of the chosen methodology within *MORE*, prior to applying GLM or PLS models, the algorithm employs filters to eliminate variables with low variation or missing values. For GLM models, variables with at least one missing value are excluded, as this methodology cannot handle them. However, in the case of PLS models, variables with up to 20% missing values will still be considered since our PLS models are constructed using the NIPALS algorithm, which can effectively handle missing values in variables.

3.2.2 Generalized Linear Models with ElasticNet in MORE

When fitting a linear regression model, the main objective is to model the response using the data in $X^{(g)}$ by $\mathbf{y}_g = \boldsymbol{\beta}_g X^{(g)}$. In contrast, when fitting a generalized linear model, the objective is to model $\mathbf{y}_g$ using a link function $\boldsymbol{\eta}_g$ such that $\boldsymbol{\eta}_g = \boldsymbol{\beta}_g X^{(g)}$.

Even if these models are nowadays and have been one of the most widely used predictive models in many fields, they are not without challenges. Due to the nature of multi-omic data, one of the challenges that we will need to face in GLM models is multicollinearity. The presence of multicollinearity in GLM models can make the coefficients of the regression models become very sensitive to small changes and, therefore, the interpretation of the model incorrect. Since *MORE* is focused on the interpretation of the coefficients for the construction of the models, we introduce a multicollinearity filter previous to the construction of the GLM model, with which we aim to reduce the number of predictors in the model and mitigate multicollinearity problems in the regression model.

The multicollinearity filter that we introduce in *MORE* is based on the correlation between the variables in $X^{(g)}$. Its procedure is outlined in Algorithm 1.

Algorithm 1 Multicollinearity filter structure

```
procedure
Compute the correlation for each pair of regulators:
    Continuous regulators → Pearson's coefficient
    Binary regulators → Phi's coefficient
    Binary vs Continuos regulator → Point biserial coefficient
Generate networks connecting regulators with correlation higher than a
threshold specified by the user
Obtain connected components of the network
    for each connected component do
        if only one regulator then
            Include in the GLM model
        else
            while there are connected components do
                if it is a fully connected graph then
                    Take as representative the one with highest correlations
                else
                    Take as representative the one with highest degree
                    (in case of tie, the one with highest correlations)
                    Remove regulators associated to it from the network
```

where Pearson's, Phi's, and point biserial coefficients for different variables $\mathbf{x}_i^{(g)}$ and $\mathbf{x}_j^{(g)}$ are calculated by the equations 1, 2 and 3 respectively.

$$\rho_{\mathbf{x}_i^{(g)},\mathbf{x}_j^{(g)}} = \frac{cov(\mathbf{x}_i^{(g)}, \mathbf{x}_j^{(g)})}{\sigma_{\mathbf{x}_i^{(g)}} \cdot \sigma_{\mathbf{x}_j^{(g)}}} \tag{1}$$

where *cov* denotes the covariance and $\sigma_{\mathbf{x}_i^{(g)}}$ and $\sigma_{\mathbf{x}_j^{(g)}}$ the standard deviation of

$\mathbf{x}_i^{(g)}$ and $\mathbf{x}_j^{(g)}$ variables respectively.

$$\phi_{\mathbf{x}_i^{(g)},\mathbf{x}_j^{(g)}} = \frac{(ad-bc)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{2}$$

where a, b, d and c are the counts of the contingency table.

$$r_{\mathbf{x}_i^{(g)},\mathbf{x}_j^{(g)}} = \frac{\overline{\mathbf{x}}_{i1}^{(g)} - \overline{\mathbf{x}}_{i0}^{(g)}}{\sigma_{\mathbf{x}_i^{(g)}}}\sqrt{pq} \tag{3}$$

where p is the proportion of 1's in the binary variable $\mathbf{x}_j^{(g)}$, $q = 1-p$ is the proportion of 0's in the binary variable $\mathbf{x}_j^{(g)}$, $\overline{\mathbf{x}}_{i1}^{(g)}$ and $\overline{\mathbf{x}}_{i0}^{(g)}$ are the conditional means of the continuous variable $\mathbf{x}_i^{(g)}$ when the binary variable $\mathbf{x}_j^{(g)}$ is 1 or 0, respectively and $\sigma_{\mathbf{x}_i^{(g)}}$ is the standard deviation of $\mathbf{x}_i^{(g)}$.

However, several authors have pointed out that the correlation calculation can lead to spurious relationships. For that reason, we include a second option for the multicollinearity filter in which we modify the calculation of the correlations by the calculation of the partial correlation (Equation 4).

$$\rho_{\mathbf{x}_i^{(g)}\mathbf{x}_j^{(g)}.\mathbf{X}\setminus\{\mathbf{x}_i^{(g)},\mathbf{x}_j^{(g)}\}} = -\frac{p_{\mathbf{x}_i^{(g)}\mathbf{x}_j^{(g)}}}{\sqrt{p_{\mathbf{x}_i^{(g)}\mathbf{x}_i^{(g)}}p_{\mathbf{x}_j^{(g)}\mathbf{x}_j^{(g)}}}} \tag{4}$$

where $(p_{\mathbf{x}_i^{(g)}\mathbf{x}_j^{(g)}}) = \Sigma^{-1}$ is computed by the inverse of the correlation matrix computed by the coefficients that we introduced above.

Even if the multicollinearity filters introduced here aim to mitigate the multicollinearity present in the models and reduce the number of variables of the model, and do so in a sense, they do not solve the whole problem (there could still be problems with collinear variables and it could still be a high-dimensionality scenario). It is for that reason that in *MORE*, we apply GLM models together with ElasticNet penalization (Equation 5) introduced by [Zou and Hastie, 2005]. ElasticNet can better cope with multicollinearity but still faces some problems with it, so there is a need to maintain the multicollinearity filter.

$$EN(\hat{\boldsymbol{\beta}}_g) \equiv argmin_{\boldsymbol{\beta}\in\mathbb{R}^G}(\parallel \boldsymbol{\eta}_g - \boldsymbol{\beta}_g X^{(g)} \parallel^2 + \lambda[(1-\alpha) \parallel \boldsymbol{\beta}_g \parallel^2 + \alpha \parallel \boldsymbol{\beta}_g \parallel_1]) \tag{5}$$

where λ is the shrinkage parameter and α controls the degree of mixing between Ridge regression and Lasso. In *MORE* both parameters can be optimized, λ using cross-validation technique and α by a grid search.

Given the nature of multi-omics data, it is unsurprising that data from different omics are measured in different scales. In ElasticNet models, it is crucial to scale the variables before applying the penalization so that all the coefficients of the models are equally interpretable and penalized indifferently of their range

of values. In *MORE*, we introduce three different scaling types for the user to select the one more appropriate for its data. Indeed, the auto-scaling, Pareto-scaling, and block-scaling (Equations 6,7 and 8, respectively).

$$auto(\mathbf{x}_i^{(g)}) = \frac{\mathbf{x}_i^{(g)}}{\sigma_{\mathbf{x}_i^{(g)}}} \tag{6}$$

$$pareto(\mathbf{x}_i^{(g)}) = \frac{\mathbf{x}_i^{(g)}}{\sigma_{\mathbf{x}_i^{(g)}} \sqrt[4]{m_i}} \tag{7}$$

$$block(\mathbf{x}_i^{(g)}) = \frac{\mathbf{x}_i^{(g)}}{\sigma_{\mathbf{x}_i^{(g)}} \sqrt{m_i}} \tag{8}$$

where $\sigma_{\mathbf{x}_i^{(g)}}$ is the standard deviation of $\mathbf{x}_i^{(g)}$ variable and m_i is the number of regulators that belong to the same omic as $\mathbf{x}_i^{(g)}$.

3.2.3 GLMs with Iterative Sparse Group Lasso in MORE

In recent years, a large amount of work has been done in the field of GLM models with penalized approaches to improve their results. In the previous section, we included how we used ElasticNet penalization and tried to avoid the problems it presents with the multicollinearity. In order to avoid problems with multicollinearity and the application of previous filters to apply the methodology to guarantee the results, we realized a revision of the literature looking for a methodology that could handle the presence of multicollinearity. In that search, we focused on Iterative Sparse Group Lasso (ISGL) penalization, which could handle multicollinearity problems and has shown successful results with multi-omics data [Laria et al., 2019].

ISGL methodology is a coordinate descent algorithm for Sparse Group Lasso (SGL) [Simon et al., 2013] hyper-parameter selection which overcomes the overfitting of SGL hyper-parameter selection in a high-dimensional scenario. We selected this methodology since, due to its nature, it really fits the problematic present in multi-omics data.

For the readers complete understanding, we remember that the SGL penalization methodology is a linear combination of Lasso and Group Lasso penalizations which provides solutions both between and within group sparse (Equation 9).

$$SGL(\hat{\boldsymbol{\beta}}_g) \equiv \underset{\boldsymbol{\beta} \in \mathbb{R}^G}{argmin}(\parallel \boldsymbol{\eta}_g - \boldsymbol{\beta}_g X^{(g)} \parallel^2 + \lambda[(1-\alpha) \sum_{l=1}^{L} \sqrt{p_l} \parallel \boldsymbol{\beta}_g^{(l)} \parallel_2 + \alpha \parallel \boldsymbol{\beta}_g \parallel_1]) \tag{9}$$

where L denotes the number of created groups and p_l the number of variables that belong to that group.

The methodology, considers then that variables belong to some group. By the nature of the data that we are considering, one can think about the omics as the groups. However, even if this methodology couples better than ElasticNet with multicollinearity, we decided to create the groups of the SGL taking into account the collinearity of the variables and reduce in that way any problematic that could still face the methodology. To do so, we propose two different methodologies to create those groups. The first, creates the groups inferring the clusters obtained by the application of steps 2-7 of Algorithm 1. So that, it creates groups of correlated variables and isolates the variables that did not show correlations higher than a threshold imposed by the user. The second, creates first a Principal Component Analysis (PCA) model with the variables. Variables are then assigned to the group linked with the component in which they showed the highest loading. So that, we are grouping variables that have been considered to be co-actuacting in PCA model.

In this methodology also it would be necessary to scale variables before introducing them into the ISGL penalization, in order to do so variables where scaled using by Equations 6.

3.2.4 Partial Least Squares Regression in MORE

Due to the complexity of the multi-omics data and the intrinsic multicollinearity of the data, even if we were able to create a methodology that can handle this type of data, we wanted to try another approach that can benefit from the multicollinearity of the data. In that sense PLS models seems to be very appropriate.

PLS multivariate regression models are part of dimensionality reduction models. PLS models try to explain the variability in $\mathbf{X} \in \mathcal{M}_{NxK}$ and in $\mathbf{Y} \in \mathcal{M}_{NxG}$ and the relation between them, in bref, what it tries is to maximize the covariance between $\mathbf{X}$ and $\mathbf{Y}$. PLS models, apart from, having benefit from multicollinealy in data, they can handle high-dimensional scenarios, another problem present in our data. Using linear combinations of original variables in $\mathbf{X}$, these models construct new variables called latent variables or components that synthesize the information and reduce the dimensionality.

To construct PLS models, considering M components, we create a matrix of projections of the observations on the components (the so-called Scores matrix), $\mathbf{T} \in \mathcal{M}_{NxM}$ by another matrix $\mathbf{W} \in \mathcal{M}_{KxM}$ (the so-called Weighting matrix), so that:

$$\mathbf{T} = \mathbf{X}\mathbf{W} \tag{10}$$

Similarly is done for $\mathbf{Y}$ matrix by $\mathbf{U} = \mathbf{C}\mathbf{Y}$. Making use of the mathematic terminology, the objective of the PLS models would be easily translated into:

$$\max_{\forall i} \ \mathrm{Cov}(\mathbf{t}_i, \mathbf{u}_i) = \max_{\forall i} \ \sigma_{\mathbf{t}_i} \sigma_{\mathbf{u}_i} \, \mathrm{Corr}(\mathbf{t}_i, \mathbf{u}_i) \tag{11}$$

where i denotes the number of the component, Cov the covariance and Corr the correlation.

As we already commented, PLS models are nothing more than regression models, so it is clear that they can be written as:

$$\mathbf{Y} = \hat{\boldsymbol{\beta}}_{PLS}\mathbf{X} + \mathbf{F} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{C}^T\mathbf{X} + \mathbf{F} \tag{12}$$

where $\mathbf{F}$ denotes error matrix.

However, due to their construction, PLS models only can handle continuous variables. Therefore, we convert every categorical variable into a dummy variable for their incorporation into the model. Simultaneously, it is crucial to scale the variables using one of the three scaling approaches (6, 7, or 8). This ensures that variables are not unduly penalized due to their intrinsic range of values.

Additionally, it is important to note that PLS models are primarily designed for dimensionality reduction rather than variable selection. In the context of our approach, *MORE* focuses on detecting variables relevant to gene expression variations. We achieve this by combining the calculation of variable significance with their importance in constructing the components, as measured by the VIP score (Variable Importance in Projection).

For the calculation of variable significance, we formulate the following hypothesis test for each of the K variables in the model:

$$\begin{cases} H_0 : \hat{\beta}_k = 0 \\ H_1 : \hat{\beta}_k \neq 0 \end{cases} \tag{13}$$

To address this hypothesis test, *MORE* introduces two approaches to create the distribution of the coefficients under the null hypothesis.

The first approach uses Jack-Knife resampling methodology, as expressed by Equation 14.

$$\hat{\beta}_k = \frac{b_k}{(\frac{N-1}{N}\sum_{i=1}^{N}(b_{k(i)} - b_k)^2)^{1/2}} \tag{14}$$

where b_k denotes the estimation of the coefficient of k-th variable and $b_{k(i)}$ the estimation of the coefficient of the k-th variable without considering i-th observation.

The second approach involves calculating the estimation of b_k values by computing their p-values relative to the distribution of coefficients under the null hypothesis. This distribution is obtained through the permutation of the Y response variable R times.

In our methodology, we integrate the significance of the variables with a threshold applied to VIP scores, as described in Equation 15, for variable selection.

$$VIP_k^2 = K \frac{\sum_{m=1}^{M} \mathbf{w}_{mk}^2 SSY_m}{SSYT} \tag{15}$$

where SSY_m denotes the explained $\mathbf{Y}$ variance by the m-th component, $SSYT$ the total explained variance of $\mathbf{Y}$, K the total number of variables in $\mathbf{X}$ and M the number of components of the PLS model.

The mathematical explanation provided above is applicable in cases where biological prior knowledge is not available, resulting in the creation of a unique model for $(\mathbf{X}, \mathbf{Y})$. In *MORE*, this option is denoted as PLS2. However, when biological prior knowledge is provided, and G different models need to be created for partitions $(X^{(1)}, \mathbf{y}_1), \ldots, (X^{(G)}, \mathbf{y}_G)$, equations 10, 11, and 12 can be applied in the same way by restricting the $\mathbf{X}$ matrix as needed. This scenario is denoted as PLS1.

3.2.5 MORE regulatory networks

As introduced in this paper, *MORE* is designed to model gene expression using GLM or PLS models to construct multi-omic regulatory networks for each biological system under study. So far, however, our discussion has focused exclusively on the modelling aspect, neglecting the elucidation of regulatory network construction. In this section, we will explain the tools that *MORE* relies on to construct these networks and the complementary functions it incorporates to improve their biological interpretation.

So far, in this paper, we have emphasized the importance of calculating the significance of the coefficients and their impact on gene expression, both in GLM and PLS models. The emphasis on these coefficients is deliberate, as we will use the extracted regression coefficients to construct the regulatory networks.

In *MORE*, we use interaction terms to construct separate regulatory networks for each biological condition under investigation. If the user does not use them, the constructed regulatory network will not consider the possible differences in the regulatory networks of the different conditions.

Depending on the significance of the coefficients of the regulators and the interactions with the conditions, these are the three scenarios that could take place and their biological meaning:

- The coefficient of the regulator is significant, but the coefficient of the interaction term is not. Which biologically means that the regulator has the same effect on the gene expression under all conditions under study.

- The coefficient of the regulator is not significant, but the coefficient of some of the interaction terms is. Biologically, this means that the regulator has no effect on the gene expression, but it does under some specific experimental conditions.

- The coefficient of the regulator is significant, and so it is the coefficient of the interaction term. Biologically, this means that the regulator affects the gene expression, but the magnitude of the effect is not the same for all the groups.

MORE connects in an undirected graph the regulators to the genes for which their regression coefficients are different from zero. Those with a negative coefficient are considered to be repressors of gene expression and those with a positive coefficient activators.

Visualization of regulatory networks on different experimental conditions looking for the difference between them could be quite hard when considering large amount of regulators of the gene expression. In order to simplify it and reduce all the information in a unique network, in *MORE* we introduce the differential regulatory networks. Taking one of the experimental groups as reference, to compute this differential regulatory networks, we simply subtract the coefficients of the groups under study to the reference one (as introduced previously in [Class et al., 2017]).

3.2.6 Functional enrichment of relevant regulations

Even if *MORE* provide differential regulatory networks to facilitate the comparison of regulatory mechanisms between experimental groups or phenotypes, depending on the complexity of such networks, it can be hard to interpret them. For that reason, *MORE* provides the way to connect relevant regulatory relationships to biological pathways through different functional enrichment analysis described next.

Functional enrichment of genes regulated by a global regulator. Over-representation analysis (ORA) is a statistical methodology employed to discern the prevalence of specific biological themes within a given set of genes. Essentially, it aims to determine whether these themes are present at a frequency higher than expected. The analysis compares a target set of genes to a reference set of genes. The objective is to assess whether particular biological annotations exhibit a significant enrichment in the target set relative to the background represented by the reference set. In order to carry out this analysis, Fisher's exact test is used. In *MORE*, the ORA is applied to those genes pointed out as global regulators during the analysis.

Functional enrichment analysis of hub genes. Originally, gene set enrichment analysis (GSEA) evaluates whether a set of genes shows statistically significant differences in expression between two experimental conditions. The

analysis ranks all genes based on their expression changes and then determines whether genes within a predefined set are enriched at the top or bottom of the ranked list. However, instead of using gene expression to perform this analysis in *MORE*, as the gene expression has already been analyzed by the method, we rank all genes based on the number of regulators they are regulated by and perform the analysis in that way. Even if it is included in *MORE*, this analysis is performed using the R package clusterProfiler [Wu et al., 2021].

3.2.7 Comparing MORE to other methods

In the revision of the literature we did, we saw how there is a big gap in models creating multi-omics networks that consider all omics at the same time or that are not tailored to specific omics. In this section, we introduce a little summary of the methods that we found comparable to *MORE* and therefore were selected to assess their performance.

KiMONo. The method was designed to integrate and analyze multi-omics data, enabling the user to better understand the biological processes and diseases under study. The method uses statistical models and prior biological knowledge to infer multi-level networks. Specifically, KiMONo creates a unique model for each gene under investigation and applies a SGL penalization to create these models. It can also be used with low sample sizes.

mixOmics. Currently, mixOmics is one of the most used available tools for multi-omics data analysis and graphical representation. The tool incorporates several different approaches for the user to analyse their data. Due to the natural configuration of mixOmics the more similar approach that would be comparable to *MORE* would be the sparse block PLS [Lê Cao et al., 2008]. As the name introduces, the sparse block PLS methodology, makes a variable selection approach using PLS models but considering that variables are grouped into blocks.

3.2.8 Performance metrics

To evaluate the performance of *MORE* for its various settings and compare it to other available softwares, we used several metrics.

In one hand, given that *MOSim* simulator provides the user with the information about the real regulators of the simulated genes, we were able to compute the F1-score (equation 16) for the different runs.

$$F1\text{–}score = 2\frac{precision \cdot recall}{precision + recall}, \quad precision = \frac{TP}{TP+FP} \quad recall = \frac{TP}{TP+FN} \tag{16}$$

where $TP = \sum_{g=1}^{G} card(RR_g \cap SR_g)$, $FP = \sum_{g=1}^{G} card(\overline{RR_g} \cap SR_g)$ and $FN = \sum_{g=1}^{G} card(RR_g \cap \overline{SR_g})$ considering that RR_g represents the set of the real regulators of gene g (what is known thanks to *MOSim*) and that SR_g represents the set of regulators identified by the software under study for gene g.

On the other hand, in order to evaluate the goodness of fit of the different tools, we computed the coefficient of determination, R^2, to measure the explained variance by the model for each of the genes. To do so, we compute $R_g^2 = 1 - \frac{RSS_g}{TSS_g} = 1 - \frac{\sum_{i=1}^{N}(y_{g_i} - \hat{y}_{g_i})}{\sum_{i=1}^{N}(y_{g_i} - \overline{y}_g)^2}$ for each genes.

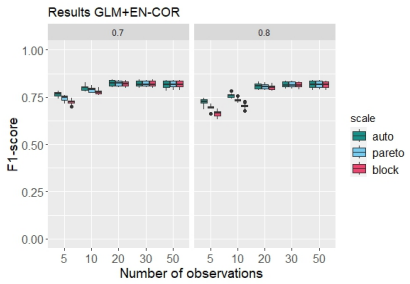
4 Results

4.1 MORE performance across settings

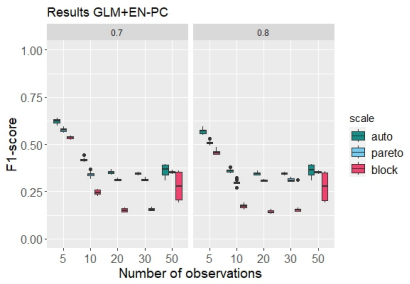
Before comparing *MORE* results to the ones obtained by other software, it was essential to select the best approaches that *MORE* could give us. In that sense, we tested the various variable selection approaches we included in *MORE* with their various settings. To test *MORE*'s performance, as introduced before, apart from considering the F1-score for comparing the different approaches, it would also be interesting to compare the computational time of the different approaches. As different settings were tried for the different variable selection methodologies and the joint visualization could be hard, we first introduce the results for each variable selection methodology and take the settings that showed the best results for each of the methodologies for their joint visualization.

Analyzing results from GLM models with ElasticNet penalization, figures 2a and 2b, at first sight, we saw how considering the partial correlation in comparison to the correlation itself worsened the results. Having a deeper look at the results of the correlation multicollinearity filter 2a, it could be seen how, at a low number of samples, the auto-scaling performed better than the rest of the scaling types. In addition, the threshold to consider variables to be collinear also affected the results at low sample sizes, showing better results with lower thresholds; however, with a higher number of observations, the differences were not so obvious. The same tendency could be seen in the partial correlation approach2b. The huge difference in the performance of partial correlation compared to the correlation itself could be due to the use of full-order partial correlation. As some authors point out [Zuo et al., 2014], considering full-order partial correlation can lead to spurious relations, and low-order partial correlation should be considered in which only the effect of one or two other variables should be considered for the correlation calculation. However, we tried to implement it, but the computational time was ridiculously increased.

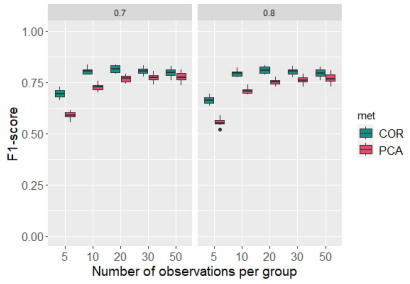
In figure 2c, we introduce the results for GLM models with ISGL penalization. In the image, we clearly saw how the grouping by correlation presented better results than those obtained by grouping of PCA loadings. For the grouping by the correlation, we saw once again how the lowest threshold for considering variables correlated presented better results. In the case of the grouping by PCA, the results did not differ too much. One of the future considerations would be to lower the percentage of variability explained by the model as it



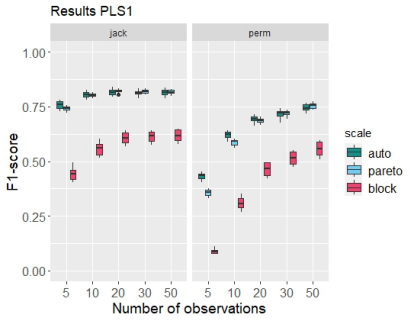
(a) F1-score of GLM models with ElasticNet penalization with correlation multicollinearity filter for different settings



(b) F1-score of GLM models with ElasticNet penalization with partial correlation multicollinearity filter for different settings



(c) F1-score of GLM models with ISGL for different settings

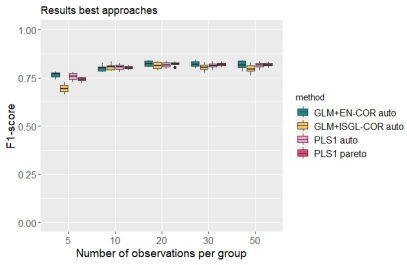


(d) F1-score of PLS1 models for different settings

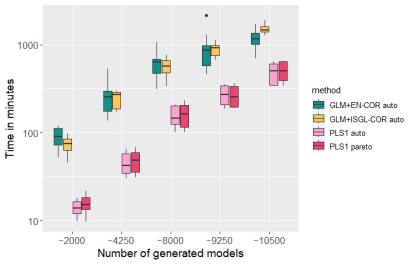
Figure 2: F1-scores of all methodologies of MORE for different settings

could be a little bit high for multi-omics data to think about a model explaining 80% of the variability.

Finally, in figure 2d, when it comes to p-value significance estimation, we saw how the Jack-Knife methodology showed better results than the ones obtained by permutation. However, we think that these results are worse just because of the number of permutations we asked for (100 permutations). We think that this number of permutations is a little bit low to obtain good, even comparable results. However, we asked for that number of permutations in order to create models that computationally had comparable results to the ones created by the Jack-Knife methodology. Moreover, we saw how results become worse when variables are too penalized by the number of variables in the omics to which they belong. However, penalizing them less showed comparable results in a low number of observations and even better results when the number of observations was elevated.



(a) F1-score of the approaches with the best settings



(b) Time consumption in minutes of the approaches

Figure 3: Comparison of MORE for the models with the best-performing settings

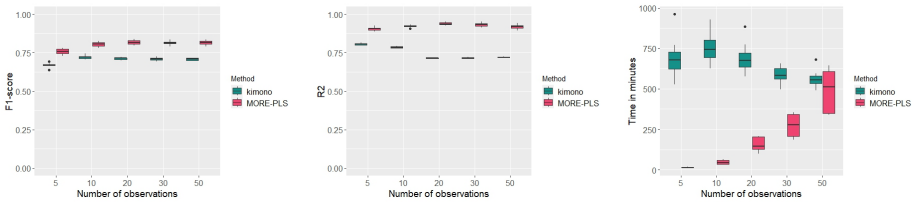
In image 3a, we compare the results of the approaches that best results showed in each of the settings. So, we introduced the GLM model with ElasticNet penalization with auto-scaling for correlation multicollinearity filter, the GLM model with ISGL penalization with auto-scaling using the correlation for the grouping, and finally, the PLS1 models both for auto-scaling and Pareto scaling as they both presented promising results.

Comparing all these methodologies, we saw varying results depending on the number of observations considered. For all number of observations, PLS methods resulted in one of the methodologies with the best results. However, GLM models with ISGL penalization, even if they did not present the best results for five observations, presented competitive results for a higher number of observations. Even if the results are comparable, we recommend the PLS method as they showed more stable results through different numbers of observations. Moreover, looking Figure 3b showed how PLS models are faster than GLM models.

4.2 Benchmarking MORE to existing tools

To evaluate the performance of *MORE*, we conducted a comparative analysis against KiMONo. This comparison aims to provide valuable insights into our tool's strengths and distinctive features in relation to existing methodologies. Before discussing the results, we comment that the KiMONo package, which is available in GitHub, is already deprecated due to the obsolescence of two of the packages it relies on. However, we managed to install them successfully to do the benchmarking.

In Figure 4, we compare the results of the different tools. Notably, the F1-score and R^2 metric showed an enhancement with our approach, which suggests a kind of improvement capturing significant biological regulations. We



(a) F1-score (b) R^2 (c) Time in minutes

Figure 4: Comparison of KiMONo and MORE

also point out the tool's high performance, showing F1-scores around 0.75 and being able to explain even more than 80% of the variability of the data.

Furthermore, in Figures 4a and 4b, results remained remarkably stable across the number of observations, showing competitive results even with a low number of observations. In terms of computing time, we saw how *MORE* outperforms KiMONo remarkably for a low number of samples; however, as the number of samples increases, the time consumption becomes similar. In case the number of observations continues to increase, it could happen that *MORE*'s time consumption would be significantly higher than KiMONo's. However, the results of KiMONo here presented were obtained using parallelization in 25 cores, while the results of *MORE* were not parallelized. For future work, we plan to parallelize our methodology so our results can obtain a more competitive performance time.

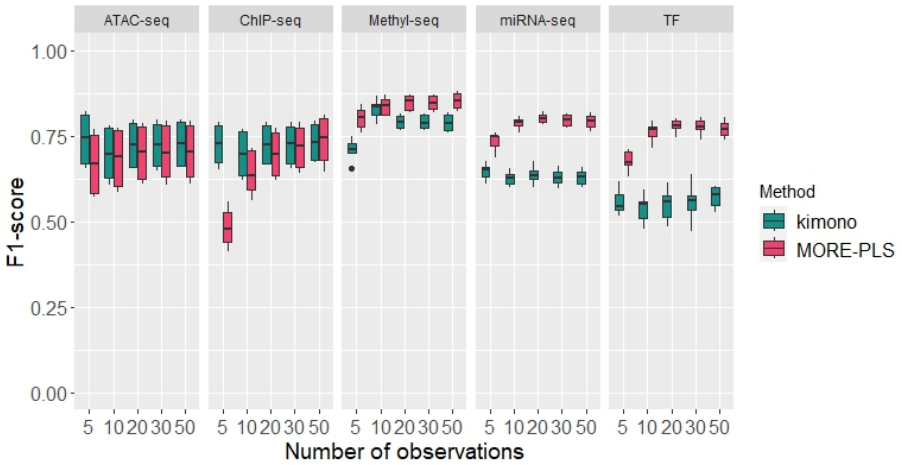


Figure 5: Comparison of KiMONo and MORE through F1-score between omics

Finally, in Figure 5, we present the results of the F1-score disaggregated by

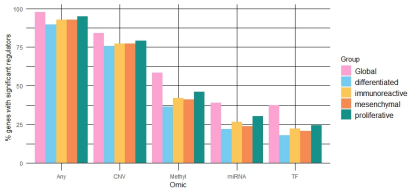
omics. There, we saw how the results for ATAC-seq and ChIP-seq were mostly equal in both methodologies, showing KiMONo even better results for a low number of observations in this second one. However, the performance of *MORE* outperformed the other in the rest of the omics, showing great differences when detecting miRNAs and TF.

Readers would have already noticed that although in section 3.2.7 we point out mixOmics as one of the most used existing tools when analyzing multi-omics data, we did not present results of that tool in the above-commented figures. In our discussion of mixOmics, we highlighted the sparse block PLS as the most comparable methodology to *MORE* among the ones that mixOmics has implemented. However, the direct quantitative comparison between *MORE* and mixOmics faced technical challenges. For that reason, we present a qualitative comparison to this tool. One of the main challenges when creating regulatory networks lies in developing an automated tool capable of correctly identifying the significant regulators of the gene expression. Our approach only requires the user to introduce the gene expression, the information about the omics, and, if available, the prior biological information. From these data, we have been able to see how *MORE* has been able to identify successfully the regulations affecting the gene expression. In mixOmics, for methodologies other than the sparse block PLS, they have implemented several functions to help users optimize their methods. When constructing PLS models, it is important to appropriately select the number of PLS components to extract. In order to do so, mixOmics for the rest of their methodologies optimize the number of components based on the goodness of the prediction. However, the results are presented visually, which makes the automation challenging, particularly when dealing with the number of datasets we are considering for benchmarking. Moreover, mixOmics offers a variable selection mechanism based on a user-provided grid. This approach also introduces a computational challenge when applied to our extensive simulated datasets, where the magnitude of the potential regulators of the gene expression is over a million. In summary, while mixOmics offers valuable functionalities for model optimization, the complexities introduced by the scale of our simulated datasets make this approach almost incomparable, and the absence of mixOmics in our quantitative comparisons is attributed to technical challenges.

4.3 MORE insights for HGSOC-TCGA Data

In this section, we delve into the application of *MORE* on the HGSOC-TCGA dataset to showcase the utility of our approach, highlighting its potential to unravel important regulatory relationships and contribute to a deeper understanding of a complex disease as HGSOC thanks to the variety of functionalities MORE offers. According to the results shown in section 4.1, we applied *MORE* PLS1 model option on the HGSOC-TCGA dataset. Regulatory omic features were auto-scaled and jack-knife resampling strategy was applied to compute the significance of PLS regression coefficients, which was combined with a VIP of 0.8, to select significant regulators.

Out of the 11700 genes considered, 696 had no potential regulators according to the prior biological knowledge available. The other 11004 genes presented a mean of 3.78 significant regulators. Figure 6a shows the percentage of genes with significant regulators decomposed by omic type and the cancer subtype in which they showed significant regulations. Patients from the *differentiated* subtype showed fewer regulations in any of the omics than patients from other gene expression subtypes. On the other hand, Figure 6b displays the percentage of significant regulations with regard to the initial potential regulations. It is worth noting that CNV regulations are predominant, highlighting the importance of considering CNV regulation as one of the most important drivers in cancer. Interestingly, while other genomic alterations such as gene mutations have been widely studied and associated with cancer, fewer studies exist in the literature devoted to CNVs. This figure shows again how the *differentiated* subtype presents fewer regulations than the other subtypes.



(a) Percentage of genes with significant regulations globally and at group level for any or specific omics

(b) Percentage of significant regulations at group level for specific omics

Figure 6: Summary plots of MORE results

Hub genes in regulatory networks are defined as genes with a high number of significant regulators and, therefore, key genes for the phenotype considered. Figure 7 compares the hub genes found by MORE for each subtype. Not surprisingly, one of the hub genes shared by all subtypes is BRCA1, which is known to play a crucial role in the DNA damage repair pathway and is closely involved in the development of ovarian cancer [Tsibulak et al., 2018; Kuchenbaecker et al., 2017; Antoniou et al., 2003; Chen and Parmigiani, 2007]. Other interesting hub genes are those associated with a specific subtype, since they can help to characterize the regulation behind that subtype. For example, BRIP1 is a hub gene only in the *differentiated* subtype and is also involved in DNA damage repair, particularly in ovarian cancer predisposition [Lilyquist et al., 2017; Suszynska et al., 2019, 2020]. BRIP1 binds to BRCA1 for homologous DNA double-strand repair and high BRIP1 expression is related to a higher risk of recurrence and shorter disease-free survival in ovarian cancer patients treated with platinum [Moes-Sosnowska et al., 2019], and with an increased cell proliferation [Tu et al., 2011], potentially contributing to earlier tumor recurrence [Liu et al., 2023]. In the *immunoreactive* subtype, three unique hub genes were

identified: CDK12, RNF24, and RAD51. CDK12, part of the cyclin-dependent kinase family, regulates gene expression by phosphorylating RNA polymerase II, impacting DNA repair and genomic stability [Blazek et al., 2011]. Inactivating CDK12 mutations in various cancers result in genomic instability, tandem focal duplications, and high fusion-induced neoantigens. This unique pattern promotes lymphocytic infiltration, indicating a better prognosis and potential benefits of immunotherapy, as seen in a trial with metastatic castration-resistant prostate cancer [Alva et al., 2022]. RNF24, encoding an integral membrane protein with a RING-type zinc finger, correlates with immune cell infiltration in various cancers, including OC [Yu et al., 2021]. RAD51, involved in homologous recombination and DNA repair, is crucial for platinum chemotherapy response and survival in ovarian cancer [Compadre et al., 2023], and defects in RAD51 trigger a STING-mediated innate immune response, emphasizing its role in genomic stability [Bhattacharya et al., 2017]. In the *mesenchymal* subtype, the transmembrane signaling enzyme PLCG2 is identified by MORE as a hub gene. This enzyme is a crucial signaling molecule in both the innate and adaptive immune systems, especially in macrophages. Therefore, a deficiency in PLCG2 expression could be related to a reduction in proinflammatory and antitumor action induced by IgE, with a strong impact on therapy response [Buttarelli et al., 2022]. Lastly, in the case of the *proliferative* subtype, the IRF2BP2 hub gene is a transcription co-repressor which its encoded protein has been identified as a nuclear protein interacting with IRF-2. While initially described as a co-repressor, IRF2BP2 exhibits diverse functions, particularly in cancer. Studies demonstrate its involvement in p53-mediated stress response, promoting an antiapoptotic role in breast cancer and contributing to transcriptional repression of pro-apoptotic factors [Koeppel et al., 2009], characteristic attributes of the *proliferative* subtype. Another hub gene for both *mesenchymal* and *proliferative* subtypes, the two subtypes with worse prognostic, is ZFHX4, which has been previously described as a prognostic biomarker for HGSOC [Millstein et al., 2020; Zong et al., 2022].

For a better understanding of the relevance of hub genes identified by MORE in HGSOC, we performed a Gene Set Enrichment Analysis (GSEA) with MORE on the whole list of input DEGs. The gene score for GSEA was the difference in the number of MORE significant regulators between the two cancer subtypes with markedly different overall survival prognoses: the *immunoreactive* subtype, having with the best overall survival outcome, and the *mesenchymal* subtype, with the worst overall survival [Network et al., 2011; Konecny et al., 2014]. Functional annotation was obtained from the Gene Ontology (GO) database. The functional categories that showed significant suppression were "chromatin binding" and "chromatin organization". These functions are crucial to cancer development through the regulation of gene expression and maintenance of genomic stability and are associated with resistance to chemotherapy in OC [Vaicekauskaitė et al., 2022]. The observed downregulation suggests a potential role in influencing the overall survival disparities between the *immunoreactive* and *mesenchymal* subtypes. Conversely, "angiogenesis", "positive regulation of cell migration", and "collagen binding" appear to be significantly activated.

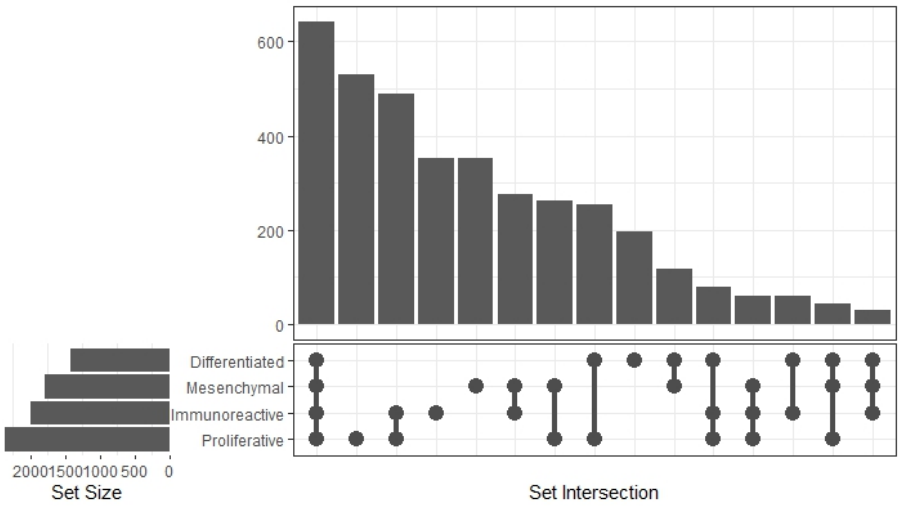


Figure 7: Number of hub genes shared by the different cancer subtypes

These functions are characteristic of processes involved in tumor growth, invasion, and metastasis. The upregulation of these pathways aligns with the reported worse overall survival in the *mesenchymal* subtype, indicating a potential association with the aggressive behavior of this subtype in OC.

Additionally, we defined a *global regulator* as an omic feature significantly regulating 10 or more genes, thus playing an important role in system regulation. We focused on the set of global regulators specific to each of the four cancer subtypes. A total of 163 global regulators were identified for the *differentiated* subtype, 199 for *immunoreactive*, 191 for *mesenchymal*, and 215 for *proliferative*. The miRNAs were, by far, the most common type of global regulators in all subtypes, with more than 98% representation in all of them. In contrast, only 3 TFs were identified as global regulators in the four subtypes: NKX2-3, NKX2-5 and NKX3-1. Nonetheless, the number of genes regulated by these 3 TFs varies considerably, ranging from 60-93 for NKX2-3, increasing to 167-318 for NKX2-5, and peaking at 1830-2465 for NKX3-1. There were also some differences between cancer subtypes, as greater numbers were found in the *immunoreactive* and *mesenchymal* subtypes. While the number of significant regulations obviously depends on the number of input initial associations given to MORE, these differences point out the ability of the model to distinguish different regulatory mechanisms across phenotypes since the initial associations are the same for all subtypes. When focusing on global regulators found only in certain cancer subtypes, which could enhance the characterization of their molecular profiles, MORE solely identified them for the *immunoreactive* and *proliferative* subtypes. It is worth mentioning the case of hsa-miR-940, a global regulator in the *immunoreactive* subtype, which has been proven to suppress

cell proliferation and induce apoptosis in ovarian cancer OVCAR3 cells [Wang et al., 2017]. Genes significantly regulated by hsa-miR-940 in the *immunoreactive* subtype revealed associations with genes involved in the immune response in cancer, including the Immunity Related GTPase Q (IRGQ) gene and the Polymeric Immunoglobulin Receptor (PIGR) gene, as well as genes involved in transcriptional regulation (ZSCAN29, ZNF490, and ZC3H4) or protection from induced-DNA damage by base-excision repair (FAM168A). These findings align with the results from the functional enrichment analysis conducted on the set of genes regulated by hsa-miR-940, indicating a substantial presence of enriched functional categories related to the immune response, transcriptional regulation, and DNA damage repair. In the case of the *proliferative* subtype, noteworthy miRNAs include hsa-miR-328-3p and hsa-miR-373-5p. Firstly, hsa-miR-328-3p has been identified as a tumor suppressor in previous research, with elevated expression commonly observed in ovarian cancer, and its inhibition in mouse orthotopic ovarian xenografts has demonstrated a hindrance to tumor growth and the prevention of metastasis [Srivastava et al., 2019]. The enrichment analysis of genes regulated by this miRNA revealed GO terms as "negative regulation of cell population proliferation" and "positive regulation of focal adhesion assembly". On the other hand, hsa-miR-373-5p has been described as a stimulator for cancer cell migration and invasion both in vitro and in vivo [Huang et al., 2008]. Some of the genes significantly regulated by this miRNA play key roles in processes related to tumor proliferation, such as cell cycle regulation, including CLSPN [Chini and Chen, 2003] and TFDP2 [Zheng et al., 1999], as well as cell differentiation and proliferation, involving ZBT7A [Xiao et al., 2019]. Conversely, when exploring the *mesenchymal* and *differentiated* subtypes, which did not exhibit unique global regulators, we found that, firstly, three members of the miR-200 family were detected as global regulators in the *mesenchymal* subtype: hsa-miR-200a-3p, hsa-miR-200b-3p, and hsa-200c-3p, regulating a higher number of genes in this subtype compared to other subtypes. The miR-200 family is well-known for its implications in the diagnosis and prognosis of OC [Pal et al., 2015] and has been recognized as a master suppressor of epithelial-mesenchymal transition (EMT) [Park et al., 2008]. The functional enrichment analysis of target genes of hsa-miR-200a-3p in this subtype revealed categories related to cell migration and regulation of mesenchymal cell proliferation, which suggests that the expression of this regulator plays a key role in the molecular environment of this cancer subtype. Finally, when delving into the global regulators of the *differentiated* subtype, we observed a parallel pattern with hsa-miR-135b-5p, mirroring the dynamics observed in the miR-200 family within the *mesenchymal* subtype. While this miRNA emerges as a global regulator across all four subtypes, its impact is more pronounced in the *differentiated* subtype, significantly regulating a greater number of genes involved in functions such as "positive regulation of endothelial cell proliferation".

Given the large number of significant regulatory connections found by MORE in this HGSOC dataset (see multi-layer regulatory networks for each cancer subtype in the Supplementary materials, Figures 8, 9,10,11), the complementary

analysis of hub genes and global regulators provided in the MORE package turns out to be extremely useful to interpret the biological intricacies inherent to this type of studies. By exploring these regulatory networks from different aspects, we gained a deeper understanding of the underlying molecular processes in HGSOC.

5 Conclusions

We have presented *MORE* a multi-omics regulation network inference method. The presented method constructs the networks using PLS or GLM models with an ElasticNet or ISGL penalization. The method is freely available as R code language at https://github.com/ConesaLab/MORE.git where a vignette can be found to its correct application.

Our tool not only allows the incorporation of prior biological information into network construction but can also infer relationships de novo in the absence of such information. The results demonstrate a highly competitive performance, even with low sample sizes. Moreover, we effectively addressed multicollinearity issues inherent in such data types, ensuring precise and reliable inference of regulatory networks when performing GLM models. We do not address it in PLS models as they take advantage of it.

In our comparison to KiMONo, our tool exhibited superior evaluation metrics, including F1-score, R2, and computational efficiency.

Finally, applying our tool to a real ovarian cancer dataset yielded intriguing and biologically meaningful results. This successful application underscores our methodology's practical utility and biological relevance in dissecting complex regulatory networks in a disease context.

In summary, our developed methodology represents a versatile and powerful multi-omic regulatory network inference approach, demonstrating good performance and applicability to real-world biological datasets.

6 Availability of data and software

The TCGA data is publicly available via TCGA data portal (downloaded November, 2023).

The implementation of *MORE* was done using R code language and it is freely available at https://github.com/ConesaLab/MORE.

The results were run in a cluster with the following specifications: x86_64 architecture processor, Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz 40c and

1500TB of RAM.

Bibliography

Alva, A. S., Li, J., Chou, J., Reimers, M. A., McKay, R. R., Zhang, J., Daignault-Newton, S., Palmbos, P. L., Reichert, Z. R., Cieslik, M., et al. (2022). Phase 2 trial of immunotherapy in tumors with cdk12 inactivation (impact): Results from cohort a of patients (pts) with metastatic castration resistant prostate cancer (mcrpc) receiving dual immune checkpoint inhibition (ici).

Alvarez, A. and Woolf, P. J. (2011). Regnetb: Predicting relevant regulator-gene relationships in localized prostate tumor samples. *BMC Bioinformatics*, 12(1):243.

Antoniou, A., Pharoah, P. D., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., Loman, N., Olsson, H., Johannsson, O., Borg, Å., et al. (2003). Average risks of breast and ovarian cancer associated with brca1 or brca2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *The American Journal of Human Genetics*, 72(5):1117–1130.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*, 14(6):e8124.

Bhattacharya, S., Srinivasan, K., Abdisalaam, S., Su, F., Raj, P., Dozmorov, I., Mishra, R., Wakeland, E. K., Ghose, S., Mukherjee, S., et al. (2017). Rad51 interconnects between dna replication, dna repair and immunity. *Nucleic acids research*, 45(8):4590–4605.

Blazek, D., Kohoutek, J., Bartholomeeusen, K., Johansen, E., Hulinkova, P., Luo, Z., Cimermancic, P., Ule, J., and Peterlin, B. M. (2011). The cyclin k/cdk12 complex maintains genomic stability via regulation of expression of dna damage response genes. *Genes & development*, 25(20):2158–2172.

Buttarelli, M., Ciucci, A., Palluzzi, F., Raspaglio, G., Marchetti, C., Perrone, E., Minucci, A., Giacò, L., Fagotti, A., Scambia, G., et al. (2022). Identification of a novel gene signature predicting response to first-line chemotherapy in brca wild-type high-grade serous ovarian cancer patients. *Journal of Experimental & Clinical Cancer Research*, 41(1):1–17.

Cancer Genome Atlas Research Network (2012). Integrated genomic analyses of ovarian carcinoma. *Nature. 2011 Jun 29;474(7353):609-15. doi:*, 10.

Chen, S. and Parmigiani, G. (2007). Meta-analysis of brca1 and brca2 penetrance. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 25(11):1329.

Chen, Y.-X., Rong, Y., Jiang, F., Chen, J.-B., Duan, Y.-Y., Dong, S.-S., Zhu, D.-L., Chen, H., Yang, T.-L., Dai, Z., and Guo, Y. (2020). An integrative multi-omics network-based approach identifies key regulators for breast cancer. *Comput Struct Biotechnol J*, 18:2826–2835.

Chini, C. C. S. and Chen, J. (2003). Human claspin is required for replication checkpoint control. *Journal of Biological Chemistry*, 278(32):30057–30062.

Class, C. A., Ha, M. J., Baladandayuthapani, V., and Do, K.-A. (2017). iDINGO—integrative differential network analysis in genomics with Shiny application. *Bioinformatics*, 34(7):1243–1245.

Compadre, A. J., van Biljon, L. N., Valentine, M. C., Llop-Guevara, A., Graham, E., Fashemi, B., Herencia-Ropero, A., Kotnik, E. N., Cooper, I., Harrington, S. P., et al. (2023). Rad51 foci as a biomarker predictive of platinum chemotherapy response in ovarian cancer. *Clinical Cancer Research*, 29(13):2466–2479.

Gambardella, G., Moretti, M. N., de Cegli, R., Cardone, L., Peron, A., and di Bernardo, D. (2013). Differential network analysis for the identification of condition-specific pathway activity and regulation. *Bioinformatics*, 29(14):1776–1785.

Gosline, S. J. C., Oh, C., and Fraenkel, E. (2014). SAMNetWeb: identifying condition-specific networks linking signaling and transcription. *Bioinformatics*, 31(7):1124–1126.

Ha, M. J., Baladandayuthapani, V., and Do, K.-A. (2015). DINGO: differential network analysis in genomics. *Bioinformatics*, 31(21):3413–3420.

Hamed, M., Spaniol, C., Nazarieh, M., and Helms, V. (2015). Tfmir: A web server for constructing and analyzing disease-specific transcription factor and mirna co-regulatory networks. *Nucleic acids research*, 43.

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

Huang, Q., Gumireddy, K., Schrier, M., Le Sage, C., Nagel, R., Nair, S., Egan, D. A., Li, A., Huang, G., Klein-Szanto, A. J., et al. (2008). The micrornas mir-373 and mir-520c promote tumour invasion and metastasis. *Nature cell biology*, 10(2):202–210.

Kim, D.-C., Kang, M., Zhang, B., Wu, X., Liu, C., and Gao, J. (2014). Integration of dna methylation, copy number variation, and gene expression for gene regulatory network inference and application to psychiatric disorders. In *2014 IEEE International Conference on Bioinformatics and Bioengineering*, pages 238–242.

Koeppel, M., van Heeringen, S. J., Smeenk, L., Navis, A. C., Janssen-Megens, E. M., and Lohrum, M. (2009). The novel p53 target gene irf2bp2 participates in cell survival during the p53 stress response. *Nucleic acids research*, 37(2):322–335.

Konecny, G. E., Wang, C., Hamidi, H., Winterhoff, B., Kalli, K. R., Dering, J., Ginther, C., Chen, H.-W., Dowdy, S., Cliby, W., et al. (2014). Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer. *Journal of the National Cancer Institute*, 106(10):dju249.

Kuchenbaecker, K. B., Hopper, J. L., Barnes, D. R., Phillips, K.-A., Mooij, T. M., Roos-Blom, M.-J., Jervis, S., Van Leeuwen, F. E., Milne, R. L., Andrieu, N., et al. (2017). Risks of breast, ovarian, and contralateral breast cancer for brca1 and brca2 mutation carriers. *Jama*, 317(23):2402–2416.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559.

Laria, J., Aguilera-Morillo, M., and Lillo, R. (2019). An iterative sparse-group lasso. *Journal of Computational and Graphical Statistics*, 28.

Li, J., Chen, F., Liang, H., and Yan, J. (2021). MoNET: an R package for multi-omic network analysis. *Bioinformatics*, 38(4):1165–1167.

Li, Y., Liang, M., and Zhang, Z. (2014). Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLOS Computational Biology*, 10(10):1–14.

Lilyquist, J., LaDuca, H., Polley, E., Davis, B. T., Shimelis, H., Hu, C., Hart, S. N., Dolinsky, J. S., Couch, F. J., and Goldgar, D. E. (2017). Frequency of mutations in a large series of clinically ascertained ovarian cancer cases tested on multi-gene panels compared to reference controls. *Gynecologic oncology*, 147(2):375–380.

Liska, O., Bohár, B., Hidas, A., Korcsmáros, T., Papp, B., Fazekas, D., and Ari, E. (2022). Tflink: an integrated gateway to access transcription factor-target gene interactions for multiple species. *Database (Oxford) 16;2022:baac083. doi:*, 10.

Liu, Y., Wu, X., Feng, Y., Jiang, Q., Zhang, S., Wang, Q., Yang, A., et al. (2023). Insights into the oncogenic, prognostic, and immunological role of brip1 in pan-cancer: A comprehensive data-mining-based study. *Journal of Oncology*, 2023.

Lê Cao, K. A., Rossouw, D., Robert-Granié, C., and Besse, P. (2008). *A sparse PLS for variable selection when integrating Omics data*. Statistical Applications in Genetics and Molecular Biology 7, article 35.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., and Califano, A. (2006). Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(1):S7.

Millstein, J., Budden, T., Goode, E. L., Anglesio, M. S., Talhouk, A., Intermaggio, M. P., Leong, H. S., Chen, S., Elatre, W., Gilks, B., Nazeran, T., Volchek, M., Bentley, R. C., Wang, C., Chiu, D. S., Kommoss, S., Leung, S. C. Y., Senz, J., Lum, A., Chow, V., Sudderuddin, H., Mackenzie, R., George, J., AOCS Group, Fereday, S., Hendley, J., Traficante, N., Steed, H., Koziak, J. M., Köbel, M., McNeish, I. A., Goranova, T., Ennis, D., Macintyre, G., Silva De Silva, D., Ramón Y Cajal, T., García-Donas, J., Hernando Polo, S., Rodriguez, G. C., Cushing-Haugen, K. L., Harris, H. R., Greene, C. S., Zelaya, R. A., Behrens, S., Fortner, R. T., Sinn, P., Herpel, E., Lester, J., Lubiński, J., Oszurek, O., Tołoczko, A., Cybulski, C., Menkiszak, J., Pearce, C. L., Pike, M. C., Tseng, C., Alsop, J., Rhenius, V., Song, H., Jimenez-Linan, M., Piskorz, A. M., Gentry-Maharaj, A., Karpinskyj, C., Widschwendter, M., Singh, N., Kennedy, C. J., Sharma, R., Harnett, P. R., Gao, B., Johnatty, S. E., Sayer, R., Boros, J., Winham, S. J., Keeney, G. L., Kaufmann, S. H., Larson, M. C., Luk, H., Hernandez, B. Y., Thompson, P. J., Wilkens, L. R., Carney, M. E., Trabert, B., Lissowska, J., Brinton, L., Sherman, M. E., Bodelon, C., Hinsley, S., Lewsley, L. A., Glasspool, R., Banerjee, S. N., Stronach, E. A., Haluska, P., Ray-Coquard, I., Mahner, S., Winterhoff, B., Slamon, D., Levine, D. A., Kelemen, L. E., Benitez, J., Chang-Claude, J., Gronwald, J., Wu, A. H., Menon, U., Goodman, M. T., Schildkraut, J. M., Wentzensen, N., Brown, R., Berchuck, A., Chenevix-Trench, G., deFazio, A., Gayther, S. A., García, M. J., Henderson, M. J., Rossing, M. A., Beeghly-Fadiel, A., Fasching, P. A., Orsulic, S., Karlan, B. Y., Konecny, G. E., Huntsman, D. G., Bowtell, D. D., Brenton, J. D., Doherty, J. A., Pharoah, P. D. P., and Ramus, S. J. (2020). Prognostic gene expression signature for high-grade serous ovarian cancer. *Ann Oncol*, 31(9):1240–1250.

Moes-Sosnowska, J., Rzepecka, I. K., Chodzynska, J., Dansonka-Mieszkowska, A., Szafron, L. M., Balabas, A., Lotocka, R., Sobiczewski, P., and Kupryjanczyk, J. (2019). Clinical importance of fancd2, brip1, brca1, brca2 and fancf expression in ovarian carcinomas. *Cancer biology & therapy*, 20(6):843–854.

Monzø, C., Martínez-Mira, C., A, A.-L., Conesa, A., and Tarazona, S. (2024). Mosim: bulk and single-cell multi-layer regulatory network simulator. *BioRxiv*, page 421834.

Network, C. G. A. R. et al. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609.

Nguyen, N. D. and Wang, D. (2020). Multiview learning for understanding functional multiomics. *PLOS Computational Biology*, 16(4):1–26.

Ogris, C., Hu, Y., Knauer-Arloth, J., and Müller, N. (2021). Versatile knowledge guided network inference method for prioritizing key regulatory factors in multi-omics data. *Scientific Reports*, 11.

Pačínková, A. and Popovici, V. (2023). IntOMICS: A bayesian framework for reconstructing regulatory networks using Multi-Omics data. *J Comput Biol*, 30(5):569–574.

Pal, M. K., Jaiswar, S. P., Dwivedi, V. N., Tripathi, A. K., Dwivedi, A., and Sankhwar, P. (2015). Microrna: a new and promising potential biomarker for diagnosis and prognosis of ovarian cancer. *Cancer biology & medicine*, 12(4):328.

Park, S.-M., Gaur, A. B., Lengyel, E., and Peter, M. E. (2008). The mir-200 family determines the epithelial phenotype of cancer cells by targeting the e-cadherin repressors zeb1 and zeb2. *Genes & development*, 22(7):894–907.

Pearl, J. R., Bergey, D. E., Funk, C. C., Basu, B., Oshone, R., Shannon, P., Hood, L., Price, N. D., Colantuoni, C., and Ament, S. A. (2017). Genome-scale transcriptional regulatory network models of psychiatric and neurodegenerative disorders. *bioRxiv*.

Randhawa, V. and Pathania, S. (2020). Advancing from protein interactomes and gene co-expression networks towards multi-omics-based composite networks: approaches for predicting and extracting biological knowledge. *Briefings in Functional Genomics*, 19(5-6):364–376.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47.

Rohart, F., Gautier, B., Singh, A., and Le Cao, K.-A. (2017). mixomics: An r package for 'omics feature selection and multiple data integration. *PLoS computational biology*, 13(11):e1005752.

Siebert, J. C., Saint-Cyr, M., Borengasser, S. J., Wagner, B. D., Lozupone, C. A., and Görg, C. (2021). Cantare: finding and visualizing network-based multi-omic predictive models. *BMC Bioinformatics*, 22(1):80.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Srivastava, A. K., Banerjee, A., Cui, T., Han, C., Cai, S., Liu, L., Wu, D., Cui, R., Li, Z., Zhang, X., et al. (2019). Inhibition of mir-328–3p impairs cancer stem cell function and prevents metastasis in ovarian cancer. *Cancer research*, 79(9):2314–2326.

Sticht, C., De La Torre, C., Parveen, A., and Gretz, N. (2018). mirwalk: an online resource for prediction of microrna binding sites. *PloS one*, 13(10):e0206239.

Suszynska, M., Klonowska, K., Jasinska, A. J., and Kozlowski, P. (2019). Large-scale meta-analysis of mutations identified in panels of breast/ovarian cancer-related genes—providing evidence of cancer predisposition genes. *Gynecologic Oncology*, 153(2):452–462.

Suszynska, M., Ratajska, M., and Kozlowski, P. (2020). Brip1, rad51c, and rad51d mutations are associated with high susceptibility to ovarian cancer: Mutation prevalence and precise risk estimates based on a pooled analysis of˜ 30,000 cases. *Journal of ovarian research*, 13(1):1–11.

Tarazona, S., Arzalluz-Luque, A., and Conesa, A. (2021). Undisclosed, unmet and neglected challenges in multi-omics studies. *Nature Computational Science*, 1(6):395–402.

Tarazona, S., Furio-Tari, P., Turra, D., Pietro, A. D., Nueda, M. J., Ferrer, A., and Conesa, A. (2015). Data quality aware analysis of differential expression in rna-seq with noiseq r/bioc package. *Nucleic Acids Research*, 43(21):e140.

Tsibulak, I., Wieser, V., Degasper, C., Shivalingaiah, G., Wenzel, S., Sprung, S., Lax, S. F., Marth, C., Fiegl, H., and Zeimet, A. G. (2018). Brca1 and brca2 mrna-expression prove to be of clinical impact in ovarian cancer. *British journal of cancer*, 119(6):683–692.

Tu, Z., Aird, K. M., Bitler, B. G., Nicodemus, J. P., Beeharry, N., Xia, B., Yen, T. J., and Zhang, R. (2011). Oncogenic ras regulates brip1 expression to induce dissociation of brca1 from chromatin, inhibit dna repair, and promote senescence. *Developmental cell*, 21(6):1077–1091.

UCSC (2020). *xenabrowser*. https://xenabrowser.net/datapages/.

Vaicekauskaitė, I., Sabaliauskaitė, R., Lazutka, J. R., and Jarmalaitė, S. (2022). The emerging role of chromatin remodeling complexes in ovarian cancer. *International Journal of Molecular Sciences*, 23(22):13670.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12):i237–45.

Wang, F., Wang, Z., Gu, X., and Cui, J. (2017). mir-940 upregulation suppresses cell proliferation and induces apoptosis by targeting pkc-in ovarian cancer ovcar3 cells. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics*, 25(1):107–114.

Werhli, A. V. and Husmeier, D. (2007). Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., and Zhan, L. (2021). Fu x. *Liu S, Bo X, Yu G (2021). clusterProfiler 4. 0: A universal enrichment tool for interpreting omics data. The Innovation*, (3):100141.

Xiao, X., Shen, Y., Yin, L., He, J., Ni, X., Luo, G., Chen, X., Zhu, W., Zhong, J., Liu, J., et al. (2019). Knockdown of zbtb7a inhibits cell proliferation of breast cancer through regulating the ubiquitination of estrogen receptor alpha. *Life sciences*, 239:117042.

Yu, B. et al. (2021). Rnf24 is a novel biomarker of prognosis and immunological cell infiltration in cancers.

Yuan, L., Guo, L.-H., Yuan, C.-A., Zhang, Y., Han, K., Nandi, A. K., Honig, B., and Huang, D.-S. (2019). Integration of multi-omics data for gene regulatory network inference and application to breast cancer. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(3):782–791.

Zheng, N., Fraenkel, E., Pabo, C. O., and Pavletich, N. P. (1999). Structural basis of dna recognition by the heterodimeric cell cycle transcription factor e2f–dp. *Genes & development*, 13(6):666–674.

Zong, S., Xu, P.-P., Xu, Y.-H., and Guo, Y. (2022). A bioinformatics analysis: ZFHX4 is associated with metastasis and poor survival in ovarian cancer. *J Ovarian Res*, 15(1):90.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320.

Zuo, Y., Yu, G., Tadesse, M. G., and Ressom, H. W. (2014). Biological network inference using low order partial correlation. *Methods*, 69(3):266–273.

7 Funding

This work was supported by the Spanish MICINN [PID2020-119537RB-I00] and by a grant from the Scientific Foundation of the Spanish Association Against Cancer [ERAPERMED2022-141].

8 Supplementary materials

In this section, we include the regulatory networks created for each gene expression subtype considered in the analysis.

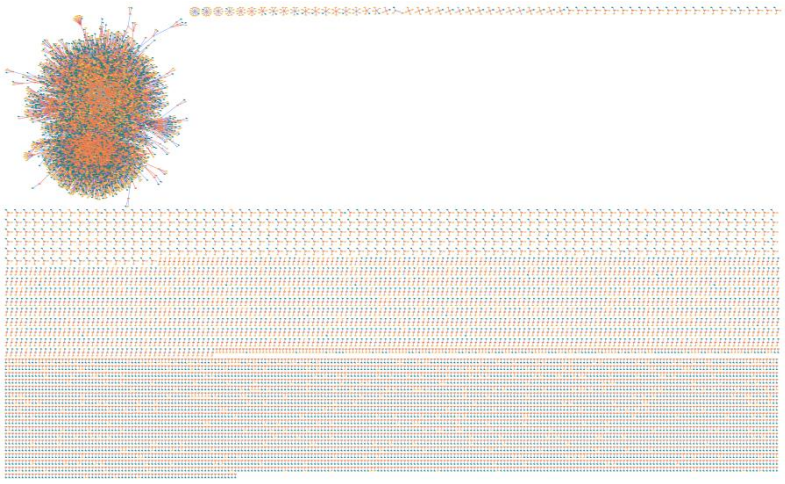


Figure 8: Network for samples of differentiated gene expression subtype

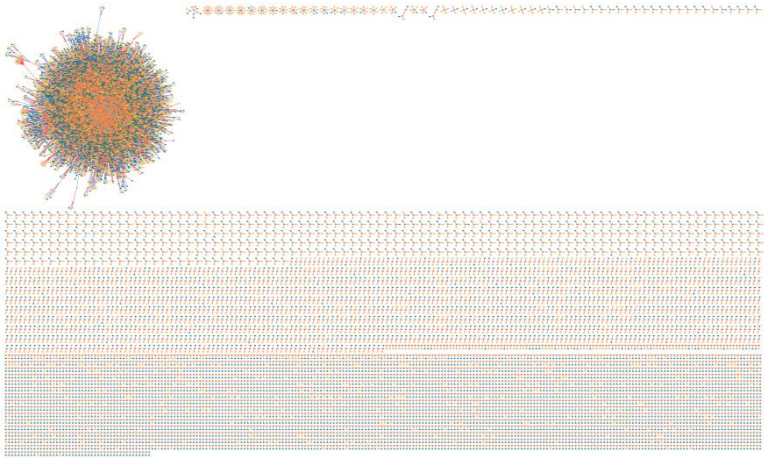


Figure 9: Network for samples of immunoreactive gene expression subtype

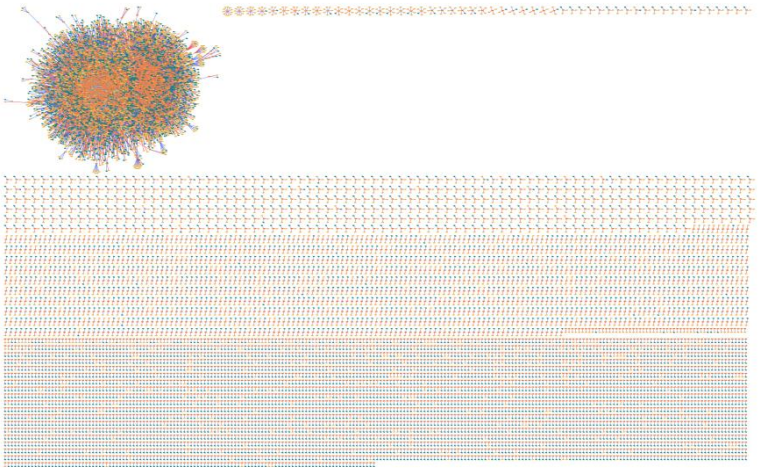


Figure 10: Network for samples of mesenchymal gene expression subtype

Figure 11: Network for samples of proliferative gene expression subtype