PAPER

MORE interpretable multi-omic regulatory networks to characterize phenotypes

Maider Aguerralde-Martin,[1] Mónica Clemente-Císcar,[2] Ana Conesa[3] and Sonia Tarazona[1,*]

[1]Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Valencia, 46022, Spain, [2]Igenomix, Ronda Narciso Monturiol, Parque tecnológico Paterna, 46980, Paterna, Spain and [3]Institute for Integrative Systems Biology, Spanish National Research Council, Catedràtic Agustín Escardino Benlloch, Paterna, 46980, Spain
*Corresponding author. sotacam@eio.upv.es

Abstract

The identification of phenotype-specific regulatory mechanisms is crucial for understanding the molecular basis of diseases and other complex traits. However, the lack of tools capable of constructing multi-omic, condition-specific regulatory networks remains a significant limitation. Here, we introduce MORE (Multi-Omics Regulation), a novel R package for the inference and comparison of multi-modal regulatory networks publicly available at https://github.com/BiostatOmics/MORE. MORE supports any number and type of omics layers, integrates prior regulatory knowledge, and employs advanced regression-based modelling and variable selection techniques to identify significant regulators of target features.

We evaluated MORE on simulated datasets and benchmarked it against state-of-the-art tools. Our tool exhibited superior accuracy in identifying key regulators, model goodness-of-fit, and computational efficiency. Additionally, we applied MORE to an ovarian cancer dataset to uncover tumour subtype-specific regulatory mechanisms associated with distinct survival outcomes.

By providing a comprehensive and user-friendly framework for constructing phenotype-specific regulatory networks, MORE addresses a critical gap in the field of multi-omics data integration. Its versatility and effectiveness make it a valuable resource for advancing our understanding of complex molecular interactions and regulatory systems.

Key words: multi-omics, regulatory networks, regression models, phenotype comparison

Introduction

Unveiling the regulatory mechanisms driving disease or other phenotypes of interest in a biological system is currently one of systems biology's most important goals and challenges. Many advances have been made in this field, from gene co-expression networks derived from transcriptomic data to multi-layer regulatory networks leveraging different omic perspectives [1]. Interestingly, single-cell technology has recently emerged as an unprecedented opportunity to study cell type-specific regulation. However, it is yet to be possible to generate regulatory models encompassing a wide variety of omic layers since no more than two omic modalities can be measured on the same cell, and, in addition, few omic combinations are commercially available [2]. Therefore, bulk multi-omics is still the only option to model complex regulatory mechanisms comprehensibly.

Although gene co-expression networks can provide valuable information about the biological system under study and many methods have been developed to this end [3, 4, 5, 6, 7, 8], this information is limited by the fact that gene expression data are only considered. Incorporating other regulatory layers into the model is crucial to better understanding the complex molecular mechanisms triggered by disease or other environmental factors. Therefore, a multi-omics approach is pivotal to unravelling intricate regulatory mechanisms and providing a more holistic and interdependent view of biological processes. The increasing availability of multi-omics experiments makes it feasible, but efficient and versatile tools are required to infer such multi-layer regulatory models.

Many efforts have been devoted to developing efficient multi-omic regulatory networks (MO-RN) inference methods. However, there is still room for improvement. Versatile MO-RN methodologies must consider the heterogeneity of the data across omic technologies, with specific noise related to each modality, and the high dimensionality of the problem when combining omic features from several omic data types. Strategies to compare MO-RN across two or more phenotypes are also required. Finally, the biological interpretation of the multi-modal regulatory models obtained is still a bottleneck in this type of analysis [9, 10] that must be addressed. Moreover,

some additional difficulties may be faced: (i) some methods can only analyse a limited number of omics (two or three at most); (ii) other methods are designed to study regulation between specific omic modalities (e.g., gene expression and methylation, gene expression and miRNA expression, etc.); (iii) some methods can computationally handle a reduced number of regulations and features; and (iv) some tools are no longer available because they could not be maintained by the research groups [11]. Next, we review some examples of these constrained yet valuable tools to highlight the existing gaps and underscore the critical need for novel integrative methodologies to infer MO-RN.

While the number of modalities measured in multi-omic experiments keeps increasing to uncover a wider diversity of molecular aspects of the biological system, many existing MO-RN tools are limited to the type or number of omic modalities, making them unsuitable for analysing such datasets. Some examples are DCGRN [12] and BMNPGRN [13], both designed for modelling the regulation of gene expression from DNA methylation and Copy Number Variation (CNV) data by employing a LASSO approach on multiple linear regression models to identify the most relevant regulators. Moreover, these two methods are not publicly available anymore. Other tools, even if extensible to more than two omics, are limited to their use on specific modalities. MoNET [14] is an R GitHub package that creates MO-RN with single nucleotide polymorphisms (SNPs), genes or proteins and metabolites from a protein-protein interaction (PPI) network. TFmiR [15] was a web server created to study regulatory interactions between transcription factors (TFs), microRNAs (miRNAs) and target genes involved in disease pathogenesis. The SAMNetWeb webserver [16] combines gene expression and proteomic data to generate condition-specific networks based on functional enrichment analysis. IntOmics [17], a deprecated Bioconductor R package, was designed to integrate gene expression, CNV, DNA methylation and biological prior knowledge on Bayesian networks. The RACER [18] pipeline is based on regression models to predict gene expression from CNV, DNA methylation, transcription factors and microRNAs, and it is not provided as a standalone package. The iDINGO R/CRAN package [19] generates MO-RN based on chained directed and undirected graph models and renders differential networks. However, due to potential computational limitations, the developers do not recommend using it with more than a few hundred features, which requires a very astringent preliminary feature filtering on omic datasets. In addition, up to three omics are admitted, although any omic modality is accepted.

Although not explicitly designed for generating MO-RN, multi-omic methods based on dimension reduction strategies have become very popular in finding common patterns between omic modalities and features of each modality. MOFA [20] and mixOmics [21] are two examples of such tools. In contrast to the methodologies described above, these tools usually admit any number or type of omic modalities and are helpful for visualising relationships among omics, features, or samples. However, they lack the functionality of providing specific feature-based MO-RN for each phenotype or condition under study, and neither can incorporate prior biological knowledge into the models. For instance, the MINT option in the mixOmics R/Bioconductor package, one of the most popular multi-omics approaches currently used for multi-omics data integration, applies either single or multi-block Partial Least Squares (PLS) regression models with a Lasso-based approach for variable selection and provides correlation networks for the selected omic features. Nevertheless, these networks are not condition-specific. Moreover, prior regulatory information cannot be given to the model, which may result in discovering novel regulations that could be biologically unmeaningful.

Interestingly, we found two methods that provide condition-specific MO-RNs and are flexible regarding the number or types of omics. CANTARE [22] is an R pipeline that fits pairwise regression models between each pair of omics to build the network and then applies logistic regression on the highly connected subnetworks to compare conditions. CANTARE does not admit prior knowledge about potential regulations. Therefore, although de novo regulations could be found with this method, false regulatory relationships are likely to be inferred. The other tool is KiMONo [23, 24], which builds MO-RNs from prior regulatory knowledge. KiMONo applies a sparse group LASSO penalisation on a multivariate regression model fitted separately for each gene to obtain the key regulators of gene expression from multi-omics measurements. The fitted models are aggregated in the final heterogeneous multi-omics network.

Given the limitations or unavailability of most of the published MO-RN tools, we introduce the MORE *R* package for MO-RN inference. MORE (Multi-Omics REgulation) is, to the best of our knowledge, one of the first software tools capable of generating condition-specific regulatory networks for any number or types of omic data, admits prior biological knowledge about regulations, and provides functionalities for comparing the resulting networks and guiding the biological interpretation of regulatory relationships. MORE relies on a regression framework and advanced variable selection techniques to identify global and condition-specific significant regulators. Figure 1 provides an overview of the MORE methodology.

Methodology

MORE general framework

MORE infers a different MO-RN for each biological condition or phenotype under study (e.g. control/disease, treatments, or stages of a particular disease) using a regression framework. For fitting the regression models, MORE requires a target omic with the "regulated" features (e.g., gene expression) and one or more regulatory omics, that is, the potential regulators (e.g., methylation, miRNA expression, chromatin accessibility, etc.), all of them measured on the same n patients or samples. Any omic modality is admitted, but let gene expression be the target omic for simplicity. Optionally, prior information about the potential target genes of each omic regulator can be provided. If not, MORE will consider all regulators as potential regulators of each gene.

The MORE strategy consists of fitting a regression model per gene, where the potential regulators of that gene will be the predictors in the model. As the regulatory mechanisms might differ for each phenotype or condition, such conditions and their interactions with the potential regulators will be included in the model. Again, for simplicity, let us consider just two conditions (e.g. control and disease), although the model can be applied to any number of conditions.

Let $\mathbf{y}_g$ be the expression vector of gene g ($g = 1, ..., G$) in the n samples; R_g the number of potential regulators of gene g; $\mathbf{x}_i^{(g)}$ the values of the potential regulator i of gene g for the n samples, and $\mathbf{z}$ the dummy variable indicating whether a sample belongs to the control or disease group. Equation 1 describes the regression model to be fitted by MORE:

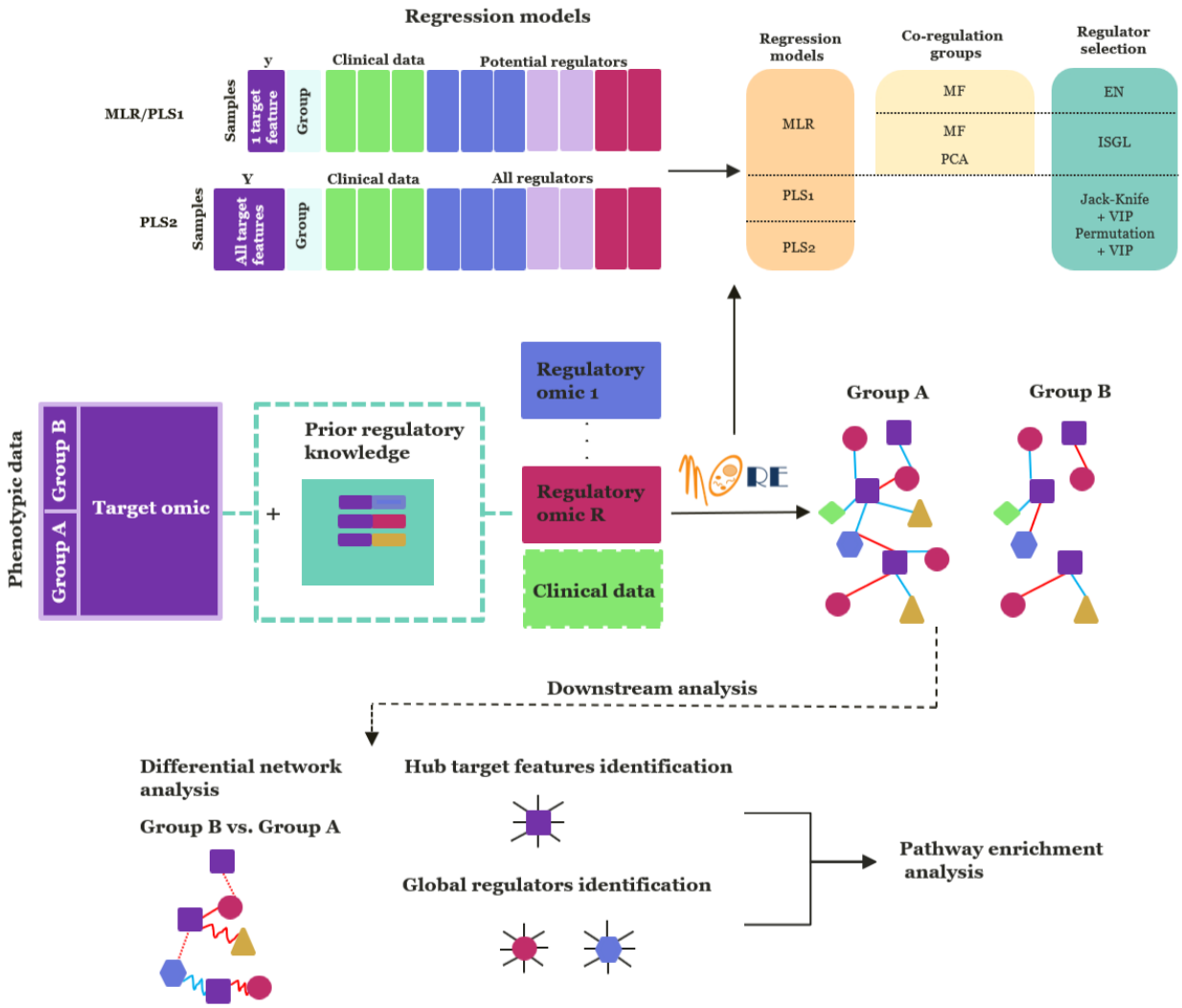

Fig. 1. Overview of the MORE R package. MORE takes as input a target or "regulated" omic (e.g. gene expression), the regulatory omics (e.g. miRNA expression, DNA methylation, etc.) or clinical data, and the phenotypic groups to be compared (e.g. healthy/disease). Optionally, prior knowledge about regulatory relationships can also be provided (e.g. miRNA target genes). MORE applies regression models and advanced variable selection techniques to obtain the phenotype-specific regulatory networks. MORE networks display activation (blue) and repression (red) regulatory relationships, and the functionalities for network biological interpretation include differential network analysis to compare phenotypes, identification of hub target features and global regulators, and functional enrichment analysis of sets of relevant features in the regulatory network.

$$E(\mathbf{y}_g) = \beta_0 + \beta_1 \mathbf{z} + \sum_{i=1}^{R_g} \gamma_i \mathbf{x}_i^{(g)} + \sum_{i=1}^{R_g} \delta_i \mathbf{z} \odot \mathbf{x}_i^{(g)}, \qquad (1)$$

where $\odot$ represents the Hadamard product, and β_0, β_1, γ_i, and δ_i are the regression model coefficients. Note that when prior knowledge about potential regulations is provided to MORE, each gene may have a different set of regulators. Consequently, this initial model equation for each gene will be different.

MORE relies on Multiple Linear Regression (MLR) or Partial Least Squares regression (PLS) [25] models to fit the regression model presented in Equation 1. In the PLS case, there are two options: fitting a PLS1 model per gene where the response matrix $\mathbf{Y}$ is the vector $\mathbf{y}_g$ or a unique PLS2 model where $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_G]$. In both cases, the predictor matrix $\mathbf{X}$ is the result of concatenating the regulators and their interactions with the studied condition. In PLS1, only the potential regulators of gene g are included in $\mathbf{X}$. In PLS2, all the potential regulators of all the G genes will constitute $\mathbf{X}$ matrix. PLS models, opposite to MLRs, take advantage of multicollinear data and can better address the high-dimensionality of multi-omics datasets as they are dimension reduction methods.

Both MLR or PLS regression models in MORE are combined with advanced variable selection strategies to properly handle high-dimensionality and multicollinearity and improve the biological interpretability of the results. These variable selection strategies are specific to each type of regression model, as described next.

Variable selection in MORE MLR models

Omic regulators tend to act together, thus presenting highly correlated values. This multicollinearity situation hampers the estimation of MLR coefficients and the interpretation of the model, making it necessary to apply efficient variable

selection methods that circumvent the problem. As detailed below, we have included in MORE two different approaches for selecting relevant regulators in MLRs: (i) a combination of a multicollinearity filter (MF) and the ElasticNet (EN) strategy, namely MF-EN, and (ii) the Iterative Sparse Group Lasso (ISGL) method.

(i) Multicollinearity filter and ElasticNet

The ElasticNet (EN) regularisation approach [26] for variable selection allows for having more predictors than observations in the MLR. Under a multicollinearity scenario, it is more effective than other methods, such as stepwise regression, but still presents a far from optimal performance when correlation among predictors is high. To avoid missing relevant predictors due to multicollinearity and, at the same time, improve the biological interpretation of the models, MORE performs a multicollinearity filter (MF) before defining the initial MLR equation for each gene. This MF finds groups of co-regulators (i.e. highly correlated regulators) by first creating a correlation network and next identifying highly connected groups of regulators (the detailed procedure can be found in Supplementary Note 1). By only including appropriate representative regulators from these connected groups in the model, the multicollinearity problem and the number of predictors are reduced. The correlation measurement used depends on the nature of regulator values: Pearson's correlation for numerical regulators, Phi coefficient for binary regulators, and point-biserial correlation between numerical and binary regulators (see Supplementary Note 1).

Once the MF has been applied, MORE fits the corresponding MLR with EN penalisation for variable selection. The EN penalisation parameters (see section Supplementary Note 2) are optimised by k-fold cross-validation techniques. EN requires the predictor variables $x_i^{(g)}$ to be scaled so, given the different number of variables from each omic modality (or omic block), MORE offers three different scaling procedures [27]: auto-scaling (Equation 2), soft block-scaling (Equation 3), and hard block-scaling (Equation 4).

$$auto(\mathbf{x}_i^{(g)}) = \frac{\mathbf{x}_i^{(g)}}{\sigma_{\mathbf{x}_i^{(g)}}}, \tag{2}$$

$$softblock(\mathbf{x}_i^{(g)}) = \frac{\mathbf{x}_i^{(g)}}{\sigma_{\mathbf{x}_i^{(g)}}\sqrt[4]{m_i}}, \tag{3}$$

$$hardblock(\mathbf{x}_i^{(g)}) = \frac{\mathbf{x}_i^{(g)}}{\sigma_{\mathbf{x}_i^{(g)}}\sqrt{m_i}}, \tag{4}$$

where $\sigma_{\mathbf{x}_i^{(g)}}$ is the standard deviation of $\mathbf{x}_i^{(g)}$ and m_i is the number of potential regulators of gene g included in the regression model belonging to the same omic block as $\mathbf{x}_i^{(g)}$.

(ii) Iterative Sparse Group Lasso

MORE implements another penalisation-based approach for variable selection, the Iterative Sparse Group Lasso (ISGL) penalisation [28]. As in the Sparse Group Lasso (SGL) [29] strategy, ISGL estimates a vector of regression coefficients ($\vec{\beta}_g = [\beta_1, \{\gamma_i, i \in \{1, ..., R_g\}\}, \{\delta_i, i \in \{1, ..., R_g\}\}]^t$) that minimises the objective function in Equation 5, which penalises the coefficients both individually and at group level.

$$\| \mathbf{y}_g - \hat{\mathbf{y}}_g \|^2 + \lambda[(1-\alpha)\sum_{l=1}^{L}\sqrt{p_l} \| \vec{\beta}_g^{(l)} \|_2 + \alpha \| \vec{\beta}_g \|_1], \tag{5}$$

where L denotes the number of groups of predictors (regulators, in our case), p_l the number of predictors in group l, λ the shrinkage parameter, and α the parameter controlling the weight of group and individual penalisation. The ISGL method uses a reformulation of the SGL problem to optimise penalisation parameters by applying a gradient-free coordinate descent algorithm, which overcomes potential overfitting issues in high-dimensional scenarios.

The L groups of predictors in the SGL are meant to be groups of related predictors. MORE implements two different strategies to create such groups: (i) obtaining clusters of highly correlated regulators by following the MF strategy (Supplementary Note 1), or (ii) applying a Principal Component Analysis (PCA) on the predictors and assigning each predictor to the principal component (group) where it presents the highest loading [30]. As ISGL also requires scaled variables, MORE again offers the options described in Equations 2, 3 and 4.

Variable selection in MORE PLS models

First, we introduce the notation for PLS models to explain variable selection strategies better. Equation 6 defines the matrix form of a PLS model:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{F}, \tag{6}$$

where $\mathbf{Y}$ and $\mathbf{X}$ denote the response and the predictor matrices respectively, as previously defined, $\mathbf{B}$ denotes the coefficient matrix, and $\mathbf{F}$ denotes the error matrix.

As a dimension reduction method, the idea behind PLS is to create M components or latent variables for both $\mathbf{Y}$ and $\mathbf{X}$ matrices in such a way that the covariance between $\mathbf{Y}$ and $\mathbf{X}$ components is maximised. These components are linear combinations of the original $\mathbf{X}$ variables and the projections of the n observations into these components are collected in the score matrix $\mathbf{T}$ ($n \times M$), defined in Equation 7:

$$\mathbf{T} = \mathbf{XW}^*, \tag{7}$$

where $\mathbf{W}^* = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}$ is a transformed version of the weight matrix $\mathbf{W}$, and $\mathbf{P}$ is the loading matrix that gives the contribution of each $\mathbf{X}$ variable to each $\mathbf{X}$ component.

Similarly, the score $\mathbf{U}$ and weigth $\mathbf{C}$ matrices are defined for $\mathbf{Y}$, such as $\mathbf{U} = \mathbf{YC}$. An iterative procedure is applied to estimate all these matrices so that the covariance between each $\mathbf{X}$ and $\mathbf{Y}$ component is maximised, and the regression coefficients matrix $\mathbf{B}$ can be computed as in Equation 8:

$$\hat{\mathbf{B}} = \mathbf{W}^*\mathbf{C}^T = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{C}^T \tag{8}$$

Variable scaling is also required in PLS models, so one of the three scaling approaches (Equations 2, 3, or 4) should be applied. Interestingly, MORE PLS models can handle missing data thanks to the nonlinear iterative Partial Least Squares (NIPALS) [31] algorithm applied for inferring the model.

The number of PLS components is estimated through the k-fold cross-validation procedure implemented in the ropls R package[32]. Once the PLS model is estimated, predictors are selected when having statistically significant regression coefficients and a high value of the Variable Importance in Projection (VIP) statistic. In MORE, the statistical significance of regression coefficients ($H0 : b_{kl} = 0$) can be computed either by a Jack-Knife resampling methodology or a permutation technique.

The Jack-Knife strategy removes one observation at a time; hence, n different models with $n-1$ observations are fitted. Let $\hat{b}_{kl(i)}$ be the regression coefficient estimation for the k-th predictor and the l-th response variable when removing the i-th observation. The Jack-Knife statistic J to test H_0 is computed as in Equation 9 and approximately follows a Student's t distribution with one degree of freedom.

$$J = \frac{\hat{b}_{kl}}{\hat{\sigma}_{Jack\beta_k}} = \frac{\hat{b}_{kl}}{\sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{b}_{kl(i)} - \hat{b}_{kl})^2}} \sim t_1, \qquad (9)$$

where $\hat{b}_{kl}$ denotes the estimation of the regression coefficient for the k-th predictor ($k = 1, ..., p$) and the l-th response variable ($l = 1$ for PLS1 and $l = 1, ..., G$ for PLS2).

For the permutation approach, the n observations in $\mathbf{Y}$ are randomly permuted R times. The estimated coefficients for each run ($\hat{b}_{kl}^{(r)}$, for $r = 1, ..., R$) are used to create the null distribution needed to compute the statistical significance of each coefficient.

Regarding the VIP statistic (Equation 10), predictors with VIP values above a certain threshold (usually around 1) can be selected as relevant predictors in PLS models. In MORE, we combine VIP selection with either Jack-Knife or permutation strategies and select variables with significant predictors and a minimum VIP of 0.8.

$$VIP_k = \sqrt{p\frac{\sum_{m=1}^{M} w_{km}^2 SSY_m}{SSYT}}, \qquad (10)$$

where p is the number of predictors in $\mathbf{X}$, M is the number of PLS components, w_{km} is the weight of the k-th predictor in the m component, SSY_m denotes the $\mathbf{Y}$ variance explained by the m-th component, and $SSYT$ is the total $\mathbf{Y}$ variance explained by the M components.

In summary, MORE will select predictors with a significant regression coefficient and a VIP value above the established threshold.

MORE regulatory networks

The models and variable selection strategies implemented in MORE allow for identifying the significant regulators of each gene to build the MO-RN. As the significance of the interactions between regulators and phenotypes or conditions is also computed, we may have different regulatory situations for each phenotype or condition that, in turn, will produce a different network per condition:

- When the regulator is significant, but the interaction term is not, the regulator has the same effect on gene expression for all the conditions under study.
- When any interaction term is significant, the regulator has a different effect on gene expression for each condition.

MORE connects the significant regulators under each condition with the regulated genes in an undirected graph, where negative regression coefficients correspond to inhibitors of gene expression and positive coefficients to activators. With this information, the MO-RN for each condition is generated and can be automatically visualised through the Cytoscape platform [33]. When comparing two conditions or phenotypes (e.g. healthy and diseased), MORE can compute the differential regulatory network between both conditions to highlight the condition-specific regulations.

As the generated networks' dimensionality may hamper the interpretability of the regulatory mechanisms, we implemented additional functionalities in MORE to assist researchers in such interpretation by connecting the relevant features in the network to biological functions or pathways. On the one hand, MORE can display the regulatory subnetwork associated with a given biological pathway. On the other hand, we implemented several options for the functional enrichment analysis of the regulatory network, as described next.

Functional enrichment of global regulators

In MORE, a regulator is defined as "global" when significantly regulating more than ten genes and having a network degree exceeding the third quartile of all network degrees. Global regulators are meant to play a critical role in regulatory mechanisms since their malfunction could trigger essential changes in the biological system. An important question about a given global regulator is whether it controls the regulation of a particular biological pathway; in other words, if the corresponding regulated genes are enriched in any biological pathway. To answer this question, MORE applies an Over-Representation Analysis (ORA) to the set of regulated genes versus the rest under study. In particular, we have implemented a Fisher's exact test to perform the ORA.

However, there are significant regulators that, by nature, will never be identified as global regulators. For instance, although mutations or methylation sites rarely regulate many genes, they may still be relevant to check if genes regulated by mutations or methylation can be associated with specific functions or pathways. Therefore, MORE also offers the possibility of applying ORA to the set of genes regulated by the same omic modality.

Functional enrichment analysis of regulated genes

Other questions to be answered from the MO-RN network are (i) whether the highly-regulated genes (genes with many significant regulators) are involved in specific biological pathways or (ii) whether genes whose regulations change between conditions have a particular biological function. MORE addresses these two questions through a Gene Set Enrichment Analysis (GSEA) performed with the clusterProfiler R package [34]. For question (i), genes are ranked by the number of significant regulators in a given condition. For question (ii), genes are ranked according to the score S_g defined for each gene g in Equation 11) when comparing conditions A and B.

$$S_g = \log_2\left(\frac{R_g^A + 1}{R_g^B + 1}\right), \qquad (11)$$

where R_g^A and R_g^B represent the number of significant regulators of gene g in conditions A and B, respectively.

MORE can also perform an ORA on the hub genes specific to a given condition.

Benchmarking MORE

As several regression models and variable selection strategies have been implemented in MORE, we first tested them on the complete simulated data (see next section) to evaluate their performance and provide some advice on the best choices for each biological or experimental scenario.

We selected the MORE model with the best overall performance from the previous simulation study and compared it with other state-of-the-art tools for MO-RN inference. As discussed in the Introduction, we found few available tools

comparable to MORE in terms of MO-RN generation with any number or types of omic modalities and for specific conditions, one of them being KiMONo [23, 24]. We also compared MORE to mixOmics [21], a reference tool in multi-omic analysis not explicitly developed for MR-RN inference.

KiMONo. Similarly to MORE, KiMONo is based on linear regression models to identify significant regulators of gene expression, applies SGL penalisation for variable selection where each omic modality is considered to be a group, and can (optionally) incorporate prior knowledge about potential regulations to infer the MO-RNs. KiMONo also allows the inclusion of the phenotype in the model but does not create phenotype-specific MO-RNs. KiMONo uses the igraph R package [35] to plot the inferred MO-RNs and identifies network hubs by computing node betweenness centrality. In addition, KiMONo integrates various Lasso-based strategies to handle missing data. We used version 2.0.1.0006 of KiMONo in GitHub, which depends on the hmlasso R package. This package is deprecated, so we installed it from the GitHub repository to make KiMONo work.

mixOmics. mixOmics incorporates different model options to analyse multi-omic data. Although none of them is directly comparable to MORE or KiMONo, we chose the sparse PLS (sPLS) [36] model for our benchmarking since it could be better adapted to the aim of this study. The sPLS methodology applies a Lasso variable selection approach to the PLS model and optimises the number of variables to be selected via cross-validation procedures. When comparing mixOmics to MORE or KiMONo, some limitations were found: (i) mixOmics does not accept prior knowledge about potential regulations. (ii) From the sPLS2 results, it is not possible to retrieve which genes are regulated by each specific significant regulator. (iii) Variables with low variability are not admitted. Considering these limitations and aiming to perform a fair comparison, we manually generated an sPLS1 model for each gene with all the omic variables as predictors to recover the specific significant regulators of each gene. We excluded regulators with a standard deviation lower than 2 to overcome the computational errors reported by the algorithm. The number of PLS components was optimised by minimising the mean absolute error (MAE).

Despite the adaptations made to the mixOmics PLS model, it is still not comparable with MORE PLS1 and KiMONo because mixOmics cannot incorporate prior regulatory knowledge. Consequently, we also included MORE PLS2 in the benchmarking, which presents the same limitation. Moreover, as variable selection in mixOmics requires a grid with the potential numbers of regulators to be considered and the optimum is selected by a 2-fold cross-validation procedure with five repetitions and MAE minimisation, we adapted this grid to overcome computational limitations. For sample sizes smaller than 20 observations per group, the grid included numbers in increments of 25 regulators, 75 for 30 observations, and 150 for 50 observations.

The comparison of MORE, KiMONo and mixOmics was performed on the same simulated data but with a reduced number of regulatory omics, given the high computational requirements of mixOmics (see Data Section).

Data

Simulation experiments

The performance of MORE or other MO-RN methods can be defined as their capacity to detect "true" regulations. For that, we simulated a wide range of multi-omic datasets with the MOSim R package [37] where gene expression is the target "regulated" omic and the other omic modalities act as potential regulators. MOSim can simulate different omic data types and their regulatory connections to gene expression data. These simulated regulatory effects will provide the set of "true" regulators to measure the model performance.

We simulated 20 different biological or experimental scenarios with MOSim. The varying factors across scenarios were:

- Sample size per phenotype (group): 5, 10, 20, 30 and 50.
- Percentage of differentially expressed genes (DEGs): 30% and 50%.
- Percentage of significant (active) regulators among the total number of potential regulators: 40% and 60%. In all cases, we established the same percentage of activator and repressor regulators but miRNA, where all significant regulators were set to be repressors.

In all of these scenarios, two phenotypes (cases and controls) and five omic modalities were simulated: gene expression (RNA-seq), transcription factor (TF) binding activity (ChIP-seq), TF expression, microRNA (miRNA) expression (miRNA-seq), DNA methylation (Methyl-seq), and chromatin accessibility (ATAC-seq). Three MO datasets were simulated for each scenario, thus resulting in the 60 MO simulated datasets we used for benchmarking.

Due to the computational cost of the mixOmics tool and its predictor variability restrictions, we used a reduced subset of the complete simulated data to compare mixOmics to MORE and KiMONo. Specifically, we only considered miRNAs and TFs as potential regulatory omics.

We first filtered out genes with an average expression of less than a count per million. Next, we applied the weighted trimmed mean of M-values (TMM) normalisation with NOISeq R package [38], voom-transformed [39] the TMM values, and computed differentially expressed genes (DEGs) between cases and controls with limma R package[40]. Genes with Benjamini and Hochberg adjusted $p - value < 0.05$ were declared DEGs.

As MOSim provides both the pairs of potential regulator-gene per omic modality and the significant regulatory pairs (let P be the number of significant regulations), we can consider these P "true" regulations as the "positive" instances and the significant regulations detected by a given method as the predicted "positives". Therefore, we can compute a confusion matrix for each method and simulated dataset and use the F1-score metric to assess the method's performance. In addition, as the selection of significant regulators is based on regression models for all the benchmarked approaches, the coefficient of determination R^2 was also computed to evaluate the goodness of fit of the models obtained by each method.

TCGA Ovarian Cancer data

A high-grade serous ovarian adenocarcinoma (HGSOC) dataset from The Cancer Genome Atlas Ovarian Cancer Cohort [41] was used to showcase MORE functionalities. The HGSOC dataset consists of 4 omic modalities measured on 292 patients: gene expression, Copy Number Variations (CNV), DNA Methylation, and microRNA expression. Data for the three first omics was downloaded from the UCSC Xena data portal [42], and normalised miRNA array data was obtained from the Broad Institute FireBrowse data portal [43].

Gene expression data was measured with RNA-seq and was already normalised and log-transformed. The NOISeq R package [38] was applied to filter out low-count genes, which resulted in keeping 20530 genes. TFs were extracted from this RNA-seq dataset and used as another regulatory layer (657 TFs). CNVs were aggregated at the gene level using the GISTIC2 method [44], resulting in CNV measurements for 24776 genes. DNA methylation was measured with Illumina microarrays, with 21666 probes, after excluding those with missing values. Methylation probes with a coefficient of variation lower than 20% were also filtered out.

We obtained the prior regulatory knowledge for MORE as follows. TF-target genes associations were obtained from the TF Link database [45]; CNVs information was already at gene level; methylation probes were associated with genes according to information provided by Illumina [46]; and target genes for the 723 miRNAs were obtained with the mirWalk web tool [47]. Specifically, we only considered associations collected in three databases: TargetScan [48], mirDB [49] and mirTarBase [50].

The phenotypes to be compared with MORE were given by the four cancer subtypes defined in [41]: *differentiated*, *immunoreactive*, *mesenchymal* and *proliferative*, also retrieved from the UCSC data portal. DEGs among subtypes were obtained with the limma R package [51] (adjusted p-value < 0.01), so 12146 DEGs were finally used as targets to generate the regulatory networks with MORE. Regulatory omic features were autoscaled, and the PLS1 model with the jack-knife resampling strategy was applied to compute the significance of PLS regression coefficients combined with a VIP value of 0.8 to select significant regulators.

Results and Discussion

Comparing models and strategies for variable selection in MORE

Before benchmarking MORE to other MO-RN tools, we compared the methodological options implemented in the package with different settings, as summarised in Table 1. This comparison was focused on the F1-score and considered the sample size (SS) per condition (Supplementary Figures S1 and S2). ANOVA models were used to assess significant differences in the average F1-score between models and their hyperparameters.

For single-response methods (MLR and PLS1, Supplementary Figure S1), the first expected conclusion was the superior performance of all methods for higher SS, that is, with at least 20 observations per group (ANOVA p-val $< 2e-16$). Interestingly, auto-scaling worked better than block-scaling options for all methods (p-val $< 2e-16$), probably because, as prior regulatory knowledge was considered, the numbers of potential regulators per omic for a given gene are not different enough in magnitude to need the block correction.

For PLS2 models (Supplementary Figure S2), as all the omic variables are included as predictors in the model, we had to reduce the omics used in this analysis only to miRNAs and TFs as MOSim generated data from over a million regulators making it unfeasible from the memory usage point of view. In this case, we got a very poor performance in terms of the F1-score, which highlights the importance of providing specific prior regulatory information for each gene to the regression models to avoid spurious associations between target features and potential regulators. Moreover, as all the omic variables are included as predictors in the PLS2 model and each omic block dimensions were still very different (130 TFs versus more than 1000 miRNAs in the simulated data used for this comparison), the hard-block scaling presented the best performance (p-val $< 2e-16$), especially in low SS scenarios. Regarding variable selection, the permutation approach presented significantly better results than Jack-knife (p-val = 0.0062).

Regarding the rest of the settings for the auto-scaling option in MLR and PLS1 methods, the optimal value for the correlation threshold was $r = 0.7$ for both MLR+EN-MF (average F1-score $\bar{F}_1 = 0.805$) and MLR+ISGL-MF ($\bar{F}_1 = 0.811$), where the difference between $r = 0.7$ and $r = 0.8$ was only significant at low SS (p-val $< 2e-16$, Supplementary Figure S1 A and C). In the case of MLR+ISGL-PCA (Supplementary Figure S1 B), the optimal value for the explained variance in PCA was 20% ($\bar{F}_1 = 0.813$). Opposite to PLS2, the Jack-Knife variable selection strategy was selected for PLS1 (Supplementary Figure S1 D, $\bar{F}_1 = 0.802$), probably due to the higher dimensionality of predictor matrix in PLS2, as discussed in [52]. Finally, in Supplementary Figure S3, we introduce the F1-score results of single-response approaches across omic types using the best performing settings in each case mentioned above.

Table 1. MORE models and variable selection strategies compared on simulated data. Optimal options according to the F1-score are highlighted in bold font.

Models	Variable Selection Methods	Parameters	Scaling type
MLR	**EN + MF**	r = **0.7,** 0.8	**auto** soft block hard block
	ISGL + MF		
	ISGL + PCA	R2 = **0.2,** 0.5, 0.7, 0.8	
PLS1	**Jack-Knife** + VIP	$\alpha = 0.05$ VIP = 0.8	
	Permutation + VIP		
		# perm = 100	
PLS2	Jack-Knife + VIP	$\alpha = 0.05$ VIP = 0.8	auto soft block **hard block**
	Permutation + VIP		
		# perm = 100	

Following the optimisation of MORE strategies described above, we performed a final comparison of the resulting optimal combinations for MLR and PLS1 models in terms of F1-score, coefficient of determination (R^2), and computational time (Figure 2) across different sample sizes per condition (SS), percentage of significant regulations among potential regulations (PSR) and percentage of DEGs (PDEG).

As expected, the SS value is critical for the performance of any methodology (p-val $< 2e-16$) (Figure 2A). As previously noted, the differences in the F1-score were more pronounced (p-value < 0.0001) for low SS ($n = 5$ or 10) than for higher SS, where we found no significant differences among compared methods (p-val = 0.168). MLR+ISGL obtained the best performance for low SS (for both MF and PCA options), while no differences were found between MLR+EN-MF and PLS1 approaches (p-val = 0.751). The R^2 was also affected by the SS. In this case, MLR+ISGL approaches presented lower

R^2 than other approaches, concluding that those approaches have high predictive performance but can not entirely explain the variability of the target omic features with the selected regulators. Regarding computational time, PLS1 was the fastest alternative, closely followed by MLR+EN-MF. The results for these two models were obtained by parallelisation with 16 cores, while the MLR+ISG strategy could not be parallelised, which explains the differences in running time.

Figure 2B shows the effect of PSR on the results. Interestingly, all methods except MLR+EN-MF present a higher F1-score and R^2 when the PSR increases. A similar behaviour is observed when increasing PDEG (Figure 2C). The lower the PSR or PDEG, the lower the number of simulated significant regulatory relationships. Therefore, the poorer performance of MLR+EN-MF at higher levels of PSR and PDEG could indicate the difficulty of this variable selection approach when dealing with highly multicollinearity scenarios.

The benchmarking results indicate that no single method within MORE consistently outperforms the others across all scenarios. For the comparison with KiMONo, we selected the PLS1 strategy due to its minimal computational time and robust overall performance, particularly regarding R^2. The PLS2 approach was benchmarked against mixOmics, as neither method incorporates prior regulatory knowledge. It is important to emphasise that choosing the most suitable methodology within MORE should be guided by the specific characteristics of the analysed dataset. To assist users in making this decision, we provide practical recommendations in Table 2, outlining the strengths and applicability of each approach.

Table 2. General recommendations to choose the most appropriate MORE settings for each experimental scenario.

	MLR + EN	MLR + ISGL	PLS1	PLS2
Target omic NOT normally distributed	✗	✗	✓	✓
Missing data	✗	✗	✓	✓
High number of target features	✓	More time consuming	✓	✓
High number of regulators	✓	✓	✓	High memory usage*
Interpretation of regulations	Against a reference group		Against the global average	
Allows for prior regulatory knowledge	✓	✓	✓	✗
Identifies groups of co-regulators	✓	✗	✓ (graphically)	✓ (graphically)
High multicollinearity	✗	✓	✓	✓
Scaling with prior regulatory knowledge for all omics	auto			block
Scaling without prior regulatory knowledge for any omic	block			

* ∼25000 regulators require more than 30GB of memory to create the model. The memory usage grows exponentially according to the number of regulators and target features included in the model. The number of phenotypes or conditions also affects.

Benchmarking MORE to other MO-RN methods

We compared the MORE-PLS1 model (with Jack-Knife coefficient significance and auto-scaling) to KiMONo, including prior regulatory knowledge in both cases. We also compared the MORE-PLS2 model to the PLS option in the mixOmics package, as none consider prior regulatory associations. We evaluated the ability of the methods to identify the significant regulations with the F1-score, the goodness of fit of the final models with the R^2 coefficient, and the computational efficiency (Figure 3). This benchmarking analysis was performed on the MOSim simulated datasets with only TFs and miRNAs as regulatory omics due to the computational limitations of methods that cannot incorporate prior regulatory knowledge. The results for MORE-PLS1 and KiMONo with all the regulatory omics are shown in Supplementary Material 4.

The F1-score results (Figure 3) show a significant difference between KiMONo and MORE-PLS1 compared to mixOmics and MORE-PLS2. This difference is basically due to incorporating prior regulatory knowledge, highlighting that such information is essential to avoid spurious regulations, as previously noted. Interestingly, MORE-PLS1 outperforms KiMONO at all sample sizes (p-val $< 2.74e-6$) but $n = 5$, where the difference between them is not statistically significant (p-val = 0.311). Both methods presented an overall excellent performance: MORE-PLS1 got a $\bar{F}_1 = 0.787$, while KiMONo obtained a $\bar{F}_1 = 0.741$. In contrast, although MORE-PLS2 significantly outperformed mixOmics (p-val = 3.2e-06), the F1-score was very low for both (below 0.06).

The superiority of MORE-PLS1 in terms of R^2 (values above 0.8) is again observed for all sample sizes but $n = 5$, where the difference is not significant. Surprisingly, the R^2 values for MORE-PLS2 are very high (close to 1 for higher sample sizes), while the R^2 for mixOmics is below 0.1.

Finally, the most time-consuming method is mixOmics (average time of more than 9200 minutes), while MORE-PLS1, followed by MORE-PLS2, requires less time to compute the models (average time lower than 5 minutes).

When considering all the regulatory omics (Supplementary Figure S4), the differences between MORE-PLS1 and KiMONO are more evident, and MORE-PLS1 outperforms KiMONO in terms of F1-score (above 0.75 in all cases), R^2 (higher than 0.9) and computational time. Therefore, we can conclude that MORE-PLS1 outperformed other state-of-the-art tools in all the evaluated aspects. Moreover, in the next section, we showcase the unique functionalities of MORE to aid users in interpreting the inferred MO-RNs.

Ovarian cancer case study

We applied MORE-PLS1 to the High-Grade Serous Ovarian Cancer (HGSOC) dataset described in the Data Section, as this model presented the best performance for sample sizes of 50 observations per group, which is approximately the sample size for the four cancer subtypes we will compare in this study.

MORE fitted 11286 models, one for each gene having potential regulators in our prior regulatory information. These genes presented an average of 70.18 significant regulators. However, from now on, we will focus only on models with $R^2 > 0.5$, corresponding to genes whose regulation was well-described. This filter reduced the number of genes in the regulatory network to 5317, with an average of 102.36 significant regulators. The percentage of genes with significant regulations by any omic modality was very similar across tumour subtypes (Figure 4A), close to 50%, and genes are regulated mainly by TFs (around 40%), closely followed by CNVs. Interestingly, the percentage of genes significantly regulated by CNVs or TFs is very similar across tumour subtypes. Still, the *proliferative* subtype presents a higher rate of regulated genes by methylation or miRNAs. Considering the initial regulatory associations provided to MORE (Figure 4B), CNVs and Methyl were naturally the omics with the

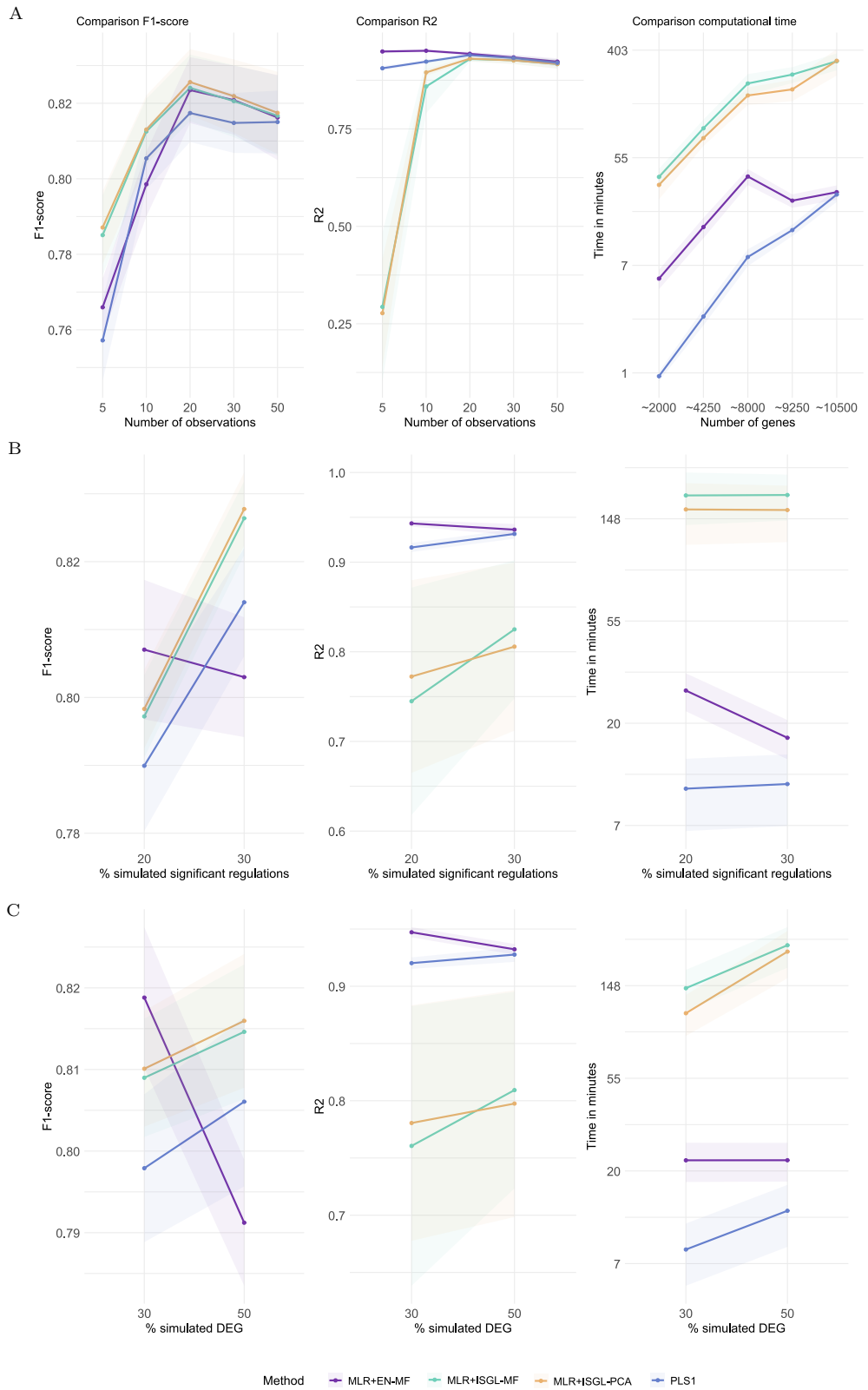

Fig. 2. F1-score, coefficient of determination (R^2), and computational time for MORE models with optimised settings. The compared models are Multiple Linear Regression with ElasticNet regularisation and multicollinearity filter (MLR+EN-MF), Multiple Linear Regression model with Iterative Sparse Group Lasso regularisation and Multicollinearity filter grouping (MLR+ISGL-MF), Multiple Linear Regression model with Iterative Sparse Group Lasso regularisation and Principal Component Analysis grouping (MLR+ISGL-PCA), and Partial Least Squares regression (PLS1) with Jack-Knife variable selection. **A)** Results across different sample sizes per condition (SS). **B)** Results for the two percentages of simulated significant regulations (PSR). **C)** Results for the two percentages of simulated differentially expressed genes (PDEG).

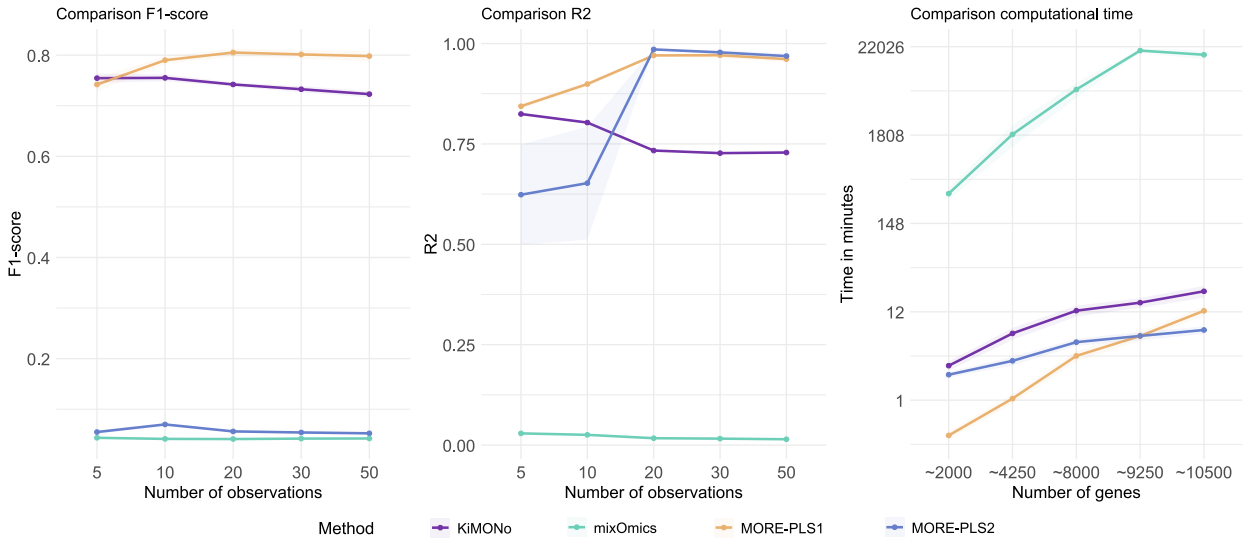

Fig. 3. Comparison of mixOmics, KiMONo, MORE PLS1 with auto-scaling and Jack-Knife and PLS2 with hard block-scaling and permutation approaches on simulated datasets. Performance results include F1-score, coefficient of determination (R2), and computational efficiency.

lowest numbers of initial associations (10860 and 10594, respectively), but CNV presents the highest percentage of significant regulations (> 40%), followed by TFs (> 35%), and methylation (around 25%). In all omics but CNV, the percentage of significant regulations tends to increase for tumour subtypes with worse prognosis, especially for the *proliferative* subtype in methylation and TFs.

Figure 5A shows the complete MO-RN for the *differentiated* subtype. The interpretation of such networks becomes unfeasible due to the vast number of analysed genes and significant regulators identified by MORE models. Therefore, providing users with additional functionalities to dive into the network and extract valuable biological insights is mandatory.

A relevant question in regulatory networks is the identification of *hub genes*, defined as genes with a high number of significant regulators in the studied phenotype. Supplementary Figure S5 compares the hub genes identified by MORE for each subtype. One of the hub genes shared by all subtypes is *ZMYND8*, which has been related to a DNA damage response factor involved in regulating transcriptional responses, DNA repair activities at DNA double-strand breaks, and plays essential roles in regulating transcription during normal cellular growth [53]. Other interesting hub genes are those associated with a specific subtype since they characterise the regulation of that subtype. For example, *BRCA1* is a hub gene only in the *differentiated* subtype. *BRCA1* is known to play a crucial role in the DNA damage repair pathway and is closely involved in the development of ovarian cancer [54, 55, 56, 57]. *BRIP1* is another interesting hub gene only present in the *differentiated* subtype, also involved in DNA damage repair, particularly in ovarian cancer predisposition [58, 59, 60]. *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

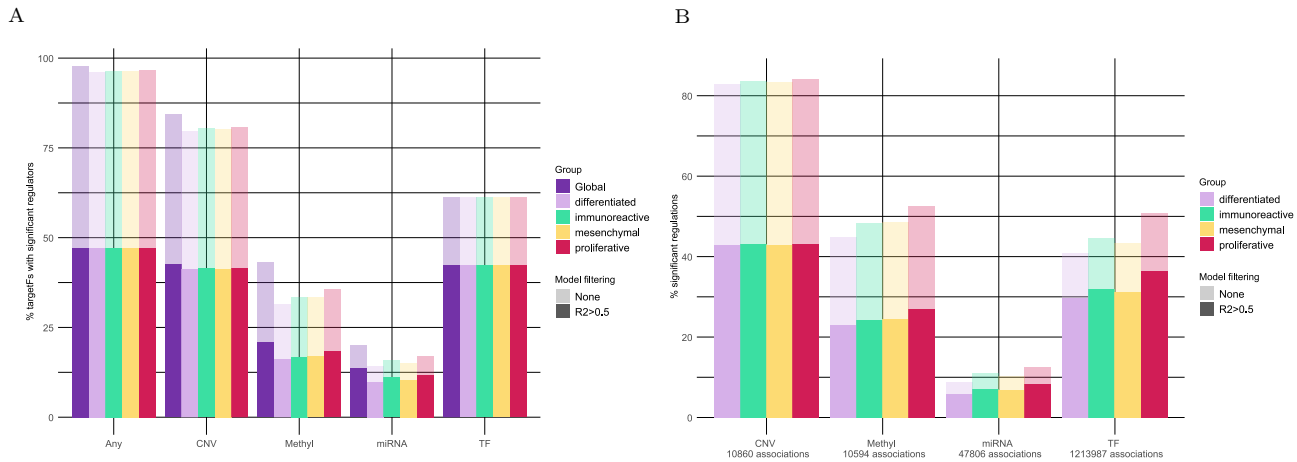

Fig. 4. Summary of significant regulations identified by MORE-PLS1 models on the ovarian cancer data. Transparent colours correspond to results derived from all fitted models. Solid colours correspond to models with $R^2 > 0.5$. A) The first set of bars shows the percentage of genes with significant regulations by any omic, globally or per tumour subtype. The following sets of bars refer to each omic modality and show the percentage of genes with significant regulations for that omic, globally or per tumour subtype. B) The percentage of significant regulations per omic over the initial potential regulations (associations) is displayed for each tumour subtype.

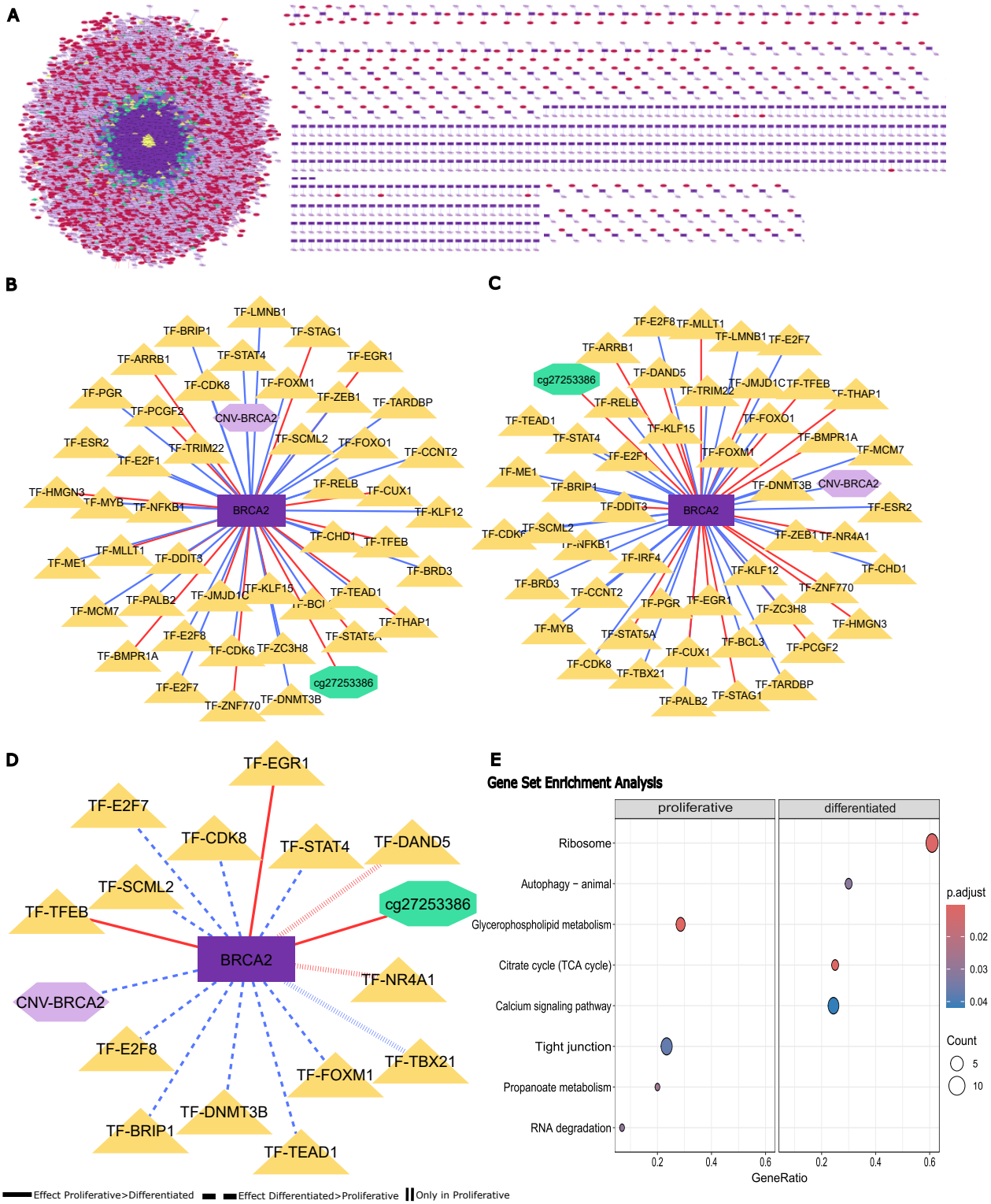

Fig. 5. MORE results on the ovarian cancer data. **A** Regulatory network for the *differentiated* subtype. **B, C** Regulatory networks for *BRCA2* gene in the *differentiated* and *proliferative* subtypes, respectively. Red lines represent repressor effects on BRCA2 expression, and blue lines activation. **D** Differential network between *differentiated* and *proliferative* subtypes. Straight lines represent a stronger regulatory effect in the *proliferative* subtype compared to the *differentiated*. Dashed lines represent a stronger regulatory effect in the *differentiated* subtype compared to the *proliferative*. Vertical lines represent a regulatory effect only significant for the *proliferative* subtype. **E** Gene Set Enrichment Analysis results on genes ranked by the ratio between the number of MORE significant regulations in the *proliferative* subtype over the *differentiated* subtype.

patients treated with platinum [61], and with an increased cell proliferation [62], potentially contributing to earlier tumour recurrence [63]. The *differentiated* subtype shares the hub genes *CCNE1* and *NF1* with the *immunoreactive* subtype. Cyclin E1 (*CCNE1*) gene amplification has been reported in 15%-20% of high-grade serous ovarian tumours and has been linked to aggressive disease courses and poor outcomes [64], while the biallelic inactivation of the neurofibromatosis type 1 (*NF1*) gene has been reported as an early event in HGSOC tumorogenesis [65]. *NF1* and *MYC* are also shared with the *mesenchymal* subtype. Amplifications of MYC oncogene have been reported to occur in nearly half of HGSOC tumours and are linked to maintaining ovarian cancer cells' oncogenic growth [66]. In the *mesenchymal* subtype, we found Fibrillin-1 (*FBN1*), which has been reported as a key factor in chemo-resistance of ovarian cancer cells by participating in the process of glycolysis and angiogenesis [67]. Finally, in the*proliferative* subtype, we found *CKB* gene. Creatine kinase B (*CKB*) is a cytosolic isoform of creatine kinase which has been reported as showing upregulated expression in several cancers at the protein level [68, 69]. This isoform has also been identified as a marker for early stages of an ovarian cancer diagnosis as its enzyme activity is significantly elevated in ovarian cancer patients, including early-stage patients [70].

For a better understanding of the differences between the *differentiated* and the *proliferative* subtypes, the two tumour subtypes with the best and worst overall survival outcome [71, 72], respectively, we performed a Gene Set Enrichment Analysis (GSEA) with MORE, where the gene score was computed as the ratio between the number of MORE significant regulations in each subtype. Functional annotation was retrieved from the Gene Ontology (GO) database [73, 74] with the AnnotationDbi Bioconductor package [75]. The functional categories more enriched in the *differentiated* subtype were "Ribosome", "Autophagy -animal", "Citrate cycle (TCA cycle)" and "Calcium signaling pathway", while "Glycerophospholipid metabolism", "Propanoate metabolism", "Tight junction", and "RNA degradation" were found for the *proliferative* subtype (Figure 5E). These biological functions are related to maintaining cellular structure, metabolism, energy production, and adaptation to stress, being critical in cell survival, growth, and functional regulation in multicellular organisms. Most of the altered pathways have been reported to play a crucial role in ovarian cancer; the autophagy pathway has even emerged as a therapy target in ovarian carcinoma, even if the clinical application is limited [76]. Interestingly, the Citrate cycle (TCA cycle) has been extensively reported to be an altered pathway in ovarian cancer, where the citric acid metabolite is significantly downregulated and considered to be a potential biomarker in this type of cancer[77, 78].

Other key features in regulatory networks are the *global regulators*, defined in MORE as significant regulators with many target genes, which play an essential role in the system regulation. As expected, TFs were the most abundant global regulators in all subtypes, with more than 80% representation. When focusing on the subtype-specific global regulators (Supplementary Figure S5), we found that the four subtypes shared 563 global regulators, while 15 were specific of *proliferative*, 2 of *mesenchymal*, 4 of *immunoreactive* and no specific regulators were found for the *differentiated* subtype. Remarkably, among these subtype-specific regulations, miRNAs are more prevalent: 2 of 2 in *mesenchymal*, 3 of 4 in *immunoreactive*, and 11 of 15 in *proliferative*. In the *proliferative* subtype, noteworthy miRNAs include hsa-miR-302c-3p and hsa-miR-186-5p. Interestingly, hsa-miR-302c-3p has been identified as a suppressor in the proliferation of cervical carcinoma cells, and it has been shown to inhibit the epithelial-mesenchymal transition and promote apoptosis in endometrial carcinoma cells [79]. As for hsa-miR-186-5p, this miRNA promotes the proliferation, invasion and stemness of ovarian cancer cells and plays an essential role in the regulatory effect of TUG1 on ZEB1 expression. ZEB1 has been reported to be overexpressed in multiple types of cancers and is a key transcriptional factor closely associated with epithelial-mesenchymal transition [80].

It is worth mentioning the case of the let-7 miRNA family, which has global regulators in all four subtypes. A low expression of this family has been associated with epithelial state and stemness, and these miRNAs are considered valuable predictors of HGSOC proliferation [81], as they have been directly related to migratory ability and invasiveness. We also found three members of the miR-200 family being global regulators in all the subtypes: hsa-miR-200a-3p, hsa-miR-200b-3p, and hsa-200c-3p. The miR-200 family is well-known for its implications in the diagnosis and prognosis of OC [82] and has been recognised as a master suppressor of epithelial-mesenchymal transition (EMT) [83].

Finally, it may also be interesting to study genes whose regulation was well-explained by MORE models. In particular, we focused on genes with models presenting an explained variance (R^2) over 80%. These genes included *BRCA1*, *BRCA2* or *IRF2BP2*, which have been reported to play important roles in ovarian cancer [84, 85]. Figures 5B and C and D show the subnetworks of *BRCA2* for the *differentiated* and *proliferative* subtypes, respectively, while Figure 5D represents the differential subnetwork "proliferative-differentiated". Among the differential regulators, we found BRIP1, FOXM1 and STAT4, previously reported as having relevant roles in ovarian cancer [86, 87, 88]. The three present an activator effect on *BRCA2*, stronger in the *differentiated* subtype. Notably, NR4A1 is a repressor of BRCA2 in the *proliferative* subtype, and high expression levels of this gene have been related to ovarian tumours with worse outcomes [89]. Supplementary Figure S6 shows the profiles of these four regulators and the *BRCA2* gene expression values.

Conclusions

Understanding regulatory mechanisms from a multi-omic perspective is essential for characterising the molecular circuits driving diseases and other phenotypes of interest. However, no publicly available tools currently exist to generate phenotype-specific, multi-layered regulatory networks in an integrative manner.

In this work, we present MORE, a versatile R GitHub package designed to infer multi-omic regulatory networks. MORE enables the generation and comparison of condition-specific regulatory networks across any number or type of omics layers and facilitates the incorporation of prior regulatory knowledge into the network construction process. By leveraging regression models and advanced variable selection strategies, MORE identifies significant regulators of target features (e.g., gene expression) and provides a comprehensive suite of functionalities to aid in the interpretation of the resulting regulatory networks.

We evaluated the performance of MORE using simulated datasets, demonstrating its high accuracy and robustness,

even with limited sample sizes. Additionally, we benchmarked MORE against established tools, such as KiMONo and mixOmics, in terms of identifying significant regulators, model goodness-of-fit, and computational efficiency. MORE consistently outperformed these tools across all evaluated criteria, highlighting its effectiveness and reliability.

To demonstrate its practical application, we applied MORE to an ovarian cancer dataset to study the regulation of gene expression by CNVs, transcription factors, miRNAs, and DNA methylation. We generated and compared networks for four tumour subtypes with distinct survival outcomes. The analysis revealed subtype-specific regulatory mechanisms which can explain their underlying prognostic variability. These findings emphasise the importance of capturing regulatory mechanisms not only at a global level but also in a phenotype-specific manner to better understand disease heterogeneity. This successful application of MORE highlights its utility as a powerful tool for dissecting complex regulatory networks and advancing the understanding of phenotype-specific molecular mechanisms.

Key points

- MORE is a flexible R package designed to infer and compare condition-specific regulatory networks across multiple omics layers, addressing the lack of publicly available tools for this purpose.
- Our tool relies on a regression framework and robust variable selection techniques to identify significant regulations, ensuring accurate network inference, and can incorporate prior regulatory information.
- Evaluations on simulated data and comparisons with state-of-the-art tools demonstrate MORE's superior performance in identifying regulators, model fit, and computational efficiency.
- MORE functionalities for downstream network analysis make it a versatile tool for interpreting complex multi-omic regulatory mechanisms.

Authors' contributions

S.T. and A.C. envisioned the study and conceptualised the MORE idea. S.T. supervised the work. M.C.-C. and S.T. designed and implemented a preliminary version of MORE. M.A.-M. developed and implemented the current version of MORE and performed all the analyses on simulated and experimental data. M.A.-M. and S.T. drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors thank Carolina Monzó for her support in simulating data with the MOSim R package.

Part of the analyses were performed on the high-performance computing cluster Garnatxa at the Institute for Integrative Systems Biology (I2SysBio), I2SysBio is a mixed research centre formed by the University of Valencia (UV) and the Spanish National Research Council (CSIC).

Funding

This work was supported by the Scientific Foundation of the Spanish Association Against Cancer through the project PERME224336TARA and by Instituto de Salud Carlos III through the project AC22/00058 (co-funded by the European Union as part of the Next Generation EU programme and the Recovery and Resilience Mechanism, MRR), both projects under the frame of ERA PerMed (ERAPERMED2022-141 - OVA-PDM). This work also received funding from the FP7 STATegra project (agreement no. 306000), the Spanish MINECO (BIO2012-40244), and the Spanish MICIN (PID2020-119537RB-100).

Data availability 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/Biostat-Omics/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.

References

1. Yunjin Li, Lu Ma, Duojiao Wu, and Geng Chen. Advances in bulk and single-cell multi-omics approaches for systems biology and precision medicine. *Briefings in Bioinformatics*, 22(5):bbab024, 2021.
2. Jongsu Lim, Chanho Park, Minjae Kim, Hyukhee Kim, Junil Kim, and Dong-Sung Lee. Advances in single-cell omics and multiomics for high-resolution molecular profiling. *Experimental & Molecular Medicine*, 56(3):515–526, 2024.
3. Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, Dec 2008.
4. Adam A. Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(1):S7, Mar 2006.
5. Angel Alvarez and Peter J. Woolf. Regnetb: Predicting relevant regulator-gene relationships in localized prostate tumor samples. *BMC Bioinformatics*, 12(1):243, Jun 2011.
6. Gennaro Gambardella, Maria Nicoletta Moretti, Rossella de Cegli, Luca Cardone, Adriano Peron, and Diego di Bernardo. Differential network analysis for the identification of condition-specific pathway activity and regulation. *Bioinformatics*, 29(14):1776–1785, 06 2013.
7. Adriano V Werhli and Dirk Husmeier. 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), 2007.
8. Min Jin Ha, Veerabhadran Baladandayuthapani, and Kim-Anh Do. DINGO: differential network analysis in genomics. *Bioinformatics*, 31(21):3413–3420, 07 2015.
9. Sonia Tarazona, Angeles Arzalluz-Luque, and Ana Conesa. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nature Computational Science*, 1(6):395–402, Jun 2021.
10. Nam D. Nguyen and Daifeng Wang. Multiview learning for understanding functional multiomics. *PLOS Computational Biology*, 16(4):1–26, 04 2020.

11. Luis Pedro Coelho. For long-term sustainable software in bioinformatics. *PLOS Computational Biology*, 20(3):e1011920, 2024.
12. Dong-Chul Kim, Mingon Kang, Baoju Zhang, Xiaoyong Wu, Chunyu Liu, and Jean Gao. 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, 2014.
13. Lin Yuan, Le-Hang Guo, Chang-An Yuan, Youhua Zhang, Kyungsook Han, Asoke K. Nandi, Barry Honig, and De-Shuang Huang. 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, 2019.
14. Jin Li, Feng Chen, Hong Liang, and Jingwen Yan. MoNET: an R package for multi-omic network analysis. *Bioinformatics*, 38(4):1165–1167, 10 2021.
15. Mohamed Hamed, Christian Spaniol, Maryam Nazarieh, and Volkhard Helms. Tfmir: A web server for constructing and analyzing disease-specific transcription factor and mirna co-regulatory networks. *Nucleic acids research*, 43, 05 2015.
16. Sara J. C. Gosline, Coyin Oh, and Ernest Fraenkel. SAMNetWeb: identifying condition-specific networks linking signaling and transcription. *Bioinformatics*, 31(7):1124–1126, 11 2014.
17. Anna Pačínková and Vlad Popovici. IntOMICS: A bayesian framework for reconstructing regulatory networks using Multi-Omics data. *J Comput Biol*, 30(5):569–574, March 2023.
18. Yue Li, Minggao Liang, and Zhaolei Zhang. Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLOS Computational Biology*, 10(10):1–14, 10 2014.
19. Caleb A Class, Min Jin Ha, Veerabhadran Baladandayuthapani, and Kim-Anh Do. iDINGO—integrative differential network analysis in genomics with Shiny application. *Bioinformatics*, 34(7):1243–1245, 11 2017.
20. R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*, 14(6):e8124, 2018.
21. F Rohart, B Gautier, A Singh, and K-A Le Cao. mixomics: An r package for 'omics feature selection and multiple data integration. *PLoS computational biology*, 13(11):e1005752, 2017.
22. Janet C. Siebert, Martine Saint-Cyr, Sarah J. Borengasser, Brandie D. Wagner, Catherine A. Lozupone, and Carsten Görg. Cantare: finding and visualizing network-based multi-omic predictive models. *BMC Bioinformatics*, 22(1):80, Feb 2021.
23. Christoph Ogris, Yue Hu, Janine Knauer-Arloth, and Nikola Müller. Versatile knowledge guided network inference method for prioritizing key regulatory factors in multi-omics data. *Scientific Reports*, 11, 03 2021.
24. Juan D Henao, Michael Lauber, Manuel Azevedo, Anastasiia Grekova, Fabian Theis, Markus List, Christoph Ogris, and Benjamin Schubert. Multi-omics regulatory network inference in the presence of missing data. *Briefings in Bioinformatics*, 24(5):bbad309, 09 2023.
25. Herman O. A. Wold. Nonlinear estimation by iterative least squares procedures. In F. N. David, editor, *Research Papers in statistics, Festschrift for J*, page 411–414. New York, New York, 1966.
26. Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005.
27. Lennart Eriksson, Erik Johansson, Henrik Antti, and Elaine Holmes. Multi-and megavariate data analysis: finding and using regularities in metabonomics data. *Metabolomics in Toxicity Assessment, Taylor & Francis Group, Boca Raton*, pages 263–335, 2005.
28. Juan Laria, M. Aguilera-Morillo, and Rosa Lillo. An iterative sparse-group lasso. *Journal of Computational and Graphical Statistics*, 28, 02 2019.
29. Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.
30. Juan C. Laria, M. Carmen Aguilera-Morillo, Enrique Álvarez, Rosa E. Lillo, Sara López-Taruella, María del Monte-Millán, Antonio C. Picornell, Miguel Martín, and Juan Romo. Iterative variable selection for high-dimensional data: Prediction of pathological response in triple-negative breast cancer. *Mathematics*, 9(3), 2021.
31. HERMAN WOLD. Nonlinear iterative partial least squares (nipals) modelling: Some current developments. In PARUCHURI R. KRISHNAIAH, editor, *Multivariate Analysis–III*, pages 383–407. Academic Press, 1973.
32. Etienne A. Thévenot, Aurélie Roux, Ying Xu, Eric Ezan, and Christophe Junot. Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and opls statistical analyses. *Journal of Proteome Research*, 14(8):3322–3335, Aug 2015.
33. Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, November 2003.
34. T Wu, E Hu, S Xu, M Chen, P Guo, Z Dai, T Feng, L Zhou, W Tang, and L Zhan. Fu x. *Liu S, Bo X, Yu G (2021). clusterProfiler 4. 0: A universal enrichment tool for interpreting omics data. The Innovation*, (3):100141., 2021.
35. Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.
36. Kim-Anh Lê Cao, Debra Rossouw, Christèle Robert-Granié, and Philippe Besse. A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol*, 7(1):Article 35, November 2008.
37. C Monzó, C Martínez-Mira, Arzalluz-Luque A, A Conesa, and S Tarazona. Mosim: bulk and single-cell multi-layer regulatory network simulator. *BioRxiv*, page 421834, 2024.
38. Sonia Tarazona, Pedro Furio-Tari, David Turra, Antonio Di Pietro, Maria Jose Nueda, Alberto Ferrer, and Ana Conesa. Data quality aware analysis of differential expression in rna-seq with noiseq r/bioc package. *Nucleic Acids Research*, 43(21):e140, 2015.
39. Charity W. Law, Yunshun Chen, Wei Shi, and Gordon K. Smyth. voom: precision weights unlock linear model

analysis tools for rna-seq read counts. *Genome Biology*, 15(2):R29, Feb 2014.

40. Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.
41. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature. 2011 Jun 29;474(7353):609-15. doi:*, 10., 2012.
42. UCSC. *xenabrowser*. https://xenabrowser.net/datapages/, November 2020.
43. Broad Institute of MIT and Harvard. *FireBrowse*. http://firebrowse.org/, October 2019.
44. Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, and Gad Getz. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*, 12(4):R41, April 2011.
45. O Liska, B Bohár, A Hidas, T Korcsmáros, B Papp, D Fazekas, and E Ari. Tflink: an integrated gateway to access transcription factor-target gene interactions for multiple species. *Database (Oxford) 16;2022:baac083. doi:*, 10., September 2022.
46. Illumina. *HumanMethylation27 Product Support Files*. https://support.illumina.com/downloads/humanmethylation27_product_support_files.html, August 2013.
47. Carsten Sticht, Carolina De La Torre, Alisha Parveen, and Norbert Gretz. mirwalk: an online resource for prediction of microrna binding sites. *PloS one*, 13(10):e0206239, 2018.
48. Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microrna target sites in mammalian mrnas. *eLife*, 4:e05005, aug 2015.
49. Yuhao Chen and Xiaowei Wang. miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res*, 48(D1):D127–D131, January 2020.
50. Sheng-Da Hsu, Feng-Mao Lin, Wei-Yun Wu, Chao Liang, Wei-Chih Huang, Wen-Ling Chan, Wen-Ting Tsai, Goun-Zhou Chen, Chia-Jung Lee, Chih-Min Chiu, Chia-Hung Chien, Ming-Chia Wu, Chi-Ying Huang, Ann-Ping Tsou, and Hsien-Da Huang. miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res*, 39(Database issue):D163–9, November 2010.
51. Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.
52. Alwyn Young. Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results*. *The Quarterly Journal of Economics*, 134(2):557–598, 11 2018.
53. Fade Gong and Kyle M. Miller. Double duty: Zmynd8 in the dna damage response and cancer. *Cell Cycle*, 17(4):414–420, 2018. PMID: 29393731.
54. Irina Tsibulak, Verena Wieser, Christine Degasper, Giridhar Shivalingaiah, Sören Wenzel, Susanne Sprung, Sigurd F Lax, Christian Marth, Heidelinde Fiegl, and Alain G Zeimet. Brca1 and brca2 mrna-expression prove to be of clinical impact in ovarian cancer. *British journal of cancer*, 119(6):683–692, 2018.
55. Karoline B Kuchenbaecker, John L Hopper, Daniel R Barnes, Kelly-Anne Phillips, Thea M Mooij, Marie-José Roos-Blom, Sarah Jervis, Flora E Van Leeuwen, Roger L Milne, Nadine Andrieu, et al. Risks of breast, ovarian, and contralateral breast cancer for brca1 and brca2 mutation carriers. *Jama*, 317(23):2402–2416, 2017.
56. Anthony Antoniou, Paul DP Pharoah, Steven Narod, Harvey A Risch, Jorunn E Eyfjord, John L Hopper, Niklas Loman, Håkan Olsson, O Johannsson, Åke Borg, et al. 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, 2003.
57. Sining Chen and Giovanni Parmigiani. Meta-analysis of brca1 and brca2 penetrance. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 25(11):1329, 2007.
58. Jenna Lilyquist, Holly LaDuca, Eric Polley, Brigette Tippin Davis, Hermela Shimelis, Chunling Hu, Steven N Hart, Jill S Dolinsky, Fergus J Couch, and David E Goldgar. 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, 2017.
59. Malwina Suszynska, Katarzyna Klonowska, Anna J Jasinska, and Piotr Kozlowski. 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, 2019.
60. Malwina Suszynska, Magdalena Ratajska, and Piotr Kozlowski. 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, 2020.
61. Joanna Moes-Sosnowska, Iwona K Rzepecka, Joanna Chodzynska, Agnieszka Dansonka-Mieszkowska, Lukasz M Szafron, Aneta Balabas, Renata Lotocka, Piotr Sobiczewski, and Jolanta Kupryjanczyk. Clinical importance of fancd2, brip1, brca1, brca2 and fancf expression in ovarian carcinomas. *Cancer biology & therapy*, 20(6):843–854, 2019.
62. Zhigang Tu, Katherine M Aird, Benjamin G Bitler, Jasmine P Nicodemus, Neil Beeharry, Bing Xia, Tim J Yen, and Rugang Zhang. Oncogenic ras regulates brip1 expression to induce dissociation of brca1 from chromatin, inhibit dna repair, and promote senescence. *Developmental cell*, 21(6):1077–1091, 2011.
63. Yongru Liu, Xi Wu, Yunlu Feng, Qingwei Jiang, Shengyu Zhang, Qiang Wang, Aiming Yang, et al. Insights into the oncogenic, prognostic, and immunological role of brip1 in pan-cancer: A comprehensive data-mining-based study. *Journal of Oncology*, 2023, 2023.
64. J. A. Park, B. Bethany, M. Frimer, K. Cheng, and V. S. John. The role of ccne1 amplification in refractory ovarian and endometrial cancer. *Gynecologic Oncology*, 159:216, Oct 2020.
65. Eric J. Norris, Wendell D. Jones, Marius D. Surleac, Andrei J. Petrescu, Darla Destephanis, Qing Zhang, Issam Hamadeh, Jeffrey S. Kneisl, Chad A. Livasy, Ram N. Ganapathi, David L. Tait, and Mahrukh K. Ganapathi. Clonal lineage of high grade serous ovarian cancer in a patient with neurofibromatosis type 1. *Gynecologic Oncology Reports*, 23:41–44, 2018.

66. Mei Zeng, Nicholas P Kwiatkowski, Tinghu Zhang, Behnam Nabet, Mousheng Xu, Yanke Liang, Chunshan Quan, Jinhua Wang, Mingfeng Hao, Sangeetha Palakurthi, Shan Zhou, Qing Zeng, Paul T Kirschmeier, Khyati Meghani, Alan L Leggett, Jun Qi, Geoffrey I Shapiro, Joyce F Liu, Ursula A Matulonis, Charles Y Lin, Panagiotis A Konstantinopoulos, and Nathanael S Gray. Targeting MYC dependency in ovarian cancer through inhibition of CDK7 and CDK12/13. *Elife*, 7, November 2018.
67. Ziliang Wang, Wei Chen, Ling Zuo, Mi-Die Xu, Yong Wu, Jiami Huang, Xu Zhang, Yongheng Li, Jing Wang, Jing Chen, Husheng Wang, and Huizhen Sun. The fibrillin-1/vegfr2/stat2 signaling axis promotes chemoresistance via modulating glycolysis and angiogenesis in ovarian cancer organoids and cells. *Cancer Communications*, 42, 03 2022.
68. Xu-Hui Li, Xiang-Jun Chen, Wen-Bin Ou, Qian Zhang, Zhi-Rong Lv, Yi Zhan, Long Ma, Tao Huang, Yong-Bin Yan, and Hai-Meng Zhou. Knockdown of creatine kinase B inhibits ovarian cancer progression by decreasing glycolysis. *Int J Biochem Cell Biol*, 45(5):979–986, February 2013.
69. Ghofraan Abdulsalam Atallah, Nor Haslinda Abd. Aziz, Chew Kah Teik, Mohamad Nasir Shafiee, and Nirmala Chandralega Kampan. New predictive biomarkers for ovarian cancer. *Diagnostics*, 11(3), 2021.
70. Heather G Huddleston, Kwong-Kwok Wong, William R Welch, Ross S Berkowitz, and Samuel C Mok. Clinical applications of microarray technology: creatine kinase B is an up-regulated gene in epithelial ovarian cancer and shows promise as a serum marker. *Gynecol Oncol*, 96(1):77–83, January 2005.
71. Cancer Genome Atlas Research Network et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609, 2011.
72. Gottfried E Konecny, Chen Wang, Habib Hamidi, Boris Winterhoff, Kimberly R Kalli, Judy Dering, Charles Ginther, Hsiao-Wang Chen, Sean Dowdy, William Cliby, et al. Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer. *Journal of the National Cancer Institute*, 106(10):dju249, 2014.
73. Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, Midori A Harris, David P Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C Matese, Joel E Richardson, Martin Ringwald, Gerald M Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.
74. The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 03 2023.
75. Pagès H, Carlson M, Falcon S, and Li N. Annotationdbi: Manipulation of SQLite-based annotations in Bioconductor. *R package version 1.68.0*, 2024.
76. Lei Zhan, Yu Zhang, Wenyan Wang, Enxue Song, Yijun Fan, Jun Li, and Bing Wei. Autophagy as an emerging therapy target for ovarian carcinoma. *Oncotarget*, 7(50):83476–83487, December 2016.
77. Xiaogang Wang, Yuxin Yin, Wei Qian, Chen Peng, Sunan Shen, Tingting Wang, and Shuli Zhao. Citric acid of ovarian cancer metabolite induces pyroptosis via the caspase-4/TXNIP-NLRP3-GSDMD pathway in ovarian cancer. *FASEB J.*, 36(6):e22362, June 2022.
78. Nicole M Anderson, Patrick Mucka, Joseph G Kern, and Hui Feng. The emerging role and targetability of the TCA cycle in cancer metabolism. *Protein & Cell*, 9(2):216–237, February 2018.
79. Yibing Li, Jianing Huo, Xin Pan, Cuicui Wang, and Xiaoxin Ma. Microrna 302b-3p/302c-3p/302d-3p inhibits epithelial–mesenchymal transition and promotes apoptosis in human endometrial carcinoma cells. *OncoTargets and therapy*, 11:1275 – 1284, 2018.
80. Fu-Liang Zhan, Chun-Fang Chen, and Mei-Zhen Yao. LncRNA TUG1 facilitates proliferation, invasion and stemness of ovarian cancer cell via miR-186-5p/ZEB1 axis. *Cell Biochem Funct*, 38(8):1069–1078, May 2020.
81. Evgeny Chirshev, Nozomi Hojo, Antonella Bertucci, Linda Sanderman, Anthony Nguyen, Hanmin Wang, Tise Suzuki, Emmanuel Brito, Shannalee R Martinez, Christine Castañón, Saied Mirshahidi, Marcelo E Vazquez, Pamela

Wat, Kerby C Oberg, Yevgeniya J Ioffe, and Juli J Unternaehrer. Epithelial/mesenchymal heterogeneity of high-grade serous ovarian carcinoma samples correlates with miRNA let-7 levels and predicts tumor growth and metastasis. *Mol Oncol*, 14(11):2796–2813, August 2020.

82. Manish K Pal, Shyam P Jaiswar, Vinaya N Dwivedi, Amit K Tripathi, Ashish Dwivedi, and Pushplata Sankhwar. Microrna: a new and promising potential biomarker for diagnosis and prognosis of ovarian cancer. *Cancer biology & medicine*, 12(4):328, 2015.
83. Sun-Mi Park, Arti B Gaur, Ernst Lengyel, and Marcus E Peter. 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, 2008.
84. D Bell, A Berchuck, M Birrer, J Chien, D W Cramer, F Dao, R Dhir, P DiSaia, H Gabra, P Glenn, A K Godwin, J Gross, L Hartmann, M Huang, D G Huntsman, M Iacocca, M Imielinski, S Kalloger, B Y Karlan, D A Levine, G B Mills, C Morrison, D Mutch, N Olvera, S Orsulic, K Park, N Petrelli, B Rabeno, J S Rader, B I Sikic, K Smith-McCune, A K Sood, D Bowtell, R Penny, J R Testa, K Chang, C J Creighton, H H Dinh, J A Drummond, G Fowler, P Gunaratne, A C Hawes, C L Kovar, L R Lewis, M B Morgan, I F Newsham, J Santibanez, J G Reid, L R Trevino, Y-Q Wu, M Wang, D M Muzny, D A Wheeler, R A Gibbs, G Getz, M S Lawrence, K Cibulskis, A Y Sivachenko, C Sougnez, D Voet, J Wilkinson, T Bloom, K Ardlie, T Fennell, J Baldwin, R Nichol, S Fisher, S Gabriel, E S Lander, L Ding, R S Fulton, D C Koboldt, M D McLellan, T Wylie, J Walker, M O'Laughlin, D J Dooling, L Fulton, R Abbott, N D Dees, Q Zhang, C Kandoth, M Wendl, W Schierding, D Shen, C C Harris, H Schmidt, J Kalicki, K D Delehaunty, C C Fronick, R Demeter, L Cook, J W Wallis, L Lin, V J Magrini, J S Hodges, J M Eldred, S M Smith, C S Pohl, F Vandin, E Upfal, B J Raphael, G M Weinstock, E R Mardis, R K Wilson, M Meyerson, W Winckler, G Getz, R G W Verhaak, S L Carter, C H Mermel, G Saksena, H Nguyen, R C Onofrio, M S Lawrence, D Hubbard, S Gupta, A Crenshaw, A H Ramos, K Ardlie, L Chin, A Protopopov, Juinhua Zhang, T M Kim, I Perna, Y Xiao, H Zhang, G Ren, N Sathiamoorthy, R W Park, E Lee, P J Park, R Kucherlapati, D M Absher, L Waite, G Sherlock, J D Brooks, J Z Li, J Xu, R M Myers, P W Laird, L Cope, J G Herman, H Shen, D J Weisenberger, H Noushmehr, F Pan, T Triche, Jr, B P Berman, D J Van Den Berg, J Buckley, S B Baylin, P T Spellman, E Purdom, P Neuvial, H Bengtsson, L R Jakkula, S Durinck, J Han, S Dorton, H Marr, Y G Choi, V Wang, N J Wang, J Ngai, J G Conboy, B Parvin, H S Feiler, T P Speed, J W Gray, D A Levine, N D Socci, Y Liang, B S Taylor, N Schultz, L Borsu, A E Lash, C Brennan, A Viale, C Sander, M Ladanyi, K A Hoadley, S Meng, Y Du, Y Shi, L Li, Y J Turman, D Zang, E B Helms, S Balu, X Zhou, J Wu, M D Topal, D N Hayes, C M Perou, G Getz, D Voet, G Saksena, Junihua Zhang, H Zhang, C J Wu, S Shukla, K Cibulskis, M S Lawrence, A Sivachenko, R Jing, R W Park, Y Liu, P J Park, M Noble, L Chin, H Carter, D Kim, J Samayoa, R Karchin, P T Spellman, E Purdom, P Neuvial, H Bengtsson, S Durinck, J Han, J E Korkola, L M Heiser, R J Cho, Z Hu, B Parvin, T P Speed, J W Gray, N Schultz, E Cerami, B S Taylor, A Olshen, B Reva, Y Antipin, R Shen, P Mankoo, R Sheridan, G Ciriello, W K Chang, J A Bernanke, L Borsu, D A Levine, M Ladanyi, C Sander, D Haussler, C C Benz, J M Stuart, S C Benz, J Z Sanborn, C J Vaske, J Zhu, C Szeto, G K Scott, C Yau, K A Hoadley, Y Du, S Balu, D N Hayes, C M Perou, M D Wilkerson, N Zhang, R Akbani, K A Baggerly, W K Yung, G B Mills, J N Weinstein, R Penny, T Shelton, D Grimm, M Hatfield, S Morris, P Yena, P Rhodes, M Sherman, J Paulauskis, S Millis, A Kahn, J M Greene, The Cancer Genome Atlas Research Network, (Participants are arranged by area of contribution and then by institution.), Disease working group and tissue source sites, Genome sequencing centres: Baylor College of Medicine, Broad Institute, Washington University in St Louis, Cancer genome characterization centres: Broad Institute/Dana-Farber Cancer Institute, Harvard Medical School, HudsonAlpha Institute/Stanford University, University of Southern California/Johns Hopkins University, Lawrence Berkeley National Laboratory, Memorial Sloan-Kettering Cancer Center, University of North Carolina at Chapel Hill, Genome data analysis centres: Broad Institute, Johns Hopkins University, University of California Santa Cruz/Buck Institute, The University of Texas MD Anderson Cancer Center, Biospecimen core resource, and Data coordination centre. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, June 2011.
85. Paula Punzón-Jiménez, Victor Lago, Santiago Domingo, Carlos Simón, and Aymara Mas. Molecular management of High-Grade serous ovarian carcinoma. *Int J Mol Sci*, 23(22), November 2022.
86. Malwina Suszynska, Magdalena Ratajska, and Piotr Kozlowski. 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. *J Ovarian Res*, 13(1):50, May 2020.
87. Cassie Liu, Carter J Barger, and Adam R Karpf. FOXM1: A multifunctional oncoprotein and emerging therapeutic target in ovarian cancer. *Cancers (Basel)*, 13(12), June 2021.
88. L Zhao, G Ji, X Le, Z Luo, C Wang, M Feng, L Xu, Y Zhang, W B Lau, B Lau, Y Yang, L Lei, H Yang, Y Xuan, Y Chen, X Deng, T Yi, S Yao, X Zhao, Y Wei, and S Zhou. An integrated analysis identifies STAT4 as a key regulator of ovarian cancer metastasis. *Oncogene*, 36(24):3384–3396, January 2017.
89. Evan Delgado, Michelle M Boisen, Robin Laskey, Rui Chen, Chi Song, Jad Sallit, Zachary A Yochum, Courtney L Andersen, Matthew J Sikora, Jacob Wagner, Stephen Safe, Esther Elishaev, Adrian Lee, Robert P Edwards, Paul Haluska, George Tseng, Mark Schurdak, and Steffi Oesterreich. High expression of orphan nuclear receptor NR4A1 in a subset of ovarian tumors with worse outcome. *Gynecol Oncol*, 141(2):348–356, March 2016.