Identification of cancer genes that are independent of dominant proliferation and lineage programs

Laura M. Selfors[a,b], Daniel G. Stover[a,b,c,1], Isaac S. Harris[a,b], Joan S. Brugge[a,b,2], and Jonathan L. Coloff[a,b,2]

[a]Department of Cell Biology, Harvard Medical School, Boston, MA 02115; [b]Ludwig Center at Harvard, Harvard Medical School, Boston, MA 02115; and [c]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115

Contributed by Joan S. Brugge, November 10, 2017 (sent for review August 22, 2017; reviewed by Christos Hatzis and Matthew G. Vander Heiden)

Large, multidimensional cancer datasets provide a resource that can be mined to identify candidate therapeutic targets for specific subgroups of tumors. Here, we analyzed human breast cancer data to identify transcriptional programs associated with tumors bearing specific genetic driver alterations. Using an unbiased approach, we identified thousands of genes whose expression was enriched in tumors with specific genetic alterations. However, expression of the vast majority of these genes was not enriched if associations were analyzed within individual breast tumor molecular subtypes, across multiple tumor types, or after gene expression was normalized to account for differences in proliferation or tumor lineage. Together with linear modeling results, these findings suggest that most transcriptional programs associated with specific genetic alterations in oncogenes and tumor suppressors are highly context-dependent and are predominantly linked to differences in proliferation programs between distinct breast cancer subtypes. We demonstrate that such proliferation-dependent gene expression dominates tumor transcriptional programs relative to matched normal tissues. However, we also identified a relatively small group of cancer-associated genes that are both proliferation- and lineage-independent. A subset of these genes are attractive candidate targets for combination therapy because they are essential in breast cancer cell lines, druggable, enriched in stem-like breast cancer cells, and resistant to chemotherapy-induced down-regulation.

breast cancer | bioinformatics | tumor biology | gene expression | oncogene

Decades of work have led to the identification of oncogenes and tumor suppressors that, when mutated, are capable of driving the development of cancer (1). Recent large-scale sequencing efforts have validated the importance of these genes in human cancer and have expanded the panel of known recurrent somatic alterations for many tumor types (2, 3). The observation that cancer cells are often addicted to their driver oncogenes (4) has motivated the development of personalized therapies targeted directly at such genetically altered gene products, and many such therapies have improved both survival and the overall quality of cancer care (5, 6). However, although most targeted therapies are initially effective, they typically have limited efficacy because of the development of resistance. In addition, many tumors are driven by the loss of tumor suppressors or undruggable oncoproteins, further highlighting the need to develop strategies to identify additional drug targets to enhance the efficacy of cancer treatments (7).

One alternative strategy is to identify genes that are not directly genetically altered in tumors, but are nonetheless required for tumor development and/or maintenance. These so-called nononcogene addictions (8) are also targets of numerous existing therapeutic agents and continue to be the subject of intense investigation (9). Nononcogene targeted therapies, which include several standard-of-care therapies (10), are often effective, but their efficacy is usually context-dependent, making it critical to identify biomarkers that predict which patients will respond to therapy.

Recently assembled large multidimensional genome-wide datasets have facilitated multiplexed analyses of tumor molecular features (3, 11, 12). In particular, genetic and transcriptomic data have been extensively used to identify functional genetic alterations, mechanisms by which tumor driver alterations induce phenotypic change, or molecular signatures for the classification of tumors (12–28). However, in most cases, the biological implications of alteration-associated gene expression and whether or not associated genes might represent potential vulnerabilities has not been extensively studied. Here, we analyzed The Cancer Genome Atlas (TCGA) data to identify important contributors to variability in breast cancer gene expression programs, with the goal of identifying potential vulnerabilities that are specific to subgroups of breast tumors. Initially, we used an unbiased approach to identify genes differentially expressed in tumors with specific genetic alterations. This approach highlighted the overwhelming importance of molecular subtype in breast tumor gene expression profiles, and refocused our attention on proliferation rate and tumor lineage, which are two critical underlying factors of the subtype designation (29, 30). We found that genetic alterations are not independently associated with broad gene expression programs when proliferation and lineage are taken into account. Building on these findings, we explored the contribution of proliferation to differences in gene expression in matched pairs of normal and tumor tissue from diverse cancer types, and found that most genes differentially expressed in tumors correlated with the proliferative signature of the tumors. However, we also identified cancer-associated genes from multiple tumor lineages

Significance

Large, multidimensional "landscaping" projects have provided datasets that can be mined to identify potential targets for subgroups of tumors. Here, we analyzed genomic and transcriptomic data from human breast tumors to identify genes whose expression is enriched in tumors harboring specific genetic alterations. However, this analysis revealed that two other factors, proliferation rate and tumor lineage, are more dominant factors in shaping tumor transcriptional programs than genetic alterations. This discovery shifted our attention to identifying genes that are independent of the dominant proliferation and lineage programs. A small subset of these genes represents candidate targets for combination cancer therapies because they are druggable, maintained after treatment with chemotherapy, essential for cell line survival, and elevated in drug-resistant stem-like cancer cells.

Author contributions: L.M.S. and J.L.C. designed research; L.M.S., D.G.S., and J.L.C. performed research; L.M.S., D.G.S., I.S.H., J.S.B., and J.L.C. analyzed data; and L.M.S. and J.L.C. wrote the paper.

Reviewers: C.H., Yale School of Medicine; and M.G.V.H., Koch Institute at Massachusetts Institute of Technology.

The authors declare no conflict of interest.

[1]Present address: Department of Internal Medicine, The Ohio State University, Columbus, OH 43210.

[2]To whom correspondence may be addressed. Email: joan_brugge@hms.harvard.edu or jonathan_coloff@hms.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1714877115/-/DCSupplemental.

whose expression is not linked to the proliferative state of the tumor, suggesting that these genes are both proliferation- and lineage-independent. The expression of many of these genes is maintained after chemotherapy treatment and is enriched in stem-like cancer cells, leading us to propose that the subset of these genes that are essential and druggable are candidate targets for combination therapies.

Results

Gene Expression Associated with Genetic Alterations, Lineage, and Proliferation. To identify transcriptional programs associated with recurrent genetic alterations in breast tumors, we analyzed DNA sequence, copy number variation, and RNAseq data from the TCGA breast cancer dataset (11). We defined genetic alterations as nonsilent somatic mutations, amplifications, or deletions, and performed differential expression analyses independently for each alteration. Any gene with significantly higher or lower mean expression in tumors with a genetic alteration relative to tumors without that alteration were considered associated with the alteration (termed gene:alteration association). We did not require gene:alteration associations to be specific to a given alteration, allowing for the identification of shared transcriptional programs among alterations with similar predicted modes of action (e.g., *PIK3CA* amplifications, *PTEN* deletions, and *PIK3CA* mutations). To reduce false positives resulting from coamplification or codeletion with copy number variations, we eliminated associations if a

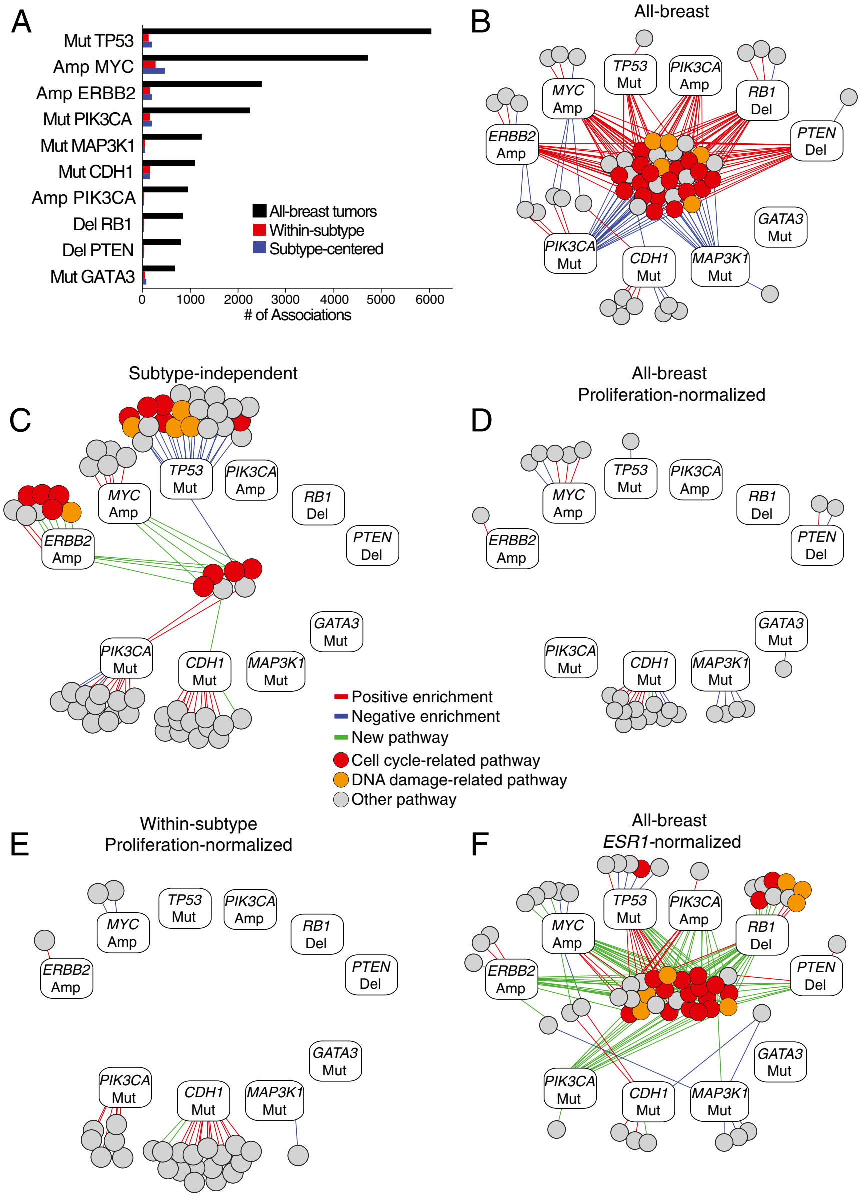


Fig. 1. Genes and pathways associated with common genetic alterations in breast tumors. (*A*) Bar graph of the number genes associated with the 10 most common genetic alterations found in the entire TCGA breast dataset ("all-breast," black), within any of the four major breast molecular subtypes ("within-subtype," red) or after subtype-centering ("subtype-centered," blue). (*B*–*F*) GeneGo Pathway Maps (circles) and genetic alterations (rounded rectangles) are connected by lines representing significant positive (red lines) or negative (blue lines) enrichment in mRNA levels in tumors bearing the indicated genetic alteration. Green lines indicate new pathways not identified in the all-breast enrichment. Pathways related to proliferation or DNA damage are indicated in red or orange, respectively. Threshold for statistical significance was false discovery rate corrected; $P < 0.05$, hypergeometric test. Pathway enrichment results are depicted from the all-breast (*B*), subtype-independent (*C*), proliferation-normalized all-breast (*D*), proliferation-normalized within-subtype (*E*), and the *ESR1*-normalized all-breast (*F*) analyses.

gene was on the same chromosome arm as the associated alteration and coamplified or codeleted in more than one third of the samples. Because the different subgroups of breast tumors can be a confounding factor in gene expression association studies (10, 31), we performed this analysis both on the entire TCGA breast cancer dataset as well as within each of the four major molecular subtypes of breast tumors (HER2$^+$, basal, luminal A, and luminal B) (29). Of the 14,209 expressed genes and 52 genetic alterations analyzed, we identified 21,890 gene:alteration associations in the entire all-breast dataset (Bonferroni-corrected $P < 0.05$, Welch's t test; Fig. 1*A* and Fig. S1*A* and Dataset S1*A*). This number was reduced by 94.8% (1,143 associations identified) when the analyses were performed within each individual molecular subtype (Fig. 1*A* and Fig. S1*A* and Dataset S1*B*). Because the within-subtype analyses also significantly reduced sample size and statistical power, we also used a subtype-centering method to normalize for subtype-dependent gene expression without reducing sample size. This revealed a similar 93.0% reduction in gene:alteration associations (Fig. 1*A* and Fig. S1*A* and Dataset S1*C*), suggesting that the reduced number of associations in the within-subtype analysis was not merely a result of decreased statistical power, but was a result of the strong confounding influence of molecular subtype on gene expression analysis of breast tumors. Five hundred seventy-nine gene:alteration associations were identified as being subtype-independent by both approaches (Fig. S1*B* and Dataset S1*C*).

To understand the biological implications of the gene:alteration association findings from the all-breast and subtype-independent analyses, we performed GeneGO enrichment analysis on the genes associated with the 10 most common alterations. In the all-breast dataset, the majority of the genetic alteration-associated pathways are found in a large cluster of shared pathways that are positively associated with a group of six alterations (*TP53* mutation; *MYC*, *ERBB2*, and *PIK3CA* amplifications; and *PTEN* and *RB1* deletions), and negatively associated with three alterations (*PIK3CA*, *CDH1*, and *MAP3K1* mutations; Fig. 1*B* and Dataset S2*A*). This shared cluster consists largely of pathways related to cell cycle control, DNA damage, and cellular proliferation. As expected, there is a significant reduction in the number of enriched pathways when the analysis is performed on the subtype-independent associations; however, this analysis revealed several interesting subtype-independent genetic alteration-associated genes/pathways (Fig. 1*C* and Dataset S2*B*). Among the most interesting is the association of *MYC* amplification with elevated levels of numerous protein folding genes (*HSP105*, *HSP90AA1*, *HSP90AB1*, *STIP1*, *PTGES3*, and *FKBP4*), suggesting that *MYC*-amplified tumors might be sensitive to agents targeting the protein folding machinery. In addition, *PIK3CA* mutant tumors show enrichment of WNT signaling genes (*TCF7L2*, *WNT5A*), and *CDH1* mutant tumors have higher HGF and IL6 pathway genes.

Although there was a significant reduction in the number of proliferation-related pathways in the subtype-independent pathway analysis compared with the all-breast analysis, several genetic alterations were associated with proliferation-related pathways within molecular subtypes (Fig. 1*C* and Dataset S2*B*). Although proliferation-dependent gene expression has been shown to play an important role in pathway enrichment analyses (10, 32, 33) and breast molecular subtype designation (29, 30, 34), there has not been a systematic analysis of the overall contribution of proliferation-dependent gene expression to genetic alteration-associated gene expression. To accomplish this, we performed a normalization for gene expression that can be attributed to differences in proliferation (32) and reanalyzed the previously identified gene:alteration associations. In the all-breast dataset, proliferation normalization resulted in a 76% reduction in associations and a dramatic 86% reduction in enriched pathways (Fig. 1*D* and Fig. S2*A* and Datasets S1*D* and S2*C*), despite only reducing the total variance in gene expression by 7% (Fig. S2*B*). Proliferation normalization also had a large, but muted, effect on the within-subtype associations and pathway enrichment (Fig. 1*E* and Fig. S2*C* and Datasets S1*E* and S2*D*). This reduced effect likely reflects the importance of proliferation in subtype designation (29, 30, 34) and the likelihood that the within-subtype analysis itself accounts for differences in gene expression resulting from proliferation. Interestingly, proliferation normalization uncovered gene and pathway associations that were not statistically significant in the uncorrected dataset (Fig. 1 *D* and *E* and Fig. S2*D* and Datasets S1*E* and S2 *C* and *D*), revealing potentially interesting associations that are masked by the dominant effects of proliferation, as we have previously observed (10). For example, proliferation normalization revealed that the O-glycan and cholesterol biosynthesis pathways are associated with *MYC* amplification and *MAP3K1* mutation, respectively (Dataset S2 *C* and *D*).

In addition to proliferation, tumor cell lineage may influence our analyses of genetic alteration associations, given that the genetic alterations we tested are disproportionally found within specific lineages and lineage is highly related to molecular subtype (11). The most fundamental way to separate breast tumors based on lineage is by distinguishing tumors that are positive or negative for expression of the estrogen receptor (ER) (35). To understand the effect of lineage on genetic alteration-associated gene expression, we performed a normalization using *ESR1* mRNA expression as a lineage marker, similar to the previously used proliferation normalization method that used a proliferation gene signature. In the all-breast dataset, *ESR1* normalization reduced gene expression variance by only 9.6% (Fig. S2*B*), but resulted in a 67% decrease in associations and an 18% reduction in enriched pathways (Fig. 1*F* and Fig. S2*A* and Datasets S1*F* and S2*E*). This suggests that genetic alteration associations are influenced by differences in gene expression caused by the distinct lineages of breast tumors and the uneven distribution of genetic alterations within these lineages, whereas enriched pathways are more dependent on differences in proliferation.

To directly assess the predictive power of proliferation, lineage, and genetic alterations on gene expression, we generated a series of linear models for each expressed gene in the breast dataset. In univariate models, we found that proliferation signature score (32), proliferation status, *ESR1* mRNA, and clinical ER status were all generally better predictors of gene expression than any of the genetic alterations, having higher mean coefficients of determination (r^2; Fig. S2*E* and Dataset S3*A*; $P < 0.05$, Student's t test). Next, we determined whether the addition of genetic alteration information improves the performance of multiple regression models that take into account both proliferation and lineage (models of gene expression based on proliferation + *ESR1* vs. proliferation + *ESR1* + genetic alteration) (Dataset S3*B*). Plotting model improvement (Δr^2) for each gene with significantly increased goodness of fit ($P < 0.05$, likelihood ratio test of nested models) reveals that the majority of models are not more accurate when genetic alteration information is added, and most of those that are improved are only minimally better (Fig. 2*A* and Dataset S3*B*). Further, the vast majority of the genes with high Δr^2 values are coamplified or codeleted with the driver gene (Fig. 2*A*, indicated in red, Dataset S3*B*). These results, which we independently validated using METABRIC data (12) (Fig. 2*B* and Dataset S4), support our finding that genetic alterations are not strong independent predictors of gene expression in breast tumors when differences in gene expression resulting from lineage, proliferation, and coamplification/codeletion are taken into account.

As an additional test of the influence of genetic alterations on gene expression, we determined the extent to which our gene:alteration association findings in breast tumors were found in 10 other tumor types (35). Only 6.9% (1,453) of the gene:alteration associations identified in analysis of the 10 most common

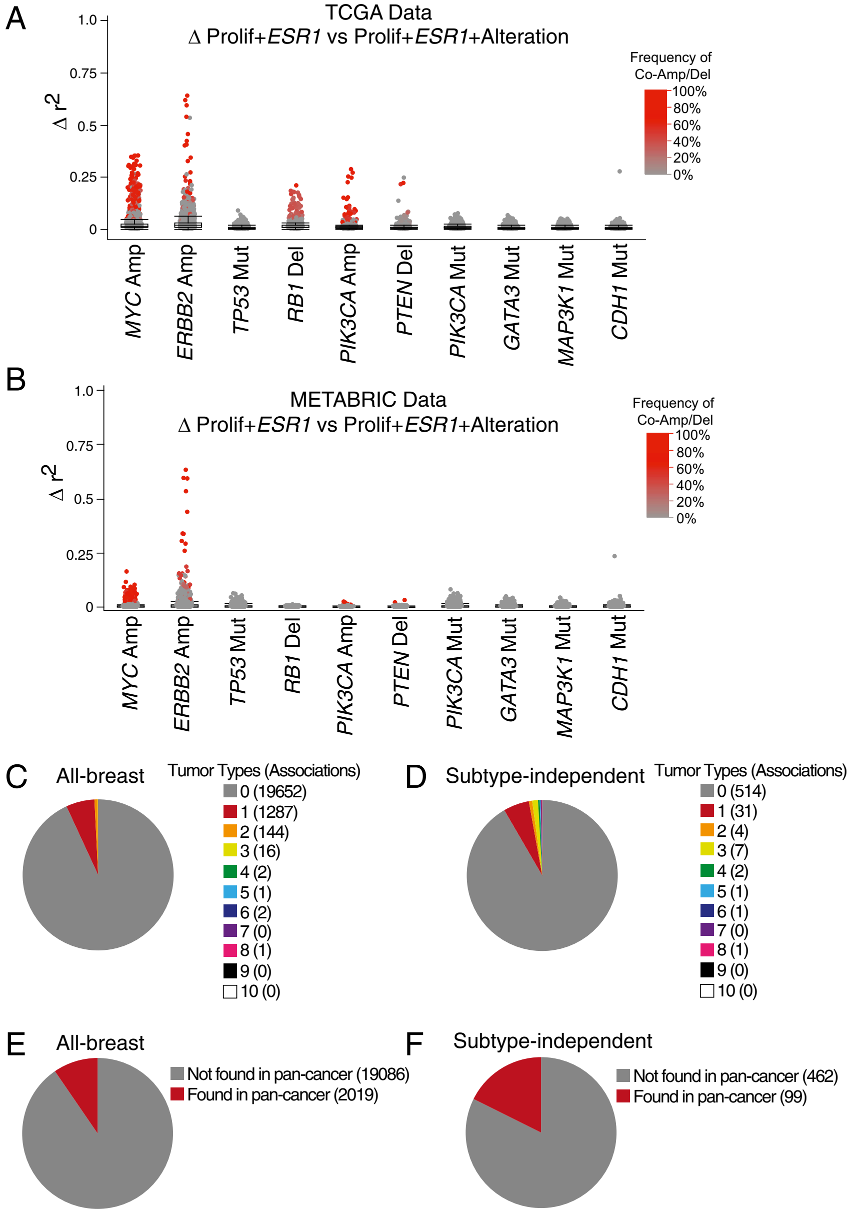


Fig. 2. The contribution of genetic alterations to gene expression beyond proliferation, tumor lineage, and in multiple tumor types. (*A* and *B*) Multiple linear regression models depicting the contribution of genetic alterations to gene expression when proliferation and *ESR1* are taken into account. Each point represents the difference in adjusted r^2 (Δr^2) between the proliferation + *ESR1* and the proliferation + *ESR1* + genetic alterations models that were found to be significantly different ($P < 0.05$, likelihood ratio test of nested models). Red coloring indicates the frequency of coamplification/deletion. Results from the TCGA breast cancer (*A*) and METABRIC (*B*) datasets are presented. (*C* and *D*) Pie charts depicting the number of other tumor types in which the all-breast (*C*) and subtype-independent (*D*) gene:alteration associations for the 10 most common alterations are found. (*E* and *F*) Pie charts depicting the number of gene:alteration associations from the all-breast (*E*) and subtype-independent (*F*) analyses of the 10 most common alterations that are found in a pan-cancer dataset.

alterations and all-breast samples were identified in one or more additional cancer types, and only 0.8% (166) were identified in two or more additional cancer types (Fig. 2*C* and Dataset S5*A*). The subtype-independent gene:alteration findings were only slightly more likely to be validated, with 8.4% (47) of the gene:alteration associations identified in one other cancer type and 2.9% (16) identified in two or more additional cancer types (Fig. 2*D* and Dataset S5*B*). Even alterations that are commonly found across tumor types (i.e., *MYC* amplification and *TP53* mutation) had fewer than 10 associated genes (6 and 8, respectively) that validated in two or more cancer types, demonstrating that alteration-associated gene expression found in breast tumors is not strongly recapitulated in other tumor types. To improve statistical power, we merged the 10 additional tumor types and analyzed this pan-cancer dataset together, where we found that 9.6% (2,019) of the all-breast and 17.6% (99) of the subtype-independent findings were validated (Fig. 2 *E* and *F* and Dataset S5 *C* and *D*). These findings demonstrate that although the vast majority of our gene:alteration associations are specific to breast tumors, subtype-independent associations are more likely to be found in other tumor types.

All these analyses support the conclusion that the transcriptional programs associated with specific genetic alterations are likely to be highly context-dependent, and most are secondary to proliferative states and/or lineages with which the genetic alterations are associated.

Identification of Proliferation- and Lineage-Independent Cancer Genes. Although the relatively small group of proliferation- and subtype-independent associations identified here warrant additional investigation, we were interested in developing a direct informatics pipeline to identify genes that are more generally associated with the cancer state independent of proliferation and cell lineage, because these may represent more attractive targets for therapeutic intervention.

To accomplish this, we first analyzed TCGA RNAseq data from 10 diverse tumor types with patient-matched tumor and normal data, where we detected the expected increase in proliferation scores in tumors relative to normal (Fig. 3 and Fig. S3 *A–E*, box plots). We next compared gene expression differences between tumor and normal samples and the degree to which the expression level of each gene correlates with proliferation scores in each of the 10 tumor types (Fig. 3 and Fig. S3 *A–E*, volcano plots and histograms). Indeed, genes elevated in tumors are more likely to be positively correlated with proliferation, whereas genes at lower levels are more likely to be negatively correlated with proliferation.

In these analyses we also identified a subgroup of genes that are exceptions to this trend and are higher in tumors but not positively correlated with proliferation (Fig. 3 and Fig. S3 *A–E*, volcano plots and histograms). These genes could be specifically up-regulated in tumors independent of their proliferative status or more highly expressed because the tumor cell of origin is not abundant in bulk normal tissue, and because their expression is elevated, their protein products might offer a potential therapeutic window for cancer therapy. To evaluate the lineage-dependence of the expression of these genes, we identified genes whose expression is elevated in tumors but not positively correlated with proliferation across multiple tumor types. Of the 2,972 genes that are elevated in tumors relative to normal in five or more cancer types (termed TA for tumor-associated, Dataset S6*A*), 363 were not positively correlated with proliferation (termed TANP for tumor-associated nonproliferation; Dataset S6*B*). Pathway enrichment of the TANP genes revealed three pathways, two of which ("Chondroitin sulfate and dermatan sulfate metabolism" and "Cell adhesion_Cell-matrix interactions") were specifically enriched in the TANP set, and one ("Transcription_Chromatin modification") that was enriched in both the TANP and TA gene sets (Fig. 4*A* and Dataset S6 *C* and *D*). The two TANP-specific pathways, which include genes involved in extracellular matrix interactions, represent tumor-associated pathways that would not be detected by methods that do not take into account the dominant influence of proliferation-associated gene expression on pathway enrichment analysis.

The TANP genes might be of particular interest as potential targets for combination cancer therapies, as slowly proliferating or quiescent tumor cells can be resistant to many chemotherapies (10, 36). On the basis of their lack of proliferation correlation, TANP genes would be predicted to be less likely to be down-regulated by chemotherapies that inhibit proliferation, and therefore may be better candidates for combination therapy. To test this idea, we analyzed gene expression data from tumor specimens that were harvested before and after one round of chemotherapy (37). Focusing on data from patients who responded to therapy to select for data in which there was effective drug delivery, we found that TANP genes are more than five times less likely to be down-regulated by chemotherapy than TA genes that are associated with proliferation (TAP genes; risk ratio = 5.1; $P = 1.9 \times 10^{-100}$, Fisher's exact test; Fig. 4 *B* and *C* and Fig. S3*F* and Dataset S6*B*). This suggests that despite the many confounding variables associated with chemotherapy treatment of patient tumors, proliferation correlations predict gene expression responses to chemotherapy. This places an emphasis on TANP genes as potential targets for combination with chemotherapy, as

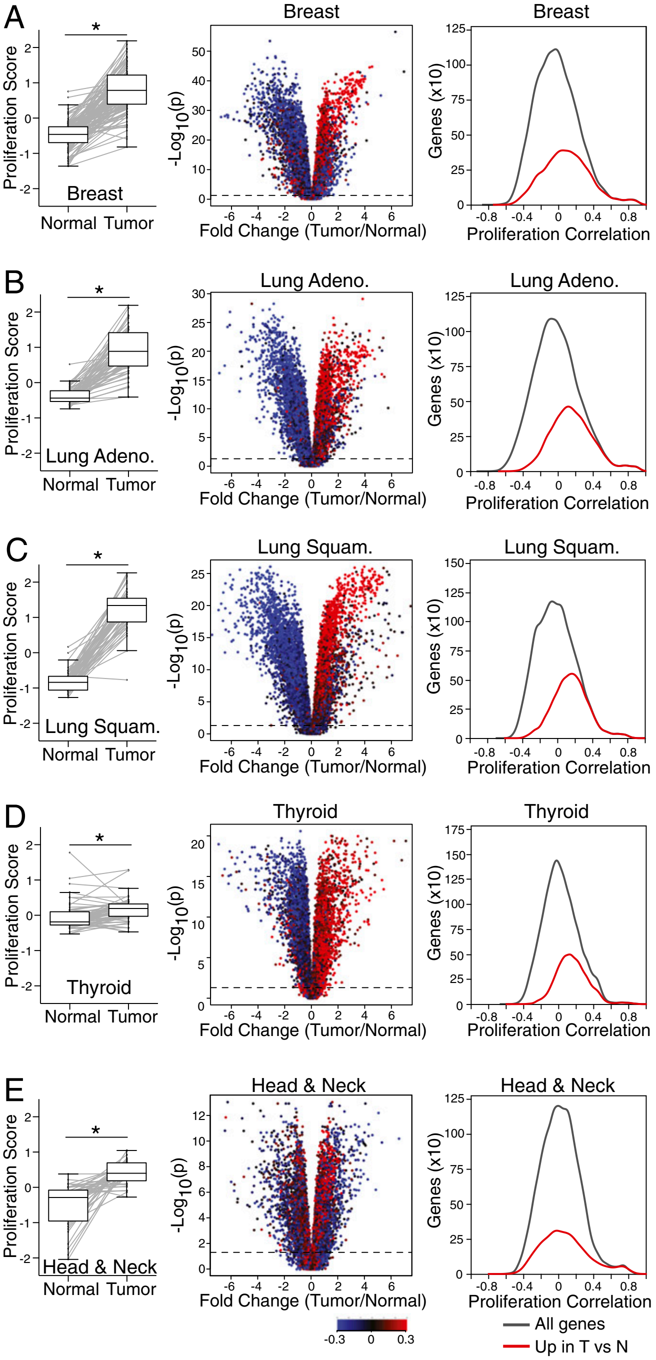
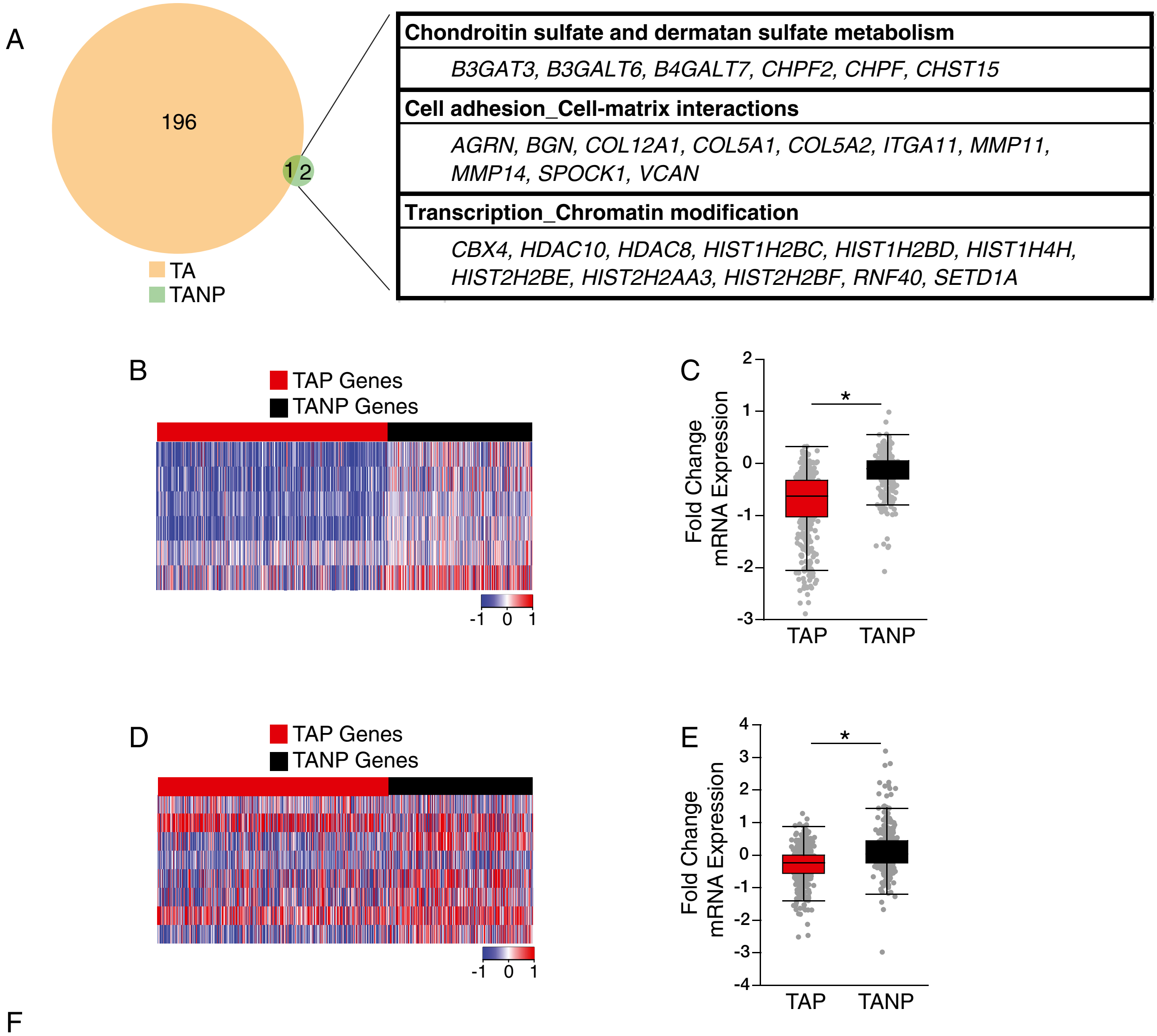


Fig. 3. Proliferation-dependent and proliferation-independent gene expression in tumors and matched normal samples. Proliferation-dependent expression in tumor and matched normal samples from TCGA breast (BRCA, *n* = 113) (*A*), lung adenocarcinoma (Adeno.; *n* = 57) (*B*), lung squamous (Squam.; *n* = 51) (*C*), thyroid (*n* = 59) (*D*), and head and neck (*n* = 42) (*E*) datasets. Box plots show the proliferation scores of matched tumor and normal samples. Lines connect samples from the same patient, and box plot summarizes the data. Volcano plots represent the gene-level log-fold change of tumor vs. normal (*x* axis) and –log(P) (*y* axis) where *P* is derived from paired Welch's *t* tests. Each data point represents a gene and is colored according to its correlation with proliferation, where red indicates a positive correlation and blue indicates a negative correlation. *P > 0.05 paired Welch's *t* test. Histograms are the degree to which genes correlate with the proliferation score. The black line is all expressed genes, and the red line is the genes that are up in tumors relative to normal.



F

Symbol	Description	Inhibitor	Class	Cancer Types
CSNK1A1	Casein kinase I isoform alpha	+	kinase	BRCA, KIRC, LUAD, LUSC, PRAD, THCA
CTSB	Cathepsin B	+	protease	BRCA, COADREAD, HNSC, KIRC, LUAD, THCA
DDR1	Epithelial discoidin domain-containing receptor 1	+	kinase	BRCA, BLCA, HNSC, LUAD, LUSC, THCA
EIF2AK3	Eukaryotic translation initiation factor 2-alpha kinase 3	+	kinase	BRCA, KIRC, LUAD, LUSC, PRAD
ERBB3	Receptor tyrosine-protein kinase erbB-3	+	kinase (inactive)	BRCA, BLCA, COADREAD, KIRC, LUAD, PRAD, THCA, UCEC
F2R	Proteinase-activated receptor 1	+	GPCR	BRCA, BLCA, COADREAD, HNSC, KIRC, PRAD, THCA
HDAC8	Histone deacetylase 8	+	chromatin remodeling	BRCA, BLCA, COADREAD, HNSC, LUAD, PRAD, THCA, UCEC
MAN1B1	Mannosyl-oligosaccharide 1,2-alpha-mannosidase	+	enzyme, N-glycan synthesis	BRCA, BLCA, HNSC, KIRC, LUAD, LUSC, PRAD, THCA, UCEC
MMP14	Matrix metalloproteinase-14	+	protease	BRCA, BLCA, COADREAD, HNSC, KIRC, LUAD, LUSC, THCA
NEU1	Sialidase-1		enzyme, N-glycan synthesis	BRCA, COADREAD, HNSC, KIRC, THCA, UCEC
PLAU	Urokinase-type plasminogen activator	+	protease	BRCA, BLCA, COADREAD, HNSC, LUAD, LUSC, THCA
TLE3	Transducin-like enhancer protein 3		chromatin remodeling	BRCA, LUAD, LUSC, PRAD, THCA

Fig. 4. Proliferation- and lineage-independent pathways and genes. (*A*) Venn diagram of GeneGo pathways that are enriched in genes elevated in tumors relative to normal in breast and four additional tumor types (TA genes, peach), and the subset of these genes that are proliferation-independent (TANP genes, green). Table lists the TANP-enriched pathways and genes. (*B*) Effect of chemotherapy on gene expression. Heat map represents the fold change ($\log_2$ ratios) in mRNA levels pre- and postchemotherapy treatment for TAP and TANP genes. Red indicates higher expression after chemotherapy treatment, and blue indicates lower expression. (*C*) Differential effect of chemotherapy on gene expression of proliferation-dependent and proliferation-independent genes. Box plot shows the fold change in mRNA ($\log_2$ ratio) before and after chemotherapy of TAP ($n = 509$) and TANP ($n = 320$) genes. *$P < 0.05$, Student's *t* test. (*D*) Gene expression in ALDH$^{+/-}$ cells. Heat map represents the fold change ($\log_2$ ratios) in TAP and TANP gene expression in ALDH$^+$ vs. ALDH$^-$ breast tumor cells. (*E*) Gene expression of proliferation-dependent and proliferation-independent genes in ALDH$^+$ cells relative to ALDH$^-$ cells. Box plot shows the fold change in mRNA ($\log_2$ ratio) of TAP ($n = 555$) and TANP ($n = 343$) genes. *$P < 0.05$, Student's *t* test. (*F*) Table lists the TANP genes that are druggable, essential in shRNA drop-out screens, not down-regulated by chemotherapy, and enriched in ALDH$^+$ stem-like cancer cells.

their expression is more likely to be maintained throughout treatment.

We next investigated the expression of the TANP genes in stem-like cancer cells, a subpopulation of tumor cells that are also implicated in chemoresistance (38, 39). We found that 93% of the TANP genes were either higher or not significantly different in ALDH$^+$ breast cancer cells relative to ALDH$^-$ cells, and were 3.7 times more likely than the TAP genes to be expressed at higher levels in ALDH$^+$ cells (risk ratio = 3.7; P = 0.000342; Fig. 4 *D* and *E* and Dataset S6*B*). These data suggest that the TANP genes are more likely to be elevated and stably expressed in cells with stem/progenitor properties, which are generally more resistant to chemotherapy. Of the 363 TANP genes, 12 are essential in shRNA screens (40), druggable (10 with currently available inhibitors) (41), unchanged by chemotherapy treatment, and enriched in stem-like cancer cells, and are therefore the most attractive potential targets for combination therapy in our study (Fig. 4*F* and Dataset S6*B*). They consist of four kinases [*CSNK1A1*, *DDR1*, *EIF2AK3*, and *ERBB3* (inactive)], three proteases (*CTSB*, *MMP14*, and *PLAU*), two enzymes involved in *N*-glycan synthesis (*MAN1B1* and *NEU1*), two enzymes implicated in chromatin remodeling (*HDAC8* and *TLE3*), and one protease-activated GPCR (*F2R*).

Discussion

The ability to directly target the protein products of mutated driver oncogenes is a landmark accomplishment in cancer research. However, single-agent targeted therapies generally are not curative because of the acquisition of drug resistance (5–7). This highlights the importance of developing combination therapies that prevent recurrence, and the need to identify additional nononcogene targets or contexts in which they might be most effective. Here, we used an unbiased computational method to analyze genes that are differentially expressed in breast tumors with distinct genetic alterations. This analysis revealed that there are relatively few genes that are specifically associated with distinct genetic alterations. Rather, gene expression variation associated with specific genetic alterations could be explained by two primary factors: proliferation and tumor lineage. Building on these findings, we explored the effect of proliferation on cancer-associated gene expression and identified putative lineage- and proliferation-independent genes that are consistently higher in diverse tumor types. Further, because these genes behave as their proliferation correlation predicts and remain highly expressed after chemotherapy treatment, they represent attractive candidate targets for combination therapies.

Previous analyses have suggested that proliferation and lineage are important contributors to the molecular subtype designation of breast tumors (29, 30). In addition, numerous reports have identified other gene expression signatures that are associated with specific genetic alterations in breast and other cancers (12, 13, 18, 22, 26, 27). Here, we performed a simplified and unbiased analysis that identified proliferation and tumor lineage as more dominant determinants of breast cancer gene expression than genetic alterations. On the basis of these analyses, we concluded that the specific effect of the most common genomic alterations is negligible for the majority of genes, other than those that are coamplified or codeleted. This conclusion is particularly surprising, considering that proliferation and lineage each account for less than 10% of the total variance in gene expression in tumors, leaving a large amount of gene expression variability unexplained by these fundamental factors. It is important to note that the small group of potential alteration-specific associated genes detected by our approach, as well as the genes identified by other groups using more complex analytical methods, are likely to be biologically important and warrant additional study.

Three of the genetic alterations analyzed here (*MYC*, *GATA3*, and *TP53*) encode transcription factors that might be predicted to have specific transcriptional signatures; however, we did not detect such enrichment in tumors carrying these genetic alterations. There are several potential explanations for this finding. First, direct transcriptional targets of MYC may not be specifically enriched in tumors with *MYC* amplification because MYC protein can be directly activated by signaling pathways induced by many oncogenes, making it likely that MYC is active in most proliferating tumor cells regardless of whether the *MYC* gene is amplified. Therefore, our proliferation correction method may "remove" many direct targets of MYC because they are activated in most highly proliferating tumors. Similarly, *TP53* mutations typically result in loss of function for p53's role in cell cycle arrest (42); thus, although mutations in *TP53* result in changes in transcription of proliferation-associated genes, these would be lost by proliferation normalization. Finally, GATA3 is an important transcription factor involved in mammary gland differentiation and is required for the growth of ER$^+$ breast cancers (43). Therefore, it is likely that our proliferation and lineage correction methods remove most genes directly regulated by these transcription factors because they are inherently linked to proliferation and/or lineage, regardless of their mutation status. In contrast, mutations in the NRF2 axis, which controls the response to oxidative stress, are associated with specific transcription signatures in lung, head and neck, and bladder tumors (44, 45). Unlike MYC and p53, however, NRF2 does not appear to be commonly activated in diverse cancer types, and only a subset of NRF2-dependent genes are redundantly regulated by other transcription factors. Therefore, association-based studies can be effective at identifying the direct oncogenic effects of some genetically altered genes.

Although our studies have revealed that the underlying genetic alterations are not broadly or independently predictive of gene expression in tumors, the genetic background of a cell has been shown to determine how cells respond to therapeutic intervention (46–48). For example, tumors harboring mutations in the *BRCA1/2* genes are more sensitive to poly(ADP ribose) polymerase inhibitors (49), an observation that is now taken advantage of in the clinic for ovarian cancer (50). Interestingly, although genetic alterations do not dramatically affect gene expression, gene expression signatures are predictive of response to chemotherapy (10). Thus, both genetic alterations and gene expression patterns can be predictive of therapeutic response even if they may not be highly related to each other.

Many current therapies are limited in effectiveness because of the reduced sensitivity of quiescent cells. For example, standard-of-care chemotherapies effectively kill proliferating tumor cells (10, 51), whereas slower cycling cells, such as stem-like cancer cells, can be resistant (38). In addition, many therapies have cytostatic effects, and tumor cells can reinitiate proliferation after treatment ends (52). In these contexts, the ideal targets for combination therapy would be proliferation-independent, druggable proteins that are not down-regulated by primary treatment and are expressed in drug-resistant cells. By removing the dominant effects of proliferation on tumor-associated gene expression, we have identified genes that meet these criteria. Although further investigation will establish their effectiveness, several of these genes (*CTSB*, *DDR1*, *ERBB3*, *HDAC8*, *MMP14*, *NEU1*, and *PLAU*) have been validated as combination therapy targets in cancer models (53–58), including two (*ERBB3* and *PLAU*) in clinical trials in combination with chemotherapy (59, 60). Notably, many of these genes are implicated in extracellular matrix remodeling (*CTSB*, *DDR1*, *MMP14*, and *PLAU*), and multiple studies have identified mechanisms by which extracellular matrix signaling mediates resistance to apoptotic stimuli (61–63), demonstrating that our method can identify attractive targets for combination therapy.

Together, our results suggest that phenotypic programs are critical determinants of gene expression patterns in tumors that confound analyses of the association of gene expression with genetic alterations. By identifying tumor-associated genes that are independent of two dominant confounding programs (proliferation and tumor lineage), we have identified genes that warrant further analysis as therapeutic targets.

Methods

Identification of Gene:Alteration Associations. Genes associated with recurrent somatic genetic alterations were identified in TCGA data (2015-02-24 datafreeze) obtained from the University of California, Santa Cruz cancer browser (https://genome-cancer.ucsc.edu/proj/site/hgHeatmap/). For copy number associations, RNAseq normalized expression counts [HiSeqV2, $\log_2(x + 1)$ transformed RNAseq by expectation maximization (RSEM) normalized count] from amplified or deleted tumors (GISTIC scores of 2 and −2, respectively) were compared with tumors diploid at that locus (GISTIC score 0). Genes on the same chromosome arm as copy number alterations were eliminated if they were coamplified or codeleted with the genetic alteration in more than 33% of tumors. For mutation associations, a tumor was considered mutated at a given locus if it had a somatic alteration call of 1 in the PANCAN AWG data. Low-expression genes, defined as >5% tumors with 0 counts or mean counts <4, were removed from the analysis. A minimum of five samples with a genetic alteration was required for all statistical tests. Statistical significance was determined using the *t*-test function in version 3.2.2. Bonferroni-corrected $P < 0.05$ was the threshold for statistical significance. A pan-cancer dataset was compiled from 10 TCGA cohorts (BLCA, COADREAD, GBM, HNSC, LGG, LUAD, LUSC, OV, PRAD, and UCEC). The RNAseq gene expression values [$\log_2(x + 1)$ transformed RSEM normalized counts] in the merged set were median centered within each cohort. For data from the Molecular Taxonomy of Breast Cancer International Consortium, normalized gene expression data, copy number data, and paired clinical feature data were obtained from the publicly available European Genome-phenome Archive (IDs EGAD00010000210 and EGAD0001000021) (12, 64).

Pathway Enrichment. Whole-genome enrichment analysis was performed using MetaCore (GeneGo; Thomson Reuters). The background list for the breast analysis was the 14,209 genes that passed a low-expression filter. The default GeneGO background list was used for enrichment of genes from multiple cancer types. Pathway depictions were generated with Cytoscape_v2.6.3.

Molecular Subtyping TCGA Breast Tumors. RNAseq normalized expression counts were subjected to the PAM50 classifying function (https://genome.unc.edu/pubsup/breastGEO/). To account for the fact that the PAM50 training set is ∼50% ER^-, confirmed ER^- samples were randomly subsampled for an equal number of ER^+ samples. The median of the subset was calculated and assigned to the PAM50 calibrationParameters argument. ER status ($ER^{+/-}$) was extracted from TCGA clinical data files.

Linear Modeling. Linear regression models were built for each gene using normalized mRNA expression values as the outcome variable and proliferation score (32), *ESR1* mRNA level, or genetic alterations [binary indicator variables representing amplified or deleted tumors (GISTIC scores of 2 and −2, respectively) and samples diploid at that locus (GISTIC score 0)] as predictors. The *lm* function from the R *stats* package was used for all modeling. The *lrtest* package was used to carry out likelihood ratio tests of nested models. The frequency of coamplification and codeletion was calculated for genes on the same chromosome arm as the recurrent breast alteration, using the *Org.Hs.eg.db* annotation package. Genes with no chromosomal location in *Org.Hs.eg.db* were withheld from the Δr^2 figure.

Proliferation and Lineage Correction. To correct for proliferation and *ESR1*, a linear model was constructed using proliferation scores or *ESR1* RNA-Seq mRNA levels for each sample fitted to the expression of each gene using the *lm* function in R. Each expression measurement was then substituted by the sum of its residual and mean expression across the dataset. This correction method is the same as the proliferation correction that has been previously published (32). Total variance in gene expression was calculated by summing the variance for each expressed gene before and after proliferation or lineage correction.

Expression Microarray Analysis. Data from the Gene Expression Omnibus were analyzed in R, using the *GEOquery* package. To assess the effect of chemotherapy on gene expression, $\log_2$ fold change values were calculated from RMA-processed expression values from pretreated (bl) and paired treated samples (c2) of patients (GSE18728). Only patients who responded to chemotherapy (response category: R) were analyzed. To evaluate expression in stem-like cancer cells, $\log_2$ fold change values ($ALDH^+/ALDH^-$) were calculated from RMA-processed expression values (GSE52327).

Box plots, correlation scatterplots, and their associated *P* values were generated in JMP Pro-11.0.0. Box plots are derived from the median (vertical line within box), the interquartile range (ends of the box), and the first or third quartile ±1.5 * interquartile range (whiskers). Clustering was performed in Cluster 3.0, using average linkage and Pearson's correlation (uncentered). Venn diagrams were generated with Venny 2.1 (bioinfogp.cnb.csic.es/tools/venny/).

Statistics. Bonferroni-corrected two-tailed Welch's *t* tests were used to identify differences in gene expression in tumors bearing specific genetic alterations. Paired two-sample *t* tests were used to compare matched tumor/normal, chemotherapy treated/untreated, and $ALDH^{+/-}$ samples. For microarray data, unpaired two-tailed Student's *t* tests were performed. Pearson's correlation coefficients and *P* values were calculated for linear correlation analysis. The statistical tests for GeneGO pathway and TA/TANP gene enrichment analysis were false discovery rate-corrected hypergeometric tests and Fisher's exact tests, respectively. For all statistical tests, the threshold for statistical significance was set at 0.05.

ACKNOWLEDGMENTS. We thank Katherine Hoadley and Charles Perou for help with molecular subtyping, and Gordon Mills, Astrid Ruefli-Brasse, and Angie Martinez Gakidis for many helpful discussions. This research was supported by Roche Pharmaceuticals (L.M.S., I.S.H., J.S.B., and J.L.C.) and a fellowship from the American Cancer Society (J.L.C.).

1. Vogelstein B, et al. (2013) Cancer genome landscapes. *Science* 339:1546–1558.
2. Forbes SA, et al. (2015) COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 43:D805–D811.
3. Cerami E, et al. (2012) The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2:401–404.
4. Weinstein IB (2002) Cancer. Addiction to oncogenes: The Achilles heal of cancer. *Science* 297:63–64.
5. Ross JS, et al. (2004) Targeted therapies for cancer 2004. *Am J Clin Pathol* 122:598–609.
6. Higgins MJ, Baselga J (2011) Targeted therapies for breast cancer. *J Clin Invest* 121:3797–3803.
7. Lackner MR, Wilson TR, Settleman J (2012) Mechanisms of acquired resistance to targeted cancer therapies. *Future Oncol* 8:999–1014.
8. Luo J, Solimini NL, Elledge SJ (2009) Principles of cancer therapy: Oncogene and non-oncogene addiction. *Cell* 136:823–837.
9. Vander Heiden MG (2011) Targeting cancer metabolism: A therapeutic window opens. *Nat Rev Drug Discov* 10:671–684.
10. Stover DG, et al. (2016) The role of proliferation in determining response to neoadjuvant chemotherapy in breast cancer: A gene expression-based meta-analysis. *Clin Cancer Res* 22:6039–6050.
11. Anonymous; Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature* 490:61–70.
12. Curtis C, et al.; METABRIC Group (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486:346–352.
13. Ding J, et al. (2015) Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nat Commun* 6:8554.
14. Alvarez MJ, et al. (2016) Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet* 48:838–847.
15. Hofree M, Shen JP, Carter H, Gross A, Ideker T (2013) Network-based stratification of tumor mutations. *Nat Methods* 10:1108–1115.
16. Shi K, Gao L, Wang B (2016) Discovering potential cancer driver genes by an integrated network-based approach. *Mol Biosyst* 12:2921–2931.
17. Gonzalez-Perez A, et al.; International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group (2013) Computational approaches to identify functional genetic variants in cancer genomes. *Nat Methods* 10:723–729.
18. Paull EO, et al. (2013) Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (TieDIE). *Bioinformatics* 29:2757–2764.
19. Jia P, Zhao Z (2016) Impacts of somatic mutations on gene expression: An association perspective. *Brief Bioinform*.
20. Ng S, et al. (2012) PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics* 28:i640–i646.
21. Shlien A, et al.; ICGC Breast Cancer Working Group, Oslo Breast Cancer Research Consortium (2016) Direct transcriptional consequences of somatic mutation in breast cancer. *Cell Rep* 16:2032–2046.
22. Bashashati A, et al. (2012) DriverNet: Uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol* 13:R124.
23. Akavia UD, et al. (2010) An integrated approach to uncover drivers of cancer. *Cell* 143:1005–1017.

24. Hou JP, Ma J (2014) DawnRank: Discovering personalized driver genes in cancer. *Genome Med* 6:56.
25. Rykunov D, et al. (2016) A new molecular signature method for prediction of driver cancer pathways from transcriptional data. *Nucleic Acids Res* 44:e110.
26. Masica DL, Karchin R (2011) Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res* 71: 4550–4561.
27. Gatto F, Schulze A, Nielsen J (2016) Systematic analysis reveals that cancer mutations converge on deregulated metabolism of arachidonate and xenobiotics. *Cell Rep* 16: 878–895.
28. Jiang T, et al. (2016) Predictors of chemosensitivity in triple negative breast cancer: An integrated genomic analysis. *PLoS Med* 13:e1002193.
29. Perou CM, et al. (2000) Molecular portraits of human breast tumours. *Nature* 406: 747–752.
30. Haibe-Kains B, et al. (2012) A three-gene model to robustly identify breast cancer molecular subtypes. *J Natl Cancer Inst* 104:311–325.
31. Hatzis C, et al. (2011) A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* 305:1873–1881.
32. Venet D, Dumont JE, Detours V (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* 7:e1002240.
33. Sotiriou C, Pusztai L (2009) Gene-expression signatures in breast cancer. *N Engl J Med* 360:790–800.
34. Gatza ML, Silva GO, Parker JS, Fan C, Perou CM (2014) An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. *Nat Genet* 46:1051–1059.
35. Hoadley KA, et al.; Cancer Genome Atlas Research Network (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 158:929–944.
36. Shah MA, Schwartz GK (2001) Cell cycle-mediated drug resistance: An emerging concept in cancer therapy. *Clin Cancer Res* 7:2168–2181.
37. Korde LA, et al. (2010) Gene expression pathway analysis to predict response to neoadjuvant docetaxel and capecitabine for breast cancer. *Breast Cancer Res Treat* 119: 685–699.
38. Moore N, Lyle S (2011) Quiescent, slow-cycling stem cell populations in cancer: A review of the evidence and discussion of significance. *J Oncol* 2011:396076.
39. Liu S, et al. (2013) Breast cancer stem cells transition between epithelial and mesenchymal states reflective of their normal counterparts. *Stem Cell Reports* 2:78–91.
40. Marcotte R, et al. (2016) Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance. *Cell* 164:293–309.
41. Griffith M, et al. (2013) DGIdb: Mining the druggable genome. *Nat Methods* 10: 1209–1210.
42. Pfister NT, Prives C (2017) Transcriptional regulation by wild-type and cancer-related mutant forms of p53. *Cold Spring Harb Perspect Med* 7:a026054.
43. Buchwalter G, et al. (2013) PDEF promotes luminal differentiation and acts as a survival factor for ER-positive breast cancer cells. *Cancer Cell* 23:753–767.
44. Goldstein LD, et al. (2016) Recurrent loss of NFE2L2 exon 2 is a mechanism for Nrf2 pathway activation in human cancers. *Cell Rep* 16:2605–2617.
45. Dwivedi S, Rajasekar N, Hanif K, Nath C, Shukla R (2016) Sulforaphane ameliorates okadaic acid-induced memory impairment in rats by activating the Nrf2/HO-1 antioxidant pathway. *Mol Neurobiol* 53:5310–5323.
46. Garnett MJ, et al. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 483:570–575.
47. Barretina J, et al. (2012) The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483:603–607.
48. Roden DM, George AL, Jr (2002) The genetic basis of variability in drug responses. *Nat Rev Drug Discov* 1:37–44.
49. Farmer H, et al. (2005) Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* 434:917–921.
50. Meehan RS, Chen AP (2016) New treatment option for ovarian cancer: PARP inhibitors. *Gynecol Oncol Res Pract* 3:3.
51. Wirapati P, et al. (2008) Meta-analysis of gene expression profiles in breast cancer: Toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res* 10:R65.
52. Magbanua MJ, et al.; I-SPY 1 TRIAL Investigators (2015) Serial expression analysis of breast tumors during neoadjuvant chemotherapy reveals changes in cell cycle and immune pathways associated with recurrence and response. *Breast Cancer Res* 17:73.
53. Shree T, et al. (2011) Macrophages and cathepsin proteases blunt chemotherapeutic response in breast cancer. *Genes Dev* 25:2465–2479.
54. Ambrogio C, et al. (2016) Combined inhibition of DDR1 and notch signaling is a therapeutic strategy for KRAS-driven lung adenocarcinoma. *Nat Med* 22:270–277.
55. Rettig I, et al. (2015) Selective inhibition of HDAC8 decreases neuroblastoma growth in vitro and in vivo and enhances retinoic acid-mediated differentiation. *Cell Death Dis* 6:e1657.
56. Ager EI, et al. (2015) Blockade of MMP14 activity in murine breast carcinomas: Implications for macrophages, vessels, and radiotherapy. *J Natl Cancer Inst* 107:djv017.
57. O'Shea LK, Abdulkhalek S, Allison S, Neufeld RJ, Szewczuk MR (2014) Therapeutic targeting of Neu1 sialidase with oseltamivir phosphate (Tamiflu®) disables cancer cell survival in human pancreatic cancer with acquired chemoresistance. *Onco Targets Ther* 7:117–134.
58. Heinemann V, et al. (2013) Phase II randomised proof-of-concept study of the urokinase inhibitor upamostat (WX-671) in combination with gemcitabine compared with gemcitabine alone in patients with non-resectable, locally advanced pancreatic cancer. *Br J Cancer* 108:766–770.
59. Liu JF, et al. (2016) Randomized phase II trial of seribantumab in combination with paclitaxel in patients with advanced platinum-resistant or -refractory ovarian cancer. *J Clin Oncol* 34:4345–4353.
60. Duffy MJ, McGowan PM, Harbeck N, Thomssen C, Schmitt M (2014) uPA and PAI-1 as biomarkers in breast cancer: Validated for clinical use in level-of-evidence-1 studies. *Breast Cancer Res* 16:428.
61. Friedland JC, et al. (2007) alpha6beta4 integrin activates Rac-dependent p21-activated kinase 1 to drive NF-kappaB-dependent resistance to apoptosis in 3D mammary acini. *J Cell Sci* 120:3700–3712.
62. Weaver VM, et al. (2002) beta4 integrin-dependent formation of polarized three-dimensional architecture confers resistance to apoptosis in normal and malignant mammary epithelium. *Cancer Cell* 2:205–216.
63. Muranen T, et al. (2012) Inhibition of PI3K/mTOR leads to adaptive resistance in matrix-attached cancer cells. *Cancer Cell* 21:227–239.
64. Pereira B, et al. (2016) The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat Commun* 7:11479.