HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2015 April 01.

Published in final edited form as:

Nat Genet. 2014 October ; 46(10): 1051–1059. doi:10.1038/ng.3073.

An integrated genomics approach identifies drivers of proliferation in luminal subtype human breast cancer

Michael L. Gatza[1,2], **Grace O. Silva**[1,2,3], **Joel S. Parker**[1,2], **Cheng Fan**[1], and **Charles M. Perou**[1,2,3,4,*]

[1]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC

[2]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC

[3]Curriculum in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC

[4]Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC

Abstract

Elucidating the molecular drivers of human breast cancers requires a strategy capable of integrating multiple forms of data and an ability to interpret the functional consequences of a given genetic aberration. Here we present an integrated genomic strategy based on the use of gene expression signatures of oncogenic pathway activity (n=52) as a framework to analyze DNA copy number alterations in combination with data from a genome-wide RNAi screen. We identify specific DNA amplifications, and importantly, essential genes within these amplicons representing key genetic drivers, including known and novel regulators of oncogenesis. The genes identified include eight that are essential for cell proliferation (*FGD5*, *METTL6*, *CPT1A*, *DTX3*, *MRPS23*, *EIF2S2*, *EIF6* and *SLC2A10*) and are uniquely amplified in patients with highly proliferative luminal breast tumors, a clinical subset of patients for which few therapeutic options are effective. Our results demonstrate that this general strategy has the potential to identify putative therapeutic targets within amplicons through an integrated use of genetic, genomic, and genome-wide RNAi data sets.

Tumorigenesis is driven by a combination of inherited and acquired genetic alterations resulting in a complex and heterogeneous disease. The ability to dissect this heterogeneity is

*Correspondence: Charles M. Perou, PhD., Lineberger Comprehensive Cancer Center, 450 West Drive, CB7295, University of North Carolina, Chapel Hill, NC, 27599, USA, Tel. 919-843-5740, cperou@med.unc.edu.

Author Contributions

MLG, JSP and CMP conceived and designed the study. MLG, GOS, and CF performed analyses. MLG and CMP wrote the manuscript. All authors have reviewed and approved the final manuscript.

URLs

TCGA data portal: https://tcga-data.nci.nih.gov/docs/publications/brca_2012/, GenePattern: http://genepattern.broadinstitute.org/gp/pages/login.jsf, COLT database: http://colt.ccbr.utoronto.ca/cancer/login.pl, European Genome-phenome Archive at the European Bioinformatics Institute: https://www.ebi.ac.uk/ega/, Gene Expression Omnibus (GEO) database: http://www.ncbi.nlm.nih.gov/geo/.

critical to understand the relevance of these alterations for disease phenotypes but also to enable the development of rational therapeutic strategies that can match the characteristics of the individual patient's tumor. Many studies, including reports from The Cancer Genome Atlas (TCGA) project, have made use of the power of multi-platform genomic analyses to identify known and novel genetic drivers of tumor phenotypes[1–3]. This has led to the identification of disease subgroups with distinct characteristics and in some instances, distinct genetic mechanisms of disease[1,2,4]. The strength of this approach relies on the integration of large-scale genomic data to reveal biological co-variation that cannot be identified when using a single technology. A weakness of this approach is in the interpretation of the underlying biology, which generally represents an inference about pathway activity based upon prior knowledge concerning an individual gene mutation/alteration.

Altered signaling pathway activity is an important determinant of the biology of a tumor and may predict therapeutic response; therefore, identifying mechanisms driving key tumorigenic pathways is essential to understand the transformation process[2,5–8]. To take advantage of the vast amounts of existing genomic data, we utilized a series of experimentally derived gene expression signatures capable of measuring oncogene or tumor suppressor pathway activity, aspects of the tumor microenvironment, and other tumor characteristics including proliferation rate, as a framework by which to integrate multiple forms of genomic data. Our results identify patterns of oncogenic signaling within each of the molecular subtypes of breast cancer, many of which directly correlate with DNA copy number aberrations. By further analyzing functional data from a genome-wide RNAi screen[9], we identified genes essential for cell viability in a pathway-dependent, and in some cases, a subtype-dependent manner. Our results identify a small number of DNA amplifications as potential drivers of proliferation in poor outcome luminal/ER+ breast cancers, and in general terms, we outline an approach that could be applied to many other tumor types where multi-platform genomic data exist.

RESULTS

Subtype-specific patterns of oncogenic signaling

To objectively identify genetic drivers of breast cancer, we examined genomic-based patterns of oncogenic pathway activity, the tumor microenvironment, and other important features in human breast tumors using a panel of 52 previously published gene expression signatures (Supplementary Tables 1 and 2)[10–32]. Each signature was applied to the breast cancer gene expression microarray data (n=476) from the TCGA project (Supplementary Table 3) for which the molecular intrinsic subtype had been determined[2]. Consistent patterns of pathway activity emerged for each subtype, as illustrated in Figure 1A, and quantitatively assessed by an ANOVA test followed by a Tukey test for pair-wise comparison (Figure 1B, Supplementary Table 4). Analyzing differences across subtypes based on these 52 features demonstrated that the strongest correlation between samples existed within each molecular subtype (Supplementary Figure 1).

Patterns of pathway activity recapitulated known characteristics of each subtype, including dysregulation of pathways that can be linked to female Hormone Receptors (HR), and/or

oncogenes, and/or tumor suppressor mutation status (Figure 1). For example, basal-like tumors, which represent ~80% of triple negative breast cancers, are characterized by low HR signaling, mutant p53 signaling and high expression of proliferation pathway activity (Figure 1). Likewise, HER2-enriched (HER2E) tumors show high expression of the Her2[11] and Her2- Amplicon (Amp)[12] signatures, while LumA tumors show high HR signaling and wild-type p53 signaling. Highly proliferative LumB tumors, which also show some HR signaling, are distinguished from less proliferative LumA samples by increased proliferation-associated pathways. Thus these data robustly recapitulate many previously published pathways versus subtype associations.

A Pearson correlation to assess the concordance between each of the 52 signatures (Supplementary Figure 2, Supplementary Table 5) identified strong relationships between independent signatures for a given pathway as well as between related pathways. For example, two Myc signatures[11,15,32] demonstrate an R value of 0.72, while a PIK3CA[18] and PTEN-deleted[27] signatures had a R value of 0.82. Signatures scoring different pathways were also concordant; for instance, Myc-mediated regulation of E2F signaling[33] was identified by the association between the RB-LOH[16] and Myc[15] signatures (R=0.79), while EGFR-mediated activation of Stat3 signaling[34] was recapitulated by the EGFR[11,32] and Stat3[11,32] (R=0.72) signatures. These results provide a measure of validity for each signature but, since differences do exist between signatures for a specific pathway, suggest that each provides an opportunity to investigate a particular pathway taking into account the genetic manipulation used to develop a given signature.

Characterization of pathway-specific copy number alterations

We next utilized DNA copy number data from the TCGA project (n=476) to identify Copy Number Alterations (CNA) associated with pathway activity (Figure 2A). We first identified genes for which CNAs were positively (or inversely) correlated with pathway activity using a Spearman Rank correlation (Bonferroni corrected to control the family-wise error rate) to assess the relationship between pathway score and gene-level DNA segment score (Supplementary Figure 3, Supplementary Tables 6–57). Secondly, we used a Fisher's exact test (Bonferroni corrected) to calculate the frequency of CNA gains (including high-level amplification and gains) or losses (including loss of heterozygosity and deletions) in samples with high (top quartile) pathway activity compared to all other samples (low activity)(Supplementary Figure 4, Supplementary Tables 6–57). To reduce potential false-positive results associated with either strategy alone, for each signature we focused on those genes that were significant in both analyses (Figure 2A); potential drivers of pathway activity had a positive correlation and a higher amplification frequency in samples with high pathway activity, while potential repressors had a negative correlation and increased frequency of copy number losses. Mapping genes that met these criteria to chromosomal loci identified pathway-specific patterns of CNA (Figure 2B). Consistent with previous studies reporting that basal-like tumors have a higher incidence and larger spectrum of CNAs[2,35], pathways associated with basal-like tumors have more complex patterns of CNA when compared to luminal-associated pathways.

To further assess the validity of this strategy, we investigated the relationship between pathway activity and a chromosomal alteration of known causative activity. We first focused on the Her2-Amp signature[12] since this signature is comprised of genes located at the 17q loci and since *ERBB2*/17q amplification is the dominant driver of this pathway. As illustrated in Figure 2C (Supplementary Table 27) *ERBB2* is amplified in 84.9% of samples with high (top quartile) pathway activity compared to 7.3% of low scoring samples ($q= 1.1 \times 10^{-55}$); likewise this relationship has a positive Spearman rank correlation ($q= 2.4\times 10^{-108}$). While several other alterations, including *MYC* amplification ($q=1.1\times 10^{-02}$, $q=6.3\times 10^{-03}$) were also associated with this signature, thus identifying a previously known relationship[36], *ERBB2* /17q amplification was the dominant alteration identified thus providing a robust positive control for this strategy. As expected, similar results were observed when analyzing the Her2 pathway using the independently developed Her2[11,32] signature (Supplementary Table 26).

This strategy was further validated by assessing the relationship between CNA and pathways associated with a more complex genomic landscape. Previous studies from our group suggest that the Her1-C2[13] signature predominantly measures the RAS/RAF/MEK arm of the EGFR pathway[13]. Consistent with this observation, we detected a correlation between the Her1-C2 signature ($q<0.01$) and *GRB2, SOS1, KRAS, BRAF, PIK3CA, PIK3CB,* and *MYC* genomic DNA amplifications as well as a negative correlation ($q<0.01$) with loss of *NF1* and PI3K repressors *INPP4B* and *PTEN* (Figure 2D, Supplementary Table 24). Finally, we analyzed CNA associated with the RB-LOH[16] signature (Figure 2E, Supplementary Table 47) and identified associations between it and CNA of known RB/E2F components including loss of *RB1* and gains of *E2F1* and/or *E2F3*. Consistent with the role of the RB/E2F pathway in mediating cell cycle progression and proliferation[37], *CCND2, CCND3, and MYC* amplification also correlated with this signature. Collectively these results demonstrate that this strategy is able to link CNA with pathway activity and does so by focusing on all aspects of the pathway, often beyond the dominant regulator, potentially allowing for the identification of novel regulatory components.

Identification of amplified genes linked to pathway activity

Given the ability of this strategy to identify known CNA of pathway activity, we next used this approach to identify novel drivers of pathway activity. Because highly proliferative luminal/ER+ tumors have a poor prognosis and poor responses to existing therapies[38,39], we sought to identify amplified genes/CNA associated with our previously published 11-gene PAM50 Proliferation signature with the hope that these might represent targetable drivers of oncogenesis.

To identify those genes that are specifically altered in highly proliferative luminal tumors, while excluding those genes that are associated with proliferation irrespective of subtype, analyses were performed on two subsets of samples: all tumors and all non-basal-like tumors (henceforth called luminal tumors). Some rationale for this binary distinction comes from recent TCGA studies where 12 tumor types were studied simultaneously and showed that breast tumors formed two groups, namely Basal-like and all other breast tumors (called

Luminal and including HER2+ tumors), suggesting that breast cancer might broadly be considered two main disease types[40].

Examining the TCGA breast cancer dataset using the PAM50 Proliferation signature[31], Basal-like, LumB, and HER2E tumors were found to have the highest proliferation levels (Figures 3A and 3B) with the top quartile (Figure 3C) comprised of Basal-like (49.6%), LumB (33.6%) and HER2E (16.8%) patients, whereas the top quartile of proliferative luminal tumors (Figure 3D) contained LumB (68.0%) and HER2E (32.0%) patients. Using the PAM50 Proliferation signature, we examined the frequency of CNA gains and losses in highly proliferative (top quartile) tumors relative to less proliferative samples irrespective of subtype using the previously discussed statistical strategies (Figure 3E–F, Supplementary Table 43). To identify those genes that are specifically amplified in highly proliferative luminal breast cancer, these analyses were repeated using the luminal tumor subset (Figures 3G–H, Supplementary Table 58). Analyzing both populations of patients identified three classes of proliferation-associated regions ($q<0.05$): (1) CNAs associated irrespective of subtype, (2) those altered in basal-like tumors, and (3) those altered in highly proliferative luminal tumors. These results allowed us to focus our analyses on those genes within regions that were uniquely altered in highly proliferative luminal tumors by censoring proliferation-associated genes altered in basal-like breast cancer (*e.g.TP53* or *INPP4B* loss), or that were altered irrespective of molecular subtype (*e.g. RB1* loss or *MYC* amplification). These analyses identified a number of regions including 3p25, 5p15, 11q13, 17q22, and 20q11-13 that were uniquely amplified in highly proliferative luminal tumors.

Identification of pathway-specific essential genes

To distinguish essential from non-essential genes in amplified regions associated with proliferation in luminal tumors, we next examined data from a genome-wide RNAi screen of multiple breast tumor-derived cell lines[9]. The 52 gene expression signatures were applied to a panel (GSE12777)[41] of breast cancer cell lines (Supplementary Figure 5, Supplementary Table 59), of which 27 had both mRNA expression data and were part of a RNAi proliferation screen in which a genome-wide shRNA library (~16,000 genes) was used to identify essential genes (Figure 4A)[9]. For each signature, a negative Spearman rank correlation was used to identify pathway-specific essential genes (Figure 4B, Supplementary Table 60) by comparing the pathway score against the normalized shRNA score across the panel of 27 cell lines. These analyses identified inverse relationships between the abundance of shRNAs targeting key regulatory genes and pathway scores. For instance, examining the ER[11,32], Her2[11,32], or Stat1[42] signatures as controls (Figures 4C–4E) showed a negative correlation between pathway score and shRNA against *ESR1* ($p=0.0143$), *ERBB2* ($p=0.0227$), and *STAT1* ($p=0.0049$) or *JAK3* ($p=0.00013$), respectively. These associations were expected for the ER and Her2 pathways given the relationship between HER2 or ERα mRNA and/or protein expression and the response of cell lines or tumors to trastuzumab or anti-estrogen therapies, respectively. These results confirm that this approach is able to identify essential genes known to be functionally associated with pathway activity, thereby suggesting that these data can serve as a biological filter to distinguish pathway-specific essential from non-essential genes.

Amplified essential genes linked to luminal tumor proliferation

We next sought to distinguish between essential and non-essential genes within regions specifically amplified in highly proliferative luminal tumors. For each subset of tumors we identified genes in amplified regions that were positively correlated with proliferation and showed an increased amplification frequency ($q<0.05$). We next examined the RNAi data in all breast cancer cell lines (Supplementary Figure 6A) and in luminal/HER2+ cell lines (Supplementary Figure 6B) in the context of the PAM50 Proliferation signature (Supplementary Table 61). Comparing the results of these four analyses (Figure 5A) identified 19 genes that were uniquely essential for cell viability in luminal cell lines and that were amplified in highly proliferative luminal tumors (Figure 5B). Two additional genes, *DNAJC5* and *SNX21*, were identified by RNAi analysis but were initially overlooked in the CNA analyses since they were located at the cusp of two segmented regions; however, since genes overlapping both 5' and 3' of these genes were amplified, these were included for further investigation. Of these 21 candidate genes, twelve showed a significant relationship between DNA copy number levels and mRNA expression in luminal tumors (Supplementary Figure 7). Interestingly, half of these genes were located at 20q11-13 (*EIF2S2, EIF6, SLC2A10, SNX21, ZBTB46* and *DNAJC5*), with two located at 3p25.1 (*FGD5* and *METTL6*), and the remaining genes located at 5p15 (*TRIO*), 11q13 (*CPT1A*), 12q13 (*DTX3*), and 17q22-23 (*MRPS23*). In contrast, permuting the data labels 1000 times for each analysis, in all samples and in luminal samples alone, identified no gene that met this statistical threshold, suggesting that the 21 candidate genes could not have been identified by chance alone.

Validation of identified candidate genes

We next confirmed that the majority of the identified genes were significantly amplified in highly proliferative luminal breast tumors by analyzing an independent breast tumor dataset (METABRIC, n=1,992) for which both mRNA expression and genomic DNA CNA data were available[3]. Of the twelve identified genes, nine (*FGD5, METTL6, TRIO, CPT1A, DTX3, MRPS23, EIFS2S, EIF6*, and *SLC2A10*) were present on both platforms used in the METABRIC study. Each of these genes (Supplementary Figure 8) showed a significant relationship between CNA status and mRNA expression in luminal breast tumors (n=1,333). Importantly, eight of the nine, the exception being *TRIO*, also showed an increased amplification frequency in highly proliferative (top quartile) luminal tumors (Supplementary Figure 9), thus recapitulating one of our main findings.

To confirm that DNA mutations of genes associated with proliferation in luminal tumors did not confound these results, we examined the relationship between the 11-gene Proliferation score and the mutation frequency of the previously identified 35 significantly mutated genes in human breast cancers reported by TCGA[2]. Using a Fisher's exact test (Bonferroni corrected), we determined that only *TP53* ($q=7.0\times10^{-10}$) and *MAP3K1* ($q=5.0\times10^{-03}$) mutations occurred at significantly different frequencies in highly proliferative (top quartile) luminal tumors compared to all other samples; *TP53* mutations occurred more frequently (51.6% vs. 18.6%) while *MAP3K1* (2.1% vs.12.4%) mutations occurred less frequently in highly proliferative luminal tumors (Supplementary Table 62). Moreover, we found no significant relationship between *MAP3K1* or *TP53* mutation status (Bonferroni corrected

Fisher's exact test, q>0.05) and the amplification status of each candidate gene (Supplementary Table 63) in highly proliferative luminal tumors.

Lastly, we investigated whether expression of the candidate genes, independent of CNA status, was associated with proliferation in luminal breast tumors. By comparing the mRNA expression patterns of each candidate gene in highly proliferative luminal tumor samples (top quartile) against all other samples, we found that tumors lacking CNA of each candidate gene fell into three categories: those that exhibited a positive relationship between mRNA expression and the PAM50 proliferation signature (*EIF2S2*, *EIF6*, *CPT1A*, and *MRPS23*), those that were anti-correlated with the signature (*DTX3*), and those that showed no correlation (*FGD5*, *METTL6*, and *SLC2A10*) between datasets (Supplementary Figure 10). These data suggest that amplification is a key mechanism driving the expression of these genes. However, our data also suggest, not surprisingly, that for some genes overall high expression may be the driver, which can be accomplished by amplification or through other unknown means.

Candidate gene amplification correlates with poor prognosis

Previous studies have shown that highly proliferative luminal tumors have a poor prognosis[38,39], therefore we investigated what impact amplification of each candidate gene had on overall survival. From the TCGA (n=388)[2] and METABRIC (n=1,333)[3] datasets we extracted the subset of LumA, LumB, or HER2E patients for which survival data was available (Supplementary Tables 64–65). We first analyzed data from the TCGA (Figure 6A–E), and despite the relatively short follow-up (median: 1.7 years), determined that amplification of *FGD5* (P<0.0001, HR: 8.0), *METTL6* (P=0.0003, HR: 5.9), *DTX3* (P=0.0387, HR: 2.6), and *MRPS23* (P=0.0078, HR: 2.9) predicted a significantly worse outcome in luminal breast cancer patients whereas *CPT1A* amplification had no effect on patient survival (P=0.3738). Extending these analyses to the METABRIC dataset (Figures 6F–J), which has a longer median survival time (7.2 years), confirmed that *FGD5* (P=0.0170, HR: 2.0), *METTL6* (P=0.0081, HR: 2.1), *DTX3* (P=0.0098, HR: 1.8), and *MRPS23* (P=0.0020, HR: 1.5) amplification correlated with a poor prognosis, while gain of *CPT1A* had no effect (P=0.099) on luminal breast cancer survival. The remaining three genes, showed no consistent effect on prognosis (Supplementary Figure 11). While it is possibly that other genes within these chromosomal loci are also prognostic, these amplified genes were associated with proliferation *in vivo*, were prognostic in multiple patient cohorts, and essential for cell viability *in vitro*.

Likewise, we determined that for most of the identified candidate genes that failed to meet all our predetermined criteria, amplification alone, without a coordinate increase in mRNA expression, was not sufficient to affect prognosis since only one (*TMEM117*) of these genes showed a consistently poor prognosis in the TCGA and METABRIC datasets (Supplementary Table 66). Lastly, we investigated whether the 12 initial candidate genes were predictive of a poor prognosis when compared with standard prognostic markers including molecular subtype, tumor stage, node status, ER status, HER2 status, age at diagnosis, and the 11-gene proliferation score when tested using a multivariate analysis (Cox model). We determined that amplification of a single candidate gene did not consistently

outperformed or improve the prognostic capacity of these clinical and genomic variables (Supplementary Table 67). However, these candidate genes were not identified to be prognostic markers, especially given that they correlate with proliferation, but instead were selected to be likely drivers of proliferation, a highly important prognostic feature.

DISCUSSION

Numerous studies, including many focused on human breast cancer, have made use of large-scale analyses to investigate the genomic landscape of human cancers and to identify molecular heterogeneity within tumor types not previously recognized[2,3,6,11]. The challenge presented by these studies, and by the enormous amount of genomic data available from resources such as the TCGA and METABRIC projects, is how to integrate multiple forms of genomic data to investigate the biology of the disease, and how to interpret the significance of identified genomic alterations without relying on inferences of “known” biology to determine the role that these alterations play in tumorigenesis.

In this study, we utilize gene expression signatures of signaling pathways to identify patterns that can distinguish the known subtypes of breast cancer. These signatures are largely developed from controlled manipulations of the relevant pathways *in vitro,* and are thus based on experimental evidence for pathway activation as opposed to extrapolations of pathway activity achieved from analyses of annotated gene lists. Therefore, use of an experimentally derived pathway signature, as opposed to the analysis of a single genomic alteration, provides a measure of pathway activity irrespective of how the pathway may have been activated. For instance, a given pathway can be active in a subset of tumors, either as a result of an activating alteration (i.e. *E2F1* or *E2F3* amplification) or an independent event that inactivates a negative regulator of the pathway (i.e. *RB1* loss and/or mutation), which nevertheless achieves the same end result (i.e. DNA replication and cell proliferation); importantly, we identified these four genetic events as being statistically associated with the RB-LOH signature[16], which is dominated by E2F-regulated genes and is a strong indicator of cell proliferation and prognosis.

Proliferation is one of the most powerful prognostic features in breast cancers, especially for ER+ cancers[38,39]. Since proliferation is so important, we utilized a gene expression signature of proliferation as a means to integrate the DNA copy number data, along with data from a genome-wide RNAi screen of luminal breast cancer cell lines, to identify luminal-specific genetic drivers of proliferation. We identified 12 genes that were uniquely amplified in highly proliferative luminal tumors in the TCGA dataset, have a correlation between mRNA expression and DNA copy number, and were shown to be essential for luminal breast cancer cell line viability; eight were validated using the independent METABRIC dataset. While *FGD5*, *METTL6*, *DTX3*, and *MRPS23* amplification was prognostic in luminal patients, these and many of the other identified genes have been previously reported to regulate tumorigenic characteristics, albeit not necessarily in human breast cancer. For example, *FGD5* has been shown to regulate the pro-angiogenic function of *VEGF* [43], potentially leading to increased proliferation. *DTX3* purportedly promotes Notch signaling[44,45], while *EIF6* is a Notch-dependent regulator of cell invasion and migration[46] and its inhibition restricts lymphomagenesis and tumor progression[47]. *MRPS23*

expression is associated with proliferation, oxidative phosphorylation, invasiveness, and tumor size in uterine cervical cancer[48]. *METTL6* has been reported to contribute to cytotoxic chemotherapy sensitivity in lung cancers[49].

Several previous studies have identified chromosomal regions specifically altered in subsets of breast cancer, including 3p25 (encompassing *METTL6, FGD5*)[2] and 11q13 (*CPT1A*)[3] in luminal breast tumors; however these studies neither discriminated between essential and non-essential genes within a specific amplicon, nor did they identified the functional consequence of these alterations. In contrast, we have shown that these regions are uniquely amplified in highly proliferative luminal tumors and importantly, we distinguish between amplified genes that are essential for cell proliferation and thus likely contribute to tumorigenesis, versus those that are amplified but not essential. For instance *SRC* (20q12-13), which is co-amplified with *EIF6*, is similarly amplified in a significant percentage of highly proliferative luminal tumors (Supplementary Tables 43 and 58), but was not identified as essential in highly proliferative luminal breast cancer cell lines in the RNAi screen (Supplementary Table 60). Interestingly, in addition to its role in regulating translation[50] and Notch signaling[46], *EIF6* has been reported to link Integrin-β4 to the intermediate filament cytoskeleton[51], potentially leading to down-stream activation of *SRC* signaling. These results may explain some of the paradoxical findings of *SRC* in that it may contribute to proliferation status, but may not be essential, whereas a gene very near it, also linked to proliferation, is essential for cell viability *in vitro*. Clearly additional experiments are needed to address this issue, but these results highlight the complex nature, and importance, of this specific amplicon.

A significant challenge to translating these findings into the clinic is the identification of genes within amplicons that are therapeutically targetable. One such event may be amplification of 11q13-14/*CPT1A*, which was recently reported[3] to be a defining feature of a high-risk ER+ subgroup (Integrative Cluster 2) and correlates with a poor prognosis in esophageal squamous cell carcinoma[52]. We identified *CPT1A* as the only gene within the amplified 11q13 locus required for cell viability within the confines of the proliferation signature and luminal cell lines, suggesting that repression of *CPT1A* could affect the proliferative phenotype of these tumors. Consistent with this hypothesis, it was recently reported that RNAi-mediated down-regulation, or drug-mediated inhibition of *CPT1A*, inhibited cancer cell line proliferation, migration and metastasis[53–55], albeit not in breast cancer cell lines. Lastly, a specific inhibitor of *CPT1A* (ST-1326) repressed tumor formation and proliferation in an Eμ-myc mouse model of Burkett's lymphoma[55].

Collectively these data demonstrate the ability of this across-platforms genomics approach to identify novel oncogenes amplified in a subset of highly proliferative luminal breast cancer patients which are essential for cell viability. These data suggest that not only are these identified genes potential drivers of oncogenesis and that an emphasis should be placed on elucidating their role in breast tumorigenesis, but also that they, or their associated pathways, may serve as novel therapeutic targets in a subset of human breast cancers for which limited therapeutic opportunities currently exist.

METHODS

Gene expression data

Agilent custom 244K whole genome gene expression microarray data for human breast cancer samples was acquired from The Cancer Genome Atlas (TCGA) project[2] data portal. Samples were filtered to include only those 476 samples for which Affymetrix SNP 6.0 data was present. As previously described, (TCGA) data were median centered for each gene. Illumina HT-29 v3 expression data for the METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) project (n=1,992 samples) was acquired from the European Genome-phenome Archive at the European Bioinformatics Institute and data were median centered for each gene[3]. Expression data for a panel of 51 breast cancer cell lines was acquired from GEO (GSE12777)[41]. Affymetrix U133+2 data were MAS5.0 normalized using Affymetrix Expression Console (ver1.2.1.20), and $\log_2$ transformed. Expression probes were collapsed using the median gene value with the GenePattern[56] module CollapseProbes.

Affymetrix SNP 6.0 data

DNA copy number values were determined in 490 TCGA primary breast tumors (of which 476 had matched mRNA expression data) and 1,992 METABRIC primary breast tumors using Affymetrix 6.0 SNP arrays as previous described[2,3]. Copy number segmentation and segment calls (i.e. NEUT, AMP, GAIN, HOMD or HETD) was performed using the Circular Binary Segmentation (CBS) algorithm as previously described[2,3]. Using the hg19 build annotation from the UCSC genome browser, genes were selected if they fell completely within a CBS identified copy number segment. Genes that were not found completely within a copy number segment across any sample were filtered out. In the METABRIC dataset, the copy number call gene-matrix was determined from genes that fell completely within a CBS identified copy number segment. Out of the 12 genes of interest, *SNX21*, *ZBTB46*, and *DNAJC5* were not found completely within a CBS identified segment among the METABRIC samples and were excluded from further analyses.

Gene expression signatures

A panel of 52 previous published gene expression signatures was used to examine patterns of pathway activity and/or microenvironmental states (Supplementary Table 1). In order to implement each signature, the methods detailed in the original studies were followed as closely as possible. The 22 signatures from Gatza *et al.*[10,11,32] were originally developed based on a Bayesian binary regression strategy and identified a signature comprised of Affymetrix probe sets with positive and negative regression weights; to translate these signatures to a form that could be applied to non-Affymetrix expression data, the original reported signatures were altered. For each signature, we excluded those probe sets with a negative correlation coefficient. The remaining probe sets with a positive coefficient were then translated to the gene level, and replicate genes were merged. To apply a given signature to a new dataset, the expression data was filtered to contain only those genes that met the previous criteria and the mean expression value was calculated using all genes within a given signature that were present in more than 80% of samples. The list of genes in each modified signature is reported in Supplementary Table 2 and the scores for the TCGA

dataset (Supplementary Table 3) and cell line dataset are provided (Supplementary Table 59).

Statistical analyses of signature scores

To quantify differences in patterns of signature scores across subtypes, an ANOVA followed by a Tukey's post-test for pairwise comparisons was used (as shown in Figure 1B). To investigate the level of concordance between each of the 52 signatures, the pathway scores calculated for each sample in the TCGA dataset (Supplementary Table 3) were analyzed. The R values calculated by a Pearson correlation are reported in Supplementary Figure 2 and Supplementary Table 5.

Identification of point mutations as a function of pathway activity

To compare the frequency of mutations, the 35 genes identified as being significantly mutated in human breast cancer[2] were assessed in the context of the 11-gene PAM50 Proliferation signature[31]. A Fisher's exact test (Bonferroni corrected) was used to compare the frequency of mutations in samples with high (top quartile) and low (all other samples) in LumA, LumB and HER2E (n=388) samples. The frequency of mutations associated with each group for each signature is summarized in Supplementary Table 60–61.

Identification of copy number alterations as a function of pathway activity

To identify copy number alterations two analysis methods were independently used. A Spearman rank correlation, both positive and negative, was used to compare gene-level segment scores with predicted pathway activity. To compare the frequency of amplification and losses, a Fisher's exact test was used to compare the frequency of either gene-specific copy number gains and amplification or deletions (both loss of heterozygosity of deletion) against non-amplified or non-deleted samples. Samples in the top quartile of calculated pathway activity were compared to those in the bottom three quartiles. For each analysis, the negative $\log_{10}$ Bonferroni-adjusted P-values are reported (Supplementary Figure 3 and Supplementary Figure 4). To identify genes that were significant across both methods, a threshold of $q<0.01$ (Bonferroni corrected) was set for validation (Figure 2) and $q<0.05$ for discovery (Figure 5). The Bonferroni corrected p-values for the positive and negative Spearman rank correlation for each gene and each signature are reported in Supplementary Tables 6 through 57. The frequency of copy number gains in the top quartile versus all other samples, as well as the Bonferroni corrected p-values calculated by a Fisher's exact test, are reported for each gene and each signature (Supplementary Tables 6–57).

Analysis of genome-wide RNAi proliferation data

In order to identify genes required for cell viability in a signature dependent manner, data from a previously published genome-wide RNAi screen carried out on a panel breast cancer cell lines were analyzed[9]. The Gene Active Ranking Profile (GARP)-normalized data were obtained from the COLT database and filtered to include only those 27 cell lines for which gene expression data (GSE12777) were also available (acquired February 2013). To identify gene essential for pathway-dependent cell proliferation, a negative Spearman correlation

was performed comparing predicted pathway activity and GARP score for each sample. A threshold of $p<0.05$ was considered significant for all analyses.

Analysis of mRNA expression in copy number neutral samples

To assess mRNA expression in luminal tumors lacking CNA of each candidate genes, luminal and HER2E samples from the TCGA (n=388) and METABRIC (n=1,333) studies were grouping to high (top quartile) and low (all other samples). Samples with copy number gains (including high-level amplification or gain) or losses (both loss of heterozygosity and homozygous deletion) were excluded and a t-test was used to examine statistical differences between the expression levels of genes in each cohort.

Survival analyses

To investigate the effect that candidate gene amplification has on disease-specific survival, clinical data for the 1,992 patients in the METABRIC study were obtained[3]. The 11-gene PAM50 Proliferation signature[31] was applied to all 1,992 samples by calculating the median value of the signature, for each sample. For survival analyses, patients that died of causes unrelated to breast cancer and patients without a date of death were censored. We extracted patients classified as Luminal A, Luminal B or HER2E for which survival data was reported (n=1,333). For survival analysis of the TCGA dataset[2], we extracted patients classified as LumA, LumB or HER2E for which clinical data were available (September 2012). Disease-specific survival was calculated comparing samples with an amplification of a candidate gene against those without. In each dataset, patients without a CNA call for a specific gene were excluded from the survival analysis. For each analysis, significance was calculated by a log rank test and the hazard ratio (HR) is reported. To compare the effect of candidate gene copy number status on common prognostic markers including proliferation (PAM50 proliferation signature), molecular subtype (PAM50), tumor stage, node status, ER status, HER2 status, or age at diagnosis, a multivariate Cox model was used.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank members of our laboratory for helpful discussion and suggestions. Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Numbers K99-CA166228-01A1 awarded to MLG, and to CMP for the Breast SPORE program grant P50-CA58223-09A1, RO1-CA148761-04, the Susan G. Komen for the Cure, and the Breast Cancer Research Foundation.

REFERENCES

1. Perou CM, et al. Molecular portraits of human breast tumors. Nature. 2000; 406:747–752. [PubMed: 10963602]
2. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490:61–70. [PubMed: 23000897]
3. Curtis C, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012; 486:346–352. [PubMed: 22522925]

4. Wood LD, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007; 318:1108–1113. [PubMed: 17932254]
5. Bild AH, et al. An integration of complementary strategies for gene-expression analysis to reveal novel therapeutic opportunities for breast cancer. Breast Cancer Res. 2009; 11:R55. [PubMed: 19638211]
6. Bild AH, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature. 2006; 439:353–357. [PubMed: 16273092]
7. Rhodes DR, et al. Molecular concepts analysis links tumors, pathways, mechanisms, and drugs. Neoplasia. 2007; 9:443–454. [PubMed: 17534450]
8. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med. 2004; 10:789–799. [PubMed: 15286780]
9. Marcotte R, et al. Essential gene profiles in breast, pancreatic, and ovarian cancer cells. Cancer Discov. 2012; 2:172–189. [PubMed: 22585861]
10. Gatza ML, et al. Analysis of tumor environmental response and oncogenic pathway activation identifies distinct basal and luminal features in HER2-related breast tumor subtypes. Breast Cancer Res. 2011; 13:R62. [PubMed: 21672245]
11. Gatza ML, et al. A pathway-based classification of human breast cancer. Proc. Nat'l. Acad. Sci. 2010; 107:6994–6999. [PubMed: 20335537]
12. Fan C, et al. Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. BMC Med Genomics. 2011; 4:3. [PubMed: 21214954]
13. Hoadley KA, et al. EGFR associated expression profiles vary with breast tumor subtype. BMC Genomics. 2007; 8:258. [PubMed: 17663798]
14. Troester MA, et al. Gene expression patterns associated with p53 status in beast cancer. BMC Cancer. 2006; 6:276. [PubMed: 17150101]
15. Chandriani S, et al. A core MYC gene expression signature is prominent in basal-like breast cancer but only partially overlaps the core serum response. PLoS One. 2009; 4:e6693. [PubMed: 19690609]
16. Herschkowitz JI, He X, Fan C, Perou CM. The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas. Breast Cancer Res. 2008; 10:R75. [PubMed: 18782450]
17. Hu Z, et al. A compact VEGF signature associated with distant metastases and poor outcomes. BMC Med. 2009; 7:9. [PubMed: 19291283]
18. Hutti JE, et al. Oncogenic PI3K mutations lead to NF-kappaB-dependent cytokine expression following growth factor deprivation. Cancer Res. 2012; 72:3260–3369. [PubMed: 22552288]
19. Oh DS, et al. Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. J. Clin. Oncol. 2006; 24:1656–1664. [PubMed: 16505416]
20. Thorner AR, et al. In vitro and in vivo analysis of B-Myb in basal-like breast cancer. Oncogene. 2009; 28:742–751. [PubMed: 19043454]
21. Thorner AR, Parker JS, Hoadley KA, Perou CM. Potential tumor suppressor role for the c-Myb oncogene in luminal breast cancer. PLoS One. 2010; 5:e13073. [PubMed: 20949095]
22. Troester MA, et al. Activation of host wound responses in breast cancer microenvironment. Clin Cancer Res. 2009; 15:7020–7028. [PubMed: 19887484]
23. Usary J, et al. Mutation of GATA3 in human breast tumors. Oncogene. 2004; 23:7669–7678. [PubMed: 15361840]
24. Harrell JC, et al. Endothelial-like properties of claudin-low breast cancer cells promote tumor vascular permeability and metastasis. Clin Exp Metastasis. 2013
25. Wong DJ, et al. Module map of stem cell genes guides creation of epithelial cancer stem cells. Cell Stem Cell. 2008; 2:333–344. [PubMed: 18397753]
26. Ji H, et al. LKB1 modulates lung cancer differentiation and metastasis. Nature. 2007; 448:807–810. [PubMed: 17676035]
27. Saal LH, et al. Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. Proc Natl Acad Sci U S A. 2007; 104:7564–7569. [PubMed: 17452630]

28. Glinsky GV, Berezovska O, Glinskii AB. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. J Clin Invest. 2005; 115:1503–1521. [PubMed: 15931389]
29. Lim E, et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nat Med. 2009; 15:907–913. [PubMed: 19648928]
30. van 't Veer LJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415:530–536. [PubMed: 11823860]
31. Parker JS, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009; 27:1160–1167. [PubMed: 19204204]
32. Chang JT, et al. SIGNATURE: a workbench for gene expression signature analysis. BMC Bioinformatics. 2011; 12:443. [PubMed: 22078435]
33. Leone G, et al. Myc requires distinct E2F activities to induce S phase and apoptosis. Mol Cell. 2001; 8:105–113. [PubMed: 11511364]
34. Grandis JR, et al. Requirement of Stat3 but not Stat1 activation for epidermal growth factor receptor- mediated cell growth In vitro. J Clin Invest. 1998; 102:1385–1392. [PubMed: 9769331]
35. Weigman VJ, et al. Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival. Breast Cancer Res Treat. 2012; 133:865–880. [PubMed: 22048815]
36. Park K, Kwak K, Kim J, Lim S, Han S. c-myc amplification is associated with HER2 amplification and closely linked with cell proliferation in tissue microarray of nonselected breast cancers. Hum Pathol. 2005; 36:634–639. [PubMed: 16021569]
37. Nevins JR. The Rb/E2F pathway and cancer. Hum Mol Genet. 2001; 10:699–703. [PubMed: 11257102]
38. Wirapati P, et al. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Res. 2008; 10:R65. [PubMed: 18662380]
39. Perreard L, et al. Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. Breast Cancer Res. 2006; 8:R23. [PubMed: 16626501]
40. Hoadley, Katherine A.; Y, C.; Wolf, Denise M.; Cherniack, Andrew D.; Tamborero, David; Ng, Sam; Leiserson, Max D.; Niu, Beifang; McLellan, Michael D.; Paull, Evan O.; Uzunangelov, Vladimir; Kandoth, Cyriac; Akbani, Rehan; Shen, Hui; Lu, Yiling; Ju, Zhenlin; van't Veer, Laura; Lopez-Bigas, Nuria; Laird, Peter W.; Raphael, Benjamin J.; Ding, Li; Byers, Lauren A.; Mills, Gordon B.; Weinstein, John; Van Waes, Carter; Chen, Zhong; Collisson, Eric A.; Benz, Christopher; Perou, Charles M.; Stuart, Joshua M. The Cancer Genome Atlas Research Network. Multi-platform integration of 12 cancer types reveals cell-of-origin classes with distinct molecular signatures. CELL. 2014
41. Hoeflich KP, et al. In vivo antitumor activity of MEK and phosphatidylinositol 3-kinase inhibitors in basal-like breast cancer models. Clin Cancer Res. 2009; 15:4649–4664. [PubMed: 19567590]
42. Rody A, et al. T-cell metagene predicts a favorable prognosis in estrogen receptor-negative and HER2-positive breast cancers. Breast Cancer Res. 2009; 11:R15. [PubMed: 19272155]
43. Kurogane Y, et al. FGD5 mediates proangiogenic action of vascular endothelial growth factor in human vascular endothelial cells. Arterioscler Thromb Vasc Biol. 2012; 32:988–996. [PubMed: 22328776]
44. Kishi N, et al. Murine homologs of deltex define a novel gene family involved in vertebrate Notch signaling and neurogenesis. Int J Dev Neurosci. 2001; 19:21–35. [PubMed: 11226752]
45. Matsuno K, Diederich RJ, Go MJ, Blaumueller CM, Artavanis-Tsakonas S. Deltex acts as a positive regulator of Notch signaling through interactions with the Notch ankyrin repeats. Development. 1995; 121:2633–2644. [PubMed: 7671825]
46. Benelli D, Cialfi S, Pinzaglia M, Talora C, Londei P. The translation factor eIF6 is a Notch-dependent regulator of cell migration and invasion. PLoS One. 2012; 7:e32047. [PubMed: 22348144]
47. Miluzio A, et al. Impairment of cytoplasmic eIF6 activity restricts lymphomagenesis and tumor progression without affecting normal growth. Cancer Cell. 2011; 19:765–775. [PubMed: 21665150]

48. Lyng H, et al. Gene expressions and copy numbers associated with metastatic phenotypes of uterine cervical cancer. BMC Genomics. 2006; 7:268. [PubMed: 17054779]
49. Tan XL, et al. Genetic variation predicting cisplatin cytotoxicity associated with overall survival in lung cancer patients receiving platinum-based chemotherapy. Clin Cancer Res. 2011; 17:5801–5811. [PubMed: 21775533]
50. Gandin V, et al. Eukaryotic initiation factor 6 is rate-limiting in translation, growth and transformation. Nature. 2008; 455:684–688. [PubMed: 18784653]
51. Biffo S, et al. Isolation of a novel beta4 integrin-binding protein (p27(BBP)) highly expressed in epithelial cells. J Biol Chem. 1997; 272:30314–30321. [PubMed: 9374518]
52. Shi ZZ, et al. Genomic alterations with impact on survival in esophageal squamous cell carcinoma identified by array comparative genomic hybridization. Genes Chromosomes Cancer. 2011; 50:518–526. [PubMed: 21484929]
53. Liu L, Wang YD, Wu J, Cui J, Chen T. Carnitine palmitoyltransferase 1A (CPT1A): a transcriptional target of PAX3-FKHR and mediates PAX3-FKHR-dependent motility in alveolar rhabdomyosarcoma cells. BMC Cancer. 2012; 12:154. [PubMed: 22533991]
54. Samudio I, et al. Pharmacologic inhibition of fatty acid oxidation sensitizes human leukemia cells to apoptosis induction. J Clin Invest. 2010; 120:142–156. [PubMed: 20038799]
55. Pacilli A, et al. Carnitine-acyltransferase system inhibition, cancer cell death, and prevention of myc-induced lymphomagenesis. J Natl Cancer Inst. 2013; 105:489–498. [PubMed: 23486551]
56. Reich M, et al. GenePattern 2.0. Nat Genet. 2006; 38:500–501. [PubMed: 16642009]

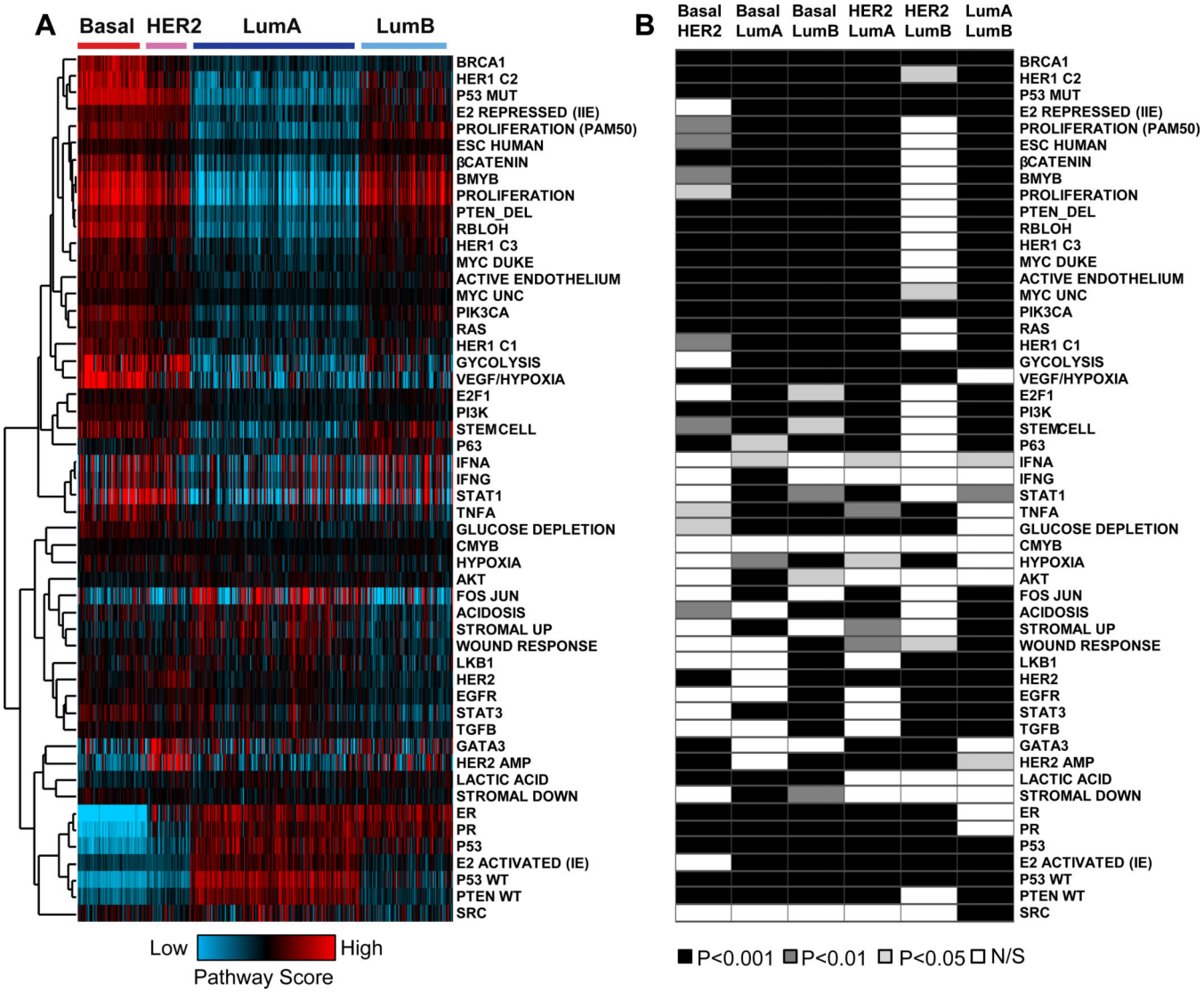

Figure 1. Patterns of genomic signature pathway activity in breast cancer

(A) Patterns of pathway activity (n=52) were determined for each sample in the published TCGA Breast Cancer cohort (n=476). Expression signature scores (y-axis) are median centered and clustered by complete linkage hierarchical clustering. (B) An ANOVA ($P<0.0001$) for all signatures according to PAM50 subtype followed by a Tukey test for pair-wise comparison demonstrates statistically significant differences in the levels of pathway expression between molecular subtypes. Box color indicates level of significance between subtypes as indicated by the legend.

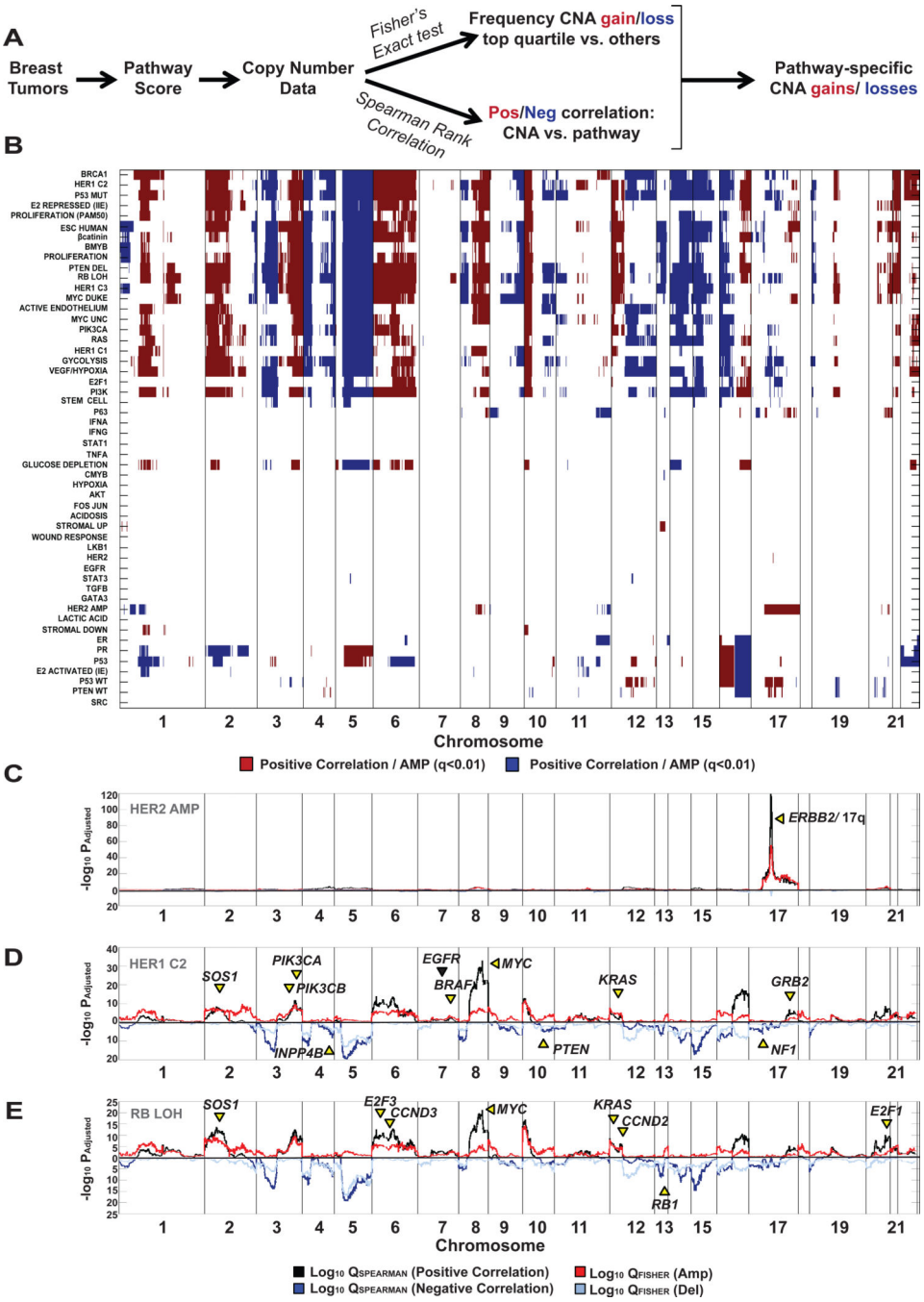

Figure 2. Identification of genomic pathway-specific copy number alterations

(A) Schematic outlining strategy used to identify CNA associated with pathway activity. (B) For each signature, significant copy number gains and losses were calculated. The plot identifies those genes that had a positive Spearman rank correlation and have increased amplification frequency (q<0.01) (red), or that have a negative Spearman rank correlation and show an increased frequency of copy number losses in the top scoring (top quartile) samples with pathway activity (q<0.01) (blue). (C–E) A Spearman rank correlation was used to identify genes positively (black line) or negatively (dark blue) associated with pathway

activity and a Fisher's exact test was used to compare the frequency of copy number gains (red) or losses (light blue) for the (C) Her2 Amp (D) Her1-C2 signature, and (E) RB-LOH signature. Yellow arrows indicate known pathway drivers with $q<0.01$ for each analysis; black arrow indicates $q<0.01$ for a single analysis. In each figure, chromosomal boundaries are indicated by vertical black lines.

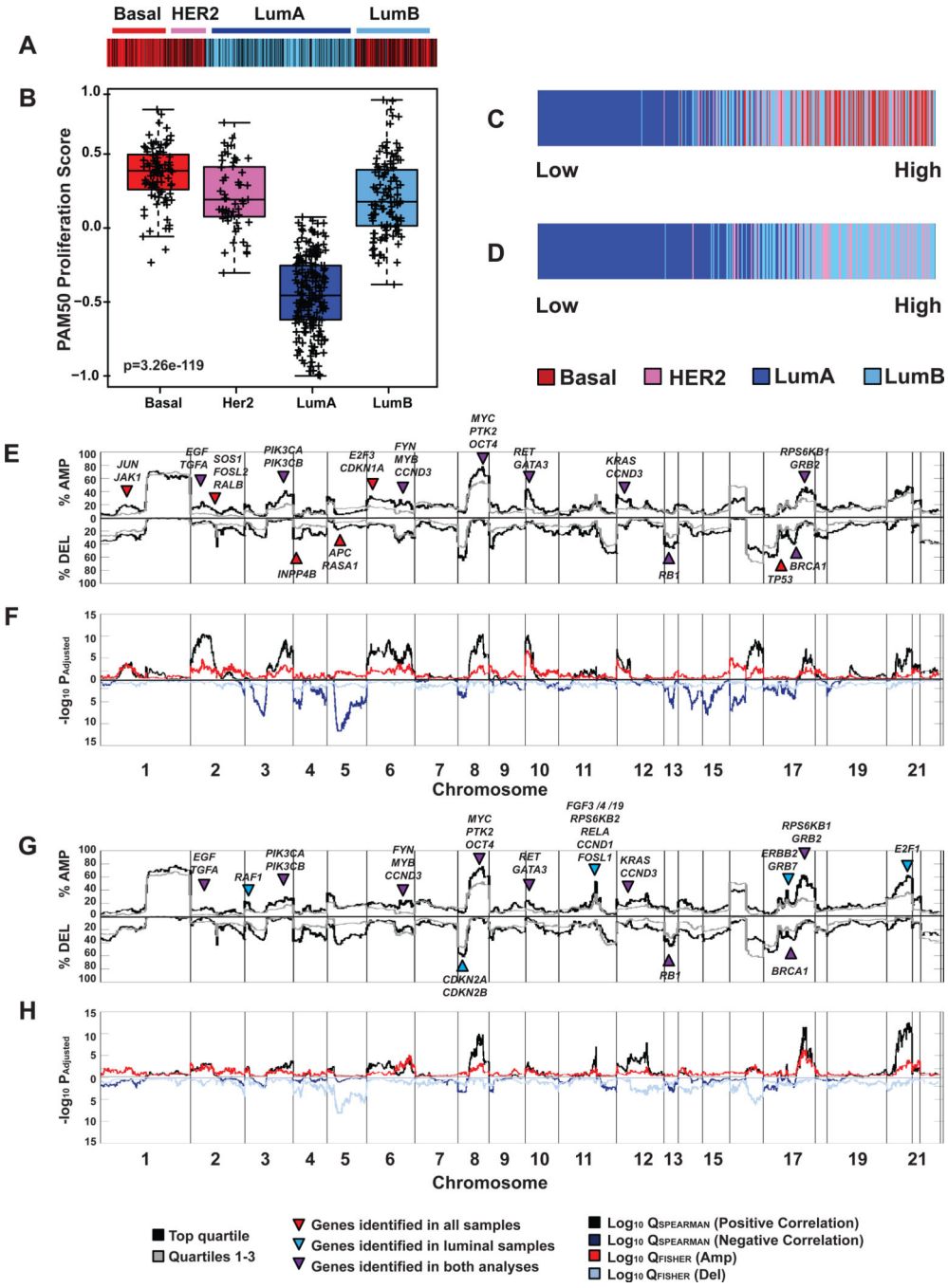

Figure 3. Identification of DNA copy number alterations in highly proliferative breast tumors

(A) Distribution of proliferation scores across all tumors and (B) by subtype. (B) Box and whisker plots indicate median score and the upper and lower quartile. Basal-like (n=88), HER2E (n=55), LumA (n=214) and LumB (n=119). (C) Highly proliferative tumors (top quartile) are comprised of Basal-like (49.6%), LumB (33.6%) and HER2E (16.8%). (D) Highly proliferative luminal tumors are restricted to LumB (68.0%) and HER2E (32.0%) samples. (E) Frequency of CNA in highly proliferative (black line) and all other samples (gray line). (F) Statistical analyses of CNA: positive correlation (black) and negative (dark

blue) Spearman rank correlation and Fisher's exact test of amplification (red) or deletion (light blue) frequency. (G) Frequency of CNA in highly proliferative luminal tumors; color key same as (E). (H) Statistical analyses of CNA in proliferative luminal tumors; color key same as (F). Chromosomal boundaries in (E–H) are defined by vertical black lines.

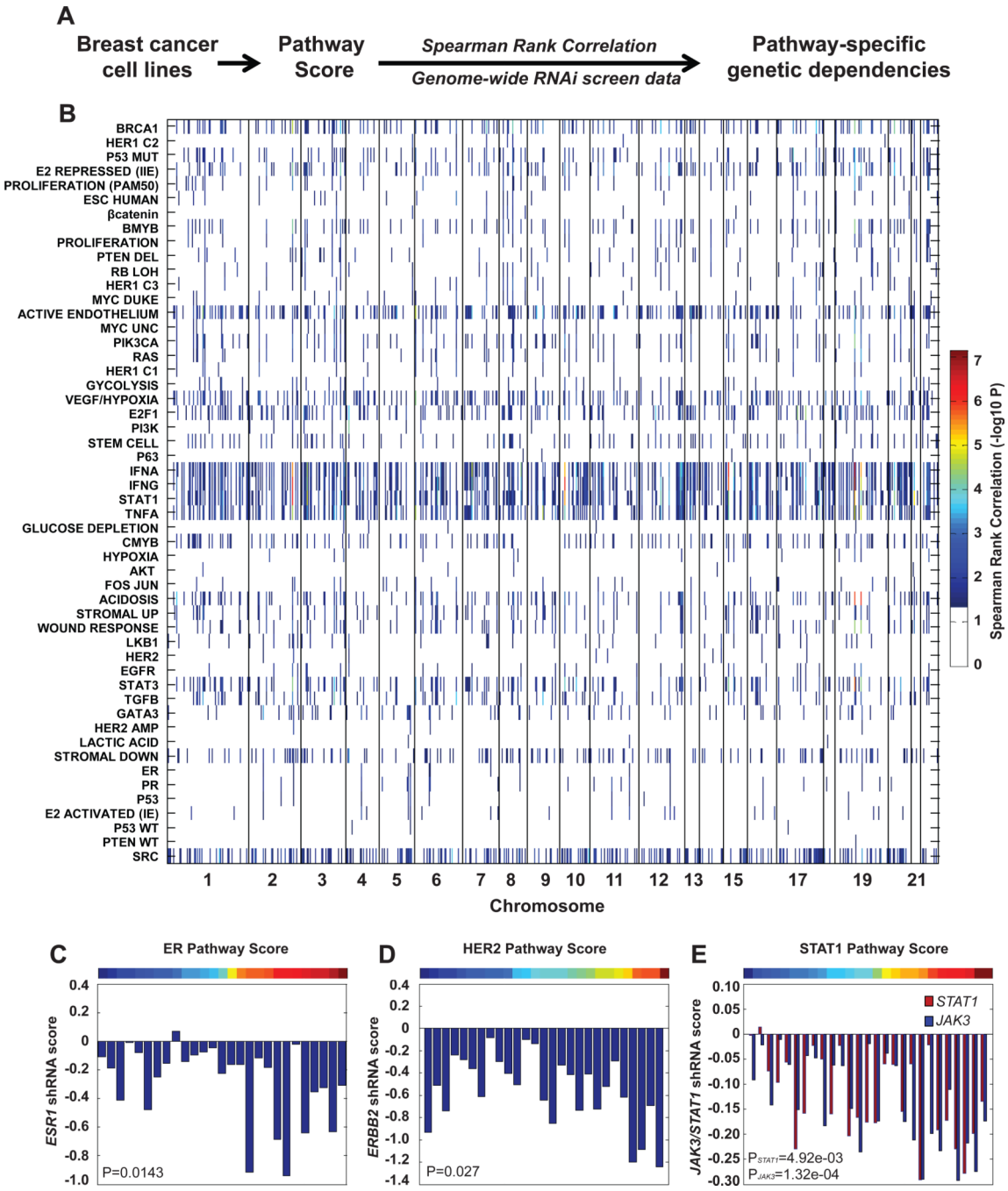

Figure 4. Identification of genomic pathway-associated essential genes in cell lines

(A) Schematic outlining strategy used to identify pathway-specific genetic dependencies. (B) A panel of 27 breast cancer cell lines with both expression data and data from a genome-wide RNAi screen was used to identify pathway-specific genes required for cell viability using a negative Spearman rank correlation (-$\log_{10}$ P-values plotted); significant genes (P<0.05) are shown according to chromosome location. Vertical black lines indicate chromosomal boundaries. (C) *ESR1* (D) *ERBB2* and (E) *STAT1* or *JAK3* shRNA levels are inversely associated with the ER, Her2 or Stat1pathway scores.

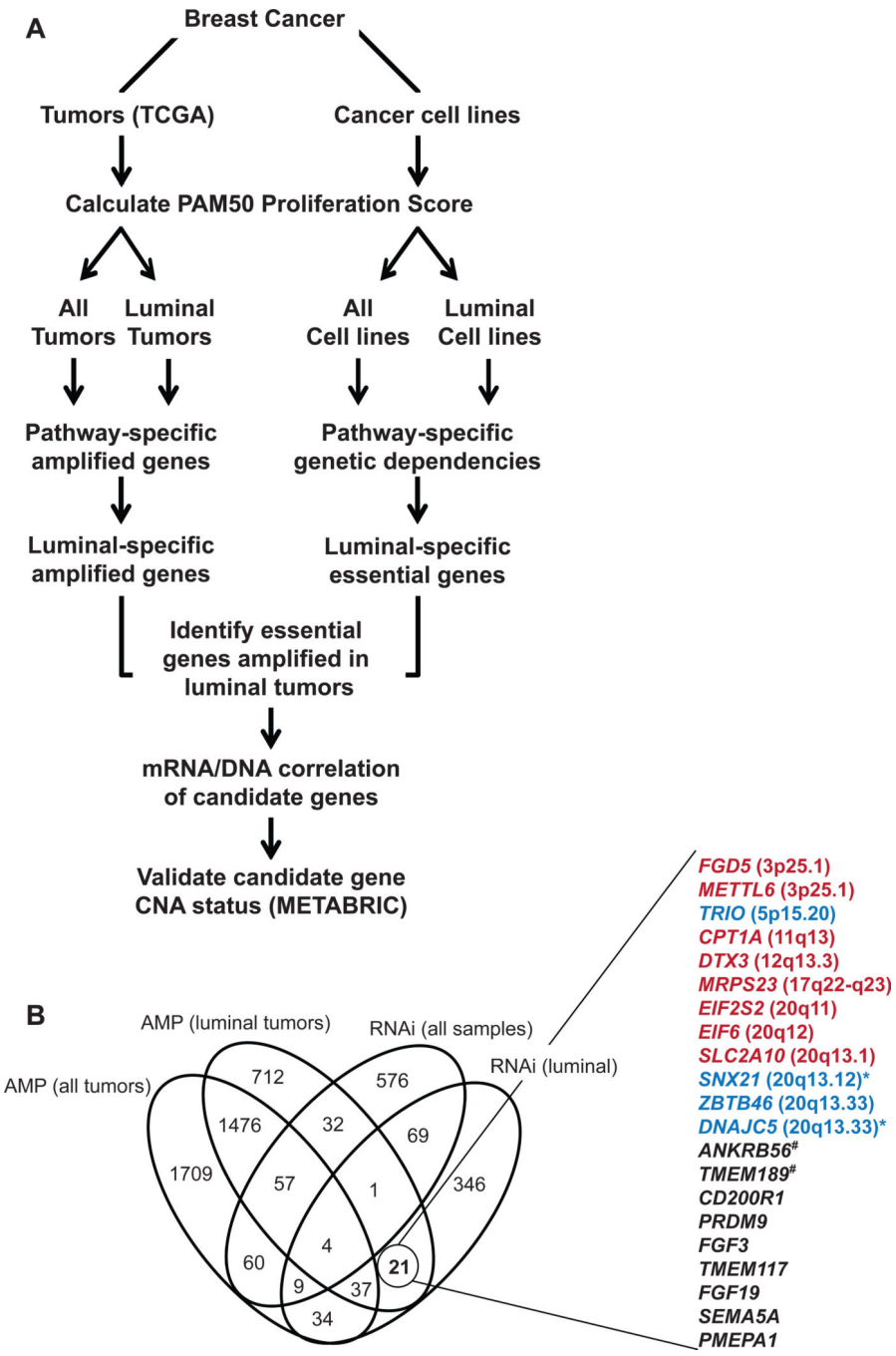

Figure 5. Identification of essential genes amplified in highly proliferative luminal tumors

(A) Schematic outlining the integrated genomic strategy to identify essential genes amplified in highly proliferative luminal breast tumors. (B) Identification of 21 genes in amplified loci that are unique to highly proliferative luminal tumors and are specifically required for luminal cell line proliferation *in vitro*. mRNA expression of genes in red and blue were significantly associated with CNA status, with the subset highlighted in red being further validated in the METABRIC dataset; genes in black do not show a significant mRNA-DNA correlation. Candidate genes demarcated by (*) are located at cusp of a CNA segment and

were originally excluded, but mentioned here. Genes identified by (#) were not included on mRNA expression microarrays, and the correlation between DNA and mRNA expression was not assessed.

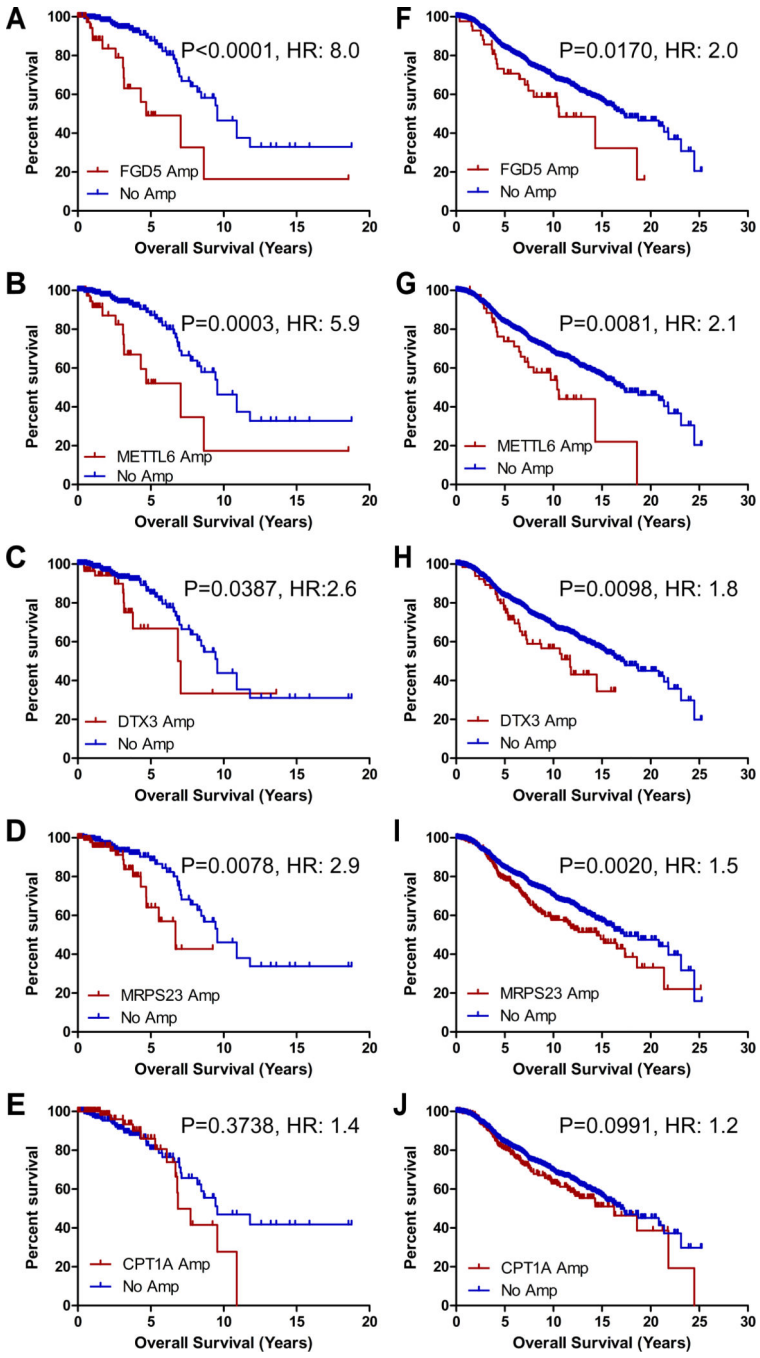

Figure 6. Candidate gene amplification correlates with a poor prognosis

Amplification of (A) *FGD5* (N_{AMP}=51, N_{NoAMP}=337), (B) *METTL6* (N_{AMP}=51, N_{NoAMP}=337), (C) *DTX3* (N_{AMP}=71, N_{NoAMP}=317) and (D) *MRSP23* (N_{AMP}=127, N_{NoAMP}=261) correlated with poor disease-specific outcome in the luminal breast cancer patients in the TCGA dataset (n=388) while (E) *CPT1A* (N_{AMP}=111, N_{NoAMP}=277) amplification had no effect on prognosis. Consistent results were observed in the METABRIC dataset (n=1,333) for (F) *FGD5* (N_{AMP}=42, N_{NoAMP}=1,218), (G) *METTL6* (N_{AMP}=44, N_{NoAMP}=1,278), (H) *DTX3* (N_{AMP}=67, N_{NoAMP}=1,266), (I) *MRPS23*

(N_{AMP}=266, N_{NoAMP}=1,062) and (J) *CPT1A* (N_{AMP}=241, N_{NoAMP}=1,029). Samples in the METABRIC dataset missing CNA calls were excluded. For each analysis, P-value determined by log-rank test and Hazard Ratio (HR) are reported.