ModulOmics: Integrating Multi-Omics Data to Identify Cancer Driver Modules

Dana Silverbush[*†1], Simona Cristea[*†2,3,4], Gali Yanovich[5], Tamar Geiger[5], Niko Beerenwinkel[‡6,7], and Roded Sharan[‡1]

[1]Blavatnik School of Computer Science, Tel Aviv University, Israel
[2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA
[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA
[4]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA
[5]Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Israel
[6]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[7]Swiss Institute of Bioinformatics, Basel, Switzerland

Abstract

The identification of molecular pathways driving cancer progression is a fundamental unsolved problem in tumorigenesis, which can substantially further our understanding of cancer mechanisms and inform the development of targeted therapies. Most current approaches to address this problem use primarily somatic mutations, not fully exploiting additional layers of biological information. Here, we describe ModulOmics, a method to *de novo* identify cancer driver pathways, or modules, by integrating multiple data types (protein-protein interactions, mutual exclusivity of mutations or copy number alterations, transcriptional co-regulation, and RNA co-expression) into a single probabilistic model. To efficiently search the exponential space of candidate modules, ModulOmics employs a two-step optimization procedure that combines integer linear programming with stochastic search. Across several cancer types, ModulOmics identifies highly functionally connected modules enriched with cancer driver genes, outperforming state-of-the-art methods. For breast cancer subtypes, the inferred modules recapitulate known molecular mechanisms and suggest novel subtype-specific functionalities. These findings are supported by an independent patient cohort, as well as independent proteomic and phosphoproteomic datasets.

[*]equal contribution
[†]corresponding author
[‡]equal contribution

Introduction

Rapid advancements in sequencing technologies led to an unprecedented increase in the generation and availability of high-resolution DNA, RNA, and protein cancer data. These large datasets are analyzed with mathematical and computational tools, unveiling mechanistic and predictive insights into cancer progression and treatment [5, 7, 8]. Key to achieving these goals is the identification of molecular alterations that drive tumorigenesis, or drivers, such as single nucleotide variants (SNVs), copy number alterations (CNAs), changes in the transcriptional activity of genes, or changes in protein concentration. Groups of such functionally connected genetic alterations, also termed cancer driver modules or pathways, activate mechanisms that drive tumorigenesis and gradually contribute to triggering the hallmarks of cancer, conferring fitness advantages to the tumors [54, 22]. Driver module elucidation can further our understanding of cancer initiation and progression, as well as inform the development of targeted therapies.

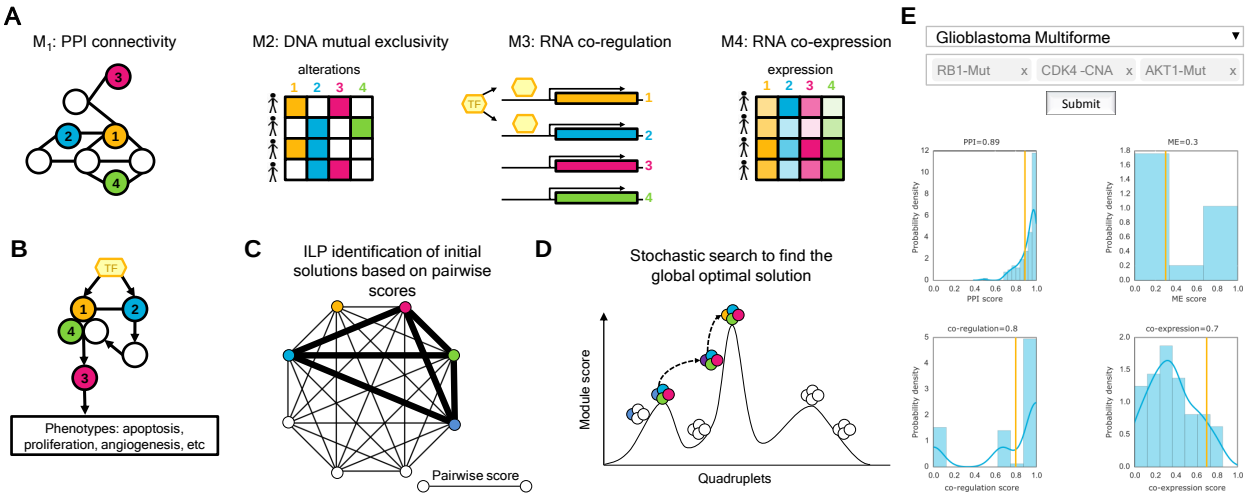

Figure 1: **Overview of ModulOmics. A)** Four different data sources, corresponding to four different models $M_1, \ldots, M_4$ (see Methods), contribute to the computation of the ModulOmics score: PPI connectivity (protein level), mutual exclusivity (DNA level), transcriptional co-regulation (regulatory connections and RNA level) and co-expression (RNA level). The four colors correspond to four different genes; full squares in the matrix for model M_2 encode the presence of alterations, while empty squares encode their absence. In M_3, genes 1 and 2 are regulated by a common transcription factor. In M_4, the different color intensities depict different expression intensities. **B)** Potential mechanism leading to a driver module exhibiting patterns of PPI connectivity, mutual exclusivity, co-regulation and co-expression. **C)** The ILP optimization identifies modules with highest sum of pairwise ModulOmics scores, computed as the average of the four scores corresponding to models $M_1, \ldots, M_4$, further z-scored and normalized to $[0, 1]$. **D)** The stochastic search optimization uses the modules identified by ILP, depicted in panel C, as seeds, and aims to improve their scores and identify the global optimal solution. The space of initial solutions is clustered and genes are exchanged between clusters in order to identify modules with high global scores. While the scores for models $M_1, \ldots, M_4$ of the modules in panel C were approximated as the average pairwise scores, here they are computed explicitly for the entire module. **E)** The webserver tool computes the ModulOmics score of any chosen gene set, based on any of the TCGA datasets analyzed in this study. For each data source employed by ModulOmics, the tool plots the single omics scores of the top 50 modules, highlighting the score of the chosen gene set.

It has been observed that members of cancer pathways often display specific alteration patterns across tumor samples, most notably co-occurrence and mutual exclusivity [12, 11, 30, 2]. Finding groups of mutually exclusive genes is an efficient way to identify cancer modules fulfilling the same biological function, since, once a single member of the group is altered, the tumor gains a significant selective advantage, and the fitness of the tumor is not expected to increase with the alteration of additional group members. However, most existing mutual exclusivity methods rely only on DNA-level data, particularly SNVs, failing to fully exploit the complex interactions involving RNA or protein molecules potentially driving tumorigenesis [14, 19]. To this end, data integration strategies have the potential to unravel previously unknown cancer driver modules.

An additional important data source for identifying interactions among cancer drivers is protein-protein interaction (PPI) networks, cataloged in databases such as HIPPIE [43], STRING [48], or BioGRID [46]. Studies that exploit this data source include HotNet2 [29], which uses PPI networks of genetic alterations to identify significantly altered subnetworks connecting recurrently mutated genes; EnrichNet [20], which identifies functional gene sets based on PPI proximity calculated similarly to HotNet2, MEMo [11], which identifies

mutually exclusive gene sets on the basis of PPI-filtered pairwise connections, and MEMCover, which integrates mutual exclusivity among genetic alterations with connectivity derived from PPI networks [27]. These methods however do not include additional layers of biological information directly into their model, such as RNA regulation or gene expression. Few approaches address the problem of integrating such additional data sources, among which TieDIE [37], which finds one large PPI subnetwork connecting DNA, RNA and regulatory signals. A separate class of data integration methods aiming to identify dysregulation in cancer focus on individual driver genes, without also connecting the drivers into modules, such as DriverNet [3], which combines mutations and gene expression, or DawnRank [25], which integrates mutations, gene expression and protein interactions.

Here, we describe ModulOmics, a method for the *de novo* identification of cancer driver modules based on the integration of PPI networks, mutual exclusivity of DNA alterations (SNVs and CNAs), and RNA-level co-regulation and co-expression, into a single probabilistic score (Figure 1). We identify modules that maximize this score by performing a two-step optimization procedure that combines Integer Linear Programming (ILP) with stochastic search. We apply ModulOmics on three large-scale TCGA datasets of breast cancer [5], glioblastoma (GBM) [6] and ovarian cancer [7], and show that it accurately identifies known cancer driver genes and pathways. Moreover, ModulOmics outperforms three state-of-the-art methods to detect cancer modules, namely the DNA-centric method TiMEx [12], the PPI-based method HotNet2 [29] and the DNA and PPI integration method MEMCover [27].

We further use ModulOmics to identify modules that characterize breast cancer subtypes. The highest scoring modules are enriched with cancer drivers, and reliably separate cancerous from normal tissues in an independent patient cohort [40, 51]. Moreover, the modules characterizing aggressive subtypes, such as Her2 and Basal, are further enriched with Gene Ontology (GO) terms related to cell proliferation. In the triple negative (TN) subtype, we identify functional connections among multiple down-regulated tumor suppressors, including *TP53*, *BRCA1*, *RB1* and *PTEN*. This pattern is also supported by reverse phase protein array (RPPA) data [5]. In Luminal A, high scoring modules containing *PTEN* suggest two potential functionalities of this protein: a canonical one as part of the PI3K pathway, and a non-canonical one as a regulator of cell proliferation.

ModulOmics is freely available in two forms, as an open-source R code for the identification of cancer driver modules from a cohort of cancer samples (`https://github.com/danasilv/ModulOmics`), and as a webserver for the evaluation of any set of genes of interest using the TCGA data processed in this study (`http://anat.cs.tau.ac.il/ModulOmicsServer/`).

Results

ModulOmics identifies driver modules on the basis on DNA and RNA cancer patient data, integrated with PPI networks and known regulatory connections. Each candidate module is scored according to the degree of mutual exclusivity among DNA alterations in its members across the patient cohort, the correlation of the RNA expression of its members across the cohort, the probability that its gene members are connected in the PPI network, and the fraction of its members that are co-regulated by a common active transcription factor. As the number of candidate groups grows exponentially with maximal group size, ModulOmics uses a heuristic two-step optimization procedure to first find good initial solutions by linearly approximating the scoring function and then refining these solutions via stochastic search (see STAR Methods). Each module is assigned an empirical p-value by comparing its score to the scores of $1,000$ random modules of altered genes of the same size. All top modules identified by ModulOmics were significant (corrected Bonferroni p-value < 0.05).

We applied ModulOmics on three large TCGA cancer datasets: breast cancer, GBM, and ovarian cancer. On these datasets, we compared ModulOmics to four simplified similar approaches, in which the score of a group is computed using only single omic data sources, namely PPI connectivity, mutual exclusivity, co-regulation or co-expression, as well as to three state-of-the-art methods for the identification of driver modules: MEMCover, HotNet2 and TiMEx. We additionally assessed the contribution of each of the single omic data sources to the ModulOmics score by only using subsets of three omics, and removing each single omic at a time.

We found that modules of sizes 3 and 4 generally have higher ModulOmics scores than pairs (Figure

S1), and no single omic data source dominates the score for any given module size (Tables S1-S3). Sensitivity analyses showed that the performance of ModulOmics is robust under different parameter choices (Supplementary Information, Figures S2-S3 and Tables S4-S5).

Driver modules are enriched with cancer drivers

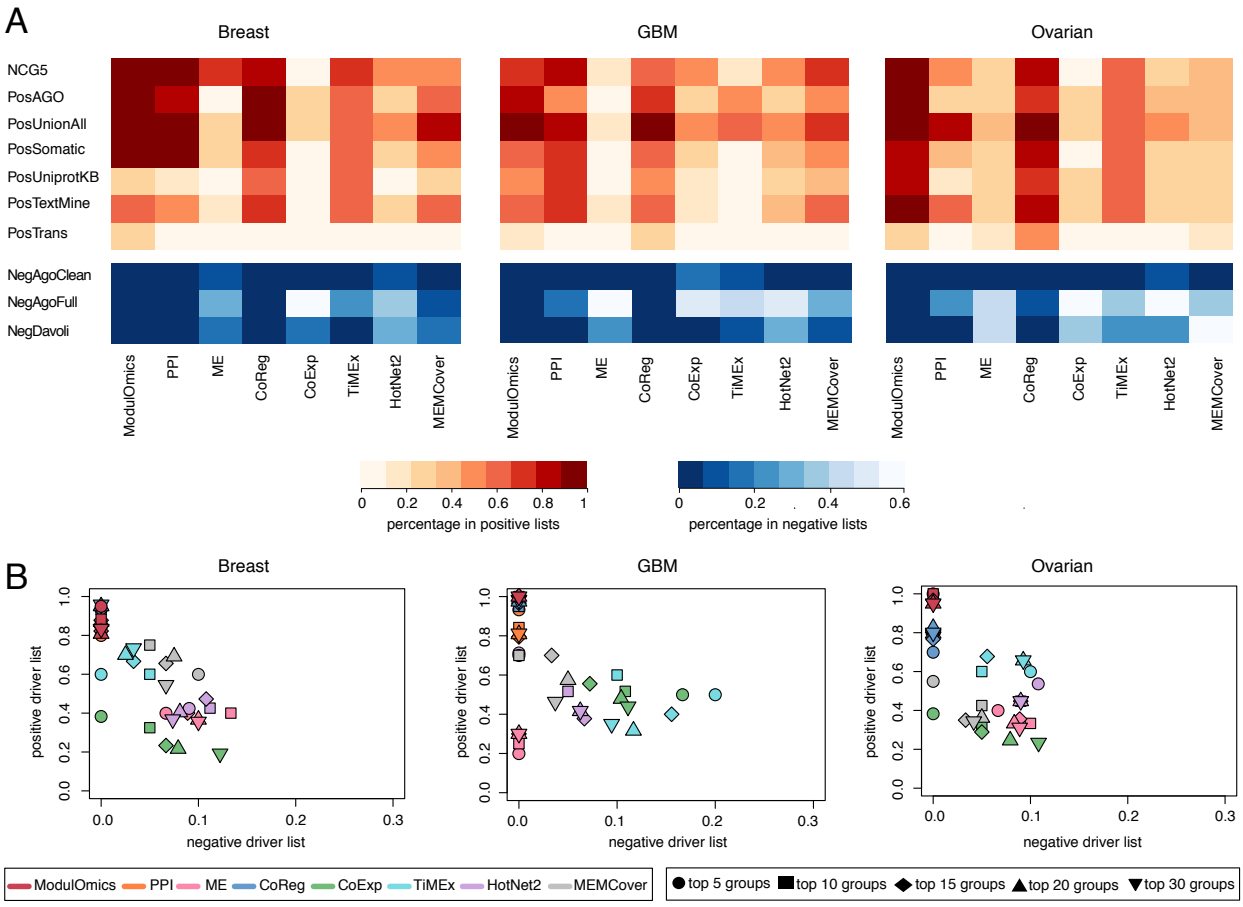

Figure 2: **The driver modules inferred by ModulOmics are enriched with cancer driver genes. A)** The average driver enrichment (red heatmaps) and non-driver enrichment (blue heatmaps) across the top 10 scoring modules inferred by each method in the three cancer types studied. The enrichment was calculated as the fraction of gene members in each module that are also part of each control list, averaged across the top 10 modules. The modules were ranked by their score, regardless of their sizes (the inferred groups consisted of two, three or four gene members). *ME* stands for mutual exclusivity, *CoReg* for co-regulation, and *CoExp* for co-expression single omic scores. *NCG5*, *PosAGO*, *PosUnionAll*, *PosSomatic*, *PosUniprotKB*, *PosTextMine*, *PosTrans* are the positive control lists, while *NegAgoClean*, *NegAgoFull* and *NegDavoli* are the negative control ones (see STAR Methods). **B)** Detailed driver and non-driver enrichment scores for the positive driver list *PosUnionAll* and the negative driver list *NegAGOClean* for the seven methods assessed, across the three cancer types, for the top scoring 5, 10, 15, 20 and 30 modules. Table S6 shows the scores for 2B.

To assess the performance of ModulOmics, we calculated the enrichment of the highest scoring driver modules with known driver genes (positive controls) and known non-driver genes (negative controls). To this end, we used the gene lists introduced in [23], complied from different sources: the Network of Cancer Genes (NCG) [1], Cancer Gene Census version 73 (CGC) [17], the Atlas of Genetics and Cytogenetics in Oncology and Hematology (AGO) [26], UniprotKB [53], DISEASES [39] and MSigDB [47] (Supplementary Information). The enrichment was calculated as the fraction of gene members in each module that were also part of each control list, averaged across the top modules considered. The top 10 modules inferred by ModulOmics generally outperformed the top 10 modules identified with the four single omic approaches and with MEMCover, HotNet2 and TiMEx across the seven positive and three negative control lists tested

(Figure 2A). Specifically, ModulOmics achieved an enrichment score of close to 1 across all three cancer types in the three largest positive control lists: the manually curated resource *NCG5*, the positive AGO list (*PosAGO*), and the Union All list (*PosUnionAll*), consisting of between 1,429 and 2,144 known drivers. Importantly, the modules inferred by ModulOmics scored close to 0 in all three negative control list assessed, namely the complete negative AGO list (*NegAgoFull*), the curated negative AGO list (*NegAGOClean*), and the negative list introduced in [15] (*NegDavoli*), consisting of between 3,272 and 9,457 known non-driver genes.

In addition, ModulOmics also outperformed the other methods when evaluating the highest scoring 5, 10, 15, 20, or 30 modules of any size (Figure 2B), or when separately evaluating modules of fixed sizes (Figure S4). Among the competing methods, PPI-based and co-regulation-based scorings exhibited good performances, MEMCover performed well only in the case of certain group sizes, while co-expression, mutual exclusivity, HotNet2 and TiMEx generally performed poorly on both positive and negative control metrics. We further evaluated the contribution of each single omic data source to the driver genes enrichment by computing reduced versions of the ModulOmics score, each time with a single omic removed. We found that integrating all four omics data sources improves the enrichment, as compared to using subsets of three omics, in 90% of the evaluated cases (92% of the positive control lists and 86% of the negative lists). Nevertheless, the performance of ModulOmics remained fairly robust when using only three omics sources, suggesting that the method can also be applied in cases when one data source is missing (Figure S5).

One of the features of ModulOmics is that each gene can participate in multiple pathways, hence the reported modules often overlap (Figure S6). Biologically, this feature is justified by the fact that the known driver genes are likely network hubs, expected to be functionally connected to multiple other less-known driver genes into different modules. In order to assess the performance of ModulOmics also in the absence of overlap among groups, we repeated the driver evaluation while considering the first 20 unique appearances of genes in the top modules. Consistent with the previous results, ModulOmics outperformed the three other competitive methods tested (Figure S7).

Driver modules are functionally coherent

An additional metric for evaluating the relevance of the inferred modules concerns their functional coherence, which we assessed via their enrichment with curated pathways from KEGG [33] (Supplementary Information). ModulOmics identified key cancer-related pathways, such as *pathways in cancer*, across all three cancer types (Figure 3A), without showing preferential enrichment for particular module sizes (Figure S8). In contrast, HotNet2 identified this pathway only in the GBM and ovarian cancer datasets, and MEMCover and TiMEx did not identify it at all. Additional highly enriched pathways included *apoptosis*, *cell cycle*, *TP53 signaling*, *mTOR signaling*, and the angiogenesis-related *VEGF pathway*. Interestingly, the set of enriched pathways also included pathways characterizing other cancers types, indicating shared mechanisms among malignancies.

To quantify the pathway enrichment performance of ModulOmics, MEMCover, TiMEx and HotNet2, we counted the number of pathways significantly enriched (Bonferroni-corrected p-value ≤ 0.05) in each of the top 5, 10 and 15 highest scoring modules (Figure 3B), and computed their average enrichment factor (Figure 3C). Enrichment p-values and factors were computed with Expander [52] (Supplementary Information). Overall, ModulOmics identified modules enriched with more general pathways and cancer-related pathways than the three competing methods, and was the only method for which all highest scoring 10 modules were enriched with at least one pathway. A high percentage of the genes identified by ModulOmics participated in known KEGG pathways, reaching an average of 78% across all three cancer types, compared to 43%, as identified by MEMCover, 39% by HotNet2, and 13% by TiMEx (Table S7). In contrast, less than 5 out of the 1,000 random modules generated for computing module significance were significantly enriched with any known pathway. Finally, we tested the contribution of each omic to pathway enrichment and found that using all four data sources improves the identification of functionally coherent modules in 92% of the tested cases (Figure S9).

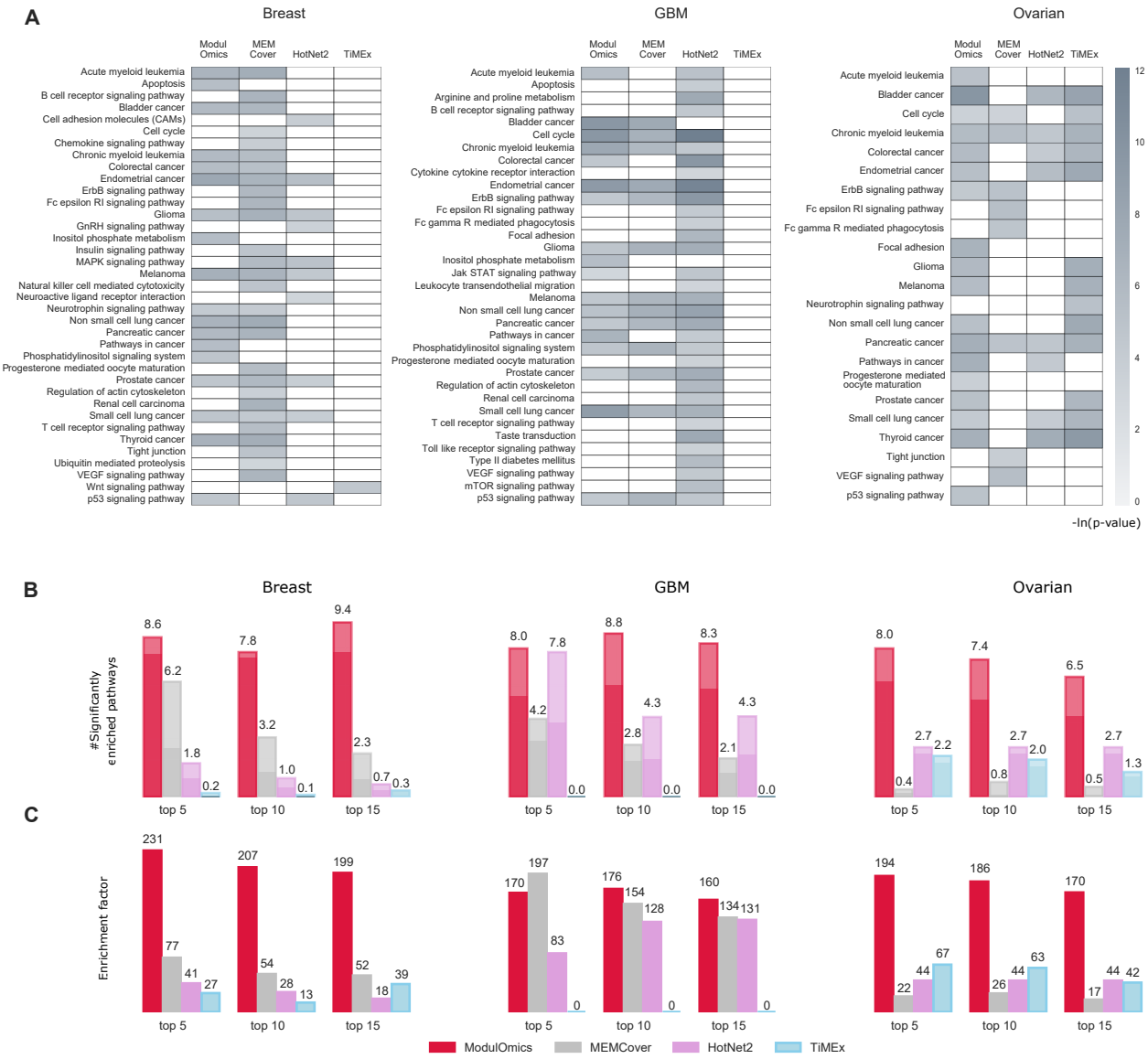

Figure 3: **The driver modules inferred by ModulOmics are enriched with cancer driver pathways. A)** Mean hyper-geometric p-value of the KEGG pathways significantly enriched in the top 10 modules identified by ModulOmics (red), HotNet2 (purple) and TiMEx (light blue). **B)** Average number of KEGG pathways significantly enriched in the top modules, indicated above the bars. The opaque bars indicate cancer-related pathways only. **C)** Average enrichment factors for top modules. The numbers displayed in panels B and C are normalized per module. Enrichment p-values and factors were computed with Expander [52].

Driver modules in breast cancer subtypes recapitulate known mechanisms and suggest novel functionalities

Next, we applied ModulOmics on molecularly defined subtypes of breast cancer, classified using the mRNA PAM50 classification [36] into Basal (125 patients), Her2 (61), Luminal A (364) and Luminal B (174) (Table S8 and Figure S10). Across all subtypes, the genes in the top 20 modules (Figure 4A) were highly enriched with cancer drivers (66% were part of the NCG5 positive control list and 70% were part of the UnionAll positive list, while only 4% were part of the AGOClean negative control list) and KEGG pathways (44 enriched pathways, 24 of which were directly related to cancer, average p-value 0.0063). The top drivers identified by ModulOmics included *TP53*, *AKT1*, *mTOR* and *PTEN*, as well as subtype-signature genes such as *BRCA1* and *BRCA2* for Basal [49, 50], *CDH1* for Luminal A and B [24], *MAP3K1* for Luminal B [5] and

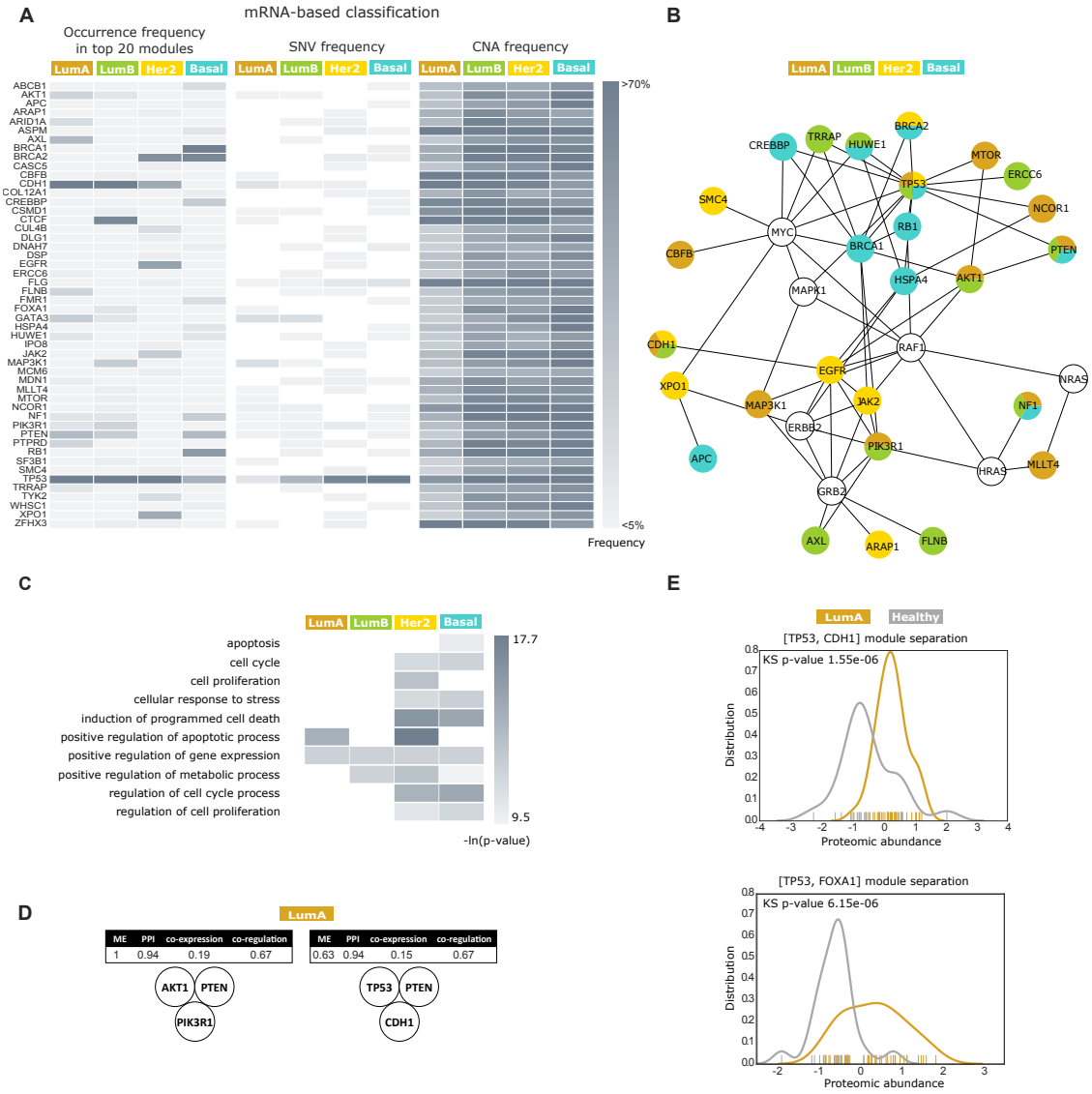

Figure 4: **Modules inferred in mRNA-classified breast cancer subtypes reflect various levels of subtype aggressiveness and separate cancerous form healthy tissues. A)** For each mRNA-based subtype and for the pooled set of genes in the top 20 modules, we computed their occurrence frequency in the top 20 modules, as well as their SNV and CNA alteration frequencies across the patient cohort. These genes are enriched with known cancer drivers and pathways, and could not have been identified if relying on SNV and CNA alteration frequencies alone. White corresponds to absent genes (0% frequency). **B)** Detailed PPI network view of the subset of genes in panel A that are either known drivers, or part of KEGG pathways. The displayed physical protein interactions underline cancer-related functional associations, such as the role of the PIK3A pathway in Luminal A tumors. **C)** Selected list of significantly enriched GO pathways across the top 20 modules (Figure S11 displays the full list), reflecting the aggressiveness of the Basal and Her2 subtypes, compared to Luminal A and Luminal B. Enrichment hyper-geometric p-values were computed with Expander [52]. White corresponds to absent pathways. **D)** Module scores for top Luminal A modules suggesting two different biological roles of the tumor suppressor *PTEN*. **E)** The highest ranking Luminal A module in an independent mass spectrometry proteomics dataset separates cancerous from healthy patient tissues. *TP53* loss is measured by its downstream regulated protein CDC2, *CDH1* loss is measured by its downstream regulated protein CTNNB1, and *FOXA1* gain is measured directly.

EGFR for Her2 [32]. An alternative strategy to ModulOmics for identifying relevant drivers would have been selecting genes with highest SNV or CNA alteration frequencies [54]. However, in that case, a substantial portion of the enriched gene list identified by ModulOmics would have been overlooked, as 34% fall below the SNV median frequency per gene and 40% fall below the CNA median frequency per gene (Figure 4A). Therefore, integrative approaches such as ModulOmics are essential to this end.

A detailed PPI network view of the genes identified by ModulOmics showed that *TP53* is a key player in

tumor progression for all subtypes, while subtype-specific key players included *EGFR* for Her2 and *BRCA1* for Basal (Figure 4B). The network view highlighted the higher rate of PPI-connected established tumor suppressors in the Basal subtype, as compared to Luminal A, matching the aggressive nature of these tumors. In addition, Luminal A modules were characterized by a higher occurrence of PI3K pathway members, such as *PIK3R1*, *AKT1*, *mTOR* and *PTEN*, as previously observed [5]. The top modules identified by ModulOmics were further highly enriched with functional relations, highlighting different GO annotations for each subtype (Figure 4C). These results captured the increased pathway activity of key pathways required for tumor progression, such as apoptosis, cell cycle process or cell proliferation, as well as the known aggressiveness of Basal and Her2 tumors, reflected in their higher pathway enrichment.

The highest ranking module for both Luminal A and Luminal B was *TP53* and *CDH1*, two known functionally associated drivers in the Luminal subtypes [5, 4]. The top Her2 modules were characterized by the recurrent appearance of the nuclear export gene *XPO1* together with *TP53*, which is one of its known targets [18, 9]. Interestingly, the mechanism of TP53 nuclear export by *XPO1* is well characterized [16], yet no pivotal role was suggested specifically in Her2 breast cancer. The highest ranking module of the Basal subtype consisted of *RB1*, *BRCA1*, *NF1*, and *CREBBP*. *CREBBP* is a *BRCA1* activator [35] and, since both *BRCA1* and *CREBBP* are involved in DNA repair, this module potentially reflects the altered DNA damage repair mechanism specific to Basal tumors [34].

One of the frequently occurring genes in the top Luminal A modules was the tumor suppressor *PTEN*, occurring both in modules reflecting its canonical PI3K pathway role, and in modules suggesting a non-canonical role (Figure 4D). The canonical *PTEN* module also included *PIK3R1* and *AKT*, thus supporting the known mutual exclusivity pattern of mutations within the PI3K pathway [42]. The module suggesting the non-canonical role of *PTEN* also included *CDH1* and *TP53*, supporting the hypothesis that *PTEN* regulates cell proliferation by increasing the binding of *CDH1* to APC\C, a complex known for its tumor-suppressive function, and by increasing *TP53* acetylation following DNA damage [44]. Indeed, according to the TRRUST database [21], *PTEN* and *CDH1* are co-regulated by two common transcription factors, namely *STAT3* and *NFKB1*.

In order to further explore the clinical relevance of the highest scoring driver modules, we examined how well they can distinguish healthy tissues from cancerous ones in an independent omic data source. To this end, we used a recently published proteomics dataset consisting of 62 samples of Luminal A and healthy tissues [40, 51], and focused on the two highest scoring Luminal A modules: *TP53* and *CDH1*, and *FOXA1* and *TP53*, respectively. These top two modules significantly separated the Luminal A cancerous tissues from the healthy ones (p-value $1.6e^{-06}$ and p-value $6.2e^{-06}$ respectively, KolmogorovSmirnov (KS) test, Figure 4E). For comparison, neither *GATA3* or *PIK3CA*, the most frequently mutated genes in Luminal A, nor *TP53*, the most frequently mutated gene in breast cancer, were able to significantly separate the two types of tissue (p-value 0.065, p-value 0.054, and p-value 0.69, respectively, KS test). Similarly, random modules of the same size did not significantly separate the tissues (p-value 0.14, averaged over $1,000$ random modules generated by sampling subsets of proteins from the proteomics dataset, KS test).

An alternative way to study breast cancer progression is by stratifying patients according to immunohistochemistry results assessing the HER2, ER and PR receptors (Supplementary Information). To this end, we separated our patient cohort into the following subtypes: TN (116 patients), Her2-enriched (30), Luminal A (477) and Luminal B (88), and used ModulOmics to infer modules for each subtype (Table S9 and Figure S12). Similarly to the mRNA-based classification, the genes part of the highest scoring 20 modules were enriched with cancer drivers (67% were part of the NCG5 positive control list and 59% were part of the UnionAll positive list, while only 2% were part of the AGOClean negative control list) and with known cancer pathways (46 enriched pathways, 25 of which were directly related to cancer, average p-value 0.01, Figure S13A). Across subtypes, the highest scoring modules highlighted a unique alteration pattern for the tumor suppressor *TP53*. In Luminal A, Luminal B and Her2-enriched, *TP53* was mutually exclusive with other tumor suppressors, such as *PTEN* and *BRCA1* in Luminal B, or *BRCA2* in Luminal A, which led to ModulOmics inferring these groups as high scoring modules. However, in TN, *TP53* was mutually exclusive with *BRCA2*, but not with other key TN drivers, such as *BRCA1*, *PTEN* or *RB1*, as both the pairwise and the group mutual exclusivity scores of *TP53* and these three drivers were 0 (Figure S13B). This suggests a TN-specific concerted down-regulation of multiple tumor suppressors, namely *TP53*, *BRCA1*, *RB1* and *PTEN*, potentially contributing to the bad prognosis of this subtype. Taken together, these results imply that the level of mutual exclusivity in tumor suppressors might reflect the aggressiveness of the tumor

subtype [38, 45, 5, 14].

Finally, we used an independent omic data source (RPPA) to further evaluate the functional connectivity among the tumor suppressors *PTEN*, *BRCA1*, *RB1*, and *TP53*. In general, evaluating protein measurements limits automatic and exhaustive analyses, since loss of function can lead to missing data, requiring the identification of downstream regulated proteins that can serve as surrogates. PTEN was found to be downregulated, while phosphorylated AKT, which is suppressed by PTEN, was upregulated. *BRCA1* loss was accounted for by its downstream regulated protein CYCLIN B1, which was highly expressed. RB1 showed an overall low expression in most samples, while the RB1-related CYCLIN-D1 was lowly expressed mostly in TN tumors (Figure S13C). *TP53* loss was accounted for by its target CDK1, which was also highly expressed. *CDK1*, *CYCLIN B1* and *AKT*, the tumor promoters regulated by *TP53*, *BRCA1* and *PTEN*, were significantly upregulated in TN tissues compared to the other subtypes (p-value $1.2e^{-16}$, KS test), while the tumor suppressors *PTEN*, *RB1*, and *BRCA2* were significantly downregulated (p-value $2.9e^{-09}$, KS test, Figure S13D). These results suggest that these two groups of genes can be used to separate TN from the other subtypes.

Discussion

ModulOmics is a novel method to *de novo* identify molecular cancer driver pathways, based on the integration of connectivity within protein-protein interaction networks, mutual exclusivity among SNV or CNA alterations, transcriptional co-regulation, and RNA co-expression, into a single probabilistic score. ModulOmics uses an efficient two-step optimization procedure to first find good initial solutions using linear approximation, and then refine these solutions with stochastic search. We demonstrate the performance of ModulOmics in identifying modules enriched with known cancer driver genes and pathways in three large-scale multi-omics TCGA datasets: breast cancer, GBM, and ovarian cancer. We further investigate breast cancer subtypes and find that some of the highest scoring modules are known to be involved in cancer-related molecular mechanisms, while others suggest novel functionalities. We evaluate these results using an independent patient cohort and independent proteomic and phosphoproteomic datasets. In addition, we show that the top modules inferred by ModulOmics can be used to reliably separate cancerous from normal tissues in Luminal A samples, as well as to distinguish TN samples from the other subtypes.

ModulOmics is a freely available open-source software. The framework is designed to be flexible, such that any of the four sources of evidence employed here can be excluded or replaced with new sources of evidence. In addition, since ModulOmics integrates independent sources of information, newly added data can also originate from different patient cohorts, such that single omic datasets from other cancer studies can be readily integrated. Moreover, the webserver implementation of ModulOmics can be used to evaluate the ModulOmics score of any user-defined gene set, on the basis of any of the TCGA datasets analyzed here. This application can be very useful in situations when gene sets were deduced from separate biological or computational analyses.

Some of the modules identified by ModulOmics may merit further experimental investigation. For example, on the basis of the highest ranking Basal module (*RB1*, *BRCA1*, *NF1*, and *CREBBP*), we propose that further validation experiments could evaluate the clinical implications of using the *CREBBP* inhibitor in *BRCA1* patients, similarly to *PARP1*, another DNA repair agent successfully used in treatment [28]. Based on the recurrent joint occurrence of *XPO1* and *TP53* in top Her2 modules, we propose to further evaluate the role of the export mechanism of *XPO1* leading to *TP53* depletion in the nucleus, thus decreasing its tumor suppression capability. The role of *XPO1* in tumor progression was previously investigated in a preclinical context of TN treatment [10, 31]. Here, we suggest it may also play a role in the Her2 subtype. Additionally, in the future, once high-throughput single-cell profiling of tumors becomes routinely performed in the clinic, ModulOmics can be used to identify functional connections derived from multiple tumor cells of single cancer patients, rather than a patient cohort. In this way, the design of personalized cancer therapies based on the tumor heterogeneity of each patient can be facilitated.

A unique feature of ModulOmics is that its scoring function not only uses different types of data, but also integrates different types of statistical tests distinctly designed for each data type (such as the mutual exclusivity test for mutational data, or the proximity-based PPI score). In contrast, previous approaches generally apply the same methodological framework to all data types [37]. Integrating different statistical

tests is empowered by online normalizing each score throughout the search, as well as by optimizing the different scores simultaneously, rather than sequentially. In addition, the weights of the different scores can be optimized from the data, with the aim of identifying groups representative of a functional phenotype of interest. To exemplify this, we inferred optimal omics weights by training a classifier to identify groups enriched with the Cell Cycle GO term from the breast cancer top modules reported by ModulOmics. The classifier assigned the weight 0.61 to the ME score, 3.61 to the PPI score, 1.83 to the co-regulation score and 1.93 to the co-expression score. All the four data sources were hence found to contribute to the classification of modules associated with the GO term, with the scores reflecting physical interactions carrying more weight, as would be expected in the case of phenotypes related to biological processes.

ModulOmics can be extended beyond finding modules in cancer. For example, the PPI, co-regulation and co-expression scores can be used to detect protein complexes. As a proof-of-concept, we collected 1,112 known biological complexes from CORUM [41] (see details in the Supplementary Information) of sizes 2, 3 and 4, as well as equal number of random protein groups and calculate their ModulOmics scores. We trained a classifier to distinguish known complexes from random groups based on the three scores, reaching an AUPR of 0.84 for all module sizes (Figure S14) and of up to 0.9 for modules of size 4. These results indicate that the ModulOmics scores are informative in identifying new biological connections outside the scope of cancer, making the tool broadly applicable.

STAR Methods

Model

Given a set $G = \{G_1, \ldots, G_n\}$ of genes and a collection $M = \{M_1, \ldots M_m\}$ of models for different data types, we are interested in computing S_G, the ModulOmics probabilistic score of the set G, reflecting how likely are the genes in G to be functionally connected. S_G is computed as the mean of m probabilistic scores $P(G \mid M_k)$. Each of these m scores represents how strongly functionally connected the genes in G are, under different models:

$$S_G = \frac{1}{m} \sum_{k=1}^{m} P(G \mid M_k) \tag{1}$$

The models we consider here are: connectivity among protein-protein interactions (M_1), mutual exclusivity among point mutations or copy number alterations (M_2), transcriptional co-regulation (M_3), and gene co-expression (M_4).

PPI Connectivity

Model M_1 assesses the functional connectivity of the set G at the protein level, by computing the probability of G being connected in the PPI network. Starting with a fully-connected literature-based PPI network (HIPPIE) and its associated interaction probabilities, we define, for each pair of genes (G_i, G_j), $\mathrm{con}(G_i, G_j)$ as the probability of the most likely path connecting G_i and G_j, *i.e.*, the product of the probabilities of the path's edges. The computation of $\mathrm{con}(G_i, G_j)$ for all $G_i, G_j \in G$ yields a complete graph on G, denoted $\mathcal{G}(G)$. If we denote the edge set corresponding to any graph H by $E(H)$, then the connectivity of the set G is defined as the sum of the probabilities over all connected subgraphs spanning G, as follows:

$$P(G \mid M_1) = \sum_{c \in C(G)} \prod_{(G_i, G_j) \in E(c)} \mathrm{con}(G_i, G_j) \prod_{(G_i, G_j) \in E(\mathcal{G}(G)) \setminus E(c)} (1 - \mathrm{con}(G_i, G_j)) \tag{2}$$

where $C(G)$ is the collection of connected subgraphs spanning $\mathcal{G}(G)$.

Mutual exclusivity

Model M_2 estimates the degree with which DNA alterations support the functional connectivity of the genes in G. Following the mutual exclusivity framework defined in the context of waiting times to alteration

introduced in TiMEx [12] and pathTiMEx [13], $P(G \mid M_2)$ is computed as the degree of mutual exclusivity of the set G, as follows:

$$P(G \mid M_2) = \begin{cases} \mu_G & \text{if p-value} \leq 0.05 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where both μ_G and p-value are reported by TiMEx. The TiMEx probabilistic graphical model estimates μ_G, which is the mutual exclusivity intensity of the group G, via a nested likelihood ratio test between an independence model and an alternative, mutual exclusivity model. The independence model assumes that the genes evolve independently during disease progression, whereas the mutual exclusivity model assumes that only the gene with the shortest waiting time in a functionally connected group of genes will fixate. The parameter μ_G represents the probability that a group of genes is perfectly mutually exclusive, *i.e.*, that no two genes in G share alterations in the same patient. Therefore, $\mu_G = 1$ corresponds to perfect mutual exclusivity, and $\mu_G = 0$ corresponds to independence. The p-value in Equation 3 is the probability of observing a given alteration pattern of the set G under the null hypothesis of independence, as described in [12] and [13].

Co-regulation

Model M_3 assesses the functional connectivity of the genes in G on the basis of their transcriptional regulation. The co-regulation score $P(G \mid M_3)$ is defined as the fraction of genes in G which are co-regulated by at least one common active transcription factor,

$$P(G \mid M_3) = \frac{|G_{\text{co-reg}}|}{|G|} \tag{4}$$

where $G_{\text{co-reg}} \subseteq G$ is the maximal set in which all genes are regulated by at least one common active transcription factor. A transcription factor is considered active if it is differentially expressed (z-score of fold change is either > 1 or < -1) in at least 25% of samples. Alternatively, other operators such as the average could be used, however choosing the maximal set reflects the co-regulation of the entire group, rather than particular subgroups.

Co-expression

Model M_4 evaluates the functional connectivity of the genes in G based on their transcriptional profiles. Let a gene be defined as expressed if its expression averaged across all samples is above the k^{th} q-quantile, and let $G_{\text{exp}} \subset G$ be the set of all expressed genes. Then, the co-expression score of G is defined as the mean among all pairwise Spearman correlations of the expression profiles of the genes in G_{exp}, and 0 corresponding to the remaining pairs, in which at least one of the genes is not expressed,

$$P(G \mid M_4) = \frac{\sum_{G_i, G_j \in G_{\text{exp}}} |\text{cor}(E_i, E_j)|}{\binom{|G|}{2}} \tag{5}$$

where E_i is the continuous expression level of gene G_i across all samples, and $\text{cor}(E_i, E_j)$ is the Spearman correlation among the expression profiles of G_i and G_j. For this application, we chose $k = 2$ and $q = 4$, *i.e.*, the 2^{nd} quartile. The choice of Spearman correlation is justified by not necessarily assuming a linear relation between expression profiles. Missing expression data can be handled by assigning the respective genes null expression profiles, leading to their consideration as unexpressed genes.

Optimization procedure

Given a large cancer dataset, identifying groups of functionally connected genes is challenging, as the number of candidate groups increases exponentially with maximal group size. Therefore, we employ a two-step procedure to optimize the global ModulOmics score in equation 1. First, in order to identify a large set of good initial solutions, we formulate the optimization problem as an ILP, and optimize a linear approximation of the global ModulOmics score. Second, we perform a stochastic search starting from these initial solutions and using the global score.

ILP

The first step of our optimization procedure linearly approximates the exact scores of the set G under each of the four models M_k, by decomposing them into pairwise scores. For each model M_k, the score of each pair of genes (G_i, G_j) is denoted by $w^k_{G_iG_j}$ and equals to $P\left((G_i, G_j) \mid M_k\right)$, further z-scored and normalized to $[0, 1]$. The goal of the optimization routine is to identify candidate subsets G with high total scores w_G, computed as:

$$w_G = \sum_{k=1}^{m} \sum_{G_i, G_{j;i<j} \in G} w^k_{G_iG_j} \tag{6}$$

The ILP retrieves sets G of fixed size K with maximal w_G score. Thus, G is the maximal weight subgraph of size K in a weighted complete graph with vertices V, corresponding to a large set of genes, and edges $E_{i,j} = \{w_{V_iV_j} \mid V_i, V_j \in V\}$. The ILP consists of the following set of binary vertex variables $V_{(i)}$ denoting the inclusion of vertex V_i in a set G, and edge variables $E_{(i,j)}$, denoting the inclusion of edge $E_{i,j}$ in G:

$$V_{(i)} \in \{0, 1\} \quad \forall\ V_i \in V \tag{7}$$

$$E_{(i,j)} \in \{0, 1\} \quad \forall\ V_i, V_j \in V,\ i < j \tag{8}$$

and the objective function:

$$\text{maximize} \sum_{V_i, V_j \in V,\ i<j} w_{V_iV_j} \cdot E_{(i,j)} \tag{9}$$

under the constraints:

$$E_{(i,j)} - V_{(i)} \leq 0 \tag{10}$$

$$E_{(i,j)} - V_{(j)} \leq 0 \tag{11}$$

$$V_{(i)} + V_{(j)} - E_{(i,j)} \leq 1 \tag{12}$$

$$\sum_{V_i \in V} V_{(i)} = K \tag{13}$$

$$\sum_{V_i, V_j \in V, i<j} E_{(i,j)} = \frac{K \times (K-1)}{2} \tag{14}$$

$\forall\ V_i, V_j \in V,\ i < j$. Constraints 10, 11, and 12 ensure that the retrieved set is a clique, and constraints 13 and 14 ensure that the clique is of size K. Let us note that identical solutions would be retrieved by discarding either constraint 13 or 14, yet we include both for efficiency considerations. With each candidate set G found, we add constraint 15 to prevent the ILP to choose the entire set G again:

$$\sum_{i \in G} V_{(i)} \leq K - 1 \tag{15}$$

Stochastic search

We use 200 high-ranking modules identified by the ILP as seeds for a stochastic search that expands the search space and optimizes directly the exact score of the modules, rather than their pairwise approximations. The stochastic search uses the seed modules as starting points and aims to find the modules with global optimal score by offering possible exchanges of module members. The seed modules are clustered into 10 clusters using k-means, and a search cycle starts independently from each cluster, in order to increase the chances of finding modules with global optimal scores. Each of these 10 cycles iterates among the modules in its cluster and tries to improve each one by suggesting 20 possible exchanges of a random module member with another random gene. If the score improves, then the exchange is accepted and the module is updated accordingly. Each cycle reports its 5 highest scoring modules. The modules reported by all 10 cycles are finally aggregated and re-ranked. Sensitivity analyses show that the performance of ModulOmics is robust under different parameter choices (Supplementary Information). Each run of the ILP followed by the stochastic search yields optimal modules of fixed size K. To retrieve the top modules in a range of sizes we run the tool with K ranges from 2 to 4, aggregate the results and retrieve the top modules regardless of their size.

References

[1] Omer An, Giovanni M. Dall'Olio, Thanos P. Mourikis, and Francesca D. Ciccarelli. Ncg 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. *Nucleic Acids Research*, 44(D1):D992–D999, 2016.

[2] Özgün Babur, Mithat Gönen, Bülent Arman Aksoy, Nikolaus Schultz, Giovanni Ciriello, Chris Sander, and Emek Demir. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biology*, 16(1):45, 2015.

[3] Ali Bashashati, Gholamreza Haffari, Jiarui Ding, Gavin Ha, Kenneth Lui, Jamie Rosner, David G. Huntsman, Carlos Caldas, Samuel A. Aparicio, and Sohrab P. Shah. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome biology*, 13(12):R124+, December 2012.

[4] Geert Berx and Frans Van Roy. The e-cadherin/catenin complex: an important gatekeeper in breast cancer tumorigenesis and malignant progression. *Breast Cancer Research*, 3(5):289, Jun 2001.

[5] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[6] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, October 2008.

[7] Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, June 2011.

[8] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404, 2012.

[9] Yan Cheng, Michael P. Holloway, Kevin Nguyen, Dilara McCauley, Yosef Landesman, Michael G. Kauffman, Sharon Shacham, and Rachel A. Altura. Xpo1 (crm1) inhibition represses stat3 activation to drive a survivin-dependent oncogenic switch in triple-negative breast cancer. *Molecular Cancer Therapeutics*, 13(3):675–686, 2014.

[10] Yan Cheng, Michael P. Holloway, Kevin Nguyen, Dilara McCauley, Yosef Landesman, Michael G. Kauffman, Sharon Shacham, and Rachel A. Altura. Xpo1 (crm1) inhibition represses stat3 activation to drive a survivin-dependent oncogenic switch in triple-negative breast cancer. *Molecular Cancer Therapeutics*, 13(3):675–686, 2014.

[11] Giovanni Ciriello, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research*, 22(2):398–406, 2012.

[12] Simona Constantinescu, Ewa Szczurek, Pejman Mohammadi, Jörg Rahnenführer, and Niko Beerenwinkel. Timex: a waiting time model for mutually exclusive cancer alterations. *Bioinformatics*, 32(7):968–975, 2015.

[13] Simona Cristea, Jack Kuipers, and Niko Beerenwinkel. pathtimex: Joint inference of mutually exclusive cancer pathways and their progression dynamics. *Journal of Computational Biology*, 24(6):603–615, 2017.

[14] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

[15] Teresa Davoli, Andrew Wei W. Xu, Kristen E. Mengwasser, Laura M. Sack, John C. Yoon, Peter J. Park, and Stephen J. Elledge. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, 155(4):948–962, November 2013.

[16] Megan Fabbro and Beric R Henderson. Regulation of tumor suppressors by nuclear-cytoplasmic shuttling. *Experimental Cell Research*, 282(2):59 – 69, 2003.

[17] Simon A. Forbes, Gurpreet Tang, Nidhi Bindal, Sally Bamford, Elisabeth Dawson, Charlotte Cole, Chai Yin Y. Kok, Mingming Jia, Rebecca Ewing, Andrew Menzies, Jon W. Teague, Michael R. Stratton, and P. Andrew Futreal. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic acids research*, 38(Database issue), January 2010.

[18] Deborah A. Freedman and Arnold J. Levine. Nuclear export is required for degradation of endogenous p53 by mdm2 and human papillomavirus e6. *Mol Cell Biol*, 18(12).

[19] Moritz Gerstung, Andrea Pellagatti, Luca Malcovati, Aristoteles Giagounidis, Matteo G Della Porta, Martin Jädersten, Hamid Dolatshad, Amit Verma, Nicholas CP Cross, Paresh Vyas, et al. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nature Communications*, 6, 2015.

[20] Enrico Glaab, Anaïs Baudot, Natalio Krasnogor, Reinhard Schneider, and Alfonso Valencia. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*, 28(18):i451–i457, September 2012.

[21] Heonjong Han, Hongseok Shim, Donghyun Shin, Jung Eun Shim, Yunhee Ko, Junha Shin, Hanhae Kim, Ara Cho, Eiru Kim, Tak Lee, et al. Trrust: a reference database of human transcriptional regulatory interactions. *Scientific Reports*, 5:11432, 2015.

[22] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, 2000.

[23] Matan Hofree, Hannah Carter, Jason F Kreisberg, Sourav Bandyopadhyay, Paul S Mischel, Stephen Friend, and Trey Ideker. Challenges in identifying cancer genes by analysis of exome sequencing data. *Nature Communications*, 7, 2016.

[24] Antoinette Hollestelle, Jord H. A. Nagel, Marcel Smid, Suzanne Lam, Fons Elstrodt, Marijke Wasielewski, Ser Sue Ng, Pim J. French, Justine K. Peeters, Marieke J. Rozendaal, Muhammad Riaz, Daphne G. Koopman, Timo L. M. ten Hagen, Bertie H. C. G. M. de Leeuw, Ellen C. Zwarthoff, Amina Teunisse, Peter J. van der Spek, Jan G. M. Klijn, Winand N. M. Dinjens, Stephen P. Ethier, Hans Clevers, Aart G. Jochemsen, Michael A. den Bakker, John A. Foekens, John W. M. Martens, and Mieke Schutte. Distinct gene mutation profiles among luminal-type and basal-type breast cancer cell lines. *Breast Cancer Research and Treatment*, 121(1):53–64, May 2010.

[25] Jack Hou and Jian Ma. DawnRank: Discovering Personalized Driver Genes in Cancer. *Genome Medicine*.

[26] J. L. Huret, S. Senon, A. Bernheim, and P. Dessen. An atlas on genes and chromosomes in oncology and haematology. *Cellular and molecular biology*, 50(7):805–807, November 2004.

[27] Yoo-Ah Kim, Dong-Yeon Cho, Phuong Dao, and Teresa M Przytycka. Memcover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics*, 31(12):i284–i292, 2015.

[28] GE Konecny and RS Kristeleit. Parp inhibitors for BRCA1/2-mutated and sporadic ovarian cancer: current practice and future directions. *British Journal of Cancer*, 115(10):1157, 2016.

[29] Mark D. Leiserson, Fabio Vandin, Hsin-Ta T. Wu, Jason R. Dobson, Jonathan V. Eldridge, Jacob L. Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael S. Lawrence, Abel Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory A. Ryslik, Nuria Lopez-Bigas, Gad Getz, Li Ding, and Benjamin J. Raphael. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 47(2):106–114, February 2015.

[30] Mark DM Leiserson, Dima Blokh, Roded Sharan, and Benjamin J Raphael. Simultaneous identification of multiple driver pathways in cancer. *PLoS Computational Biology*, 9(5):e1003054, 2013.

[31] Dilara McCauley, Yosef Landesman, William Senapedis, Trinayan Kashyap, Jean-Richard Saint-Martin, Louis Plamondon, Vincent Sandanayaka, Sharon Shechter, Doriana Froim, Raphael Nir, Jennifer Williams, Lynda Chin, Cyril Benes, Mansoor Raza Mirza, Michael Kauffman, and Sharon Shacham. Preclinical evaluation of selective inhibitors of nuclear export (sine) in basal-like breast cancer (blbc). *Journal of Clinical Oncology*, 30(15_suppl):1055–1055, 2012.

[32] Fernanda Milanezi, Silvia Carvalho, and Fernando C Schmitt. Egfr/her2 in breast cancer: a biological approach for molecular diagnosis and therapy. *Expert Review of Molecular Diagnostics*, 8(4):417–434, 2008.

[33] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1):29–34, January 1999.

[34] Kohno T Ogiwara H. CBP and p300 Histone Acetyltransferases Contribute to Homologous Recombination by Transcriptionally Activating the BRCA1 and RAD51 Genes. *PLoS ONE*, 7:12, December 2012.

[35] G. M. Pao, R. Janknecht, H. Ruffner, T. Hunter, and I. M. Verma. CBP/p300 interact with and function as transcriptional coactivators of BRCA1. *Proceedings of the National Academy of Sciences of the United States of America*, 97(3):1020–1025, February 2000.

[36] Joel S. Parker, Michael Mullins, Maggie C. Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, John F. Quackenbush, Inge J. Stijleman, Juan Palazzo, J. S. Marron, Andrew B. Nobel, Elaine Mardis, Torsten O. Nielsen, Matthew J. Ellis, Charles M. Perou, and Philip S. Bernard. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160–1167, March 2009.

[37] Evan O. Paull, Daniel E. Carlin, Mario Niepel, Peter K. Sorger, David Haussler, and Joshua M. Stuart. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics*, 29(21):2757–2764, November 2013.

[38] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A. L. Børresen-Dale, P. O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, August 2000.

[39] Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafou, Janos X. Binder, and Lars Juhl J. Jensen. DISEASES: text mining and data integration of disease-gene associations. *Methods*, 74:83–89, March 2015.

[40] Yair Pozniak, Nora Balint-Lahat, JanDaniel Rudolph, Cecilia Lindskog, Rotem Katzir, Camilla Avivi, Fredrik Pontn, Eytan Ruppin, Iris Barshack, and Tamar Geiger. System-wide clinical proteomics of breast cancer reveals global remodeling of tissue homeostasis. *Cell Systems*, 2(3):172 – 184, 2016.

[41] Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and H-Werner W. Mewes. CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic acids research*, 38(Database issue):D497–501, January 2010.

[42] Lao H. Saal, Karolina Holm, Matthew Maurer, Lorenzo Memeo, Tao Su, Xiaomei Wang, Jennifer S. Yu, Per-Olof O. Malmström, Mahesh Mansukhani, Jens Enoksson, Hanina Hibshoosh, Ake Borg, and Ramon Parsons. PIK3CA mutations correlate with hormone receptors, node metastasis, and ERBB2, and are mutually exclusive with PTEN loss in human breast carcinoma. *Cancer Research*, 65(7):2554–2559, April 2005.

[43] Martin H. Schaefer, Jean-Fred F. Fontaine, Arunachalam Vinayagam, Pablo Porras, Erich E. Wanker, and Miguel A. Andrade-Navarro. HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PLoS ONE*, 7(2), 2012.

[44] Min S. Song, Leonardo Salmena, and Pier P. Pandolfi. The functions and regulation of the PTEN tumour suppressor. *Nat Rev Mol Cell Biol*, 13(5):283–296, May 2012.

[45] Therese Sørlie, Charles M. Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B. Eisen, Matt van de Rijn, Stefanie S. Jeffrey, Thor Thorsen, Hanne Quist, John C. Matese, Patrick O. Brown, David Botstein, Per E. Lønning, and Anne-Lise Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874, September 2001.

[46] Chris Stark, Bobby-Joe J. Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34:D535–D539, 2006.

[47] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, October 2005.

[48] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian von Mering. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43:D447–D452, 2015.

[49] N. C. Turner and J. S. Reis-Filho. Basal-like breast cancer and the BRCA1 phenotype. *Oncogene*, 25(43):5846–5853, September 2006.

[50] N. C. Turner, J. S. Reis-Filho, A. M. Russell, R. J. Springall, K. Ryder, D. Steele, K. Savage, C. E. Gillett, F. C. Schmitt, A. Ashworth, and A. N. Tutt. BRCA1 dysfunction in sporadic basal-like breast cancer. *Oncogene*, 26(14):2126–2132, March 2007.

[51] Stefka Tyanova, Reidar Albrechtsen, Pauliina Kronqvist, Juergen Cox, Matthias Mann, and Tamar Geiger. Proteomic maps of breast cancer subtypes. *Nature Communications*, 7:10259, 2016.

[52] Igor Ulitsky, Adi Maron-Katz, Seagull Shavit, Dorit Sagir, Chaim Linhart, Ran Elkon, Amos Tanay, Roded Sharan, Yosef Shiloh, and Ron Shamir. Expander: from expression microarrays to networks and functions. *Nature Protocols*, 5(2):303–322, February 2010.

[53] UniProt Consortium. UniProt: a hub for protein information. *Nucleic acids research*, 43:D204–D212, 2015.

[54] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–1558, 2013.