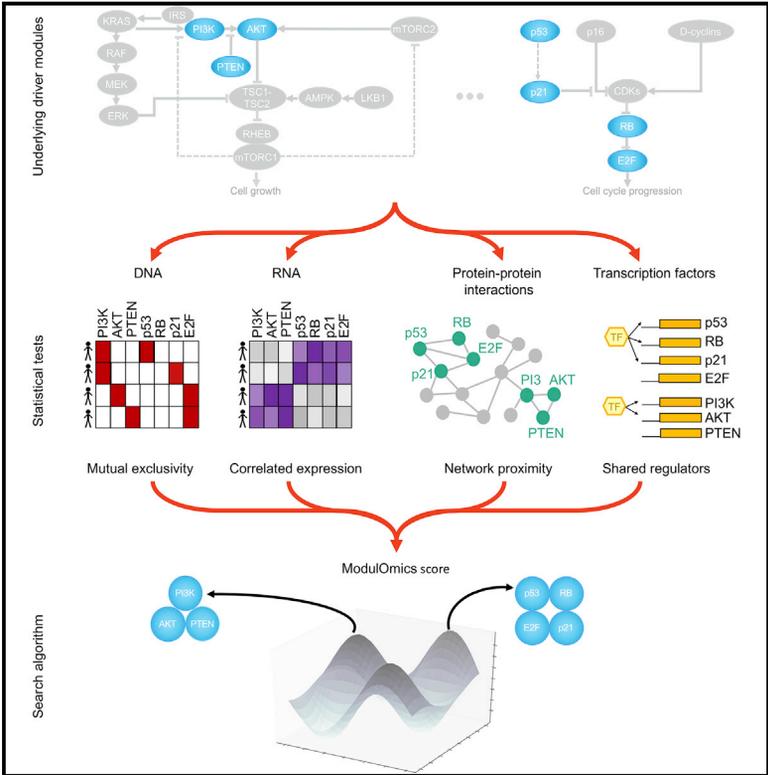


Cell Systems

Simultaneous Integration of Multi-omics Data Improves the Identification of Cancer Driver Modules

Graphical Abstract



Authors

Dana Silverbush, Simona Cristea, Gali Yanovich-Arad, Tamar Geiger, Niko Beerenwinkel, Roded Sharan

Correspondence

dsilverb@broadinstitute.org (D.S.), scristea@jimmy.harvard.edu (S.C.)

In Brief

Integrating different data types to answer biological questions is a challenging problem, which can, however, provide stronger insights than using each dataset separately. Modulomics is a statistical framework to integrate multiple omics data types and various statistical tests into one probabilistic model, with the aim of identifying functionally connected modules. It simultaneously (rather than sequentially) optimizes all tests and efficiently searches the large candidates space with a two-step optimization procedure. Across cancer types, Modulomics identifies key modules representing cancer-related mechanisms.

Highlights

- Modulomics is a framework for simultaneous omics data integration to identify modules
- Efficiently searches the space of candidate modules, using ILP and stochastic search
- Identifies pathway-enriched cancer driver modules based on DNA, RNA, and protein data
- Freely available as open-source code and webserver implementation



Simultaneous Integration of Multi-omics Data Improves the Identification of Cancer Driver Modules

Dana Silverbush,^{1,2,3,10,12,*} Simona Cristea,^{4,5,6,10,*} Gali Yanovich-Arad,⁷ Tamar Geiger,⁷ Niko Beerenwinkel,^{8,9,11} and Roded Sharan^{3,11}

¹Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

²Department of Pathology and Center for Cancer Research, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA

³Blavatnik School of Computer Science, Tel Aviv University, 69978 Tel Aviv, Israel

⁴Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215, USA

⁵Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

⁶Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA

⁷Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, 69978 Tel Aviv, Israel

⁸Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland

⁹SIB Swiss Institute of Bioinformatics, 4058 Basel, Switzerland

¹⁰These authors contributed equally

¹¹These authors contributed equally

¹²Lead Contact

*Correspondence: dsilverb@broadinstitute.org (D.S.), scristea@jimmy.harvard.edu (S.C.)

<https://doi.org/10.1016/j.cels.2019.04.005>

SUMMARY

The identification of molecular pathways driving cancer progression is a fundamental challenge in cancer research. Most approaches to address it are limited in the number of data types they employ and perform data integration in a sequential manner. Here, we describe ModulOmics, a method to *de novo* identify cancer driver pathways, or modules, by integrating protein-protein interactions, mutual exclusivity of mutations and copy number alterations, transcriptional coregulation, and RNA coexpression into a single probabilistic model. To efficiently search and score the large space of candidate modules, ModulOmics employs a two-step optimization procedure that combines integer linear programming with stochastic search. Applied across several cancer types, ModulOmics identifies highly functionally connected modules enriched with cancer driver genes, outperforming state-of-the-art methods and demonstrating the power of using multiple omics data types simultaneously. On breast cancer subtypes, ModulOmics proposes unexplored connections supported by an independent patient cohort and independent proteomic and phosphoproteomic datasets.

INTRODUCTION

Rapid advancements in sequencing technologies led to an unprecedented increase in the generation and availability of various types of high-resolution omics data. This wealth of

data requires appropriate computational models for filtering, distinguishing signals from noise, and interpretation. Key to the success of these models is the integration of different omics data types into single frameworks that can take advantage of orthogonal cellular views and construct a coherent and reliable picture of the process under study (Karr et al., 2012). A prime example for such a data-rich field in which integrative methods are called for is cancer genomics, where the use of multi-omics datasets recently emerged with the aim of better understanding cancer progression and treatment. Necessary for this goal is the identification of functionally connected groups of gene alterations that drive tumorigenesis, also termed driver modules or pathways, where genes alterations include single-nucleotide variants (SNVs), copy number alterations (CNAs), changes in the transcriptional activity of genes, and changes in protein concentration. Active driver modules contribute to triggering the hallmarks of cancer and confer fitness advantages to cancer cells (Vogelstein et al., 2013; Hanahan and Weinberg, 2000). Hence, their elucidation can substantially further our understanding of cancer development and inform optimal treatment design.

A successful single omics approach to identify cancer driver modules is the mutual exclusivity test, employed by tools such as TiMEx (Constantinescu et al., 2016), MultiDendrix (Leiserson et al., 2013), Comet (Wu et al., 2015), and others (Babur et al., 2015; Vandin et al., 2012; Jerby-Arnon et al., 2014). The biological foundation behind mutual exclusivity in cancer is that joint alterations in functionally connected genes provide similar or reduced fitness advantages than single alterations alone, such that selection favors cells with at most one alteration in a functionally connected group. Across patients, this process generates patterns of mutual exclusivity among alterations. An additional important omics test is the proximity of genetic alterations in the protein-protein interaction (PPI) network, explored by studies



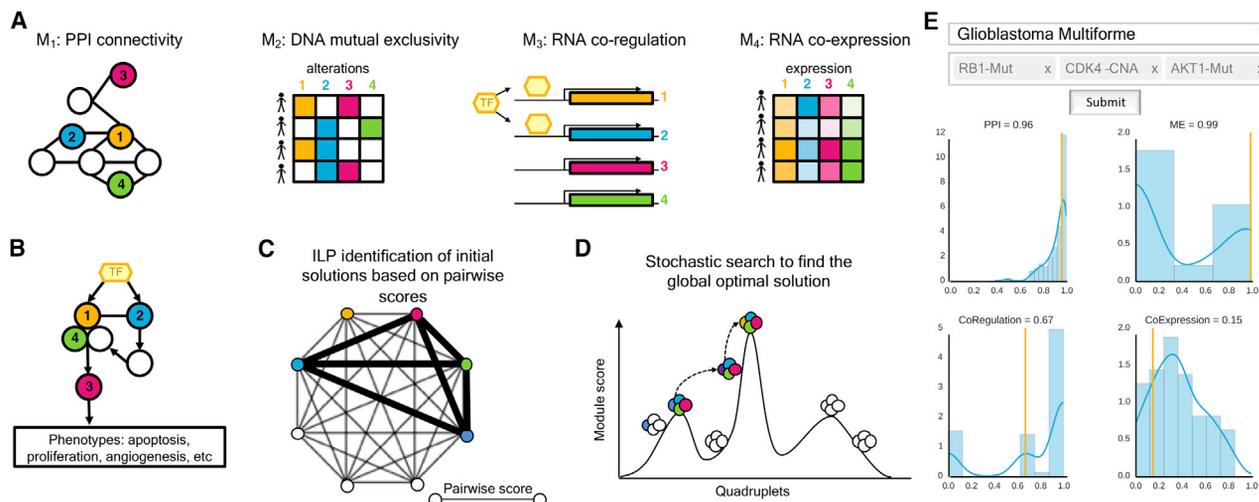


Figure 1. Overview of Modulomics

(A) Four different data types and corresponding models M_1 – M_4 (STAR Methods) contribute to the computation of the Modulomics score: PPI connectivity (protein level), mutual exclusivity (DNA level), transcriptional coregulation (regulatory connections and RNA level), and coexpression (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, coregulation, and coexpression.

(C) The ILP optimization identifies the modules with highest sum of pairwise Modulomics scores, computed as the average of the four scores corresponding to models M_1 – M_4 , further z-scored and normalized to [0,1].

(D) The stochastic search optimization uses the modules identified by the ILP, depicted in (C), as seeds, and aims to improve their scores by identifying a higher-scoring global solution. The space of initial solutions is clustered, and genes are exchanged between clusters to identify modules with high global scores. While the scores for models M_1 – M_4 , of the modules in (C) were approximated as average pairwise scores, here they are computed exactly for the entire module.

(E) The Modulomics webserver highlights any chosen gene set among the top 50 modules ranked by the single omics scores for each data type, on the basis of any of the three TCGA datasets analyzed in this study.

such as HotNet2 (Leiserson et al., 2015) or EnrichNet (Glaab et al., 2012).

Approaches to integrate different data sources generally use the omics tests in a sequential manner, such that each additional data type further refines the groups inferred on the basis of the previous types. These sequential frameworks have the advantage of efficiently pruning the huge search space, however, at the cost of losing sensitivity. As the pruning is done based on only some of the omics data types, the discarded low-scoring modules could have been informative w.r.t. a data type not yet considered. For example, the tool MEMo (Ciriello et al., 2012) extracts candidate modules based on PPI pairwise connections and then scores them using mutual exclusivity, while MEMCover (Kim et al., 2015) integrates pairwise mutual exclusivity and PPI scores between genetic alterations. Few approaches also include gene expression, such as TieDIE (Paull et al., 2013), which uses a PPI-proximity test once for DNA alterations and once for gene expression, optimizing each individually, and then crossing the two to find one large subnetwork. In contrast to the sequential approach however, simultaneously optimizing multi-omics data types via a single objective function can increase the sensitivity and specificity of module identification by yielding relevant modules simultaneously across multiple levels of genetic information.

Here, we describe Modulomics, a method for the *de novo* identification of cancer driver modules from multi-omics data. Modulomics integrates proximity in a PPI network, mutual exclu-

sivity of DNA alterations (SNVs and CNAs), and RNA level coregulation and coexpression, into a single probabilistic framework, by simultaneously optimizing over all four model components, one for each data type (Figure 1). Modulomics overcomes the computational challenge of searching the huge space of potential modules by performing a two-step optimization procedure that combines integer linear programming (ILP) with stochastic search. We apply Modulomics to three large-scale TCGA datasets, breast cancer (Cancer Genome Atlas Network, 2012), glioblastoma (GBM) (Cancer Genome Atlas Research Network, 2008), and ovarian cancer (Cancer Genome Atlas Research Network, 2011), and show that it accurately identifies known cancer driver genes and pathways. Modulomics outperforms state-of-the-art tools for driver modules identification, namely the DNA-centric method TIMEx (Constantinescu et al., 2016), the PPI-based method HotNet2 (Leiserson et al., 2015) and the DNA and PPI integration-based method MEMCover (Kim et al., 2015). Comparisons of Modulomics to alternative strategies in which either only subsets of the four omics data types are used, or the data types are integrated sequentially, demonstrate that the strength of the approach stems from scoring candidate modules simultaneously across all data types.

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 (Pozniak et al., 2016; Tyanova et al., 2016). In the most aggressive breast

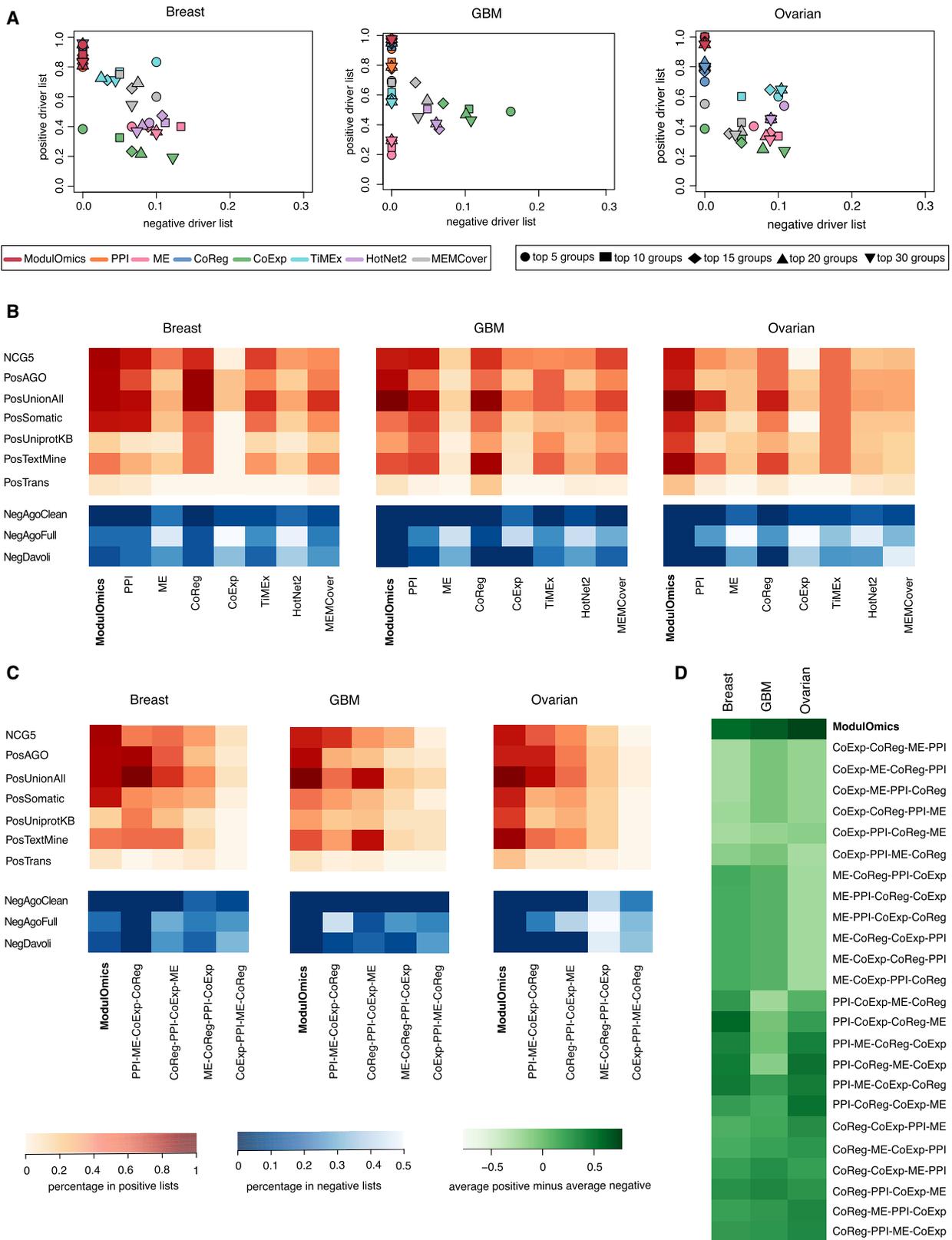


Figure 2. The Driver Modules Inferred by ModulOmics Are Enriched with Cancer Driver Genes

(A) Detailed driver and nondriver enrichment scores for the positive driver list *PosUnionAll* and the negative driver list *NegAGOClean*, for the top scoring 5, 10, 15, 20, and 30 modules. ME stands for mutual exclusivity, CoReg for coregulation, and CoExp for coexpression scores of single omics data types. The enrichment

(legend continued on next page)

cancer subtype, triple negative (TN), we identify functional connections among multiple downregulated tumor suppressors, including *TP53*, *BRCA1*, *RB1*, and *PTEN*. These patterns are also supported by reverse-phase protein array (RPPA) data (Cancer Genome Atlas Network, 2012). 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 noncanonical 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 three TCGA datasets processed in this study (<http://anat.cs.tau.ac.il/ModulOmicsServer/>).

RESULTS

ModulOmics identifies driver gene modules on the basis of DNA and RNA profiles of cancer samples, integrated with a PPI network and known regulatory connections. Each candidate module is scored according to the degree of mutual exclusivity among DNA alterations across a 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 coregulated by a common active transcription factor. As the number of candidate groups grows exponentially with maximal group size, ModulOmics employs 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 (STAR Methods).

We used ModulOmics to identify driver modules of fixed size, consisting of 2–4 gene members, from three TCGA cancer datasets (Tables S1–S3). No single omics data type dominated the ModulOmics score for any given module size. The identified modules of different sizes were pooled and ranked according to their scores. Each module was assigned a p value by comparing its score to the scores of the top 100 modules obtained from 10 instances of shuffled omics data sets that preserved the characteristics of each data type, such as mutation frequency and network degree distribution (STAR Methods). All top 30 modules identified by ModulOmics were significant (Bonferroni-corrected $p < 0.05$).

Driver Modules Are Enriched with Cancer Drivers

To assess the performance of ModulOmics, we calculated the enrichment of the top modules with known driver genes (positive controls) and known nondriver genes (negative controls). To this

end, we used the gene lists introduced in Hofree et al. (2016), compiled from the Network of Cancer Genes (NCG) (An et al., 2016), Cancer Gene Census version 73 (CGC) (Forbes et al., 2010), the Atlas of Genetics and Cytogenetics in Oncology and Hematology (AGO) (Huret et al., 2004), UniprotKB (UniProt Consortium, 2015), DISEASES (Pletscher-Frankild et al., 2015), and MSigDB (Subramanian et al., 2005) (Data S1 and S2; STAR Methods). 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. ModulOmics outperformed the four single omics approaches, as well as MEMCover, HotNet2, and TiMEx, when evaluating a variable number (top 5, 10, 15, 20, or 30) of the highest scoring modules of any size (Figure 2A and Table S4), or when separately evaluating modules of fixed sizes (Figure S1).

When focusing on the top 10 modules, ModulOmics consistently outperformed the other methods across the seven positive and three negative control lists tested (Figure 2B). 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. Complementary, 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 Davoli et al., 2013 (*NegDavoli*), consisting of between 3,272 and 9,457 known nondriver genes. Among the competing methods, PPI-based and coregulation-based scorings exhibited good performances, MEMCover, HotNet2, and TiMEx performed well only for certain group sizes, while coexpression and mutual exclusivity generally performed poorly on both positive and negative control metrics. TieDIE (Paull et al., 2013), aiming to infer one large subnetwork, identified a single module of 300 genes, containing less than 30% known driver genes and more than 10% known nondriver genes; this led us to exclude TieDIE from further comparisons. To assess whether ModulOmics identified higher-order connections in the data, as opposed to only the structure of the data (s.a., protein degree for PPI, and mutation frequency for DNA), we ran ModulOmics on shuffled data. ModulOmics and each individual omics data type performed better on the real data, as compared to the shuffled alternatives (Figure S2).

We next evaluated the specific contribution of each omics data type to driver genes enrichment by computing reduced versions of the ModulOmics score, each time with a single data type removed. We found that, in 90% of the evaluated cases (92% of the positive control lists and 86% of the negative lists), integrating all four omics data sources led to

was calculated as the average fraction of gene members in each module that are also part of the control lists. The modules were ranked by their score, regardless of their sizes. Table S4 displays the scores for Figure 2A.

(B) Average driver enrichment (red heatmaps) and nondriver enrichment (blue heatmaps) across the top 10 modules. *NCG5*, *PosAGO*, *PosUnionAll*, *PosSomatic*, *PosUniprotKB*, *PosTextMine*, *PosTrans* are the positive control lists, while *NegAgoClean*, *NegAgoFull* and *NegDavoli* are the negative control ones (STAR Methods).

(C) Average driver and nondriver enrichment scores for the best performing sequential alternatives of ModulOmics, per starting omics data type.

(D) Single enrichment scores per omics data type, computed as the difference between the average positive and the average negative scores, across all lists, for ModulOmics and all its sequential alternatives. Across all panels, ModulOmics generally outperforms competing methods, as well as simplified or sequential alternatives.

higher enrichment, as compared to using subsets of three omics. Nevertheless, the performance of ModulOmics remained fairly robust when using only three data types, suggesting that the method can also be applied in cases when one data source is missing (Figure S3).

To directly compare the simultaneous optimization employed by ModulOmics to sequential approaches, we built sequential alternatives of ModulOmics, which use one omics data type to find initial candidate modules, and each sequential omics data to further refine the identified modules (STAR Methods). We assessed all possible orderings of data types by both comparing the average driver and nondriver enrichment scores for all control lists individually (Figures 2C and S4) and by assigning a single score per omics data across all lists, as the difference between the average positive and the average negative list scores (Figure 2D). Based on these evaluations, we found that using simultaneous optimization consistently outperformed the sequential alternatives of ModulOmics.

One of the features of ModulOmics is that each gene can participate in multiple driver pathways, hence the reported modules often overlapped (Figure S5A). 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 via different modules. In order to assess the performance of ModulOmics also in the absence of overlap among groups, we repeated the driver enrichment evaluation for the first 20 unique genes in the highest-ranking modules. Consistent with previous results, ModulOmics outperformed competitive methods (Figure S5B).

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 the Kyoto Encyclopedia of Genes and Genomes, KEGG (Ogata et al., 1999). The top 10 ModulOmics modules were significantly enriched with an average of 8 pathways per module, whereas the top 10 modules generated by shuffling the data (STAR Methods) were only significantly enriched with an average of 3.6 pathways. ModulOmics identified key cancer-related pathways, such as *pathways in cancer*, across all three cancer types (Figure 3A). 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*. The set of enriched pathways also included pathways characterizing other cancer types, indicating shared mechanisms among malignancies.

To further quantify pathway enrichment performance, we counted the number of pathways significantly enriched (Bonferroni-corrected $p \leq 0.05$) in each of the top 5, 10, and 15 modules (Figure 3B) and computed their average enrichment factor with Expander (Ulitsky et al., 2010) (Figure 3C; STAR Methods), as well as the ratio of top modules enriched with at least one pathway (Figure 3D). Overall, ModulOmics identified modules enriched with more general pathways and cancer-related pathways, as well as more modules enriched with at least one pathway, than the three competing methods. A high percentage of genes identified by ModulOmics participated in known KEGG

pathways, reaching an average of 77% across all three cancer types, compared to 43% as identified by MEMCover, 39% by HotNet2, and 10% by TiMEx (Table S5). Since coregulation was the best performing single omics data type for cancer driver genes enrichment, we further assessed the top coregulation modules in terms of their diversity and pathway enrichment. We found that those modules were less diverse and mostly enriched with well-studied functional connections, whereas ModulOmics was able to extend beyond these connections (Figure S6).

Using all four data sources improved the identification of functionally coherent modules in 92% of the tested cases, as compared to using subsets of three omics data (Figure S7). As in the case of driver genes enrichment, optimizing across all omics data simultaneously, rather than sequentially, provided an advantage in terms of pathway enrichment (Figure 3E). Certain omics data, such as PPI and coregulation, were expected to be tightly coupled to pathway enrichment, since they are based on physical interactions. Indeed, the sequential versions starting with these two data types were the best performing sequential alternatives, with the sequences PPI-CoExp-ME-CoReg and CoReg-PPI-ME-CoExp performing better than ModulOmics in ovarian cancer pathway enrichment (but not cancer driver genes enrichment), yet not for the other cancers. Overall, ModulOmics was the only consistently high performing approach across all three cancer types.

Driver Modules in Breast Cancer Subtypes Recapitulate Known Mechanisms and Suggest Unexplored Functionalities

Next, we applied ModulOmics on molecularly defined subtypes of breast cancer, classified using the mRNA PAM50 classification (Parker et al., 2009) into basal (125 patients), Her2 (61), luminal A (364), and luminal B (174) (Table S6). Across all subtypes, the genes in the top 20 ModulOmics 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 (Turner and Reis-Filho, 2006; Turner et al., 2007), *CDH1* for luminal A and B (Hollestelle et al., 2010), *MAP3K1* for luminal B (Cancer Genome Atlas Network, 2012), and *EGFR* for Her2 (Milanezi et al., 2008). An alternative strategy to ModulOmics for identifying relevant drivers would have been selecting genes with highest SNV or CNA alteration frequencies (Vogelstein et al., 2013). However, in that case, a substantial portion of the enriched genes 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.

A detailed PPI network view of the genes identified by ModulOmics revealed *TP53* as a key player in tumor progression for all subtypes, while subtype-specific key players included *EGFR* for Her2 and *BRCA1* for basal (Karaayvaz et al., 2018) (Figure 4B). The network view highlighted the higher rate of

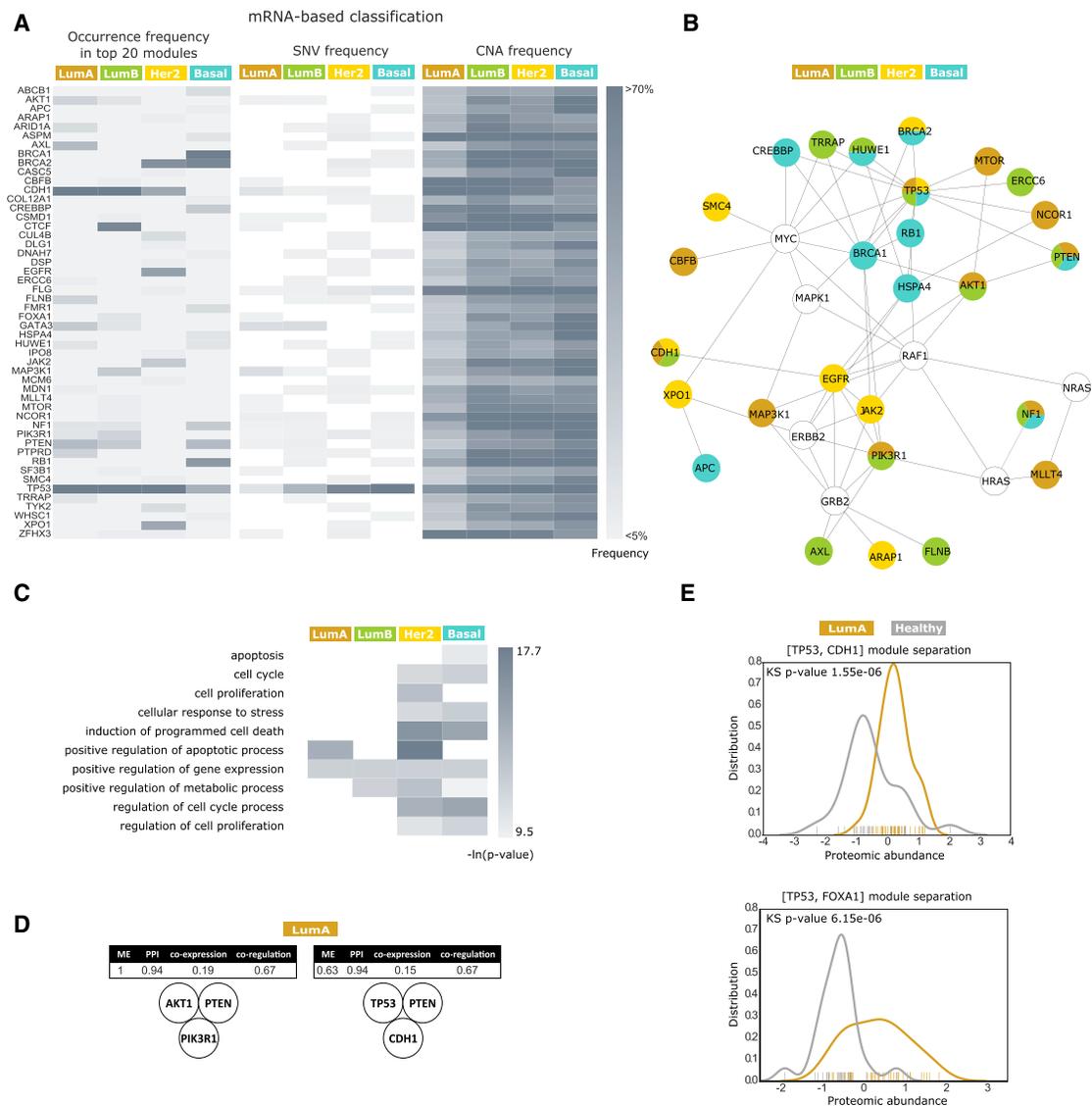


Figure 4. Modules Inferred in mRNA-Classified Breast Cancer Subtypes Reflect Various Levels of Subtype Aggressiveness and Separate Cancerous Form Healthy Tissues

(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. This gene pool is 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.

(B) Detailed PPI network view of the subset of genes in (A) that are either known drivers, or part of KEGG pathways. The displayed protein interactions underline functional associations, such as the role of PI3K pathway in luminal A.

(C) Selected list of significantly enriched GO pathways across the top 20 modules (Figure S8 displays the full list), reflecting the aggressiveness of the basal and Her2 subtypes, compared to luminal A and luminal B. Enrichment hypergeometric p values were computed with Expander (Ulitsky et al., 2010). 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 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 *CTNBN1*, and *FOXA1* gain is measured directly.

(Figures 4C and S8). These results capture 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.

Some of the highest-ranking modules may merit further experimental investigation. For example, the highest-ranking module

in the basal subtype consisted of *RB1*, *BRCA1*, *NF1*, and *CREBBP*. Since both *BRCA1* and its activator *CREBBP* (Pao et al., 2000) are involved in DNA repair, this module potentially reflects the altered DNA damage repair mechanism specific to basal tumors (Ogiwara and Kohno, 2012) and may hint at the clinical implications of using the *CREBBP* inhibitor in *BRCA1* patients, similarly to *PARP1* (Konecny and Kristeleit, 2016). The

top Her2 modules were characterized by the recurrent appearance of the nuclear export gene *XPO1* together with *TP53*, one of its known targets (Freedman and Levine, 1998; Cheng et al., 2014). The role of *XPO1* in tumor progression was previously investigated in a preclinical context of TN treatment (Cheng et al., 2014; McCauley et al., 2012), and here we suggest it may also play a role in Her2. Finally, 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 noncanonical role (Figure 4D). The canonical module *PTEN*, *AKT1*, *PIK3R1* recapitulates the known mutual exclusivity pattern of mutations within the PI3K pathway (Saal et al., 2005). In contrast, the module suggesting the noncanonical role (*PTEN*, *CDH1*, *TP53*) supports 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 (Song et al., 2012). Indeed, according to the database Transcriptional Regulatory Relationships Unraveled by Sentence-Based Text, TRRUST (Han et al., 2015), *PTEN*, and *CDH1* are coregulated by two common transcription factors, namely *STAT3* and *NFKB1*.

In order to further explore the clinical relevance of the top ModulOmics modules, we examined how well they can distinguish healthy tissues from cancerous ones in an independent omics data source. To this end, we used a recently published proteomics dataset consisting of 62 samples of luminal A and healthy tissues (Pozniak et al., 2016; Tyanova et al., 2016), and focused on the two highest scoring luminal A modules: *TP53* and *CDH1*, and *FOXA1* and *TP53*. These top 2 modules significantly separated the luminal A cancerous tissues from the healthy ones, when evaluating single module scores per patient, computed by averaging the proteomic abundance of its members (p values $1.6e^{-06}$ and $6.2e^{-06}$, respectively, Kolmogorov Smirnov (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 separate the 2 types of tissue as well (p values 0.065, 0.054, and 0.69, respectively, KS test). Similarly, random groups of the same size did not significantly separate the tissues (p value 0.14, averaged over 1,000 random groups generated by sampling subsets of proteins from the proteomics dataset, KS test).

An alternative way to study breast cancer is by stratifying patients according to immunohistochemistry results assessing the HER2, ER, and PR receptors. To this end, we separated the patient cohort into subtypes, using the TCGA classification, as follows: TN (116 patients), Her2-enriched (30), luminal A (477), and luminal B (88), and used ModulOmics to infer modules for each subtype (Table S7). Similarly to the mRNA-based classification, the genes of the highest scoring 20 modules (Figure S9A) 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 enrichment p value 0.01). 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 (*PTEN* and

BRCA1 in luminal B, and *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 S9B). These findings suggest a TN-specific concerted downregulation of multiple tumor suppressors, namely *TP53*, *BRCA1*, *RB1*, and *PTEN*, potentially contributing to the poor prognosis of this subtype. Taken together, these results imply that the level of mutual exclusivity in these tumor suppressors might indicate the aggressiveness of the tumor subtype (Perou et al., 2000; Sørlie et al., 2001; Cancer Genome Atlas Network, 2012; Curtis et al., 2012).

To further evaluate the functional connectivity among these tumor suppressors (*PTEN*, *BRCA1*, *RB1*, and *TP53*), we used an independent omics data source, RPPA (Cancer Genome Atlas Network, 2012). In general, evaluating protein measurements limits large-scale analyses, since loss of function can lead to missing data, requiring the identification of downstream-regulated proteins that can serve as surrogates. We therefore used the following surrogates: *CDK1* to account for *TP53* loss, *CYCLIN B1* for *BRCA1* loss, and phosphorylated *AKT*, which is suppressed by *PTEN*, to account for *PTEN*. Indeed, all these tumor suppressors were downregulated in the TN samples, while their anticorrelated surrogates were upregulated (Figure S9C). Using these genes, we were able to separate TN from the other subtypes (p -value $1.2e^{-16}$, KS test, for the surrogate tumor promoter proteins and p -value $2.9e^{-09}$, KS test, for the tumor suppressors, Figure S9D).

The less-frequent genes of the top ModulOmics modules (genes remaining after filtering the 10 most frequent genes in the top modules) may reveal previously unexplored functional connections (Tables S8 and S9). For example, *HUWE1* coappears in luminal B modules with *TP53*, a known target of its E3-ligase activity (Chen et al., 2005), and with the *AXL* receptor tyrosine kinase. While *AXL* is well-known for its role in metastatic breast cancer (Zhang et al., 2008; Gjerdrum et al., 2010), the role of *HUWE1* in cell motility is sparsely described in the literature (Vaughan et al., 2015), and its connection with *AXL* represents a potentially previously unexplored mechanism. Another interesting module related to cell motility is identified in TN and consists of *NID1*, *MUC16*, *PTEN*, and *VCAN*. Except *PTEN*, all three other genes are associated with the extracellular matrix (Scholler and Urban, 2007; Wight, 2002; Lee et al., 2006; Fox et al., 1991). While *MUC16* is a known biomarker for ovarian cancer (Scholler and Urban, 2007), its role in breast cancer is less characterized. Further investigating these modules experimentally may further our understanding of tumorigenesis.

DISCUSSION

ModulOmics is a method for the *de novo* identification of cancer driver pathways, based on the integration of connectivity within PPI networks, mutual exclusivity among SNV and CNA alterations, transcriptional coregulation, and RNA coexpression, 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 have demonstrated the ability of

ModulOmics to identify modules enriched with known cancer driver genes and pathways in three large-scale TCGA datasets: breast cancer, GBM, and ovarian cancer. We further investigated breast cancer subtypes and found that some of the highest scoring modules are known to be involved in cancer-related molecular mechanisms, while others suggest lesser known functionalities and may merit further experimental investigation. We evaluated these results using an independent patient cohort and independent proteomic and phosphoproteomic datasets. In addition, we showed 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 implemented as freely available and flexible open-source software, such that any of the four omics data types employed here can be excluded or replaced with new sources of evidence. The webserver implementation can be used to evaluate the ModulOmics score of any user-defined gene set, on the basis of the three TCGA datasets analyzed here. This application can be useful in situations when candidate gene sets were proposed from separate biological or computational analyses. Since ModulOmics integrates independent sources of information, newly added data can also originate from different patient cohorts. Each data type is assessed on the basis of its own tailored statistical test, with variable weight. By tuning the weights for each omics type, the ModulOmics optimization function can be further refined to identify groups representative of particular functional phenotypes. One such example is inferring optimal weights by training a classifier on known modules to identify groups enriched with specific GO terms. The inferred optimal weights can be further used when running ModulOmics on new data.

The modeling and computational framework introduced here represents a conceptual advancement over previous data integration methods, by simultaneously (rather than sequentially) optimizing the scores of the different omics data types. Throughout the paper, we have justified the superiority of simultaneous optimization, as compared not only to existing methods, but also to alternative sequential implementations of ModulOmics. The power of the proposed framework stems from its generality, namely (1) integrating multiple omics data simultaneously, not confined to a specific order; and (2) using 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). Lastly, ModulOmics may also be applied to data integration problems outside cancer genomics, where functionally connected modules are relevant. One such example is identifying functionally related protein complexes, which can be done by applying the PPI, coregulation and coexpression tests on data derived from healthy tissues. Taken together, our results indicate that the ModulOmics scores are informative in identifying biological connections, making the tool broadly applicable.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- [KEY RESOURCES TABLE](#)
- [CONTACT FOR REAGENT AND RESOURCE SHARING](#)

● METHOD DETAILS

- Model
- PPI Connectivity
- Mutual Exclusivity
- Co-regulation
- Co-expression
- Optimization Procedure
- ILP
- Stochastic Search
- Sensitivity Analyses

● QUANTIFICATION AND STATISTICAL ANALYSIS

- Alternative Data Integration Approaches
- Evaluation Metrics

● DATA AND SOFTWARE AVAILABILITY

- Data Availability
- Software Availability

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.1016/j.cels.2019.04.005>.

ACKNOWLEDGMENTS

The authors would like to thank Tovi Almozilino for building the ModulOmics frontend website, Dr. Mariya Mardamshina for helping with regards to breast cancer clinical classification, Dr. Yoo-Ah Kim for information on MEMCover, and Dr. Roni Wilentzik-Muller for her feedback on the manuscript. S.C. is supported by the Swiss National Science Foundation, project number P2EZP2_175139. D.S. is supported by the Israeli Ministry of Science, Technology and the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. N.B. is supported by ERC Synergy grant 609883.

AUTHOR CONTRIBUTION

S.C., D.S., R.S., and N.B. designed the study. S.C. and D.S. designed, implemented, and tested the model. T.G. and G.Y.-A. collected and analyzed independent data for evaluation and advised biological and clinical aspects of the project. D.S., S.C., and G.Y.-A. analyzed the results. S.C., D.S., G.Y.-A., and R.S. wrote the manuscript. All authors contributed to editing and finalizing the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 5, 2018

Revised: November 13, 2018

Accepted: April 19, 2019

Published: May 15, 2019

WEB RESOURCES

ModulOmics GitHub, <https://github.com/danasilv/ModulOmics>

ModulOmics, <http://anat.cs.tau.ac.il/ModulOmicsServer>

REFERENCES

- An, O., Dall'Olio, G.M., Mourikis, T.P., and Ciccarelli, F.D. (2016). Ncg 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. *Nucleic Acids Res.* *44*, D992–D999.
- Babur, Ö., Gönen, M., Aksoy, B.A., Schultz, N., Ciriello, G., Sander, C., and Demir, E. (2015). Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biol.* *16*, 45.

- Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70.
- Cancer Genome Atlas Research Network. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068.
- Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615.
- Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401–404.
- Chen, D., Kon, N., Li, M., Zhang, W., Qin, J., and Gu, W. (2005). Arf-bp1/mule is a critical mediator of the arf tumor suppressor. *Cell* 121, 1071–1083.
- Cheng, Y., Holloway, M.P., Nguyen, K., McCauley, D., Landesman, Y., Kauffman, M.G., Shacham, S., and Altura, R.A. (2014). Xpo1 (crm1) inhibition represses stat3 activation to drive a survivin-dependent oncogenic switch in triple-negative breast cancer. *Mol. Cancer Ther.* 13, 675–686.
- Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* 22, 398–406.
- Constantinescu, S., Szczurek, E., Mohammadi, P., Rahnenführer, J., and Beerenwinkel, N. (2016). Timex: a waiting time model for mutually exclusive cancer alterations. *Bioinformatics* 32, 968–975.
- Cristea, S., Kuipers, J., and Beerenwinkel, N. (2017). Pathtimex: joint inference of mutually exclusive cancer pathways and their progression dynamics. *J. Comput. Biol.* 24, 603–615.
- Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346–352.
- Davoli, T., Xu, A.W.W., Mengwasser, K.E., Sack, L.M., Yoon, J.C., Park, P.J., and Elledge, S.J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* 155, 948–962.
- Forbes, S.A., Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., Kok, C.Y., Jia, M., Ewing, R., Menzies, A., et al. (2010). COSMIC (the catalogue of somatic mutations in cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res.* 38, D652–D657.
- Fox, J.W., Mayer, U., Nischt, R., Aumailley, M., Reinhardt, D., Wiedemann, H., Mann, K., Timpl, R., Krieg, T., and Engel, J. (1991). Recombinant nidogen consists of three globular domains and mediates binding of laminin to collagen type iv. *EMBO J.* 10, 3137–3146.
- Freedman, D.A., and Levine, A.J. (1998). Nuclear export is required for degradation of endogenous p53 by mdm2 and human papillomavirus e6. *Mol. Cell Biol.* 18, 7288–7293.
- Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J.R., and Mann, M. (2010). Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* 7, 383–385.
- Gjerdum, C., Tiron, C., Høiby, T., Stefansson, I., Haugen, H., Sandal, T., Collett, K., Li, S., McCormack, E., Gjertsen, B.T., et al. (2010). Axl is an essential epithelial-to-mesenchymal transition-induced regulator of breast cancer metastasis and patient survival. *Proc. Natl. Acad. Sci. USA* 107, 1124–1129.
- Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., and Valencia, A. (2012). EnrichNet: network-based gene set enrichment analysis. *Bioinformatics* 28, i451–i457.
- Han, H., Shim, H., Shin, D., Shim, J.E., Ko, Y., Shin, J., Kim, H., Cho, A., Kim, E., Lee, T., et al. (2015). Trustrust: a reference database of human transcriptional regulatory interactions. *Sci. Rep.* 5, 11432.
- Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. *Cell* 100, 57–70.
- Hofree, M., Carter, H., Kreisberg, J.F., Bandyopadhyay, S., Mischel, P.S., Friend, S., and Ideker, T. (2016). Challenges in identifying cancer genes by analysis of exome sequencing data. *Nat. Commun.* 7, 12096.
- Hollestelle, A., Nagel, J.H.A., Smid, M., Lam, S., Elstrodt, F., Wasielewski, M., Ng, S.S., French, P.J., Peeters, J.K., Rozendaal, M.J., et al. (2010). Distinct gene mutation profiles among luminal-type and basal-type breast cancer cell lines. *Breast Cancer Res. Treat.* 121, 53–64.
- Huret, J.L., Senon, S., Bernheim, A., and Dessen, P. (2004). An atlas on genes and chromosomes in oncology and haematology. *Cell. Mol. Biol.* 50, 805–807.
- Jerby-Arnon, L., Pfetzer, N., Waldman, Y.Y., McGarry, L., James, D., Shanks, E., Seashore-Ludlow, B., Weinstock, A., Geiger, T., Clemons, P.A., et al. (2014). Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell* 158, 1199–1209.
- Karaayvaz, M., Cristea, S., Gillespie, S.M., Patel, A.P., Mylvaganam, R., Luo, C.C., Specht, M.C., Bernstein, B.E., Michor, F., and Ellisen, L.W. (2018). Unravelling subclonal heterogeneity and aggressive disease states in tnbc through single-cell rna-seq. *Nat. Commun.* 9, 3588.
- Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., Assad-Garcia, N., Glass, J.I., and Covert, M.W. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell* 150, 389–401.
- Kim, Y.A., Cho, D.Y., Dao, P., and Przytycka, T.M. (2015). Memcover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics* 31, i284–i292.
- Konecny, G.E., and Kristeleit, R.S. (2016). Parp inhibitors for BRCA1/2-mutated and sporadic ovarian cancer: current practice and future directions. *Br. J. Cancer* 115, 1157–1173.
- Lee, H.K., Seo, I.A., Park, H.K., and Park, H.T. (2006). Identification of the basement membrane protein nidogen as a candidate ligand for tumor endothelial marker 7 in vitro and in vivo. *FEBS Lett.* 580, 2253–2257.
- Leiserson, M.D., Blokh, D., Sharan, R., and Raphael, B.J. (2013). Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput. Biol.* 9, e1003054.
- Leiserson, M.D., Vandin, F., Wu, H.-T.T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47, 106–114.
- McCauley, D., Landesman, Y., Senapedis, W., Kashyap, T., Saint-Martin, J.-R., Plamondon, L., Sandanayaka, V., Shechter, S., Froim, D., Nir, R., et al. (2012). Preclinical evaluation of selective inhibitors of nuclear export (sine) in basal-like breast cancer (blbc). *J. Clin. Oncol.* 30, 1055.
- Milanezi, F., Carvalho, S., and Schmitt, F.C. (2008). Egfr/her2 in breast cancer: a biological approach for molecular diagnosis and therapy. *Expert Rev. Mol. Diagn.* 8, 417–434.
- Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34.
- Ogiwara, H., and Kohno, T. (2012). CBP and p300 histone acetyltransferases contribute to homologous recombination by transcriptionally activating the BRCA1 and RAD51 genes. *PLoS One* 7, e25810.
- Pao, G.M., Janknecht, R., Ruffner, H., Hunter, T., and Verma, I.M. (2000). CBP/p300 interact with and function as transcriptional coactivators of BRCA1. *Proc. Natl. Acad. Sci. USA* 97, 1020–1025.
- Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* 27, 1160–1167.
- Paul, E.O., Carlin, D.E., Niepel, M., Sorger, P.K., Haussler, D., and Stuart, J.M. (2013). Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* 29, 2757–2764.
- Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. *Nature* 406, 747–752.
- Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J.X., and Jensen, L.J.J. (2015). DISEASES: text mining and data integration of disease-gene associations. *Methods* 74, 83–89.
- Pozniak, Y., Balint-Lahat, N., Rudolph, J.D., Lindskog, C., Katzir, R., Avivi, C., Pontén, F., Ruppén, E., Barshack, I., and Geiger, T. (2016). System-wide clinical proteomics of breast cancer reveals global remodeling of tissue homeostasis. *Cell Syst.* 2, 172–184.

- Saal, L.H., Holm, K., Maurer, M., Memeo, L., Su, T., Wang, X., Yu, J.S., Malmström, P.-O.O., Mansukhani, M., Enoksson, J., et al. (2005). PIK3CA mutations correlate with hormone receptors, node metastasis, and ERBB2, and are mutually exclusive with PTEN loss in human breast carcinoma. *Cancer Res.* *65*, 2554–2559.
- Schaefer, M.H., Fontaine, J.-F.F., Vinayagam, A., Porras, P., Wanker, E.E., and Andrade-Navarro, M.A. (2012). HIPPIE: integrating protein interaction networks with experiment based quality scores. *PLoS One* *7*, e31826.
- Scholler, N., and Urban, N. (2007). Ca125 in ovarian cancer. *Biomark. Med.* *1*, 513–523.
- Song, M.S., Salmena, L., and Pandolfi, P.P. (2012). The functions and regulation of the PTEN tumour suppressor. *Nat. Rev. Mol. Cell Biol.* *13*, 283–296.
- Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* *98*, 10869–10874.
- Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* *102*, 15545–15550.
- Turner, N.C., and Reis-Filho, J.S. (2006). Basal-like breast cancer and the BRCA1 phenotype. *Oncogene* *25*, 5846–5853.
- Turner, N.C., Reis-Filho, J.S., Russell, A.M., Springall, R.J., Ryder, K., Steele, D., Savage, K., Gillett, C.E., Schmitt, F.C., Ashworth, A., et al. (2007). BRCA1 dysfunction in sporadic basal-like breast cancer. *Oncogene* *26*, 2126–2132.
- Tyanova, S., Albrechtsen, R., Kronqvist, P., Cox, J., Mann, M., and Geiger, T. (2016). Proteomic maps of breast cancer subtypes. *Nat. Commun.* *7*, 10259.
- Ulitsky, I., Maron-Katz, A., Shavit, S., Sagir, D., Linhart, C., Elkon, R., Tanay, A., Sharan, R., Shiloh, Y., and Shamir, R. (2010). Expander: from expression microarrays to networks and functions. *Nat. Protoc.* *5*, 303–322.
- UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic Acids Res.* *43*, D204–D212.
- Vandin, F., Upfal, E., and Raphael, B.J. (2012). De novo discovery of mutated driver pathways in cancer. *Genome Res.* *22*, 375–385.
- Vaughan, L., Tan, C.T., Chapman, A., Nonaka, D., Mack, N.A., Smith, D., Booton, R., Hurlstone, A.F., and Malliri, A. (2015). Huwe1 ubiquitylates and degrades the rac activator tiam1 promoting cell-cell adhesion disassembly, migration, and invasion. *Cell Rep.* *10*, 88–102.
- Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., and Kinzler, K.W. (2013). Cancer genome landscapes. *Science* *339*, 1546–1558.
- Wight, T.N. (2002). Versican: a versatile extracellular matrix proteoglycan in cell biology. *Curr. Opin. Cell Biol.* *14*, 617–623.
- Wu, H.-T., Leiserson, M.D., Vandin, F., and Raphael, B.J. (2015). Comet: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Cancer Res.* *75*, 1936–1943.
- Zhang, Y.X., Knyazev, P.G., Cheburkin, Y.V., Sharma, K., Knyazev, Y.P., Örfi, L., Szabadkai, I., Daub, H., Kéri, G., and Ullrich, A. (2008). Axl is a potential target for therapeutic intervention in breast cancer progression. *Cancer Res.* *68*, 1905–1915.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
TCGA	National cancer institute (NIH)	https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
Hippie	Andrade Lab	http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/
TRRUST	Netbio Lab	https://www.grnpedia.org/trrust/
Breast mass-spectrometry dataset	Pozniak et al. (2016) and Tyanova et al. (2016)	N/A
Software and Algorithms		
ModulOmics	this paper	https://github.com/danasilv/ModulOmics http://anat.cs.tau.ac.il/ModulOmicsServer/
HotNet2	Raphael Lab	http://compbio.cs.brown.edu/projects/hotnet2/
TiMEx	Beerenwinkel Lab	https://github.com/cbg-ethz/TiMEx
TieDie	Stuart Lab	https://sysbiowiki.soe.ucsc.edu/tiedie
MEMCover	Przytycka Lab	https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/index.cgi#memcover
Expander	Shamir Lab	http://acgt.cs.tau.ac.il/expander/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Dana Silverbush (dsilverb@broadinstitute.org).

METHOD DETAILS

Model

Given a set $G = \{G_1, \dots, G_n\}$ of genes and a collection $M = \{M_1, \dots, M_m\}$ 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 $P(G|M_k)$. Each of these m scores represents how strongly functionally connected the genes in G are, under different models:

$$S_G = \frac{1}{m} \sum_{k=1}^m P(G|M_k) \quad (\text{Equation 1})$$

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 (Schaefer et al., 2012)) and its associated interaction probabilities, we define, for each pair of genes (G_i, G_j) , $\text{con}(G_i, G_j)$ 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 $\text{con}(G_i, G_j)$ 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:

$$P(G|M_1) = \sum_{c \in \mathcal{C}(G)} \prod_{(G_i, G_j) \in E(c)} \text{con}(G_i, G_j) \prod_{(G_i, G_j) \in E(\mathcal{G}(G)) \setminus E(c)} (1 - \text{con}(G_i, G_j)) \quad (\text{Equation 2})$$

where $\mathcal{C}(G)$ 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 (Constantinescu et al., 2016) and pathTiMEx (Cristea et al., 2017), $P(G|M_2)$ is computed as the degree of mutual exclusivity of the set G , as follows:

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

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 Constantinescu et al., 2016 and Cristea et al., 2017.

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 $P(G|M_3)$ is defined as the fraction of genes in G which are co-regulated by at least one common active transcription factor,

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

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 (Figure S10A). 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,

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

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 choose $k = 2$ and $q = 4$, i.e., the 2nd quartile, based on the amount of information they provide (Figure S10B). 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, 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_{G_i, G_j}^k and equals $P((G_i, G_j)|M_k)$, 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:

$$w_G = \sum_{k=1}^m \sum_{G_i, G_j \in G} w_{G_i, G_j}^k \quad (\text{Equation 6})$$

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 $E_{ij} = \{w_{V_i V_j} | V_i, V_j \in V\}$. 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_{ij} in G :

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

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

and the objective function:

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

under the constraints:

$$E_{(i,j)} - V_{(i)} \leq 0 \quad (\text{Equation 10})$$

$$E_{(i,j)} - V_{(j)} \leq 0 \quad (\text{Equation 11})$$

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

$$\sum_{V_i \in V} V_{(i)} = K \quad (\text{Equation 13})$$

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

$\forall V_i, V_j \in V, i < j$. 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:

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

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. 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.

In a separate analysis, we show that the stochastic optimization step improves the functional connectivity of the identified modules on real data (Figure S11).

Sensitivity Analyses

Table S10 shows the default parameters used by ModulOmics. The recommended values are based on scanning a range of possible thresholds, and choosing the thresholds which resulted in a substantial amount of retained information for all the studied cohorts (Figure S10 and Table S11).

To evaluate robustness, we vary the parameters of the stochastic search as follows: 300 initial module seeds instead of the default 200, 15 clusters instead of the default 10, and 7 top results reported by each cluster instead of the default 5, and assess the following metrics: i) the repetition of gene connections, i.e. gene pairs co-residing in the same module, and ii) the repetition of the gene pool reported by the top modules, regardless of which module they belong to (Figure S12).

QUANTIFICATION AND STATISTICAL ANALYSIS

Alternative Data Integration Approaches

HotNet2

As recommended by the authors of HotNet2 (Leiserson et al., 2015), we use SNVs and CNAs as the prior set, and assign the initial score of each genetic alteration to be its alteration frequency in the data. We apply HotNet2 on the same PPI network we use with ModulOmics. To assign a p value, we use 100 permuted networks as background. To calculate a hypergeometric score for pathway enrichment with Expander (Ulitsky et al., 2010), we use modules of up to size 7, since larger modules are more likely to be unspecific from a mechanistic perspective. HotNet2 identifies only 2 modules of size 2 in ovarian cancer, only 2 modules of size 3 in GBM, and no modules in ovarian cancer, as well as no modules of size 4 in any of the three cancer types; these results do not meet the comparison criteria of top 5 modules and are not included in Figure S1.

TiME_x

We run TiME_x (Constantinescu et al., 2016) with default parameters on the same binary datasets used as input for ModulOmics, consisting of binary SNV and CNA alterations. We consider as significant all resulting mutually exclusive groups with Bonferroni-corrected p value <0.05. Even though TiME_x and the mutual exclusivity score of ModulOmics are based on the same probabilistic model, the search strategy is different for the two methods. Therefore, TiME_x and the simplified single omics data approach of mutual exclusivity (ME) are expected to identify different modules in the data. TiME_x identifies only 3 modules of size 3 in ovarian cancer, 1 module of size 4 in breast cancer, and no size 4 modules in GBM or ovarian cancer; these results do not meet the comparison criteria of top 5 modules and are not included in Figure S1.

TieDIE

ModulOmics substantially differs from TieDIE in its goal, as TieDIE detects a single subnetwork, whereas ModulOmics identifies multiple modules. We run TieDIE on the three cancer cohorts used in this study, using the same PPI and regulatory networks as for ModulOmics.

MEMCover

We run MEMCover (Kim et al., 2015) on each of the three cancer types, with default parameters. We use the same PPI as for ModulOmics, with an edge weight threshold of 0.4. The resulting modules are separated by size and further ranked by their average coverage in each cohort. MEMCover identifies only 3 modules of size 3 in breast cancer, and 1 module in GBM and ovarian cancer, as well as only 1 module of size 4 in any of the three cancer types; these results do not meet the comparison criteria of top 5 modules and are not included in Figure S1.

Sequential ModulOmics

To compare the ModulOmics simultaneous optimization scheme to sequential optimizations, we build a sequential version of ModulOmics in which each omics type is used to further refine the results obtained by the other omics types. The first omics type is used to filter 200 initial driver modules, starting with 300 seeds from pairwise approximations. Each additional omics type is then used to refine the results: each of the candidate modules is scored using the new omics type, the old and new scores are averaged, and the distribution of scores across modules is z-scored in order to combine the new omics type with the previous ones. Each additional omics type refines the modules by, each time, removing the 50 least informative groups (i.e. the set of modules are refined to the top 150, then 100 and lastly 50 modules). All parameters (Table S10) are kept as in the simultaneous optimization version.

Shuffled Controls

Each of the four data sources is shuffled to create a random benchmark: i) the edges in the PPI network are shuffled while preserving the degree of each node (shuffled PPI); ii) the mutated genes matrix is shuffled while preserving the frequency of mutations and the number of mutations in a given sample (shuffled ME); iii) the edges in the transcription factors network are shuffled while preserving the degree of each node (shuffled CoReg); iv) the gene expression profiles are shuffled by randomly switching expression profiles among genes (shuffled CoExp). We run ModulOmics using each of the shuffled data structures as described above, and compare the results with the single omics. In addition, we also run ModulOmics by integrating all four shuffled controls (shuffled ModulOmics).

Random Modules

We generate random modules by running ModulOmics on the shuffled data and considering the top 100 yielded modules. We further recalculate the ModulOmics scores for these modules using the un-shuffled data. To assign a p value for a ModulOmics module, we calculate where its score falls in the random modules distribution.

Evaluation Metrics

Known Cancer Genes Enrichment

As positive controls, we use the following lists, introduced in Hofree et al., 2016 and available as supplementary resources to this paper: i) The Cancer Gene Census (CGC) version 73 (*PosSomatic* and *PosTrans*), a set of 569 genes manually curated by The Sanger Institute, which have alterations in somatic and germline SNVs, CNVs and translocations; ii) UniprotKB (*UniProt Consortium, 2015*) (*PosUniprotKB*), a manually curated database of 412 functional proteins, classified as proto-oncogene, oncogene and tumour suppressor gene; iii) a query of DISEASES (*Pletscher-Frankild et al., 2015*) (*PosTextMine*), a database of disease-gene associations extracted mainly from text-mining, which consists of 711 genes associated with cancer; and iv) The Atlas of Genetics and Cytogenetics in Oncology and Hematology (*PosAGO*) (*Huret et al., 2004*), a list of 1,430 cancer genes manually curated by a collaborative effort spanning multiple centers. *PosUnionAll* is the union of all these positive control lists. In addition, we use the Network of Cancer Genes

(NCG5) (An et al., 2016), a manually curated list consisting of 1,571 protein-coding cancer driver genes compiled by The Sanger Institute. The gene members of the two shortest CGC lists, germline SNVs (38 genes) and CNVs (15 genes) are not shown here, since they were not identified in any high scoring module by any of the tested methods.

As negative controls, we use the following lists introduced in Hofree et al., 2016: i) a list derived from AGO (Huret et al., 2004) consisting of 9,457 genes that have no evidence of association with cancer (*NegAgoFull*); ii) a conservative version of the negative AGO list (*NegAGOClean*), created by filtering genes that are part of any cancer-related pathway from the MSigDB database (Subramanian et al., 2005), resulting in 3,272 genes, and iii) a list of known non-driver genes introduced in Davoli et al., 2013 (*NegDavoli*).

Pathway Enrichment

We use two statistical tests of module-pathway intersection, as proposed by the Expander software (Ulitsky et al., 2010): i) a hypergeometric enrichment test to calculate the occurrence probability of the intersection of a module with a random pathway when randomly drawing from all protein-coding genes, and ii) an enrichment factor designed to ease the bias towards larger modules. The enrichment factor is defined as the ratio between the sizes of the intersection of each module and each pathway and the intersection of that pathway and the set of all background genes (all protein-coding genes), normalized by the sizes of the module and background respectively:

$$\frac{|\text{module} \cap \text{pathway}|}{|\text{pathway} \cap \text{background genes}|} \times \frac{|\text{background genes}|}{|\text{module}|}$$

DATA AND SOFTWARE AVAILABILITY

Data Availability

ModulOmics identifies driver modules on the basis on DNA and RNA cancer patient data, integrated with a PPI network and known regulatory connections. In this study, we use DNA and RNA cancer patient data retrieved from the TCGA project (Cancer Genome Atlas Network, 2012; Cancer Genome Atlas Research Network, 2008; Cancer Genome Atlas Research Network, 2011) for GBM, breast cancer and ovarian cancer, downloaded from the cBio portal (Cerami et al., 2012). For breast cancer, we only use SNVs, while for GBM and ovarian cancer, we use both SNVs and CNAs. For breast cancer, we additionally evaluate each subtype individually, based on two classifications: 1) receptor classification based on immunohistochemistry of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (Her2); and 2) mRNA classification based on PAM50 signature. The PPI network used is Hippie (Schaefer et al., 2012), containing 238,165 physical interactions. The regulatory connections were downloaded from the database TRRUST Han et al., 2015 containing 8,908 Transcription Factor (TF)-target regulatory pairs of 821 human TFs. To evaluate functional connections in the TN breast cancer subtype, we use the RPPA data published by the TCGA (Cancer Genome Atlas Network, 2012). To evaluate the power of the highest ranking modules in distinguishing healthy tissues from cancerous ones, we use an independent publicly available mass-spectrometry dataset containing 62 samples of Luminal A and healthy tissue (Pozniak et al., 2016; Tyanova et al., 2016), where protein quantification was done with Super-SILAC (Geiger et al., 2010).

Software Availability

ModulOmics is freely available as an open-source R code for the identification of cancer driver modules from a cohort of cancer patients. The github link (<https://github.com/danasilv/ModulOmics>) contains the R code, a manual, a Snakefile and an example input for testing. To run ModulOmics on real data, the user may use the static data from the link <http://anat.cs.tau.ac.il/ModulOmicsServer/>, or re-create it from their favorite sources. The static data provided include:

1. Regulatory connections, based here on the TRRUST database
2. Shortest paths in the Hippie PPI network

The dynamic part of the data used in this study, namely genetic alterations and gene expression patient profiles, is available in TCGA (<https://cancergenome.nih.gov/>).

ModulOmics is also available as a webserver implementation for the evaluation of any set of genes of interest, using the three TCGA datasets processed in this study: <http://anat.cs.tau.ac.il/ModulOmicsServer/>.

All default parameters used to run ModulOmics in this study are available in Table S10.