Title

Integration of Tumor Genomic Data with Cell Lines Using Multi-dimensional Network Modules Improves Cancer Pharmacogenomics

Permalink

https://escholarship.org/uc/item/1xc1f54v

Journal

Cell Systems, 7(5)

ISSN

2405-4712

Authors

Webber, James T
Kaushik, Swati
Bandyopadhyay, Sourav

Publication Date

2018-11-01

DOI

10.1016/j.cels.2018.10.001

Peer reviewed

Published in final edited form as:
Cell Syst. 2018 November 28; 7(5): 526–536.e6. doi:10.1016/j.cels.2018.10.001.

Integration of tumor genomic data with cell lines using multi-dimensional network modules improves cancer pharmacogenomics

James T. Webber[1], **Swati Kaushik**[1], and **Sourav Bandyopadhyay**[1,*]
[1]Department of Bioengineering and Therapeutic Sciences, Institute for Computational Health Sciences, Helen Diller Family Comprehensive Cancer Center. University of California, San Francisco. San Francisco, CA, USA.

Abstract

Leveraging insights from genomic studies of patient tumors is limited by the discordance between these tumors and the cell line models used for functional studies. We integrate -omics datasets using functional networks to identify gene modules reflecting variation between tumors and show that the structure of these modules can be evaluated in cell lines to discover clinically relevant biomarkers of therapeutic responses. Applied to breast cancer, we identify 219 gene modules that capture recurrent alterations, subtype patients and quantitate various cell types within the tumor microenvironment. Comparison of modules between tumors and cell lines reveals that many modules composed primarily of gene expression and methylation are poorly preserved. In contrast, preserved modules are highly predictive of drug responses in a manner that is robust and clinically relevant. This work addresses a fundamental challenge in pharmacogenomics that can only be overcome by the joint analysis of patient and cell line data.

Graphical Abstract

[*] Lead contact: sourav.bandyopadhyay@ucsf.edu

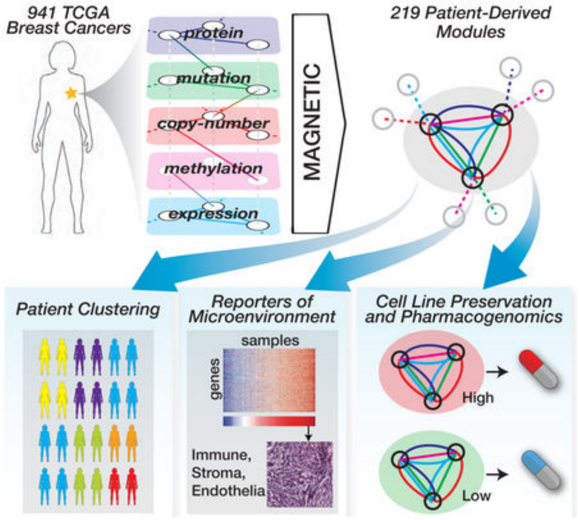
AUTHOR CONTRIBUTIONS

J.T.W. and S.B wrote the manuscript and conceived the method. J.T.W. performed the computational analysis. S.K. provided additional analysis and key datasets, and reviewed the manuscript.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

DECLARATION OF INTERESTS

S.B. is a consultant to Ideaya Biosciences and has research funding from Pfizer and Clovis Oncology.


941 TCGA Breast Cancers
protein
mutation
copy-number
methylation
expression
MAGNETIC
219 Patient-Derived Modules
Patient Clustering
Reporters of Microenvironment
samples
genes
Immune, Stroma, Endothelia
Cell Line Preservation and Pharmacogenomics
High
Low

eTOC

Webber et al. develop an approach to compress data from the multi-omic characterization of hundreds of cancers into a small set of gene modules using a modular analysis of gene networks in cancer, or MAGNETIC. Modules can be readily compared with cell lines revealing components of the tumor micro-environment and used to improve the development of biomarkers of therapeutic responses based on cell line collections.

INTRODUCTION

Cancer is caused by molecular aberrations that lead to de-regulation of cellular networks. Large-scale tumor sequencing efforts have sought to identify these molecular events in common cancer types across thousands of patients (Cancer Genome Atlas Research et al., 2013). For example, The Cancer Genome Atlas (TCGA) cataloged the molecular aberrations present in human tumors by systematically measuring somatic mutations, copy-number alterations, gene methylation, transcriptomes and proteomes in over 11,000 tumors (Hoadley et al., 2014; Network, 2012). While these studies have been highly successful, the precise function and therapeutic relevance of most of these molecular events is unclear. As a complement to tumor analysis, cell lines are established models to study cancer biology and connect genomic alterations to drug response. Recently several groups have profiled large collections of cell lines by measuring baseline molecular features as well as sensitivity to a wide range of compounds, and used machine learning methods to identify predictors of response (Barretina et al., 2012; Basu et al., 2013; Costello et al., 2014; Daemen et al., 2013; Garnett et al., 2012; Neve et al., 2006). By connecting disease biology with model systems, linking data from the comprehensive molecular characterization of thousands of tumor specimens and the pharmacogenomic profiling of thousands of cell lines has the potential to greatly inform the development of cancer therapeutics. The development of drug response predictors using cell line data has been met with mixed results and none have been validated in independent clinical cohorts (Borst and Wessels, 2010; Wang et al., 2013).

One of the challenges in integrating human tumor and cell line data stems from the lack of approaches to evaluate the similarities and differences that exist between genomic features in cell lines and patient samples (Goodspeed et al., 2016). Because of the complex and multifactorial nature of the disease, such a comparison must consider the complement of somatic mutations, copy-number, gene expression, methylation and proteomic changes found in tumors as well as the molecular features whose variation across patients is inextricably linked, necessitating a modular analysis. Differences between cell lines and patient samples stem from the altered biology and tumor environment between the two (Borst and Wessels, 2010; Gillet et al., 2013) and a failure to account for these differences are likely responsible for failures when translating cell line based biomarkers to human tumors. Computational approaches that address these factors in the biomarker discovery process have not yet been developed.

We hypothesized that a multi-omic, integrative analysis of tumor genomes would facilitate the biological interpretation of molecular programs in cancer and serve as a basis for the comparison of these programs in cancer cell lines. We present a method called Modular Analysis of Genomic NETworks In Cancer (MAGNETIC) that integrates data across molecular profiling platforms by performing functional network analysis to identify gene modules that are preserved in both cancer patients and in cell lines. As proof of concept, we applied this approach to breast cancer, where proliferative responses to the majority (69-75%) of drugs are not predictable based on standard molecular subtypes therefore requiring new methods to connect tumor genomics with therapy (Daemen et al., 2013; Heiser et al., 2011). Using profiling data from breast cancers in TCGA, we identify 219 gene modules in breast cancer that capture molecular features that are closely linked across samples and enriched for physically interacting proteins. Using MAGNETIC we uncover biological processes involved in breast cancer and present a quantitative landscape of micro-environmental factors in breast cancers. This modular analysis enables the discovery of molecular features that are limited to tumorigenesis *in-vivo* and reveals that a significant amount of gene expression and methylation data from human tumors reflects signals derived from the tumor environment that are not reflected in cell lines. Building on this finding, we show that modules preserved in cell lines can act as accurate biomarkers that are more robust than standard approaches because they are more reflective of a tumor context. This work reveals an approach for the integrative analysis of molecular programs within human tumors and provides a powerful and clinically relevant way to connect tumor genotype to therapy.

RESULTS

Pathway signals are embedded across –omics platforms.

We sought to measure the extent to which the interactions found in protein interaction databases carry information for interpreting correlations between molecular changes detected in cancer. We started with molecular profiles of primary breast cancers from TCGA obtained using gene expression, DNA methylation, copy-number alteration, exome sequencing and Reverse Phase Protein Array (RPPA) platforms covering 941 patients (Network, 2012). After normalization, we constructed a correlation network by comparing all pairs of gene features across patients both within and between platforms. To measure concordance with known interactions we compared the distribution of these correlations to a compendium of 60,194 functional protein-protein interactions (PPIs) from the HumanNet database (Lee et al., 2011).

We observed strong agreement of molecular signals between protein pairs known to interact both within and across –omics platforms in TCGA. For example, two related ubiquitin specific peptidases USP32 and USP6 were not only highly co-expressed (r=0.95, $p=1.6\times10^{-259}$) but the copy-number of USP32 was also strongly predictive of expression of USP6 (r=0.73, $p=1.9\times10^{-75}$) (Figure 1a,b). We observed that gene pairs where the copy number of one was highly correlated (r>0.75) with the expression of another were 100-fold more likely to interact than at random (Figure S1a). In another example, the expression of LCK was highly anti-correlated with the methylation of its substrate LAT (r=−0.68, $p=3.7\times10^{-66}$) (Figure 1c,d). Gene pairs with the most anti-correlation (r<−0.75) between the expression of one and the methylation of another were over 100-fold more likely to interact (Figure S1a). Other comparisons within the same platform revealed a similar trend. For example, gene pairs that were highly co-expressed (r>0.7) were 40-fold more likely to interact (Figure S1a). These results mirror previous observations that co-expressed genes are more likely to physically interact (Stuart et al., 2003; von Mering et al., 2002) and that gene copy number and expression data are often concordant within tumors (Curtis et al., 2012). Here we expand on these findings by showing that correlation networks between diverse data types encode interaction information that can be used to identify new functional relationships in cancer.

A Modular Analysis of Genomic NETworks In Cancer (MAGNETIC).

Based on the observation that molecular correlations are reflective of protein interactions we sought to integrate these gene linkages into a single network. The multi-platform correlation network can be visualized as an undirected graph with multiple layers, each representing a data type (Figure 1e). Also known as a multigraph, this structure allows for multiple edges to connect nodes and includes 15 different edge types (all pairs of five data types, with repetition). Since we observed that molecular correlations were often predictive of PPIs, but to a different degree based on the type of comparison, we next re-scored the network edges by mapping a given correlation value to the log-likelihood ratio (LLR) of it reflecting a known interaction in HumanNet and merged them into a single network (see STAR Methods). We chose HumanNet for integration of datasets as it represents the largest functional network available although this approach was robust with respect to protein-protein interaction network used. This integrated network was enriched for known PPIs, including PPIs not found in HumanNet, more so than any single correlation network type alone (Figure S1b-f), highlighting the value of integrating heterogeneous datasets to identify pathways.

We next used a network-clustering algorithm based on repeated random walks to find densely interconnected modules within the integrated network (Rosvall et al., 2009). Clustering of this network at different score cutoffs revealed an optimal overlap with known PPIs at an LLR>3 (Figure S1g-i). At this cutoff, we identified a total of 219 modules with a median number of 18 genes per module (Figure S2a-d, Table S1). We next assigned patient specific module scores for each sample in the TCGA cohort using a weighted sum of normalized values for each molecular entity representing genes in a module (see STAR Methods, Table S2). 84.5% (185/219) of the modules contained edges in the LLR>3 network which reflect data from multiple sources. For example, the module containing ERBB2 (HER2) consisted of 25 genes and included 60 co-expression and 44 co-copy number variation edges, reflecting the coordinated amplification and expression of genes in the HER2 amplicon (Figure 1f,g, Figure S2c). Module 37, based largely on gene expression, reflected the status of ER and was enriched for its direct transcriptional targets ($p=3.1\times10^{-6}$, via hypergeometric test) (Figure 1h,i) (Bhat-Nakshatri et al., 2008). In addition, consensus clustering of modules across TCGA samples revealed 10 patient groups that were significantly associated with the molecular subtypes called by PAM50, as well as the receptor status of ER, PR and HER2 (Parker et al., 2009) (Figure S2e). Highly significant molecular associations with receptor status were also seen via scoring modules and clustering of 1,966 samples from the METABRIC breast cancer cohort (Figure S2f, Table S3) (Curtis et al., 2012). Sixteen percent of modules were significantly enriched for a particular GO, KEGG, or Reactome pathway or function and 44% contained known PPIs (Table S1). Modules recovered known highly mutated genes in breast cancers such as TP53 (#181), PIK3CA (#9), GATA3 (#37), and MAP3K1 (#5) as well as published gene signatures associated with proliferation, stromal involvement and angiogenesis (Paik et al., 2004; Winslow et al., 2015) (Table S4). In comparison to other approaches that integrate multiple datasets or use single datasets to identify related gene sets, MAGNETIC modules were more enriched for known interactions, including interactions not found in HumanNet (Figure S3a-e). Simulation analysis revealed that >95% of the same modules could be identified using half the number of samples and 71% overlapped when using only 10% of samples (210/219 and 155/219 respectively) (Figure S3f,g). Independent module discovery in the METABRIC cohort resulted in approximately 80% of the same modules identified. Therefore, our approach integrates data from diverse platforms covering 48,093 unique gene and protein features into a set of 219 interaction-enriched modules that decompose complex tumor genomes into independent molecular signatures.

We next sought to detect whether particular aspects of tumor biology could be detected using MAGNETIC. We observed a module (#27) that was largely based on expression and the genes in this module were highly enriched for genes whose promoters are marked by histone H3 Lysine 27 trimethylation (H3K27me3) in ES cells, including SOX1, NEUROG1, NEUROG3, FOXB1 and FOXD3 (Figure S3h,i, Table S5) (Liberzon et al., 2011; Subramanian et al., 2005). H3K27me3 is deposited by the Polycomb Repressive Complex 2 (PRC2) histone methyltransferase complex consisting of EZH1/2, SUZ12, RBBP7/4, and EED (Kuzmichev et al., 2002; Margueron and Reinberg, 2011). This modification is associated with the repression of genes involved in development and differentiation, but its role in breast cancer is almost completely unknown. To determine if genes in this module are regulated via H3K27me3 in breast cancer, we obtained H3K27me3 ChIP-seq data for three breast cancer cell lines: SUM159PT (score=0.11), T47D (0.37) and MCF7 (1.39) (Su et al., 2015). We first verified that the module score reflected the expression level of genes within the module (Figure S3j). We found that the promoters of genes in module 27 were more likely to be marked by H3K27me3 than other genes (p<0.01 in all cell lines, Figure S3k) and that the level of H3K27me3 occupancy for each gene was correlated with the module score (r=−0.12, p= 0.004), indicating that histone modification state regulates the activity of this module (Table S5). Given the significance of PRC2-driven epigenetic regulation in other cancers (Kim and Roberts, 2016), determining the importance of regulation of this module during tumorigenesis could yield new biological insights and targets for breast cancer.

Identification of preserved gene modules in cancer cell lines.

We next determined whether analysis of MAGNETIC modules could be used to systematically separate tumor cell specific variation that might be more accurately modeled by cell lines in culture and therefore more useful in pharmacogenomics efforts. We therefore sought to exclude modules that reflect processes not captured in cell culture due to alterations in the growth environment. Since each module represents a set of relationships in an underlying molecular network, we reasoned that modules which maintain these relationships in both tumors and cell lines are likely to reflect shared biology (Figure 2a). We investigated whether the molecular features that are linked our TCGA-derived network were also linked in a panel of 82 molecularly characterized breast cancer cell lines (Daemen et al., 2013). For each module we calculated an edge preservation score reflecting the average increase in the pairwise correlation between molecular features when calculated across cell lines as compared to background (see Supplementary Methods). Based on the approximately bimodal distribution of preservation scores we chose a threshold of module preservation that identified 59 lowly preserved modules to simplify further analysis (Figure 2b). Since most histologic aspects of tumor growth cannot be recapitulated in cell culture, we next asked if the activity of these lowly preserved modules was associated with specific pathologic features present in the tumor. Across breast tumors, we found that the scores of these 59 modules were significantly more correlated with levels of necrosis and normal-cell infiltration in comparison to modules that were highly preserved (Figure 2c). In addition to pathological assessment, lowly preserved modules were also associated with tumor impurity measured by computational prediction of normal cell contamination using gene expression, copy number and methylation data from tumor specimens (Figure 2d) (Aran et al., 2015). Therefore, lowly preserved modules capture genomic features that tend to reflect aspects of tumorigenesis that are not reflected in cell culture.

The differences in module preservation led us to ask whether preserved modules were more likely to reflect data from certain types of –omics platforms. We found that modules largely associated with gene expression or methylation were much less likely to be preserved (Figure 2e,f). For example, all modules that were composed of more than 75% co-expression edges were not preserved (Figure 2g). In support, gene networks based solely on co-expression or co-methylation in TCGA samples only marginally overlapped with such networks from breast cancer cell lines whereas networks that include copy-number variation (CNV) as a component were much more robust (Figure S4a,b). CNV-expression edges were more common in highly versus lowly preserve modules, suggesting that expression data is most useful when tied to genomic events that are tumor specific. Based on the 2,596 genes present in lowly-preserved modules we estimate that the expression and methylation of at least 13% of the genome reflects differences in biology between human tumor samples and cell lines. Our analysis suggests that breast cancer biomarkers in cell lines tied to events such as copy-number variation and mutation are more likely to yield clinically translatable biomarkers because they are the most robust to issues such as tumor purity.

Modules as reporters of the tumor microenvironment.

Since lowly preserved modules were correlated with pathologic assessment of normal cells, we next asked if the influence of the microenvironment is apparent in the activity specific modules. Module #3 was highly enriched for genes related to the immune system and was not preserved (GO enrichment, $p=7.4\times10^{-156}$, Table S1) (Figure 2e). This module included the expression of B cell marker CD19, T cell marker CD4, and NK cell marker IL15, and protein abundance of the T cell kinase LCK (Figure 3a) (Heng et al., 2008; Palacios and Weiss, 2004). Comparing with gene expression data from purified cell populations we found that genes in the module were highly expressed in NK cells, ILC1 cells, T cells, monocytes and macrophages (Figure 3b, Table S6) (Heng et al., 2008). Furthermore, we found that the activity of this module was related to general lymphocyte infiltration scores based on pathological assessment in the TCGA which was reproduced in METABRIC, an independent cohort of over 2,000 patients (Figure 3c) (Curtis et al., 2012).

Other modules were reflective of non-tumor cell types as well. Module #12 enriched for a stromal gene signature and included many collagens associated with the extra-cellular matrix (ECM) ($p=6.9\times10^{-46}$, Figure 3d) (Winslow et al., 2015). This module was correlated with pathological assessments of stromal cells and indicated tumors with the presence of significant ECM involvement (Figure 3e,f). Module #16 was enriched for genes involved in vascularization and highly expressed in endothelial cells, including F10 and KDR/VEGFR2 (Figure 3g, Table S1). Tumors scoring highly for module #16 were highly vascularized and less necrotic based on pathological assessment (Figure 3h,i). Both of these modules were not preserved whereas modules reflecting the activity of oncogenes recurrently amplified in tumor cells such as HER2 and MYC were preserved (Figure 2e). Taken together, these data indicate that gene modules can be used to report on the components of the microenvironment present in a tissue sample.

A module-drug network identifies determinants of drug sensitivity that are robust to differences between patients and cell lines.

We next investigated whether the preserved modules could be used for therapeutic stratification across a panel of 82 breast cancer cell lines profiled across 90 drugs based on a GI_{50} analysis (Daemen et al., 2013). We found a total of 271 module-drug relationships covering 74 drugs and 99 modules with an FDR<5% and using the GR_{50} metric we identified 284 module-drug relationships, 64 overlapping with the GI_{50} analysis (Table S7) (Hafner et al., 2017). Approximately 36% of drug GI_{50} measurements could be predicted by PAM50 subtype and we found that modules could improve on this prediction in 50% (16/32) of cases (Figure S4c, Table S7). We next turned our attention to the 58 drugs (64%) whose response could not be predicted by PAM50 subtype. Overall, modules could improve predictions of response for 74% (43/58) of these drugs (Figure S4c) resulting in 97 module-drug connections where the module combined with subtype information was more predictive than subtype alone (Figure 4a). These subtype-independent predictions, such as the association of module 139 with oxaliplatin sensitivity, are distinct from subtype-dependent biomarkers such as the association of HER2 amplification with sensitivity to lapatinib, a HER2 inhibitor (Figure 4b-c). In the case of the DNA-damaging chemotherapy oxaliplatin, module 139 contains genes on chromosome 11q14 (amplified in 1.7% of breast cancers in TCGA) a region that has also been linked with 5-fluorouracil sensitivity, a related chemotherapy that also causes DNA damage during DNA replication (Ooyama et al., 2007). This region includes CREBZF, a transcription factor whose expression confers sensitivity to 5- fluorouracil (Lopez-Mateo et al., 2012), suggesting a potential mechanistic link between this amplicon and drug sensitivity. All together, modules could be used to improve upon PAM50 subtype for 65% of all tested drugs (59/90). In this way modules represent potential biomarkers that are largely complementary to subtype information.

Current pharmacogenomics approaches are based almost exclusively on data from cancer cell lines (Barretina et al., 2012; Basu et al., 2013; Costello et al., 2014; Daemen et al., 2013; Garnett et al., 2012). We reasoned that the highly preserved modules might constitute biomarkers that not only show strong performance as predictors of drug responses but also are more likely to translate when evaluated in human tumors. We first established that preserved modules perform comparably to genes as features used to build predictive models of drug responses using common methods of machine learning for this task (Figure 4d) (Costello et al., 2014). A central assumption in predictive modeling is that the interrelationships between input features are maintained across samples. In the case of biomarker discovery, if this structure is not preserved in the clinical setting the biomarker cannot be expected to be predictive.

Therefore, to assess if a biomarker is likely to translate when applied to human tumors we examined the relationships among its features by measuring their cross-correlation independently in cell lines and in TCGA. We found that the relationships among features based purely on cell line data were completely altered in human tumors. As an example, top molecular features correlated with imatinib sensitivity were cross-correlated in cancer cell lines (mean r^2=0.165), but these relationships are lost in tumor samples (r^2=0.015) (Figure 4e). Logically, if the gene inter-relationships used to construct a biomarker are not maintained in tumors, it cannot be predictive of its intended drug response. We performed this analysis across all drugs using both gene sets and modules, using simple rank based methods (FDR cutoff of 1% and 5%) as well as with elastic net regression. Applied to all drugs, biomarkers based on genes had a significantly reduced cross-correlation in TCGA when compared to cell lines, whereas module-based approaches maintained a consistently high cross-correlation in both cell lines and human tumors (Figure 4f). Since previously published approaches based on expression clustering or pathway information do not account for dissimilarities between cell lines and tumors they also show significantly reduced crosscorrelations (Figure S4d,e) indicating a potential for incorporating cell line preservation into other existing approaches. Therefore –omics biomarkers based on modules are more likely to maintain the same molecular relationships among features found in both cell lines and tumors and we propose that they are therefore more likely to be robust predictors of the drug response.

To test the hypothesis that module biomarkers are more likely to replicate across platforms, we compared gene versus module based biomarkers across a panel of 39 breast patient-derived xenograft (PDX) models (Bruna et al., 2016). There were a total of 13 drugs in common between the cancer cell line and PDX study. Of these, the response to 3 drugs could be significantly predicted in cancer cell lines using an elastic net regression model built on either genes or modules alone (based on a spearman correlation between predicted and observed response greater than zero) (Figure S4f). While using either genes or modules would lead to an equally predictive model for these 3 drugs in cancer cell lines, we tested the performance of these models in the separate PDX-derived dataset. For two of the three drugs we found that modules were significantly better in predicting drug responses than models built on genes alone (Figure S4g) indicating that module based biomarkers may be more robust and transferrable across different types of cancer models. An important consideration for the development of biomarkers is their applicability to human tumors, and these data reveal that an approach based on the comparison of molecular relationships found in both cell lines and tumors can aid in the development of improved, clinically relevant biomarkers.

DISCUSSION

The development of biomarkers from cell line data is a subject of intense investigation in the field. Challenges in using these data are multiple and there has been debate over the degree to which cell line profiling data is reproducible even in vitro (Haibe-Kains et al., 2013; Haverty et al., 2016). A fundamental issue is that the use of cell line genomics for translational medicine is contingent on the degree to which cell lines recapitulate the biology of the tumor sample they originate from. While cell lines mirror many genomic aspects of human tumors, they also harbor significant differences as well as an inability to model the complex tumor microenvironment. To identify molecular features that were shared between cell lines and human tumors, we placed molecular features into the context of a gene network (i.e. the module) which allowed us to compare and assess the preservation of this network in cell lines. We quantified the preservation of modules across a panel of breast cancer cell lines and found that many of the strongest co-expression and co-methylation signals in TCGA were not preserved in cell lines. This has a large impact on downstream molecular analysis and our results indicate that the expression and methylation of at least 13% of the genome in TCGA samples is specific to tumor biology and microenvironment. Our findings, combined with related reports on the influence of tumor purity on gene co-expression analysis (Aran et al., 2015), raise significant concern on the use of gene expression and methylation based biomarkers in clinical samples based purely on cancer cell line data.

In contrast to the standard approach of developing therapeutic biomarkers primarily in cell lines, we propose that first learning a set of robust biomarkers from patient data and then evaluating them in cell lines could lead to increased success in biomarker development. We provide key evidence that extant procedures used to generate pharmacogenomic biomarkers in cell lines are highly prone to error, as the relationships upon which they are built in cell lines fall apart when translated into human tumor specimens. As a solution to this significant and persistent problem in translational discovery we develop a framework to use modules as the basis for biomarker discovery that shows comparable performance but is resistant to this source of error.

Our framework, MAGNETIC, integrates genomic, transcriptomic, epigenomic, and proteomic data across breast cancers to identify a set of gene modules that have coordinated activity across patients. As opposed to previous approaches for identifying gene modules based solely on co-expression (Stuart et al., 2003; Zhang and Horvath, 2005) our approach integrates complementary data types based on comparison with known protein-protein interactions and the resulting modules show a strong enrichment for interacting proteins and shared functions, outperforming other comparable approaches at this task. Many identified modules were concordant with disease subtypes, cell surface receptors and prognostic gene signatures in breast cancer. In contrast to previous approaches that integrate diverse datasets into scaffolds of molecular networks (Vandin et al., 2012; Vaske et al., 2010), our method is not limited to known interaction networks and pathways which allowed the discovery of patterns in cancer data that are reflective of distinct cell types (e.g. immune, stromal, endothelial). Although applied to breast cancer here, future work could apply MAGNETIC to uncover patterns of interaction and module activities across other tumor types that may represent tissue specific networks that could be useful to delineate tumor and microenvironment-specific modules. We expect that computational data integration at two levels, the first between -omics platforms and the second between *in vivo* and *in vitro* samples, will ultimately aid in the translation of the cancer genome into clinical practice.

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

All code for the MAGNETIC pipeline are open source and available at:

https://github.com/bandyopadhyaylab. Further information and requests should be directed to and will be fulfilled by the Lead Contact, Sourav Bandyopadhyay (Sourav.bandyopadhyay@ucsf.edu)

METHOD DETAILS

Experimental data and processing—We used patient data from the TCGA breast cancer study (BRCA)(Network, 2012). We downloaded TCGA-curated level 3 data sets for gene expression, copy number variation (CNV), DNA methylation, mutation, and protein abundance (RPPA) (Table 1 below). We also downloaded clinical information for these patients including overall survival data, results of immunohistochemistry staining, and estimates of tumor purity.

The JWGray Breast Cancer Cell Line Panel was downloaded from Synapse (https://www.synapse.org/-!Synapse:syn2346643). We again downloaded data for gene expression, CNV, DNA methylation, RPPA, and dose-response data for 90 compounds (Table 2 below). Because the cell lines do not have matched normal samples to control for somatic mutations, we did not include mutation data in our analysis.

The gene expression, CNV, methylation and protein abundance datasets were processed in a standardized way, following the pipeline used in Wang et al (Wang et al., 2014). Data were mapped to HGNC gene IDs (hg19) and features that could not be mapped to a gene ID were discarded. One sample was discarded from the methylation dataset because more than 20%

of the values were missing. Data were normalized by individual gene to a standard normal distribution, providing gene scores. To fill in any remaining missing values the K nearest neighbors (KNN) algorithm was used for imputation, with K = 20. Non-synonymous mutations in the TCGA were recorded as binary events. Genes that were mutated in fewer than 2% of samples were discarded. Expression and CNV data from METABRIC (Curtis et al., 2012) were put through the same preprocessing pipeline as the other datasets with the exception that imputation was not necessary.

The HumanNet v1 network was downloaded from http://www.functionalnet.org/humannet (Lee et al., 2011). The iRefWeb v13.0 network was downloaded from http://www.irefindex.org (Turner et al., 2010). BioGrid Multi-Validated 3.4.146 was downloaded from http://thebiogrid.org (Stark et al., 2006). For all analyses of HumanNet, a cutoff of 2 was used (60,194 interactions in total after translation to HGNC IDs). The iRefWeb and BioGrid MV networks were translated to HGNC IDs but not otherwise filtered, and contained 26,008 and 30,828 interactions respectively. To facilitate testing we constructed separate networks representing interactions that were present in iRefWeb but not HumanNet (iRefWeb-HumanNet, 19,796 interactions) and in BioGrid but not HumanNet (BioGrid-HumanNet, 23,570 interactions).

Network Construction—The construction of the integrated multigraph involved two steps: for each pair of data types we measured the enrichment for known edges using a reference network. Next, we used these enrichment estimates to calibrate each correlation matrix before combining them into a single network. Specifically, we constructed a matrix M_{iJ} consisting of the correlation values between each gene x in dataset i and each gene y in dataset j across TCGA samples:

$$M_{ij}(x, y) = corr\left(x_i, y_j\right) \quad (1)$$

Note that only in cases where $i = j$ is the matrix M_{ij} symmetric, otherwise if $i \neq j$ then $corr(x_i, y_j) \neq corr(x_j, y_i)$. An enrichment matrix was calculated for each matrix M_{ij} as follows. We define the enrichment function $E_{ij}(r)$ as the log-likelihood that a given edge is present in the reference network N (HumanNet, iRefWeb or BioGrid MV) given its level of correlation compared to the background probability that a random edge is in N. We calculated an enrichment function from −1 to 1 for r in increments of 0.001. For cases where $r \geq 0$:

$$r \geq 0: \qquad E_{ij}(r) = \ln\left[\frac{P\left((x, y) \in N \;\middle|\; M_{ij}(x, y) \geq r\right)}{P((x, y) \in N)}\right] \quad (2)$$

And to evaluate negative correlations when $r < 0$:

$$r > 0: \qquad E_{ij}(r) = \ln\left[\frac{P((x, y) \in N \mid M_{ij}(x, y) < r)}{P((x, y) \in N)}\right] \quad (3)$$

Where $P((x, y) \in N)$ reflects the probability of any two random genes being connected in N. To obtain a stable estimate and confidence intervals, $E_{ij}(r)$ was bootstrapped by resampling the edges in N with replacement 1000 times and taking the median value. Because of low sample sizes at the extremes, this function was enforced to be monotonically increasing or decreasing when r was above or below 0, respectively.

Next, the matrix M_{ij} was converted to a matrix of log-likelihood ratios LLR_{ij} using E_{ij}, where for each gene pair (x, y) in M_{ij}:

$$LLR_{ij}(x, y) = E_{ij}\left(M_{ij}(x, y)\right) \quad (4)$$

This procedure was performed independently for each dataset pair (i, j). Therefore, $LLR_{ij}(x, y)$ represents reference-network normalized strength of the connection between gene x in dataset i and gene y in dataset j.

For the choice of network N we evaluated the networks HumanNet, iRefWeb, and the BioGrid Multi-Validated dataset. The result of evaluating the enrichment function E_{ij} for HumanNet is shown in Figure S1a. All three networks led to a similar set of enrichment functions. HumanNet resulted in the most high-scoring edges and modules and so we chose it for further analysis (see next section).

Module Identification—At a given cutoff, the LLR matrices are combined to define an integrated network, or multigraph, with multiple edge types (cnv-cnv, exp-cnv, etc.) and clustered using the Infomap algorithm (Rosvall et al., 2009). This algorithm was selected for its speed and ability to process large multigraphs. Infomap uses random walks to find neighborhoods of high density in the network. The algorithm was run with 10 trials per run and otherwise default parameters. Self-edges were not allowed. The algorithm was run 25 times with different random seeds and genes that clustered together in all runs were considered to be a member of the same module. Infomap was run with different cutoffs of LLR network to define modules and gene pairs within modules identified with a cutoff of 3.0 were the most enriched for known protein-protein interactions (Figure S1g-i). We chose a minimum module size of three genes.

After modules were identified, we developed a scoring method to evaluate modules in individual samples. Because of the possibility of anti-correlated data being included in a single module we developed an approach to classify gene features as contributing positively or negatively to a final module score (e.g. expression of a gene may be positively weighted while methylation of the same gene may be negatively weighted). We corrected the directionality of node values by multiplying a subset of node values by −1. The following was performed on a per-module basis. We define $\boldsymbol{d_t} = (d_1, d_2 \cdots d_m)$ to be the set of data points for module genes in sample *t*. Gene values from a given dataset were only included if at least one of the edges in the LLR network was associated with that gene and derived from that dataset. For each module we calculated the value:

$$\mathrm{Var}(\mathrm{diag}(\boldsymbol{u}) \cdot \boldsymbol{d_t}) \quad (5)$$

Where before optimization, ***u*** is a vector of ones of length *m*. We summed this variance over the set of *T* samples to obtain following sum:

$$\sum_{t \in T} \mathrm{Var}(\mathrm{diag}(\boldsymbol{u}) \cdot \boldsymbol{d_t}) \quad (6)$$

We use eq. 6 as an objective function for minimization by toggling the elements in ***u*** to be either 1 or −1. For small modules the optimal values of ***u*** were determined by enumeration, while large modules were optimized with simulated annealing. *Score*(*t*), the module score in sample *t*, is defined as an edge-weighted sum of node values:

$$Score(t) = \sum_{x,\, y} \sum_{i,\, j} LLR_{ij}(x, y) \cdot (u_{xi}s_{xit} + u_{yj}s_{yjt}) \quad (7)$$

Where (*x, y*) is all the gene pairs in the module and (*i, j*) is all the data types that connect them in the LLR network. The variable s_{xit} is the normalized value of gene *x* in dataset *i* in sample *t* (e.g. the normalized expression of a gene in a sample). Similarly, s_{yjt} is the value of gene *y* in dataset *j* in sample *t*. The variable u_{xi} denotes the element of ***u*** that corresponds to s_{xi} (i.e. u_{xi} is the gene and dataset-specific multiplier derived from the minimization of eq. 6).

Module scores were normalized across samples to control for the different number of genes in each module. Finally, modules were merged if they had highly correlated module scores (r > 0.765, the 99.5th percentile for inter-module correlation). This merging led to a set of 219 modules that were re-scored in the same way.

Module Characterization and Comparison—To associate modules with known biological processes and gene sets we used the g:Profiler server (Reimand et al., 2016) to identify enrichment for GO biological process, KEGG pathway or Reactome pathways at a significance cutoff of $p<1\times10^{-6}$ via hypergeometric test after Bonferroni correction for the number of pathways tested. Modules were associated with PAM50 and receptor subtypes by Pearson correlation and random bootstrap sampling was used to filter based on an FDR cutoff of 5% in Table S1.

To determine robustness of module identification, the modules were identified using smaller subsets of TCGA samples as well as the entire METABRIC cohort. Each new module was tested to examine if it had at least one match with the original set of modules identified from the entire TCGA cohort via hypergeometric test of gene overlap. Two modules were considered matching if they overlapped by at least two genes and the hypergeometric test p-value was <0.05 for overlap after Bonferroni correction. For percent overlap based studies, a new module was considered to match with a TCGA module if it contained more than the indicated percentage of the genes in the TCGA module.

Cell Type and H3K27me3 Analysis—Normalized RNAseq data on 227 cell types was downloaded from the Immunological Genome Project (immgen.org). Each specific cell type sample was grouped into a broader category based on annotation. The mean expression of all genes in a module was tested against 0 using a t-test and the $-\log_{10}$ of the p-value is shown. For clarity, only positive enrichments (i.e. over-expression) are shown. Gene Set Enrichment Analysis was performed via the website http://software.broadinstitute.org/gsea (Liberzon et al., 2011; Subramanian et al., 2005). H3K27me3 occupancy data was obtained from GEO GSE38788. Reads were mapped to the human genome hg18 with annotations +/− 5kb from the annotated TSS using BEDOPS (Neph et al., 2012).

Consensus Clustering—Patient clusters in TCGA and METABRIC datasets were identified by consensus clustering using the scikit-learn package (Pedregosa et al., 2011). First, the matrix of module scores across patients was constructed, consisting of 219 modules in TCGA. An 80% subsample was taken and clustered using hierarchical clustering (average method, Spearman correlation distance). This was repeated 5000 times and the rate of co-clustering was recorded for each pair of patients. The resulting co-clustering matrix was again clustered using the Ward method and cluster quality was assessed using silhouette scores, Dunn Index and cophenetic correlation coefficient. This was done with k=3 through 20, with k=10 having mostly consistent high scores across clustering quality metrics.

Evaluation of Preservation in Cell Lines—A module consists of an underlying network, and we reasoned that only modules that maintain their network structure would translate to a new dataset. Therefore, for a module to be preserved between tumors and cell lines the correlation network between gene pairs should be similar in both datasets. Because the number of cell lines was much smaller than the number of TCGA samples, molecular correlations cannot be directly compared between the two. To determine if correlations in cell lines were significantly above background we calculated a z-score for each edge, based on comparison to the total distribution of correlation values for that edge type. Let M'_{ij} be the correlation matrix for data types (i, j) in the collection of cell lines, as opposed to M_{ij} which is the correlation matrix in the TCGA as in eq. (1). Then $Z'_{ij}(x, y)$ is the normalized matrix giving a z-score for each edge between gene x and y between data types i and j:

$$Z'_{ij}(x, y) = \frac{M'_{ij} - \mu_{ij}}{\sigma_{ij}} \qquad \text{where } \mu_{ij} = \text{mean}\left(M'_{ij}\right) \text{and } \sigma_{ij} = \text{std}\left(M'_{ij}\right) \quad (8)$$

The z-score for each edge is multiplied by the sign of the corresponding edge in TCGA, $M_{ij}(x, y)$, to reward preserved directionality of correlation. The preservation score for a module is defined as the average of these modified z-scores over all edges in the module.

Evaluating the preservation of smaller modules was more difficult for two reasons. Differences between data platforms can lead to missing genes that cannot be evaluated in cell lines, and for small modules this can have a significant impact. Three modules of size 3 could not be evaluated in cell lines for this reason. More subtly, the preservation score of a smaller module will be noisier than that of a larger module due to the smaller number of edges and consequentially higher variance of the mean score. There is a slight trend of smaller modules being less preserved but we found no significant association between module size and preservation status.

Module-Drug Association—Cancer cell lines were scored in same manner as TCGA samples, using the same module nodes, edge and node weights. These scores were then tested for association with drug response by performing a bootstrapped correlation against the GI_{50} values for a panel of 90 drugs. For a given module-drug association, a single bootstrap iteration was performed by resampling cell lines with replacement and comparing this correlation with that of a randomly shuffled set of module scores. The false discovery rate was estimated from the proportion of times out of 1,000 iterations the real correlation value was greater than that of the random set (or less than in the case of a negative correlation). To associate modules or drugs with subtypes an ordinary least squares (OLS) model was used to predict module score or drug GI_{50} from an indicator matrix of subtypes. Random resampling was used to calculate a false discovery rate as above. Finally, we assessed whether modules improved drug prediction beyond subtype. In this case the OLS model was trained on a matrix that contained the subtype indicator variables as well as module score, and fit to a vector of drug sensitivity. In each bootstrap iteration the performance of this model was compared to a model based on subtype data alone. A module is considered to perform better than a subtype only model if it increases performance compared to subtype in >95% of the bootstrapped samples (corresponding to a <5% FDR).

Evaluation of machine learning methods—Module scores were compared to other biomarker discovery methods in two ways: as features for drug response prediction and as a mechanism for feature selection. To evaluate predictive performance, the matrix of 219 module scores in cell lines was compared to a matrix of expression, methylation, and copy-number variation data, totaling 48,235 features. These two matrices were used as input for elastic net, random forest, and support vector regression (SVR) models, trained on drug GI_{50} values. To evaluate the predictive performance we performed 100-fold cross validation using random 90/10 splits of the data for training and testing and recorded the median absolute error (MAE). The scikit-learn package was used for all methods(Pedregosa et al., 2011). Elastic net was run with $\lambda = 0.5$ and α chosen by internal crossvalidation during training. Random forest models were built with 1001 estimators, a maximum tree depth of 4, and considering a random third of features for each tree. SVR was run with default parameters. For feature selection the most predictive individual modules or genes were selected based on an FDR cutoff calculated by comparison of the real correlation with drug GI_{50} against a randomly shuffled background using the bootstrapping procedure described above. Feature selection was also performed using elastic net by training 200 models against bootstrapped samples of cell lines and using modules or genes that were selected by ≥50% of models.

Evaluation of related approaches—MAGNETIC integrates –omics data from multiple datatypes and also allows for the discovery of new pathways (Table 5). One class of comparable methods include PARADIGM (Vaske et al., 2010) and Omics Integrator (Tuncbag et al., 2016) that can integrate multiple data types but are limited to the annotated pathways used as input. For example, the network used for module discovery (LLR>3) covers 18,314 genes, while the union of all genes used by a leading pathway approach, PARADIGM, covers 2,581 genes. Thus, we attain roughly a sevenfold increase in coverage by expanding beyond known pathways. Beyond this advantage we note that the many of the interesting modules we identify in our analysis, the immune (#3), H3K27me3 (#27), Stromal (#12) and Endothelial (#16) modules would not be found by analysis limited to known pathway databases. Another class of comparable methods such as WGCNA (Zhang and Horvath, 2005) are able to identify new pathways based on correlations in the gene data, but are limited to a single data set. Modules identified by WGCNA were not as enriched for protein-protein interactions as they were using MAGNETIC (Figure S3a-e, see implementation of related approaches section). This was true for enrichments compared to raw PPI networks (HumanNet, iRefWeb and BioGrid) but also evaluated on interactions that were exclusively not used in the construction of MAGNETIC (iRefWeb-HumanNet and BioGrid-HumanNet).

Relatively few methods exist that both identify related genes and integrate multiple datatypes. Because of computational complexity these methods are limited to smaller gene sets (i.e. <5,000 genes) as opposed to MAGNETIC which can run on the whole genome (~20,000 genes). To compare these methods with MAGNETIC we first identified the molecular features with the highest variance across the TCGA cohort and used them as input to each of these methods. One approach (Zhang et al., 2013) uses super k-means clustering of an integrated data matrix to identify modules. Another approach, MDI (Kirk et al., 2012; Mason et al., 2016) uses a Bayesian model to capture relationships between gene sets. When the three methods were run using the most variable features from each dataset, we found that MDI and super k-means do not enrich for PPIs at the same level as MAGNETIC for raw networks as well as networks not used in the construction of MAGNETIC (Figure S3a-e, see implementation of related approaches section).

Implementation of related approaches—To compare MAGNETIC against other methods for module discover and data integration we compared against WGCNA, MDI and super k-means clustering. Due to computational limits of MDI and super k-means we could not use the entire TCGA dataset as input, and so we filtered the data to the most variable in either the copy-number, expression, or methylation datasets resulting in eight input datasets ranging. in size from 298 to 5080 genes. To facilitate comparison MAGNETIC, WGCNA, MDI, and super k-means were run on the same gene sets.

WGCNA clusters were identified using the WGCNA package from CRAN following Tutorial I using a soft threshold of 6. This method identified gene clusters using expression data only. To evaluate MDI we downloaded the code for MDI-GPU (https://warwick.ac.uk/fac/sci/systemsbiology/research/software/). On each of the test datasets the algorithm was run for 50,000 sampling iterations. Modules were identified by clustering the pairwise allocation agreement matrix. To evaluate super k-means we obtained compiled binaries from the authors and ran it with default settings on a file containing the concatenated feature from the copy-number, expression and methylation TCGA datasets. This algorithm would not successfully complete on larger inputs than 1,161 genes.

QUANTIFICATION AND STATISTICAL ANALYSIS

For boxplots, the box covers the middle two quartiles with a line at the median, and whiskers extend from the 5th to 95th percentiles. Circos plots in Figure 1f,h show genes ordered by betweenness centrality starting from the top right. All p-values are based on a two-tailed analysis and using a non-parametric t-test unless otherwise specified.

DATA AND SOFTWARE AVAILABILITY—All code for the MAGNETIC pipeline are open source and available at: https://github.com/bandyopadhyaylab.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The authors would like to thank Dvir Aran, David Quigley, Max Ranall, Nevan Krogan, and Andrej Sali for helpful comments. This work was funded by U01 CA168370 and R01 GM107671 to S.B.

REFERENCES

Aran D, Sirota M, and Butte AJ (2015). Systematic pan-cancer analysis of tumour purity. Nature Communications 6, 8971.

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607. [PubMed: 22460905]

Basu A, Bodycombe Nicole E., Cheah Jaime H., Price Edmund V., Liu K, Schaefer Giannina I., Ebright Richard Y., Stewart Michelle L., Ito D, Wang S, et al. (2013). An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules. Cell 154, 1151–1161. [PubMed: 23993102]

Bhat-Nakshatri P, Wang G, Appaiah H, Luktuke N, Carroll JS, Geistlinger TR, Brown M, Badve S, Liu Y, and Nakshatri H (2008). AKT alters genome-wide estrogen receptor alpha binding and impacts estrogen signaling in breast cancer. Mol Cell Biol 28, 7487–7503. [PubMed: 18838536]

Borst P, and Wessels L (2010). Do predictive signatures really predict response to cancer chemotherapy? Cell Cycle 9, 4836–4840. [PubMed: 21150277]

Cancer Genome Atlas Research, N., Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, and Stuart JM (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45, 1113–1120. [PubMed: 24071849]

Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, Bansal M, Ammad-ud-din M, Hintsanen P, Khan SA, et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. Nature Biotechnology advance online publication.

Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486, 346–352. [PubMed: 22522925]

Daemen A, Griffith OL, Heiser LM, Wang NJ, Enache OM, Sanborn Z, Pepin F, Durinck S, Korkola JE, Griffith M, et al. (2013). Modeling precision treatment of breast cancer. Genome Biology 14, R110. [PubMed: 24176112]

Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson IR, Luo X, Soares J, et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 483, 570–575. [PubMed: 22460902]

Gillet J-P, Varma S, and Gottesman MM (2013). The clinical relevance of cancer cell lines. Journal of the National Cancer Institute 105, 452–458. [PubMed: 23434901]

Goodspeed A, Heiser LM, Gray JW, and Costello JC (2016). Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. Molecular cancer research : MCR 14, 3–13. [PubMed: 26248648]

Hafner M, Niepel M, and Sorger PK (2017). Alternative drug sensitivity metrics improve preclinical cancer pharmacogenomics. Nat Biotechnol 35, 500–502. [PubMed: 28591115]

Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJ, and Quackenbush J (2013). Inconsistency in large pharmacogenomic studies. Nature 504, 389–393. [PubMed: 24284626]

Haverty PM, Lin E, Tan J, Yu Y, Lam B, Lianoglou S, Neve RM, Martin S, Settleman J, Yauch RL, et al. (2016). Reproducible pharmacogenomic profiling of cancer cell line panels. Nature 533, 333–337. [PubMed: 27193678]

Heiser LM, Sadanandam A, Kuo W-L, Benz SC, Goldstein TC, Ng S, Gibb WJ, Wang NJ, Ziyad S, Tong F et al. (2011). Subtype and Pathway Specific Responses to Anticancer Compounds in Breast Cancer. Proceedings of the National Academy of Sciences.

Heng TSP, Painter MW, and Consortium IGP (2008). The Immunological Genome Project: networks of gene expression in immune cells. Nature Immunology 9, 1091–1094. [PubMed: 18800157]

Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V, et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 158, 929–944. [PubMed: 25109877]

Kim KH, and Roberts CW (2016). Targeting EZH2 in cancer. Nat Med 22, 128–134. [PubMed: 26845405]

Kirk P, Griffin JE, Savage RS, Ghahramani Z, and Wild DL (2012). Bayesian correlated clustering to integrate multiple datasets. Bioinformatics 28, 3290–3297. [PubMed: 23047558]

Kuzmichev A, Nishioka K, Erdjument-Bromage H, Tempst P, and Reinberg D (2002). Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. Genes Dev 16, 2893–2905. [PubMed: 12435631]

Lee I, Blom UM, Wang PI, Shim JE, and Marcotte EM (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Research 21, 1109–1121. [PubMed: 21536720]

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, and Mesirov JP (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740. [PubMed: 21546393]

Lopez-Mateo I, Villaronga MA, Llanos S, and Belandia B (2012). The transcription factor CREBZF is a novel positive regulator of p53. Cell Cycle 11, 3887–3895. [PubMed: 22983008]

Margueron R, and Reinberg D (2011). The Polycomb complex PRC2 and its mark in life. Nature 469, 343–349. [PubMed: 21248841]

Mason SA, Sayyid F, Kirk PD, Starr C, and Wild DL (2016). MDI-GPU: accelerating integrative modelling for genomic-scale data using GP-GPU computing. Stat Appl Genet Mol Biol 15, 83–86. [PubMed: 26910751]

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, et al. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics 28, 1919–1920. [PubMed: 22576172]

Network CGA (2012). Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70. [PubMed: 23000897]

Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe J-P, Tong F, et al. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell 10, 515–527. [PubMed: 17157791]

Ooyama A, Okayama Y, Takechi T, Sugimoto Y, Oka T, and Fukushima M (2007). Genome-wide screening of loci associated with drug resistance to 5-fluorouracil-based drugs. Cancer science 98, 577–583. [PubMed: 17425594]

Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351, 2817–2826. [PubMed: 15591335]

Palacios EH, and Weiss A (2004). Function of the Src-family kinases, Lck and Fyn, in T-cell development and activation. Oncogene 23, 7990–8000. [PubMed: 15489916]

Parker JS, Mullins M, Cheang MCU, 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. Journal of Clinical Oncology 27, 1160–1167. [PubMed: 19204204]

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. (2011). Scikit-learn: Machine Learning in Python. J Mach Learn Res 12, 2825–2830.

Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, and Vilo J (2016). g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res 44, W83–89. [PubMed: 27098042]

Rosvall M, Axelsson D, and Bergstrom CT (2009). The map equation. The European Physical Journal Special Topics 178, 13–23.

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, and Tyers M (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34, D535–539. [PubMed: 16381927]

Stuart JM, Segal E, Koller D, and Kim SK (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science 302, 249–255. [PubMed: 12934013]

Su Y, Subedee A, Bloushtain-Qimron N, Savova V, Krzystanek M, Li L, Marusyk A, Tabassum DP, Zak A, Flacker MJ, et al. (2015). Somatic Cell Fusions Reveal Extensive Heterogeneity in Basal-like Breast Cancer. Cell Rep 11, 1549–1563. [PubMed: 26051943]

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102, 15545–15550. [PubMed: 16199517]

Tuncbag N, Gosline SJ, Kedaigle A, Soltis AR, Gitter A, and Fraenkel E (2016). Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. PLoS Comput Biol 12, e1004879. [PubMed: 27096930]

Turner B, Razick S, Turinsky AL, Vlasblom J, Crowdy EK, Cho E, Morrison K, Donaldson IM, and Wodak SJ (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database (Oxford) 2010, baq023.

Vandin F, Clay P, Upfal E, and Raphael BJ (2012). Discovery of mutated subnetworks associated with clinical data in cancer. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 55–66. [PubMed: 22174262]

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, and Stuart JM (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics (Oxford, England) 26, i237–245.

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, and Bork P (2002). Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417, 399–403. [PubMed: 12000970]

Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, and Goldenberg A (2014). Similarity network fusion for aggregating data types on a genomic scale. Nature Methods 11, 333–337. [PubMed: 24464287]

Wang W, Baggerly KA, Knudsen S, Askaa J, Mazin W, and Coombes KR (2013). Independent validation of a model using cell line chemosensitivity to predict response to therapy. J Natl Cancer Inst 105, 1284–1291. [PubMed: 23964133]

Winslow S, Leandersson K, Edsjó A, and Larsson C (2015). Prognostic stromal gene signatures in breast cancer. Breast Cancer Research : BCR 17.

Zhang B, and Horvath S (2005). A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4, Article17.

Zhang W, Liu Y, Sun N, Wang D, Boyd-Kirkup J, Dou X, and Han JD (2013). Integrating genomic, epigenomic, and transcriptomic features reveals modular signatures underlying poor prognosis in ovarian cancer. Cell Rep 4, 542–553. [PubMed: 23933257]

HIGHLIGHTS

- MAGNETIC integrates high-dimensional patient –omics datasets into gene modules.
- Modules connect tumor and cell line biomarkers by accounting for in vivo factors.
- Cell line pharmacogenomics is more robust using module based biomarkers.
- Modules delineate molecular and tumor microenvironmental factors in breast cancer.

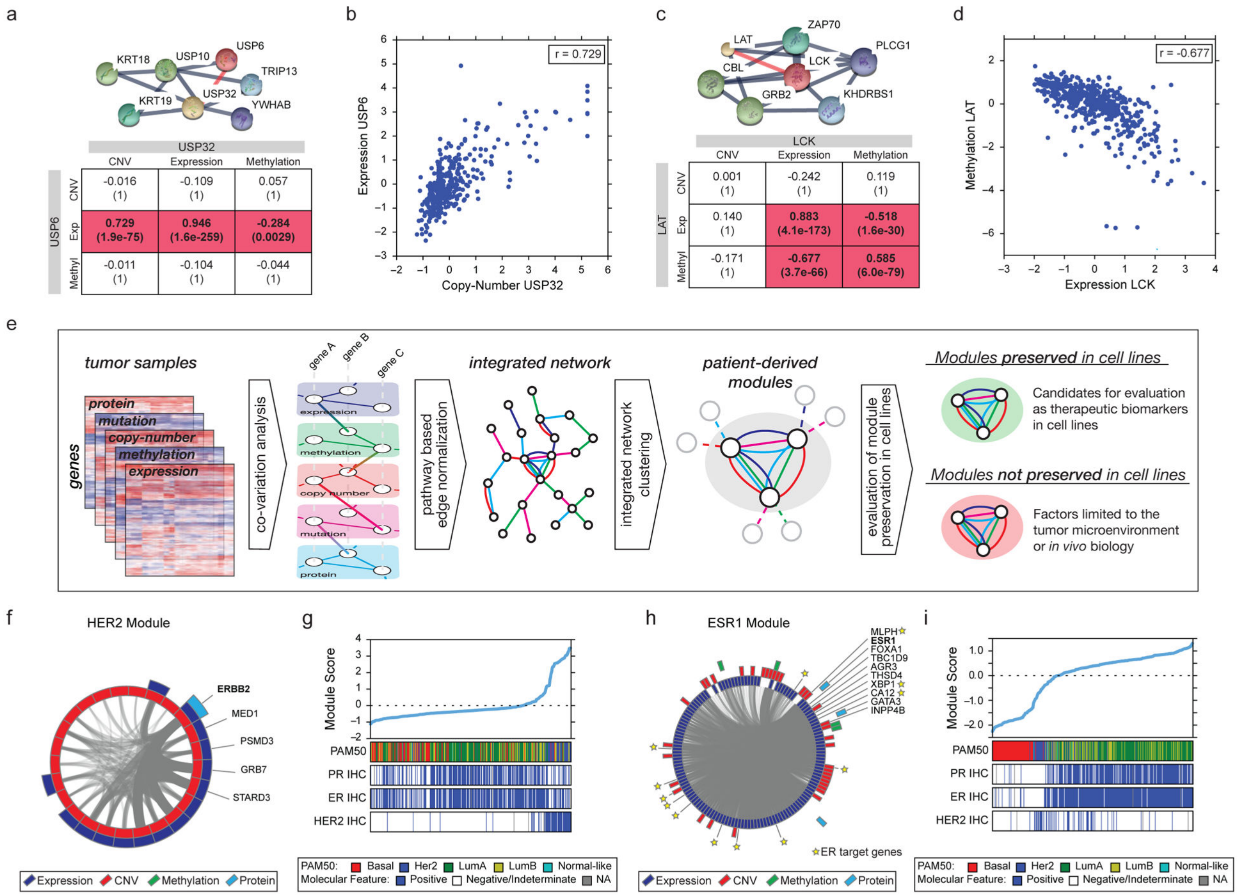


Figure 1: Data integration and module discovery using MAGNETIC.

(a) Interaction network of ubiquitin specific peptidases USP6 and USP32 from the STRING database and Pearson correlation between molecular features of USP6 and USP32 across TCGA breast cancers. P-value of association after Bonferroni correction for multiple testing are in parentheses. **(b)** Scatter of normalized USP32 copy-number and USP6 expression across TCGA. **(c)** The interaction network of the kinase LCK and its substrate LAT and relationships between their molecular profiles across platforms. **(d)** Scatter of LCK expression and LAT methylation. **(e)** MAGNETIC uses as input the normalized DNA copy-number, methylation, somatic mutations, mRNA expression and protein abundance data from a collection of tumor samples. We compute a multi-layer pairwise gene similarity network by computing the correlation between all pairs of gene features both within and between profiling platforms. Each linkage in this correlation network is normalized through comparison against a benchmark of pathways reflected in protein-protein interaction databases. Scored edges are then merged into a multigraph in which nodes represent genes and the edges between nodes represent co-incidence of different types of linkages. Clustering of this network using a random walk algorithm reveals gene modules whose components are closely related in multiple data types. **(f)** Circos plot representation of the module network containing HER2. Colors represent different data sources selected in the final integrated network for each gene and edge thickness is proportional to edge score. Top central genes are labeled. **(g)** TCGA samples sorted by HER2 module score. PAM50 subtype and molecular receptor status as determined by IHC are shown. **(h)** The module network containing the estrogen receptor, ESR1. Direct transcriptional targets of ER as assessed through ChIP analysis are marked with a star. **(i)** TCGA samples sorted by ESR1 module score. See also Figure S1-S3.

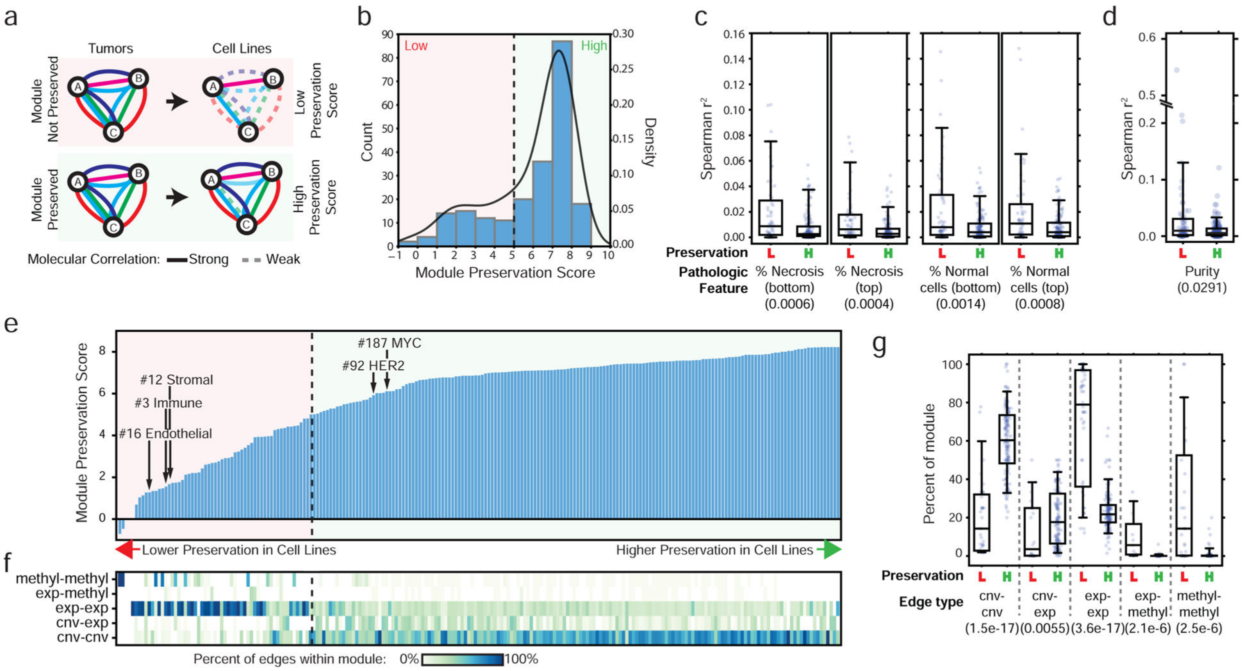


Figure 2: Many patient derived modules are not preserved in cell lines and are associated with specific data types.

(a) Overview of approach to score module preservation in cell lines. MAGNETIC takes molecular correlations present across tumor samples and determines if they remain significantly correlated across a cell line panel. Solid edges, above random background, dotted edges, below random background (see STAR Methods). Different colors represent edges derived from comparison between different molecular profiling platforms. **(b)** Histogram and kernel density estimation of the distribution of module preservation scores. The vertical dotted line shows the cutoff of 5 chosen for further evaluation. **(c)** Correlation of module scores with pathologic assessments of necrosis and normal cell infiltration for lowly (L) and highly (H) preserved modules. **(d)** Comparison of module types with computational assessment of tumor purity. **(e)** Sorted preservation scores for 219 breast cancer modules evaluated in cell lines. Lower preserved modules have a score less than 5 (dotted line). **(f)** For each module in (e), the percent of the LLR>1 network that corresponds to each edge type are shown. **(g)** Percent of each edge type for lowly and highly preserved modules in the LLR>1 network. P-values based on Mann-Whitney U-test in parenthesis. See also Figure S4.

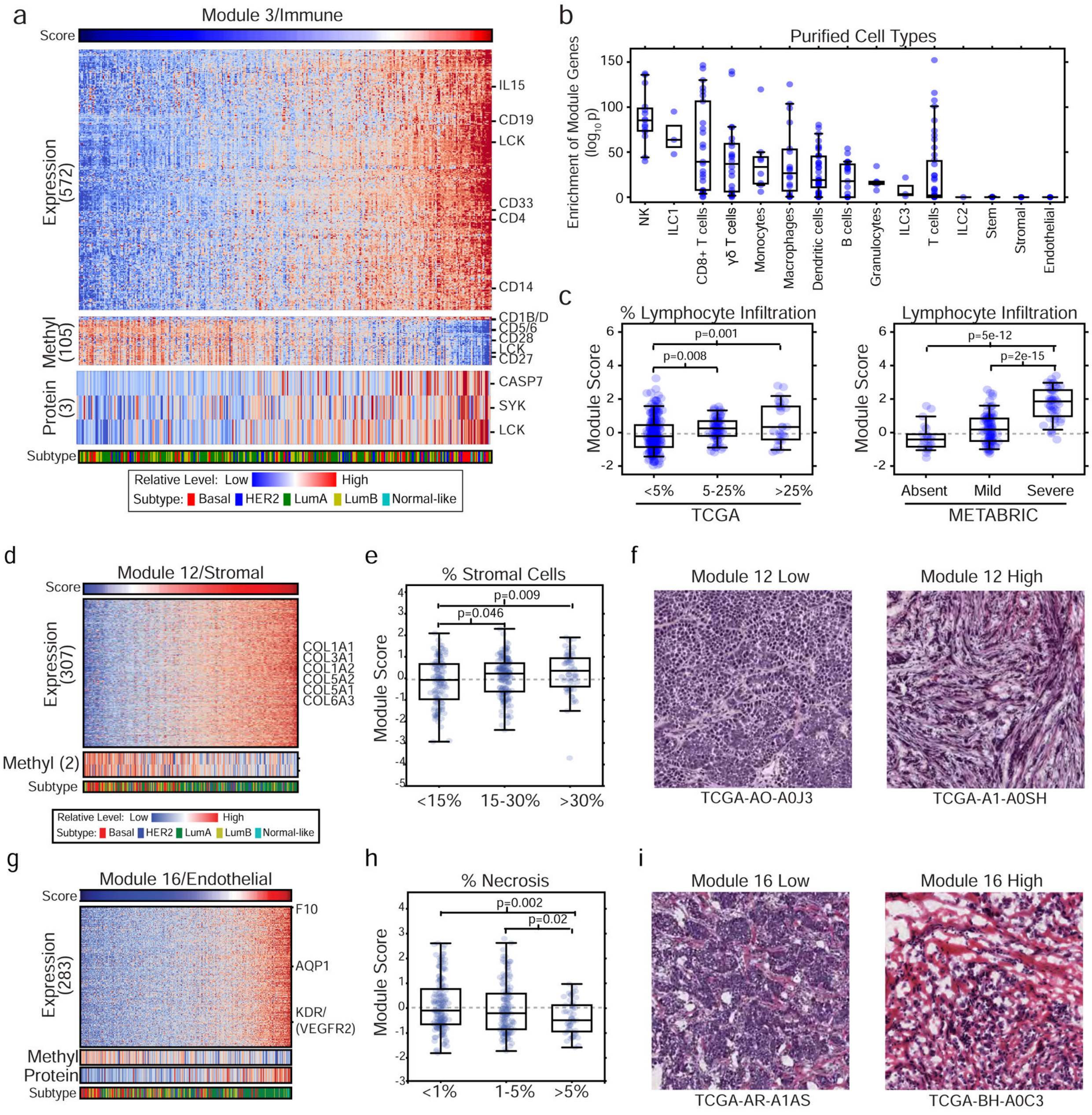


Figure 3: Modules reflect specific aspects of the tumor microenvironment.

(a) Heatmap of molecular features associated with the overall activity of the immune module (r^2>0.1). For clarity, the CNV of one gene is not shown. **(b)** Enrichment for high expression of module genes from normalized RNA-seq data in 227 purified immune cell type datasets. Cell types are categorized into 15 groups and enrichment based on a t-test. **(c)** Comparison of module scores with annotated lymphocytic infiltration values in TCGA and METABRIC datasets. **(d)** Heatmap of molecular features associated with module 12, associated with stromal cells. **(e)** Comparison of module scores with pathologic assessment of stromal cells in TCGA samples. P-values based on t-test. **(f)** Examples of samples from TCGA with low and high scores for module 12, showing the difference in stromal content. **(g)** Heatmap of molecular features associated with module 16, associated with endothelial cells. **(h)** Comparison of module scores with annotations of necrosis. **(i)** Examples of samples with low and high scores for module 16.

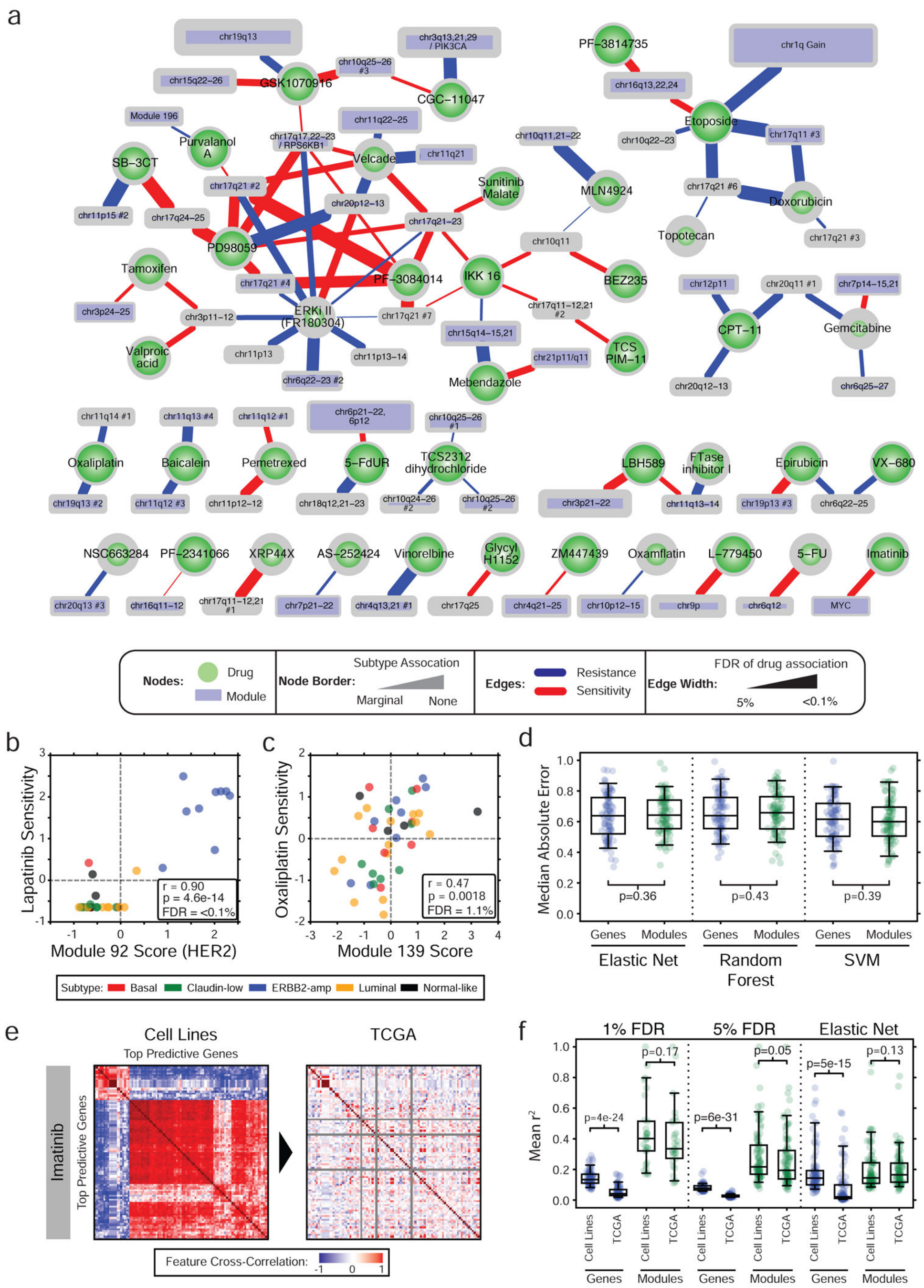


Figure 4: A module-drug network identifies high performance biomarkers that are preserved between patients and cell lines.

(a) Network of 97 module-drug associations based on breast cancer cell line modeling. Modules significantly associated with drug response are shown (FDR≤5%). Drugs are limited to those that are not associated with PAM50 subtype based on an FDR threshold of 5%. The size of each module is proportional to the number of genes within it, and the thickness of the border depicts the strength of a module or drug's association with PAM50 subtype. Edges are colored red when a module correlated with sensitivity to a drug, and blue when it correlated with resistance. Thicker edges have a lower FDR. As an example, gain of chr1q is associated with resistance of Etoposide at an FDR of <0.1%. **(b)** Scatter plot of cell line association of lapatinib response with module #92 (HER2) and **(c)** oxaliplatin with module #139 (chr11q14#1). Cell lines colored by PAM50 subtype. **(d)** Comparison of median absolute error of cross-validated predictions of drug sensitivity using single gene features or modules as input to elastic net, random forest or SVM based predictors. P-values based on Mann-Whitney U-test. **(e)** Cross-correlation for all pairs of molecular features that are the most predictive of response to imatinib in cell lines at an FDR of 1% and cross-correlation of the same features in TCGA. **(f)** The average cross-correlation (r^2) of features selected by various statistical methods (FDR, elastic net) using single genes or modules in cell lines and evaluation of cross-correlation of the same features in TCGA. Each point represents a model for a single drug. P-values based on Mann-Whitney U-test. See also Figure S4.

Table 1:

TCGA datasets used

Dataset	Platform	Number of Samples	Number of Genes
Gene expression	Agilent G4502A_07_3 array	547	16933
Copy number variation	Affymetrix Genome-Wide Human SNP Array 6.0	773	17552
DNA methylation	Illumina Infinium Human Methylation450 BeadChip Kit	939	13412
Mutation	Illumina Genome Analyzer	468	60
Protein abundance	Reverse Phase Protein Array	403	136

Table 2:

Breast cancer cell line datasets used

Dataset	Platform	Number of Samples	Number of Genes
Gene expression	Affymetrix GeneChip Human Gene 1.0 ST exon array	54	16825
Copy number variation	Affymetrix Genome-Wide Human SNP Array 6.0	77	17440
DNA methylation	Illumina Infinium Human Methylation27 BeadChip Kit	49	13970
Protein abundance	Reverse Phase Protein Array	52	53

Table 3:

Result of TCGA data processing.

Dataset	# Initial Genes	# Genes Mapped	# Values Imputed
Gene expression	17800	16963	1313
Copy number variation	20630	17568	0
DNA methylation	13957	13416	6195
Mutation	11279	60[*]	0
Protein Abundance	139	136	0

[*] In the case of mutation data, only genes mutated in ≥2% of samples were considered.

Table 4:

Result of breast cancer cell line data processing.

Dataset	# Initial Genes	# Genes Mapped	# Values Imputed
Gene expression	18619	17471	0
Copy number variation	27228	16918	273
DNA methylation	14476	13993	0
Protein Abundance	53	53	0

Table 5:

Comparison of methods related to MAGNETIC.

	Limited to a single dataset	**Integration of multiple datasets**
Limited to known pathways	Gene set enrichment (GSEA) (many others)	PARADIGM(Vaske et al., 2010) Omics-integrator[*](Tuncbag et al., 2016)
De novo pathway discovery	Hierarchical clustering[†] WGCNA(Zhang and Horvath, 2005)	MAGNETIC MDI[†](Kirk et al., 2012; Mason et al., 2016) Super k-means[†](Zhang et al., 2013)

[†] Because of computational complexity, the standard implementation requires pre-filtering input data to a small set of genes.

[*] Omics-Integrator analyzes differences compared to a control sample