Published online 3 May 2019 *Nucleic Acids Research, 2019, Vol. 47, No. 14* ***e80***
doi: 10.1093/nar/gkz324

MOSClip: multi-omic and survival pathway analysis for the identification of survival associated gene and modules

Paolo Martini[1,*], Monica Chiogna[2], Enrica Calura[1,†] and Chiara Romualdi[1,*,†]

[1]Department of Biology, University of Padova, Via U.Bassi 58B, 35121 Padova, Italy and [2]Department of Statistical Sciences 'Paolo Fortunati', University of Bologna, via delle Belle Arti 41, 40126 Bologna, Italy

Received September 22, 2018; Revised March 29, 2019; Editorial Decision April 11, 2019; Accepted April 29, 2019

ABSTRACT

Survival analyses of gene expression data has been a useful and widely used approach in clinical applications. But, in complex diseases, such as cancer, the identification of survival-associated cell processes - rather than single genes - provides more informative results because the efficacy of survival prediction increases when multiple prognostic features are combined to enlarge the possibility of having druggable targets. Moreover, genome-wide screening in molecular medicine has rapidly grown, providing not only gene expression but also multi-omic measurements such as DNA mutations, methylation, expression, and copy number data. In cancer, virtually all these aberrations can contribute in synergy to pathological processes, and their measurements can improve a patient's outcome and help in diagnosis and treatment decisions. Here, we present MOSClip, an R package implementing a new topological pathway analysis tool able to integrate multi-omic data and look for survival-associated gene modules. MOSClip tests the survival association of dimensionality-reduced multi-omic data using multivariate models, providing graphical devices for management, browsing and interpretation of results. Using simulated data we evaluated MOSClip performance in terms of false positives and false negatives in different settings, while the TCGA ovarian cancer dataset is used as a case study to highlight MOSClip's potential.

INTRODUCTION

Cancer is a disease of the genome. The genome defects impact the transcriptome and the methylome and their combined effects are observed in tumor cell processes (1).

In the last decade, next-generation sequencing technologies have boosted cancer genomic studies, complemented histology-based classification, improved the definition of clinical outcomes and suggested tailored treatments (2).

The multi-omics dimensions of cells, in healthy as well as pathological conditions, interact to complement each other. Thus, improving our ability to study multi-omic signals is essential to understand biological processes. However, while the statistical analyses of a single-omic dataset is straightforward, the integration of multi-omic data is still challenging (3).

The approaches to identify survival-associated markers currently used in medical protocols (4) go through univariate or multivariate survival models one molecular variable at a time, and do so separately for each omic dataset. However, this approach has limits, including missing interactions among genes and among different layers of gene deregulation, and missing context-specific cellular mechanisms involved.

Large-scale cancer genomics projects, such as The Cancer Genome Atlas, have generated terabytes of matched omic data on hundreds—and sometimes thousands—of patients, shifting the main challenge from data collection to data analysis. In this scenario, multi-omic data integration is emerging as a promising approach to prioritize findings and generate a more comprehensive view of the mechanism disrupting cellular functions (3,5,6).

In recent years many efforts have been dedicated to multi-omic data integration (7–12), see (13) and (14) for comprehensive reviews. Different strategies have been proposed, from machine learning algorithms to correlation and penalized linear models. However, these are focused mainly for two-class comparison and for identification of cancer subtypes (not necessarily different prognoses) (15–21). A

*To whom correspondence should be addressed. Tel: +39 0498277401; Email: chiara.romualdi@unipd.it
Correspondence may also be addressed to Paolo Martini. Tel: +39 0498276319; Email: paolo.martini@unipd.it
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

few approaches have been developed for survival pathway analysis (8,22–25), but still they do not combine multi-omic datasets (8,24,25) or do not consider interactions among genes (22,23).

In an attempt to fill in these gaps, here we present MOSClip: a multi-omic statistical approach based on pathway topology. MOSClip (Multi-Omics Survival Clip) exploits the topology of pathway annotations and integrates multi-omics data to identify pathways or pathway modules associated with right-censored survival data.

MOSClip is implemented as an R package. It is highly flexible, accepting from one to many omics datasets, also allowing the use and combination of different data reduction strategies. It furthermore contains several tools to graphically summarize the results and help the user during data interpretation. These features make MOSClip a useful and versatile tool for omics integration analyses.

MATERIALS AND METHODS

The multi-omic model and tests

MOSClip analysis is summarized in Figure 1. Given the structure of a pathway converted into a graph structure, MOSClip can perform survival multi-omic tests either on a pathway (Figure 1A), or at the module level (connected components of the pathway graph, Figure 1B). The pathway and the module tests can be independently performed.

Briefly, MOSClip applies omic-specific dimensionality reduction techniques on each pathway or module, and then uses the multivariate survival model to identify those associated with survival. This approach is highly flexible, allowing the use of several omic datasets together, even those containing different data distributions.

In the next paragraph, a brief description of MOSClip functionalities is reported for the case of four omics: expression, methylation, copy number variation and mutation. The generalization of these functions and adaptation for fewer or more datasets is straightforward. Let $X^T_{v\times p_e}$, $M^T_{v\times p_m}$, $C^T_{v\times p_m}$ and $U^T_{v\times p_u}$ be respectively the gene expression, methylation, copy number variation and mutational matrices of the genes belonging to graph G across v samples. Expression data is expected to be normalized and log transformed, methylation data should be a β value matrix, while mutational or CNV data should be binary matries (presence/absence of mutation/CNV or GISTIC thresholded data for CNV). It is not required to have all the omics for all genes, but patient matching across the different omics is required.

Dimensionality reduction strategies. To reduce the dimensions of numerical matrices such as X or M (gene expression or methylation) we propose principal component (PC) analysis using the information on pathway topology as represented by the graphical model described below or by cluster analysis.

Principal component analysis. Let $G = (P, E)$ be a directed acyclic graph (DAG) with P nodes (genes) and E edges representative of a specific pathway topology and for instance $X_{p\times v}$ the expression matrix of the P genes belonging to G across v samples (columns). Then we can model the data with a graphical model as follows:

$$M(G) = \{X \sim N_p(\mu, \Sigma),\ K = \Sigma^{-1} \in S^{+}(G)\}, \qquad (1)$$

where N is a Normal distribution, p is the number of genes (nodes of the graph), K is the concentration matrix (inverse of the covariance matrix) of the model and $S^{+}(G)$ is the set of symmetric positive definite matrices with null elements corresponding to the missing edges of G. Without loss of generality, we assume $\mu = 0$. In this model, the maximum likelihood estimate of Σ (hereafter $\hat{\Sigma}_{IPS}$) can be obtained by using the Iterative Proportional Scaling (IPS) algorithm with the sample covariance matrix ($\hat{\Sigma} = X'X/(n-1)$) as starting value. The IPS guarantees that $\hat{\Sigma}^{-1}_{IPS} \in S^{+}(G)$.

Classic principal component analysis is based on the spectral decomposition of the sample covariance matrix:

$$\hat{\Sigma} = VLV^{T}, \qquad (2)$$

where L is a diagonal matrix with eigenvalues arranged in decreasing order and V is the matrix of corresponding eigenvectors. The eigenvectors are the principal directions of the data. The jth principal component is given by j-th column of XV.

Here, we propose using the spectral decomposition of $\hat{\Sigma}_{IPS}$, instead of $\hat{\Sigma}$, to calculate the PCs. Given that IPS takes sample covariance matrices as starting values, in case of small sample sizes with respect to the number of variables, we will use a shrinkage approach to estimate $\hat{\Sigma}$. The number k of PCs to be selected for the final model is estimated using a cross-validation approach (26). When the dimension reduction is performed on modules of graph G (fully connected component), IPS is not needed and a sparse PCA (R package *elasticnet*) is implemented (see Prognostic module identification paragraph for more details).

Cluster analysis. Hierarchical cluster analysis is applied on X or M matrices. *NbClust* R package (27) was used to identify the optimal number of clusters. *NbClust* computes 30 different validity measures along with a consensus estimation in the case that all the measures are selected. By default, MOSClip uses the Silhouette index, but any other validity measure can be selected. On the basis of the optimal number of clusters identified, patients are classified into groups. Then, the numerical matrix is summarized with a vector reporting the cluster in which the patients are assigned. The same cluster analysis strategy is used for both pathway and module analyses.

To reduce the dimensions of binary matrices such as C or U (copy number variation or mutations) we propose a binary summary or a vote counting strategy. Differently from mutational events with CNV data, amplifications and deletions are considered as separate events and thus as two different covariates in the model. We marked a CNV event when a severe amplification/deletion was found (following GISTIC thresholded data a severe amplification is 2 and a severe deletion is –2).

Binary summary. We summarize the binary matrix with a sample binary vector having 1 if at least one gene in the

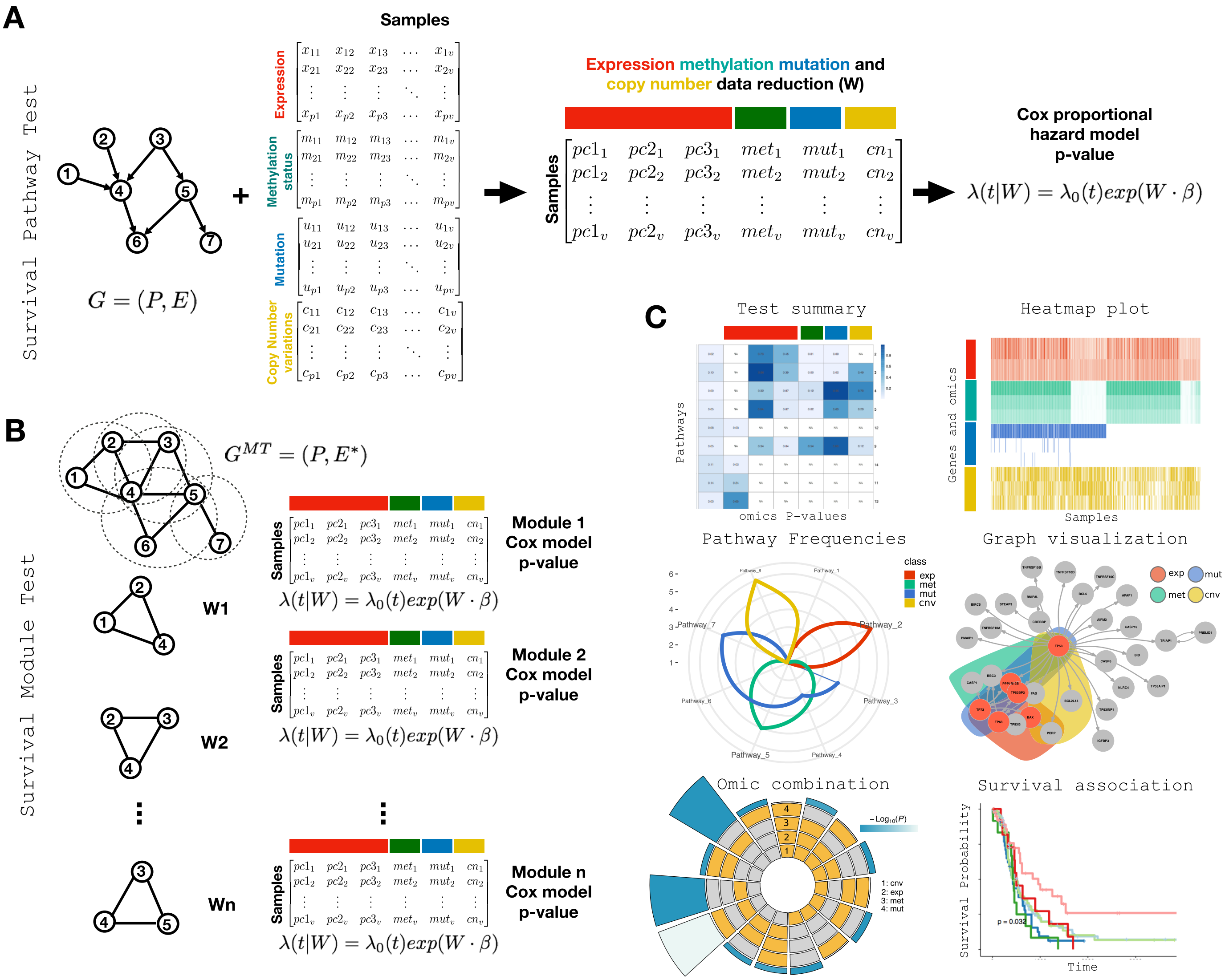

Figure 1. A schematic overview of the computational strategy of MOSClip. (**A**) Survival Pathway Test. Given the graph topology *G* and gene expression, methylation, copy number and mutational matrices, dimension reduction is applied to generate matrix *W*. *W* is composed by reduced omic vectors with patient classes obtained from PCA, hierarchical clustering and binary/vote counting as described in the Methods. Then a multivariate Cox proportional hazard model is applied using the *W* matrix as covariates. The full model *P-value* is returned. (**B**) Survival Module Test. After graph moralization and if necessary triangularization, modules (maximal cliques) are identified. For each module the same analysis as reported in panel (A) is applied. (**C**) Panoramic view of MOSClip graphical tools. MOSClip package provides: a summary to rank modules/pathways considering their *P*-values; an heatmap with top genes for each omic along with sample annotations; a radial plot for pathway frequencies by omic; a graph visualization to highlight the modules in the whole pathway; a summary of omic combination as frequency distributions by pathways/modules and by omics; Kaplan–Meyer curves and log-rank test of sample groups classified according to omic combination.

pathway/module is mutated, amplified or deleted and 0 otherwise. The same strategy is used for both pathway and module analyses.

Vote counting summary. We summarize the binary matrix with a numeric vector with values from 0 to the number of altered genes within the pathway/module. The same strategy is used for both pathway and module analyses.

Pathway test. The multivariate Cox proportional hazard model is used to perform survival analysis. In the model, the omic reduced vectors are included as covariates, while survival measures (e.g. overall and/or progression-free survival) as response variables. The *P*-value of the likelihood ratio test of the model and of the coefficient *P*-values of the models are returned to give insights into the association between each covariate and the survival measure. In addition to multi-omic data the survival test implemented in MOSClip can support non-omic covariates, such as clinical variables or known confounders. Moreover, the package optionally supports a robust proportional hazard model using a smooth modification of the partial likelihood as implemented in coxrobust R package.

Module test. Using graphical theory after appropriate transformations (moralization and if necessary triangularization), G can be decomposed into small, overlapping, connected components (maximal cliques in graph theory), hereafter called modules. Modules are identified using the *maxClique* function implemented in *RBGL* R package. Then, for each module, MOSClip applies the same strategy as for the whole pathway test. The dimensionality reduction strategies using clusters and binary/vote counting summaries are the same as described above, while the estimation of the PCs is based on penalized regression (sparse PCA). As modules are completely connected, IPS is not necessary. If the sample size is small compared to the dimensions of the module, shrinkage covariance estimation is used in place of sample covariance. The use of penalized PCA is meant to enhance the signal of the most important genes, as a penalization is used in the PC regression model estimation step. Finally, for each module the multivariate Cox proportional hazard model is applied. As well as in pathway test, in addition to multi-omic data the survival module test can support clinical variables and known confounders and a robust proportional hazard model can be optionally used.

Gene prioritization. Given a prognostic module or pathway, MOSClip implements different strategies to identify the genes most associated with survival. To this end, we include a prioritization system specific for each method of data reduction: the absolute value of gene loadings is used in PCA, then the Kruskall–Wallis test is used to compare measurements across patient groups for cluster analyses, and the three genes with the highest number of events are reported for binary data.

Resampling strategy. Genes in pathways, as well as in modules, are highly redundant. Thus *P-values* obtained are not independent, violating the assumption of FDR-based methods for *P*-value correction. Thus, to control false positives and to improve the robustness of pathway/module selection, we implemented a re-sampling strategy. From the original cohort, we created 100 sub-cohorts of patients by randomly removing 1% of the patients. We ran MOSClip on all these cohorts to identify a list of significant pathways and modules. Finally, these results are checked to see how many times pathways/modules are significant. In the analysis of TCGA data, we chose the re-sampling success threshold of 80% in both the pathway test and the module test (pathways or modules significant in at least 80 re-sampled cohorts).

Visualization and graphical summaries. MOSClip provides several graphical tools to browse, manage and help interpretation of results (Figure 1 C). A brief description is reported below:

- Test summary. Pathways/modules heat-map of *P*-values. The *P*-value of the Cox model along with those of the omic coefficients (PCs, methylation, CNV and mutation variables) are reported for each pathway and for modules within a pathway. These plots visualize pathway or module rank and are useful to evaluate omic contributions.
- Heatmap plot. In this plot, we performed sample clustering with prioritized genes. Prioritized genes for each omic are reported along with clinical annotations. The heatmap reports the prioritized gene measurements in each omic across different patients.
- Pathway frequencies. This radial plot frequency distribution shows the frequency distribution of pathways aggregated into macro-categories, using the Reactome or KEGG hierarchical structure separately for each omic combinations. This plot suggests prognostic biological processes that may be impacted by the omics and their cross-talk.
- Network visualization. The pathway network is reported with the module genes highlighted in red, along with the impact of the different omics as colored areas.
- Omic combination. This feature implements a Super Exact test and multi-set multi-omics visualization to provide efficient computation of the statistical distributions of multi-omic pathway/module set intersections, for this the theoretical framework implemented in *SuperExactTest* R package was used (28). A circle plot is returned with the frequency of all significant omic combinations and their significance levels.
- Survival Annotations. Kaplan–Mayer curves and log-rank tests are used to stratify patients according to the combination of pathway/module omic variables.

Simulations

To assess MOSClip performance in terms of the rate of false positives and false negatives, we used simulated data under H_0 (no pathway/module association with survival) and under H_1. Under H_1 data was simulated so that a selection of gene measurements were associated with the prognosis, while under H_0 data was simulated so that no genes were associated with survival.

Simulations were planned in order to test a single or combination of omics within a pathway and a module. Specifically, given a graph structure G obtained from the *graphite* package (29,30), expression, methylation, mutation and CNV have been simulated 1000 times to obtain the following scenarios (Figure 2 A): (i) one module (module 27) with one differentially methylated gene (*IGF1R*), (ii) one module (module 21) with one differentially expressed gene *JAK1*, (iii) two modules (modules 12 and 23) with two mutated genes (*PDGFB*, *PDGFA*, and *ERBB2*, *EGFR*), (iv) one module (module 7) with three significantly deregulated genes, one with altered methylation (*IGF1R*) and two mutated genes (*ERBB2*, *EGFR*) and (v) one module (module 8) with four deregulated genes with respective methylation (*IGF1R*), expression (*JAK1*) and mutation alterations (*ERBB2*, *EGFR1*).

Right-censored survival data (status, follow-up and right-censoring) was simulated using the *survsim* package (31) (see Supporting material). Distributions used to simulate omic datasets were selected according to the nature of the data: Gaussian for log-expression, uniform (between 0 and 1) for methylation β values and Bernullian for mutation/CNV data; details are reported briefly below. For both the whole pathway and the module tests we used n = 300.

Expression alteration. Gene expression was generated using *simPATHy* (32). We used the graph structure reported in Figure 2 A. The graph contains 79 genes, with 299 edges distributed in 34 modules. Under H_0 a single matrix with

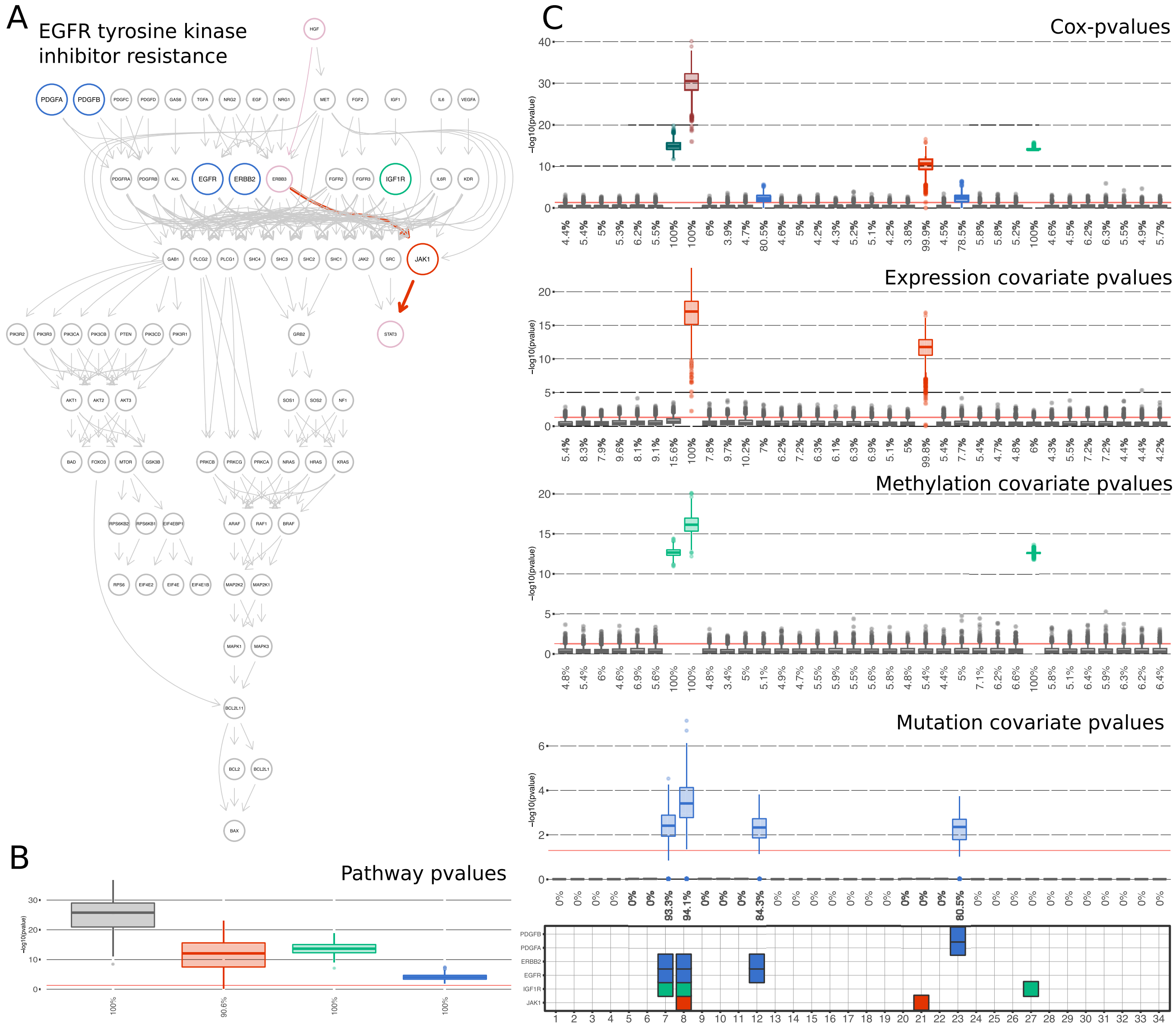

Figure 2. Simulation results under H_1. (**A**) Pathway graph Pathway graph used for simulations. (**B**) Pathway results Distributions of the log-transformed likelihood ratio test (LRT) *P-values* of the full-model (gray) and of the model coefficients (red for expression, green for methylation and blue for mutation) over 1000 simulated datasets under H_1. Red line represents the 0.05 threshold, while the number on the *x*-axis are the true positives rate. (**C**) Module results Distributions, across all pathway modules, of the log-transformed LRT *P-values* of the full-model (sub-panel 'Cox-pvalues') and of the model coefficients (red for expression, green for methylation and blue for mutation) over 1000 simulated datasets under H_1. Red line represents the 0.05 threshold, while the number on the *x*-axis are the true positives rate. The grid in the lower part of the panel highlights the module numbers, along with the presence and the type of alterations. Modules without colored boxes are composed of genes that were not altered in the simulations.

$n_1 + n_2$ rows and 79 columns was simulated with default settings. Under H_1 a first matrix with n_1 rows and 79 columns was simulated with default settings and a second matrix with n_2 rows and 79 columns was simulated to have the path from *HGF*, to *STAT3* (in pink in Figure 2A) perturbed. The node *JAK1* (in red in Figure 2A) was selected to have the maximum significant difference (mean > 2) between classes, i.e. the expected true positive.

Methylation alteration. Denote with $M_j^{(k)}$ the random variable representing the methylation alteration for gene j, $j = 1, \ldots, 79$, in condition k, $k = 1, 2$. We assume that $M_j^{(k)}$ is a uniform random variable centred at $\theta_j^{(k)}$ with range 2ϵ, with $\theta_j^{(k)} \in (\epsilon, 1 - \epsilon)$ and $\epsilon \in (0, 1/2)$. The hypothesis of absence of dysregulation corresponds to the hypothesis H_0 : $\theta_j^{(1)} = \theta_j^{(2)} = \theta_j$. Under H_0, $M_j^{(k)} \sim U(\theta_j - \epsilon; \theta_j + \epsilon)$, $k = 1, 2$. In the simulations, we set $\epsilon = 0.1$ For each non dysregulated gene j, θ_j is fixed to a number randomly chosen in the interval (0.1,0.9). For each dysregulated gene j, we set $\theta_j^{(1)} = 0.8$ and $\theta_j^{(2)} = 0.4$.

Mutation/CNV alteration. Mutations and CNV events are simulated in the same way. Under H_0, genes are simulated to have a random alteration rate of 10^{-3} across all n_1 + n_2 patients. Under H_1, the set of patients with poor prognosis defined by the *survsim* package were stratified to have 5% of patients with all selected genes altered (solid blue in Figure 2A), 10% with three out of four, 20% with two altered, and the remaining samples with only one altered gene. For the remaining patients, the genes are simulated to have a random alteration rate of 10^{-3}.

Once omic datasets were simulated, clinical data were matched accordingly.

TCGA ovarian cancer data and pathway preprocessing

TCGA multi-omic ovarian cancer (OVC) data (expression, methylation, CNV and mutation) were used as a case study. Datasets were downloaded using *TCGAbiolinks* R package (version 2.9.2), and all the data referred to human genome version 38. Clinical annotations were downloaded and used to select only primary tumor samples. All data were processed following TCGA guidelines as described below.

Copy number variation. We downloaded CNV matrix using *getGistic* function with type = 'thresholded' (TCGABiolinks R package). The matrix is in the form 'gene per patients'. It contains integer values from –2 to 2: positive and negative numbers respectively indicate amplification and deletion; ±2 represents severe amplification/deletion, ±1 mild amplification/deletion and 0 means no defects detected.

Somatic mutations. We downloaded the somatic mutations called by *mutect2* pipeline. Following the mutation impact defined by VEP software (33), all the variants with high and moderate impact were included in the analysis. We excluded those with low impact (assumed to be mostly harmless or unlikely to change protein behavior), and those defined as modifier (non-coding variants or variants affecting only non-coding genes, or predictions without evidence of impact) (for more details, see https://docs.gdc.cancer.gov/Data/). Note that summarizing at the gene level, multiple mutations on the same gene are counted as one. The data was transformed into a boolean sparse matrix of genes per patients, representing the presence/absence of mutations.

Gene (promoter) methylation. Methylation profiles from the Illumina Human Methylation 27K and 450K platforms were downloaded. We summarized CpG islands into clusters using the *methylMix* R package (version 2.10.0 used with default parameters). A gene may have more than one cluster of CpG. Probes with no detectable values in more than 60% of patients were excluded. Within a single cell, any cluster of CpGs can be methylated or unmethylated. However, when measuring methylation for a population of cells, it has to be interpreted as the proportion of cells showing the site methylated. Different sites are then collapsed and associated to a gene promoter as β values, which are defined as the percentage of methylated sites of the promoter of a gene (34). After filtering of our data, missing values were imputed using *knn* (knn R package, $k = 5$).

Expression profiles. We used gene expression quantification obtained via NGS technology. We then filtered out those genes with <100 raw counts in at least one patient, normalized data with upper-quantile (*EDAseq* R package, version 2.14.0) and used the $\log_2$-transformed pseudocounts for the analyses.

Survival data. Progression Free Survival (PFS) was defined according to (35). PFS times were defined as the days from surgery to tumor relapse or tumor death, or alternatively to the last contact date.

Patient cohort. All patients with survival annotations and profiled for expression, methylation, CNV and mutation were included in the analysis. Analysis was conducted with a total of 266 patients.

Pathways. Reactome biological pathways were downloaded using the *graphite* R package (version 1.26.1). We filtered the pathway genes to include those with an expression profile. For pathway analysis, we further filtered pathways to include those with >10 genes. For module analysis, pathways were filtered to only include pathways containing 20–100 genes. These cutoffs were empirically selected to remove reaction redundancies and cut the Reactome pathway hierarchy. After filtering, we ended up with 1283 pathways of which 728 where used for module analyses. The Reactome hierarchy was downloaded from the Reactome website.

RESULTS AND DISCUSSION

MOSClip is a new tool to perform survival pathway analyses on multi-omic datasets. Both pathways and their graphical model decompositions (modules) are tested to find biological processes impacting the patient's survival. In the MOSClip model, multi-omic gene measurements are tested as covariates of a Cox proportional hazard model after a dimensionality reduction step. The significance of the modules could be driven by a single or by a combination of prognostic elements, and graphical tools can help the user interpret the results accordingly.MOSClip has a modular structure, allowing the use of one or multiple omics with different data distributions, and the choice of different reduction strategies. Furthermore, specific graphical tools have been implemented to browse, manage and provide help in result interpretation.

MOSClip implementation

MOSClip is available as an R package on GitHub (cavei.github.io/MOSClip/). Thanks to compatibility with the *Bioconductor graphite* package, MOSClip allows survival analysis using many of the major pathway databases such as KEGG (36) and Reactome (37). To enhance and simplify its usage, MOSClip is distributed along with tutorials on how to (i) download TCGA data using the *TCGAbiolinks* package, (ii) pre-process data for the analysis, (iii) perform analyses both at the pathway and module level and (iv) result visualizations. Finally, we also provide an advanced tutorial for the creation of multi-omic networks that can be visualized with Cytoscape (38). These tutorials guide users in the reproduction of all data and plots contained in this publication.

MOSClip on simulated data

Simulated data were used to (i) compare methods of dimension reduction for expression, methylation and binary data and (ii) evaluate MOSClip statistical power and control of type I errors.

We tested PC and cluster analysis summaries on both expression and methylation data. Results show that at the level of the pathway, PCA and cluster analysis are the best choices respectively for expression (Supplementary Figure S1A, power of 90%) and methylation data (Supplementary Figure S2A, power 100%), while at the level of modules both methods show excellent performance (Supplementary Figures S1B and S2B).

Then we tested binary and vote counting for mutation/CNV data. Results show that, although at the module level both methods reach a power grater than 90% (Supplementary Figure S3B), at the pathway level the vote counting method is more powerful (Supplementary Figure S3A; 100% for vote counting and 88% for binary). As a general conclusion, although the user might select the dimension reduction strategy that best fits the biological question of the study, we suggest the use of PCA for expression data, clustering for methylation data and vote counting for mutational/CNV data.

Regarding MOSClip performance in case of multiple deregulated omics within the same pathway/module, our simulations show a good control of the type I error under the null hypothesis both at the level of pathway (Supplementary Figure S4) and at the level of modules (Supplementary Figure S5). In particular mutation and methylation data show an error rate very close to the nominal value, while expression data seems to have a slightly higher rate of false positives. On the other hand, under the alternative hypothesis MOSClip shows an excellent power (greater than 90%) at the level of the pathway (Figure 2B) and modules (Figure 2C).

MOSClip and TCGA data analyses

We used MOSClip to find progression-free survival (PFS) associated pathways and modules in the TCGA multi-omic dataset of ovarian cancer (OVC). The analyses have the main purpose of demonstrating MOSClip feasibility and usefulness. The OVC results have been discussed in light of published literature which is considered a benchmark in the field. We summarized the state-of-the-art knowledge on expression, CNV and methylation alterations and mutations in OVC in Supplementary Table S1.

Specifically, OVC is characterized by ubiquitous *TP53* mutations (39), alterations of *PI3K/AKT* signaling, loss of *E-cadherin* expression, mutations or epigenetic loss of *RB1*, *NF1* and *PTEN* (40,41), and deregulation of *TGF-β/SMAD* signaling resulting in the promotion of epithelial to mesenchymal transition (42,43). More aggressive OVC phenotypes show matrix metalloproteinase (*MMP*) signal deregulations influencing cellular migration and tissue invasion (44–50). OVC tumors of long term survivors had increased somatic mutations and frequent *BRCA1/2* biallelic inactivation through mutation and loss of heterozygosity (51). Characteristics of short term survivors included focal copy number gain of *CCNE1*, lack of a *BRCA* mutation signature, and low homologous recombination deficiency scores (51). Copy number alterations dominate the landscape of OVC genomes (41); specific copy number changes have been found in *CCNE1*, *MECOM*, *MYC* (CNV gain), *PTEN* and *RB1* (CNV loss) (52). Copy number alterations deregulate cellular senescence pathways, cell cycle, interleukin, *PI3K/AKT*, *RAS* and *WNT* signalings and Toll Like Receptor cascades (53).

Pathway results. MOSClip identified 33 significant pathways out of 1283 tested (2.6%, P-value $\leq$ 0.05 and re-sampling successes $\geq$ 80%, Figure 3 A, Supplementary Table S2). Consistent with OVC literature, many pathways involved OVC known oncogenes and processes, such as *PI3K/AKT* signaling, *PTEN* and *TP53* regulation and Toll-like receptor cascades. Considering the contribution of each omic on these pathways, we found that 23 are lead by gene expression, 6 by changes in methylation, 16 by cumulative gene mutations and 7 by CNVs. Moreover, 21 pathways show PFS association guided by the combination of at least two (Figure 3B, Supplementary Table S3) or three (CNVs, expression and mutation: ‘Nucleobase catabolism’ and ‘Purine catabolism’) different omics. Other than being building blocks for DNA and RNA, purine metabolites provide fuel for cell survival and proliferation; this result tantalizingly links OVC to therapeutic strategies to reprogram cancer adenosine metabolism (54,55) and re-activate the anti-cancer immune response (56,57). The combination of expression and mutation (E–M), and expression and CNVs (E–C) are over-represented (P-values $\leq 1 \times 10^{-7}$, Figure 3B) if compared to all the other omics combinations (Supplementary Table S3), suggesting causal relationships. Specifically, pathways guided by the combination of E–M and E–C are involved in signaling and disease pathways and cell-cell communication, respectively (Figure 3 C). Pathway guided by expression alone are involved in immune system and development, those guided by methylation are involved in developmental processes, while those guided by mutation are mainly disease pathways (Figure 3C and Supplementary Table S4).

Module results. MOSClip tested 4931 modules as part of 728 pathways, finding 213 modules significantly associated with PFS (4.3%, P-value $\leq$ 0.05 and re-sampling successes $\geq$ 80%, Supplementary Table S5). Gene expression guides survival association in 110 modules, methylation in 70, mutation in 16 and CNV in 88 modules (Figure 4A). Gene expression leads metabolism, signaling pathways and extracellular matrix organization, while methylation changes are mainly involved in metabolism of proteins and developmental modules, and CNVs mainly hit immune system modules (Figure 4B, Supplementary Table S6). The significant co-occurrence of different omics is observed for (i) expression and mutation (9 modules, P-values $\leq 1 \times 10^{-7}$) in modules associated with metabolism, (ii) expression and methylation (18 modules P-values $\leq 1 \times 10^{-7}$) in modules associated with metabolism of proteins and the immune system and (iii) expression and CNVs (34 modules

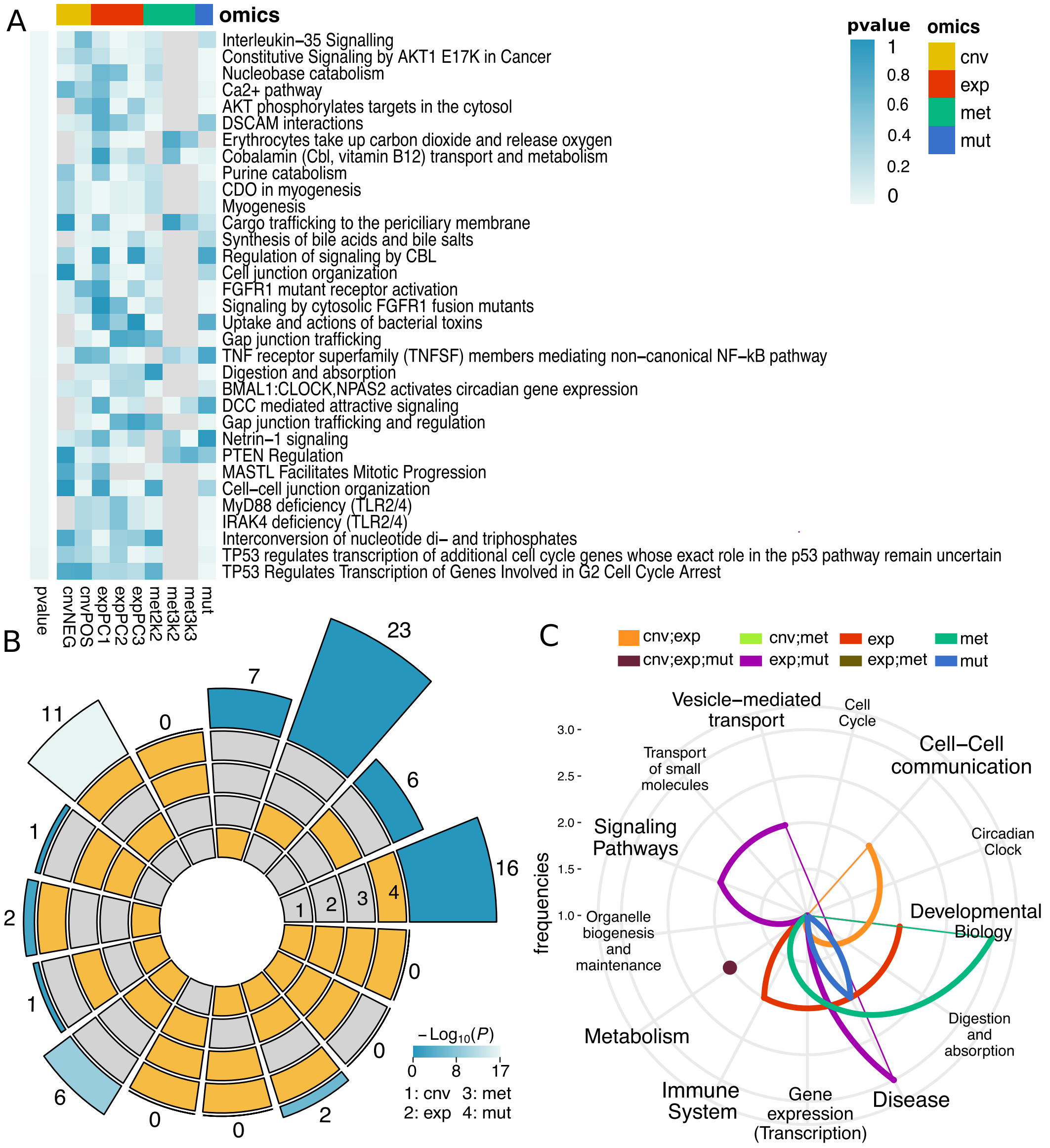

Figure 3. Pathway Visual Inspection. (**A**) Summary of the significant pathways (*P-value* < 0.05 and resampling > 80%). The heatmap has pathways on the rows ordered by full model Likelihood ratio test *P-values* (white=0; blue=1) and omics on the columns (yellow = CNV; red = expression; green = methylation; blue = mutation). The full model LRT test along with coefficient specific *P-values* are reported within each cell with color degree. (**B**) Circle plot representing the frequency of significant pathways whose survival association is guided by a single or a combination of omic data. The four innermost layers represent the combination of omics (the yellow sector means 'on'), while the external layer represents the number of pathways. *P-values* of the overlap between omic-specific sets of pathway were calculated using the *SuperExactTest* R package. (**C**) Radial plot showing the frequency distribution of significant pathways according to each omic. The categories are obtained by mapping the pathways to broader categories given by Reactome hierarchy.

P-values $\leq 1 \times 10^{-7}$) in modules focused on metabolism. Only four survival modules were found to be guided by the contribution of three omics (CNVs, methylation and expression) (Figure 4A–C, Supplementary Tables S6 and S7). Consistent with OVC literature, many PFS associated modules involved OVC known oncogenes and processes, such as *TP53*, *PI3K/AKT* pathway and *MMPs* (CNVs, expression, methylation and mutation), *TGF-β* (expression and mutation), *SMADs* (expression), *WNT* (CNVs and expression).

To provide examples of the MOSClip visualization tools, we focused on 'Activation of Matrix Metalloproteinases'. This pathway describes the turnover of extra-cellular matrix components by metalloproteinases. It is associated with PFS via expression (PC1 *P*-value = 0.02) and methylation (*P*-value = 0.01). The majority of metalloproteinase substrates are cytokines, growth factor binding proteins and receptors; deregulated expression of metalloproteinases and their epigenetic control play a role in tumor cell invasion and metastasis (58). Among the most significant modules in the 'Activation of Matrix Metalloproteinases' pathway we found the number 2 (Figure 4D), in which modulation of expression (PC1) and methylation well predicts patients' PFS.

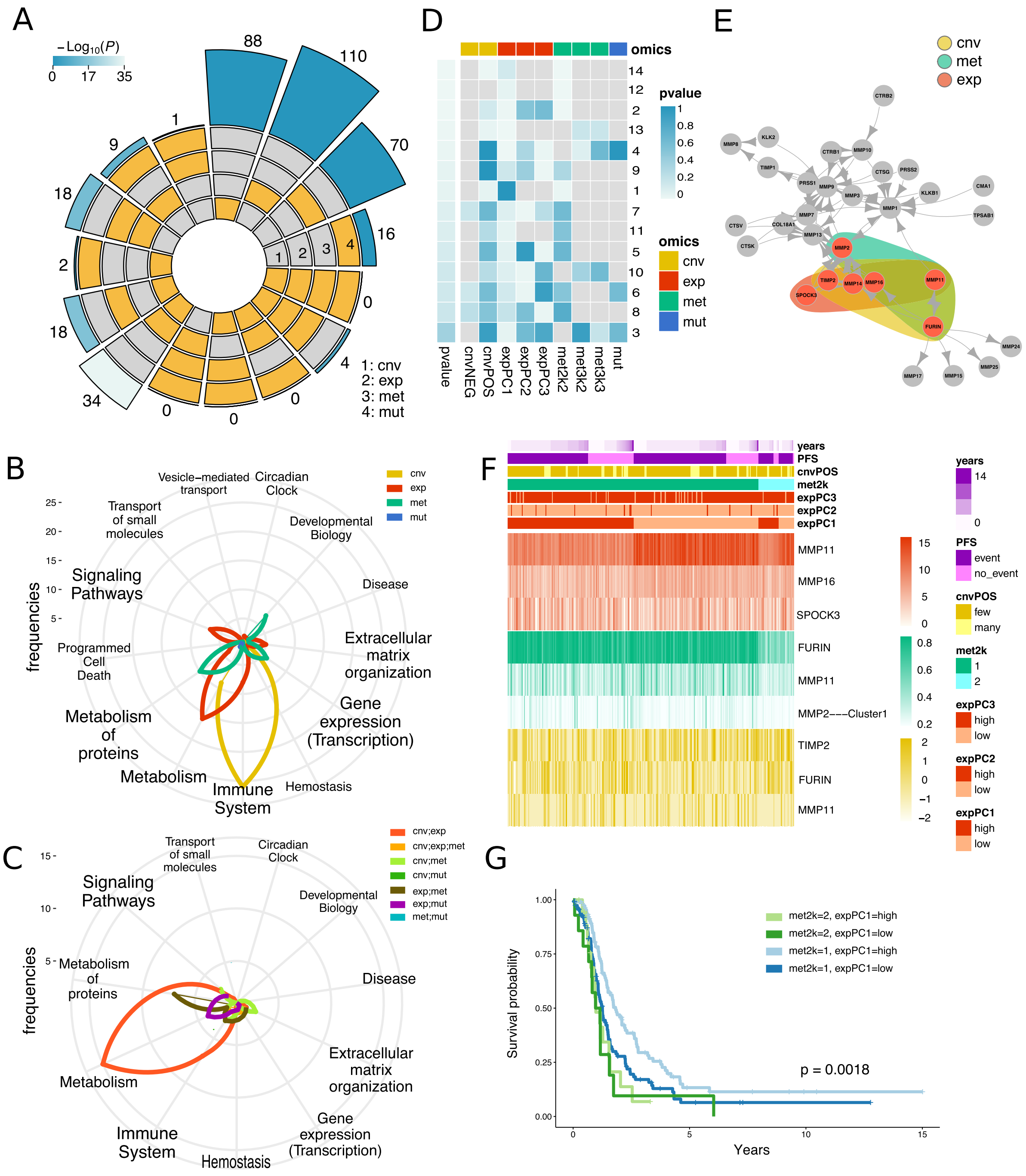

Figure 4. Module results. (**A**) Circle plot representing the frequency of modules with a single or a combination of significant omic variables. The four innermost layers represent the combination of omics (the yellow sector means 'on'), while the external layer represents the frequency of the combination. (**B**) Radial plot showing the frequency distribution of modules with a single significant omic. The categories were obtained by mapping the pathways to broader categories given by Reactome hierarchy. (**C**) Radial plot showing the frequency distribution of modules multiple significant omics. The categories were obtained as described in panel B. (**D**) 'Activation of Matrix Metalloproteinases' module summary. The heatmap has modules on the rows ordered by full model LRT *P-values* (white = 0; blue =1) and omics on the columns (yellow= CNV; red = expression; green = methylation; blue = mutation). The full model LRT test along with coefficient specific *P*-values are reported within each cell with color degree (white = 0 to blue = 1). (**E**) Network of 'Activation of Matrix Metalloproteinases' pathway. Genes belonging to module 2 are coloured in red while the omic impact is highlighted with colored areas (red = expression; green = methylation, yellow=CNV). (**F**) Heatmap of module 2 of 'Activation of Matrix Metalloproteinases' pathway. The heatmap shows the profiles of prioritized genes for each omic. On top sample annotations are reported. (**G**) Module 2 Kaplan–Meier curves. Patient groups were defined using the combination of expression and methylation classes.

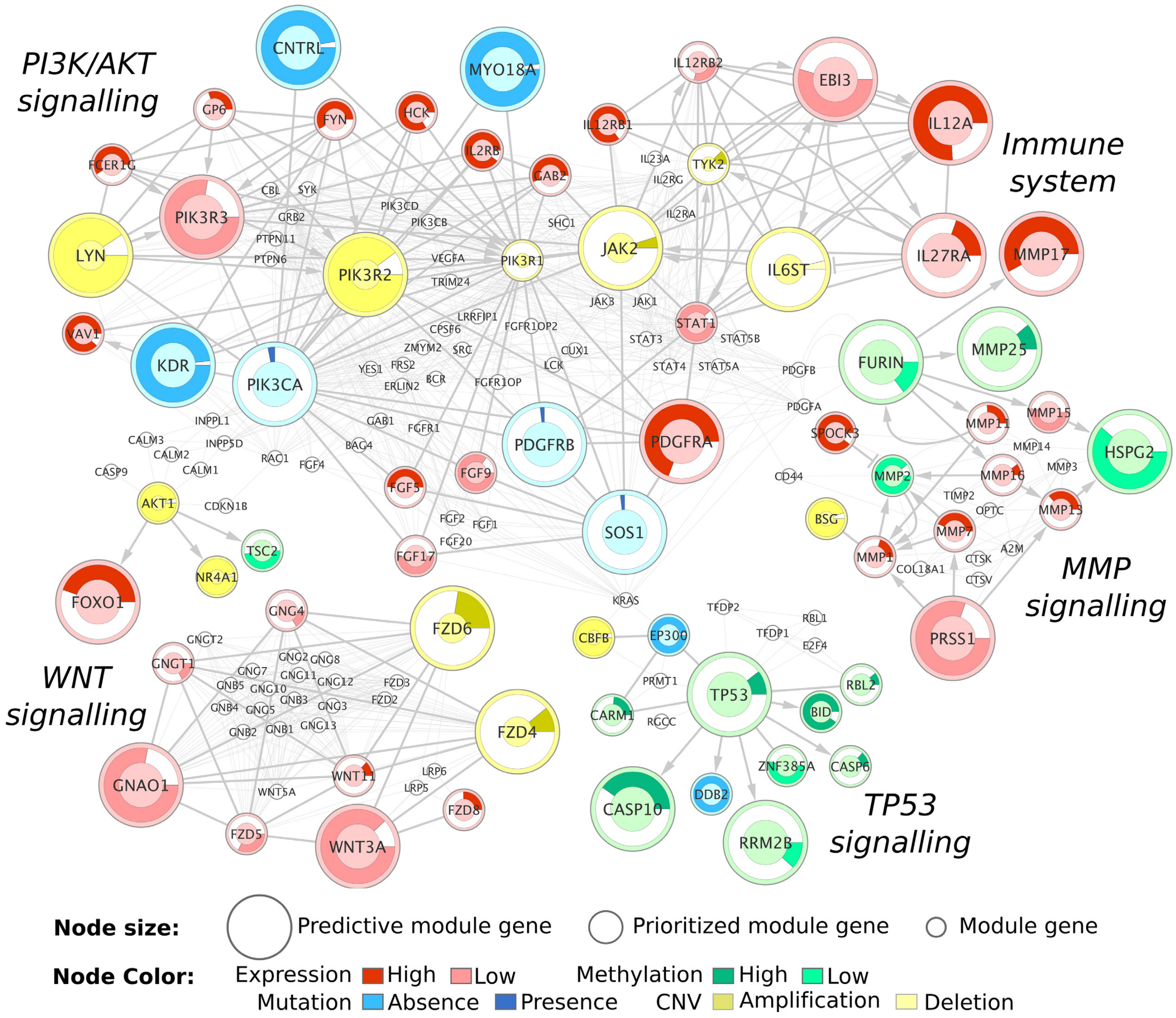

Figure 5. Selection of OVC related modules from the module combination. Node colors represent the omic variable with survival association: red for expression, green for methylation, yellow for CNV, and blue for mutation. The node sizes represent the strength of the survival association. Nodes with the biggest size represent those selected by a penalized Cox model. Within each node, a ring chart delineates the portion of the cohort showing the deregulation (the colored slice). The colored ring slice represents a negative prognosis. Dark and light colors specify the type of aberration associated with the negative prognosis (e.g. high or low expression/methylation; amplification of deletion for CNV; absence or presence for mutation). For example, methylation of *TP53* is strongly associated with survival, the hypermethylation is associated with a negative prognosis, which is in 10% of the cohort.

Figure 4E shows the pathway graph with module 2 (nodes in red) and the influence of the different omics over different genes. *MMP11* and *FURIN* genes are clearly the most representative for expression (PC1) and methylation (Figure 4F) respectively. Combining the expression and methylation variables, we see that patients characterized by high levels of PC1 and low level of methylation (hyper-methylation of *FURIN* promoters and low expression of *MMP11*) have a significantly better prognosis with respect to the other patients (Figure 4G, for the risk table see Supplementary Table S8).

Redundancy reduction through module combination. Pathways are usually characterized by gene and reaction redundancy, i.e. the same genes are often found in different pathways. To remove these redundancies from results, we can combine pathways or module into a unique non-redundant network.

Here, we chose to merge all the significant modules, in order to offer a comprehensive survival network. This network, provided in Supplementary Figure S6, has 681 genes (5252 connections), the genes are colored by the omics with the best survival association, and node size is proportional to the gene prognostic power (73 genes with the biggest

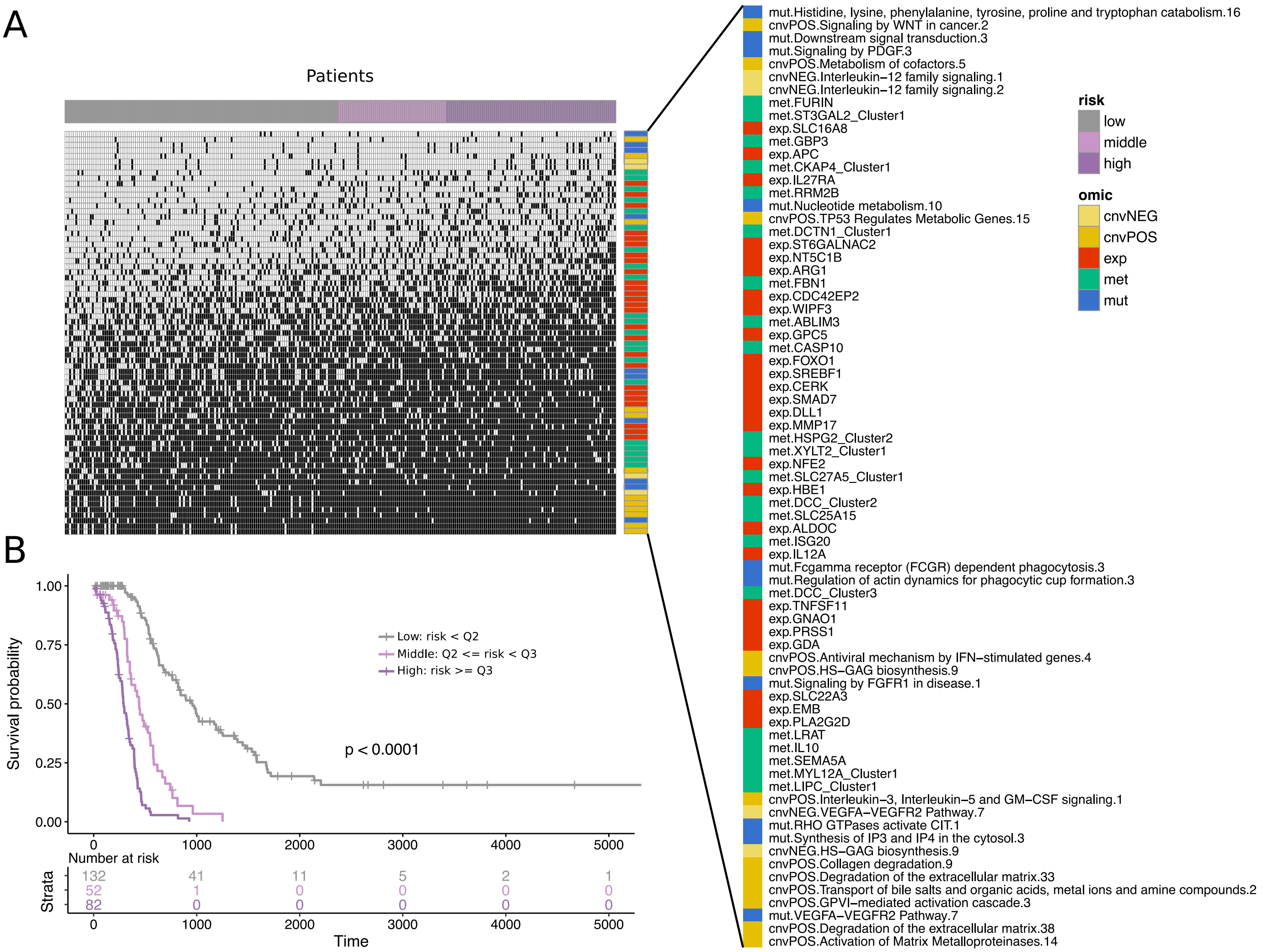

Figure 6. Signature for patient prognosis prediction. (**A**) Heatmap of the number of biomarkers predicting the poor prognosis. Each column represents a patient, while each row is a prognostic multi-omic feature. Black cells represent a prognostic feature in the poor prognosis configuration, and white cells represent the opposite. The feature annotations on the right represent the omic. Patients are ordered according to the number of poor prognostic feature, quantiles are used to cut the population into classes of risk. Mutations and CNVs are grouped by pathway module. For methylation variables, the name of the gene and the CpG cluster number are reported. (**B**) Kaplan–Meyer, log-rank test *P*-value and risk table on the three classes of patients obtained through the stratification of (A).

size resulted as significant with Cox penalized analysis). The network in Figure 5 offers a simplified view of this network with a selection of 25 genes, and their neighbours, known to be involved in OVC prognosis and thus considered as our benchmark results.

The processes highlighted in Figure 5 show the connection across the PI3K/AKT signaling pathway, *TP53* and *WNT* signals, the activation of matrix metalloproteinases (*MMPs*) and immune system regulation. This multi-omic circuit shows only a subset of the MOSClip results, but is useful to appreciate the overview provided by our tool of the survival associated cell circuits.

Possible application of MOSClip results: the identification of predictive survival signatures. A common task in the analysis of omic datasets is the identification of prognostic signatures. Using penalised Cox model on the molecular features of Supplementary Figure S6, we selected 73 prognostic features (26 expression and 21 methylation profiles, 10 modules of mutated genes and 16 modules with CNV alterations, see Supplementary Table S1). Based on this signature, patients were sorted by the number of features predicting a negative prognosis (Figure 6A) and stratified using quantiles of this distribution (low risk: score $< Q2$, middle risk: $Q2 \leq$ score $< Q3$, high risk: score $> Q3$). Figure 6B shows the differences in survival time of patient classes with Kaplan–Meier curves (P-value < 0.0001). However, the identification of a predictive signature is somehow tricky due to the large dimensionality of the data and the presence of many unknown confounding factors (59). We tested the significance of our signature using both random sampling signatures and 189 known signatures from MSigDB oncogenic gene sets C6 proposed by Venet *et al.* (59) using the *sigCheck* Bioconductor package. We found that in both cases MOSClip signature predicted patient prognosis significantly better than random signatures (P-value $\leq$ 0.001, Supplementary Figure S7) and known oncogenic signatures (P-value $\leq$ 0.001, Supplementary Figure S8).

CONCLUSION

In the last 10 years we have witnessed a dramatic change in the clinical treatment of patients thanks to molecular and personalized medicine. As the amount of genome wide data grows, we need to adapt and improve our methods to cope with the higher complexity and multi-level structure of available information, thus integrating multi-omic dimensions. Pathway topology allows the switching from a gene-focused view to a model/pathway view. The advantages of this switch are twofold: firstly, we are able to consider the cooperation among neighbor genes; secondly, we gain statistical power and we better contextualize genes and their functions.

MOSClip can deal with this complexity, allowing multi-omic data integration through survival pathway and module analyses.

The use of modules turned out to be a good choice for multi-omic integration. As methylation and mutation events may have direct effects on the expression of target genes, the use of a network structure allows researchers to model this situation. Overall, a gene by gene approach is often poorly predictive, while the simultaneous analysis of multidimensional data can overcome this issue and increase predictive power. A possible application of the knowledge provided by MOSClip is the suggestion of putative prognostic signature.

MOSClip is freely available as an R package, with tutorials, guidelines and tips for best practice provided to fully exploit the potential of MOSClip and to guide the user in data analysis and interpretation of results. Effort was dedicated to visualization tools, thus giving the user the opportunity to dissect every single aspect of their results.

We have tested MOSClip with the OVC dataset from TCGA using four omics (19430, 16590, 11906 and 24776 genomic features for expression, methylation, CNV and mutation respectively) on 266 patients, identifying a survival associated circuit whose combination allows for survival prognostication.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Prof. Hautaniemi and the members of his lab at the University of Helsinki for their support and useful comments during MOSClip development.

FUNDING

Italian Association for Cancer Research (AIRC) [IG17185 and IG21837 to C.R.]; European Molecular Biology Organization (EMBO) Short-Term Fellowship [7592 to E.C.]. Funding for open access charge: Italian Association for Cancer Research (AIRC) [IG21837 to C.R.].

Conflict of interest statement. None declared.

REFERENCES

1. Vogelstein,B., Papadopoulos,N., Velculescu,V.E., Zhou,S., Diaz,L.A. and Kinzler,K.W. (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.
2. Esplin,E.D., Oei,L. and Snyder,M.P. (2014) Personalized sequencing and the future of medicine: discovery, diagnosis and defeat of disease. *Pharmacogenomics*, **15**, 1771–1790.
3. Werner,H.M., Mills,G.B. and Ram,P.T. (2014) Cancer Systems Biology: a peek into the future of patient care? *Nat. Rev. Clin. Oncol.*, **11**, 167.
4. Mehta,S., Shelling,A., Muthukaruppan,A., Lasham,A., Blenkiron,C., Laking,G. and Print,C. (2010) Predictive and prognostic molecular markers for cancer medicine. *Therap. Adv. Med. Oncol.*, **2**, 125–148.
5. Lightbody,G., Haberland,V., Browne,F., Taggart,L., Zheng,H., Parkes,E. and Blayney,J.K. (2018) Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. *Brief. Bioinformatics*, **1**, 17.
6. Rabbani,B., Nakaoka,H., Akhondzadeh,S., Tekin,M. and Mahdieh,N. (2016) Next generation sequencing: implications in personalized medicine and pharmacogenomics. *Mol. BioSyst.*, **12**, 1818–1830.
7. Rohart,F., Gautier,B., Singh,A. and Lł Cao,K.-A. (2017) mixOmics: an R package for omics feature selection and multiple data integration. *PLOS Comput. Biol.*, **13**, 1–19.
8. Mankoo,P.K., Shen,R., Schultz,N., Levine,D.A. and Sander,C. (2011) Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLOS ONE*, **6**, 1–12.
9. Wang,B., Mezlini,A.M., Demir,F., Fiume,M., Tu,Z., Brudno,M., Haibe-Kains,B. and Goldenberg,A. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333.
10. 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**, i237–i245.
11. Ruffalo,M., Koyutürk,M. and Sharan,R. (2015) Network-Based Integration of Disparate Omic Data To Identify 'Silent Players' in Cancer. *PLoS Comput. Biol.*, **11**, e1004595.
12. Nguyen,T., Tagett,R., Diaz,D. and Draghici,S. (2017) A novel approach for data integration and disease subtyping. *Genome Res.*, **27**, 2025–2039.
13. Huang,S., Chaudhary,K. and Garmire,L.X. (2017) More is better: recent progress in multi-omics data integration methods. *Front. Genet.*, **8**, 84.
14. Rappoport,N. and Shamir,R. (2018) Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.*, **46**, 10546–10562.
15. Rappoport,N. and Shamir,R. (2019) NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, doi:10.1093/bioinformatics/btz058.
16. Singh,A., Shannon,C.P., Gautier,B., Rohart,F., Vacher,M., Tebbutt,S.J. and Lê Cao,K.A. (2019) DIABLO: an integrative approach for identifying key molecular drivers from multi-omic assays. *Bioinformatics*, doi:10.1093/bioinformatics/bty1054.
17. Ramazzotti,D., Lal,A., Wang,B., Batzoglou,S. and Sidow,A. (2018) Multi-omic tumor data reveal diversity of molecular mechanisms underlying survival. *Nat. Commun.*, **9**, 4453.
18. Wang,B., Mezlini,A.M., Demir,F., Fiume,M., Tu,Z., Brudno,M., Haibe-Kains,B. and Goldenberg,A. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333.
19. Le Van,T., van Leeuwen,M., Carolina Fierro,A., De Maeyer,D., Van den Eynden,J., Verbeke,L., De Raedt,L., Marchal,K. and Nijssen,S. (2016) Simultaneous discovery of cancer subtypes and subtype features by molecular data integration. *Bioinformatics*, **32**, i445–i454.
20. Lock,E.F. and Dunson,D.B. (2013) Bayesian consensus clustering. *Bioinformatics*, **29**, 2610–2616.
21. Mo,Q., Wang,S., Seshan,V.E., Olshen,A.B., Schultz,N., Sander,C., Powers,R.S., Ladanyi,M. and Shen,R. (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 4245–4250.

22. Chaudhary,K., Poirion,O.B., Lu,L. and Garmire,L.X. (2018) Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.*, **24**, 1248–1259.
23. Zhu,B., Song,N., Shen,R., Arora,A., Machiela,M.J., Song,L., Landi,M.T., Ghosh,D., Chatterjee,N., Baladandayuthapani,V. *et al.* (2017) Integrating clinical and multiple omics data for prognostic assessment across human cancers. *Scientific Rep.*, **7**, 16954.
24. Zhang,W., Ota,T., Shridhar,V., Chien,J., Wu,B. and Kuang,R. (2013) Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLOS Comput. Biol.*, **9**, 1–16.
25. Chuang,H.-Y., Lee,E., Liu,Y.-T., Lee,D. and Ideker,T. (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.
26. Josse,J. and Husson,F. (2012) Selecting the number of components in principal component analysis using cross-validation approximations. *Comput. Stat. Data Anal.*, **56**, 1869–1879.
27. Charrad,M., Ghazzali,N., Boiteau,V. and Niknafs,A. (2012) NbClust package: finding the relevant number of clusters in a dataset. *UseR! 2012*, doi:10.18637/jss.v061.i06.
28. Wang,M., Zhao,Y. and Zhang,B. (2015) Efficient test and visualization of multi-set intersections. *Scientific Rep.*, **5**, 16923.
29. Sales,G., Calura,E., Cavalieri,D. and Romualdi,C. (2012) graphite - a bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*, **13**, 20.
30. Sales,G., Calura,E. and Romualdi,C. (2018) metaGraphite—a new layer of pathway annotation to get metabolite networks. *Bioinformatics*, **35**, 1258–1260.
31. Morina,D. and Navarro,A. (2014) The R package survsim for the simulation of simple and complex survival data. *J. Stat. Softw.*, **59**, 1–20.
32. Salviato,E., Djordjilovic,V., Chiogna,M. and Romualdi,C. (2017) simPATHy: a new method for simulating data from perturbed biological PATHways. *Bioinformatics*, **33**, 456.
33. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R., Thormann,A., Flicek,P. and Cunningham,F. (2016) The ensembl variant effect predictor. *Genome Biol.*, **17**, 122.
34. Gevaert,O. (2015) MethylMix: an R package for identifying DNA methylation-driven genes. *Bioinformatics*, **31**, 1839–1841.
35. Liu,J., Lichtenberg,T., Hoadley,K.A., Poisson,L.M., Lazar,A.J., Cherniack,A.D., Kovatich,A.J., Benz,C.C., Levine,D.A., Lee,A.V. *et al.* (2018) An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, **173**, 400–416.
36. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
37. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R., Matthews,L. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.
38. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
39. Erickson,B.K., Kinde,I., Dobbin,Z.C., Wang,Y., Martin,J.Y., Alvarez,R.D., Conner,M.G., Huh,W.K., Roden,R.B., Kinzler,K.W. *et al.* (2014) Detection of somatic TP53 mutations in tampons of patients with high-grade serous ovarian cancer. *Obst. Gynecol.*, **124**, 881.
40. Weberpals,J.I., Koti,M. and Squire,J.A. (2011) Targeting genetic and epigenetic alterations in the treatment of serous ovarian cancer. *Cancer Genet.*, **204**, 525–535.
41. Patch,A.-M., Christie,E.L., Etemadmoghadam,D., Garsed,D.W., George,J., Fereday,S., Nones,K., Cowin,P., Alsop,K., Bailey,P.J. *et al.* (2015) Whole–genome characterization of chemoresistant ovarian cancer. *Nature*, **521**, 489.
42. Chou,J.-L., Chen,L.-Y., Lai,H.-C. and Chan,M. W.Y. (2010) TGF-beta: friend or foe? The role of TGF-beta/SMAD signaling in epigenetic silencing of ovarian cancer and its implication in epigenetic therapy. *Expert Opin. Therap. Targets*, **14**, 1213–1223.
43. Marchini,S., Fruscio,R., Clivio,L., Beltrame,L., Porcu,L., Nerini,I.F., Cavalieri,D., Chiorino,G., Cattoretti,G., Mangioni,C. *et al.* (2013) Resistance to platinum-based chemotherapy is associated with epithelial to mesenchymal transition in epithelial ovarian cancer. *Eur. J. Cancer*, **49**, 520–530.
44. Kessenbrock,K., Wang,C.-Y. and Werb,Z. (2015) Matrix metalloproteinases in stem cell regulation and cancer. *Matrix Biol.*, **44**, 184–190.
45. Curran,S. and Murray,G.I. (2000) Matrix metalloproteinases: molecular aspects of their roles in tumour invasion and metastasis. *Eur. J. Cancer*, **36**, 1621–1630.
46. Kamat,A.A., Fletcher,M., Gruman,L.M., Mueller,P., Lopez,A., Landen,C.N., Han,L., Gershenson,D.M. and Sood,A.K. (2006) The clinical relevance of stromal matrix metalloproteinase expression in ovarian cancer. *Clin. Cancer Res.*, **12**, 1707–1714.
47. Takahashi,Y., Hamasaki,M., Aoki,M., Koga,K., Koshikawa,N., Miyamoto,S. and Nabeshima,K. (2018) Activated EphA2 processing by MT1-MMP is involved in malignant transformation of ovarian tumours in vivo. *Anticancer Res.*, **38**, 4257–4266.
48. Ma,R., Tang,Z., Ye,X., Cheng,H., Chang,X. and Cui,H. (2018) Low levels of ADAM23 expression in epithelial ovarian cancer are associated with poor survival. *Pathology-Res. Pract.*, **214**, 1115–1122.
49. Li,X., Bao,C., Ma,Z., Xu,B., Liu,X., Ying,X. and Zhang,X. (2018) Perfluorooctanoic acid stimulates ovarian cancer cell migration, invasion via ERK/NF-κB/MMP-2/-9 pathway. *Toxicol. Lett.*, **294**, 44–50.
50. Manders,D.B., Kishore,H.A., Gazdar,A.F., Keller,P.W., Tsunezumi,J., Yanagisawa,H., Lea,J. and Word,R.A. (2018) Dysregulation of fibulin-5 and matrix metalloproteases in epithelial ovarian cancer. *Oncotarget*, **9**, 14251.
51. Yang,S.C., Lheureux,S., Karakasis,K., Burnier,J.V., Bruce,J.P., Clouthier,D.L., Danesh,A., Quevedo,R., Dowar,M., Hanna,Y. *et al.* (2018) Landscape of genomic alterations in high-grade serous ovarian cancer from exceptional long-and short-term survivors. *Genome Med.*, **10**, 81.
52. Wang,Y.K., Bashashati,A., Anglesio,M.S., Cochrane,D.R., Grewal,D.S., Ha,G., McPherson,A., Horlings,H.M., Senz,J., Prentice,L.M. *et al.* (2017) Genomic consequences of aberrant DNA repair mechanisms stratify ovarian cancer histotypes. *Nat. Genet.*, **49**, 856.
53. Macintyre,G., Goranova,T., De Silva,D., Ennis,D., Piskorz,A.M., Eldridge,M., Sie,D., Lewsley,L.-A., Hanif,A., Wilson,C. *et al.* (2018) Copy-number signatures and mutational processes in ovarian carcinoma.*Nature Genetics 50 1262 1270*,
54. Yin,J., Ren,W., Huang,X., Deng,J., Li,T. and Yin,Y. (2018) Potential mechanisms connecting purine metabolism and cancer therapy. *Front. Immunol.*, **9**, 1697.
55. Pedley,A.M. and Benkovic,S.J. (2017) A new view into the regulation of purine metabolism: the purinosome. *Trends Biochem. Sci.*, **42**, 141–154.
56. Vijayan,D., Young,A., Teng,M.W. and Smyth,M.J. (2017) Targeting immunosuppressive adenosine in cancer. *Nat. Rev. Cancer*, **17**, 709.
57. Ohta,A. (2016) A metabolic immune checkpoint: adenosine in tumor microenvironment. *Front. immunol.*, **7**, 109.
58. Chernov,A.V. and Strongin,A.Y. (2011) Epigenetic regulation of matrix metalloproteinases and their collagen substrates in cancer. *Biomol. Concepts*, **2**, 135–147.
59. Venet,D., Dumont,J.E. and Detours,V. (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.*, **7**, e1002240.