nature genetics

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

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

**Elucidating the molecular drivers of human breast cancers requires a strategy that is 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 RNA-mediated interference screen. We identify specific DNA amplifications and essential genes within these amplicons representing key genetic drivers, including known and new 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. This general strategy has the potential to identify therapeutic targets within amplicons through an integrated use of genomic 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 critical to understanding 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 multiplatform genomic analyses to identify known and new 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 covariation 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 on prior knowledge concerning an individual gene mutation or protein alteration.

Altered signaling pathway activity is an important determinant of the biology of a tumor and may predict therapeutic response; therefore, identifying the mechanisms driving key tumorigenic pathways is essential to understanding the transformation process[2,5–8]. To take advantage of the vast amounts of existing genomic data, we used a series of experimentally derived gene expression signatures that are 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 correlate directly with DNA copy number aberrations. By further analyzing functional data from a genome-wide RNA-mediated interference (RNAi) screen[9], we identified genes that are essential for cell viability in a pathway-dependent and, in some cases, subtype-dependent manner. Our results identify a small number of DNA amplifications as potential drivers of proliferation in poor-outcome luminal breast cancers, and in general terms, we outline an approach that could be applied to many other tumor types for which multiplatform 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]. We applied each signature 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 **Fig. 1a**), and we quantitatively assessed these patterns using an analysis of variance (ANOVA) test followed by Tukey's test for pairwise comparison (**Fig. 1b** and **Supplementary Table 4**). Analyzing differences across subtypes on the basis of these 52 features demonstrated that the strongest correlation between samples existed within each molecular subtype (**Supplementary Fig. 1**).

The patterns of pathway activity recapitulated known characteristics of each subtype, including dysregulation of pathways that can be linked to female hormone receptors, oncogenes and/or tumor suppressor mutation status (**Fig. 1**). For example, basal-like tumors,

[1]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [2]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [3]Curriculum in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [4]Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. Correspondence should be addressed to C.M.P. (cperou@med.unc.edu).

Received 26 February; accepted 29 July; published online 24 August 2014; doi:10.1038/ng.3073

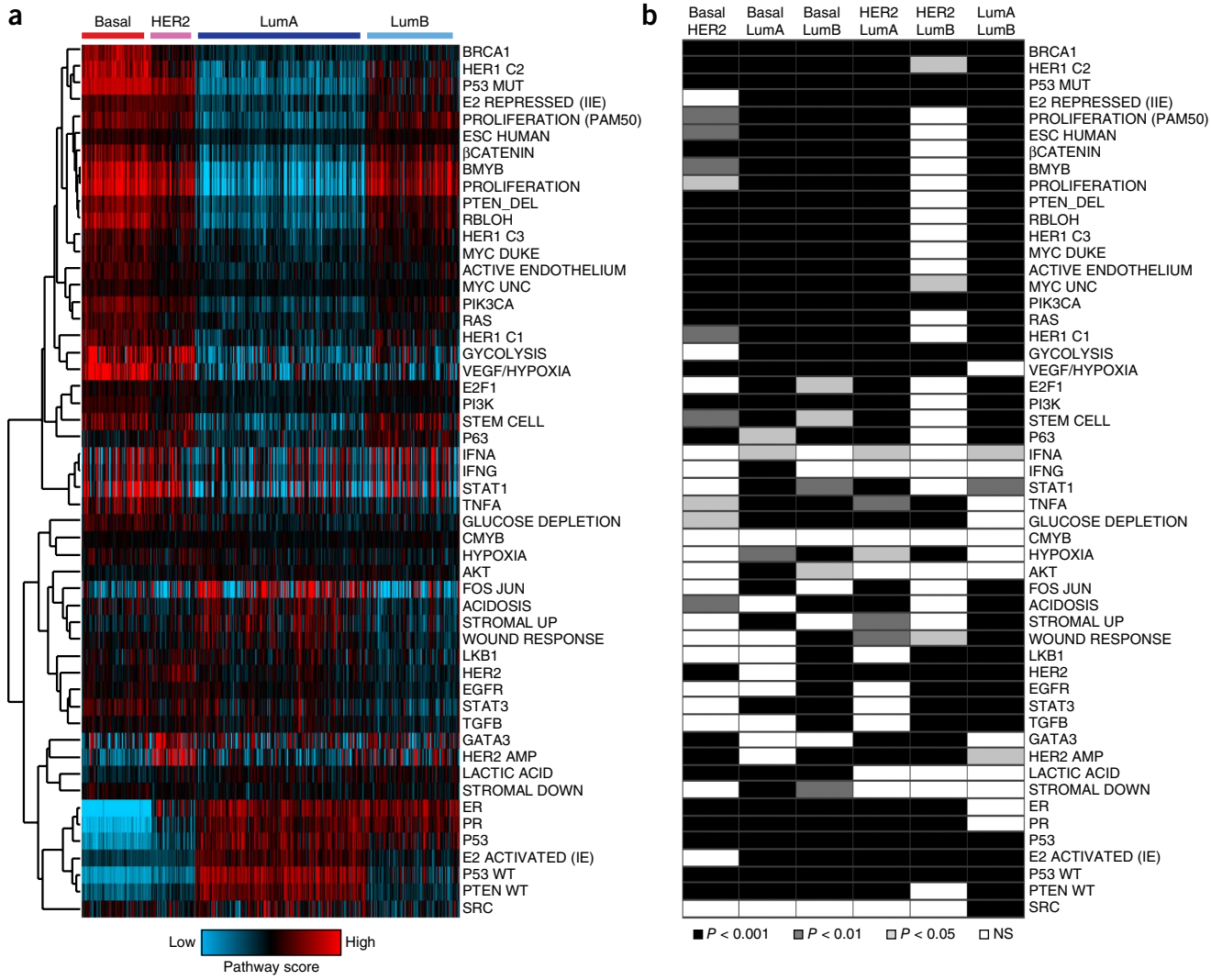


**Figure 1** Patterns of genomic signature pathway activity in breast cancer. (**a**) Patterns of pathway activity ($n = 52$) were determined for each sample in the published TCGA breast cancer cohort ($n = 476$). Expression signature scores ($y$ axis) are median centered and clustered by complete linkage hierarchical clustering. (**b**) ANOVA ($P < 0.0001$) for all signatures according to PAM50 subtype followed by Tukey's test for pairwise comparison demonstrates statistically significant differences in the levels of pathway expression between molecular subtypes. Box colors indicate the level of significance between subtypes, as indicated in the legend. NS, not significant.



which represent ~80% of triple-negative breast cancers, are characterized by low hormone receptor signaling, mutant p53 signaling and high expression of proliferation pathway activity (**Fig. 1**). Likewise, HER2-enriched (HER2E) tumors show high expression of the HER2 (ref. 11) and HER2 amplicon (HER2-AMP)[12] signatures, whereas luminal A (LumA) tumors show high hormone receptor signaling and wild-type p53 signaling. Highly proliferative LumB tumors, which also show some hormone receptor signaling, are distinguished from less proliferative LumA samples by increased proliferation-associated pathways. Thus, these data robustly recapitulate many previously published pathway and subtype associations.

Calculating a Pearson correlation coefficient to assess the concordance between each of the 52 signatures (**Supplementary Fig. 2** and **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] demonstrated an $R$ value of 0.72, whereas PIK3CA[18] and PTEN-deleted[27] signatures had an $R$ 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 loss of heterozygosity (RB-LOH)[16] and MYC[15] signatures ($R = 0.79$), whereas EGFR-mediated activation of STAT33 signaling[34] was recapitulated by the EGFR[11,32] and STAT3 (refs. 11,32) ($R = 0.72$) signatures. These results provide a measure of validity for each signature, but because differences do exist between signatures for a specific pathway, the results suggest that each signature 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 used DNA copy number data from the TCGA project ($n = 476$) to identify copy number alterations (CNAs) associated with pathway activity (**Fig. 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 familywise error rate) to assess the relationship between pathway

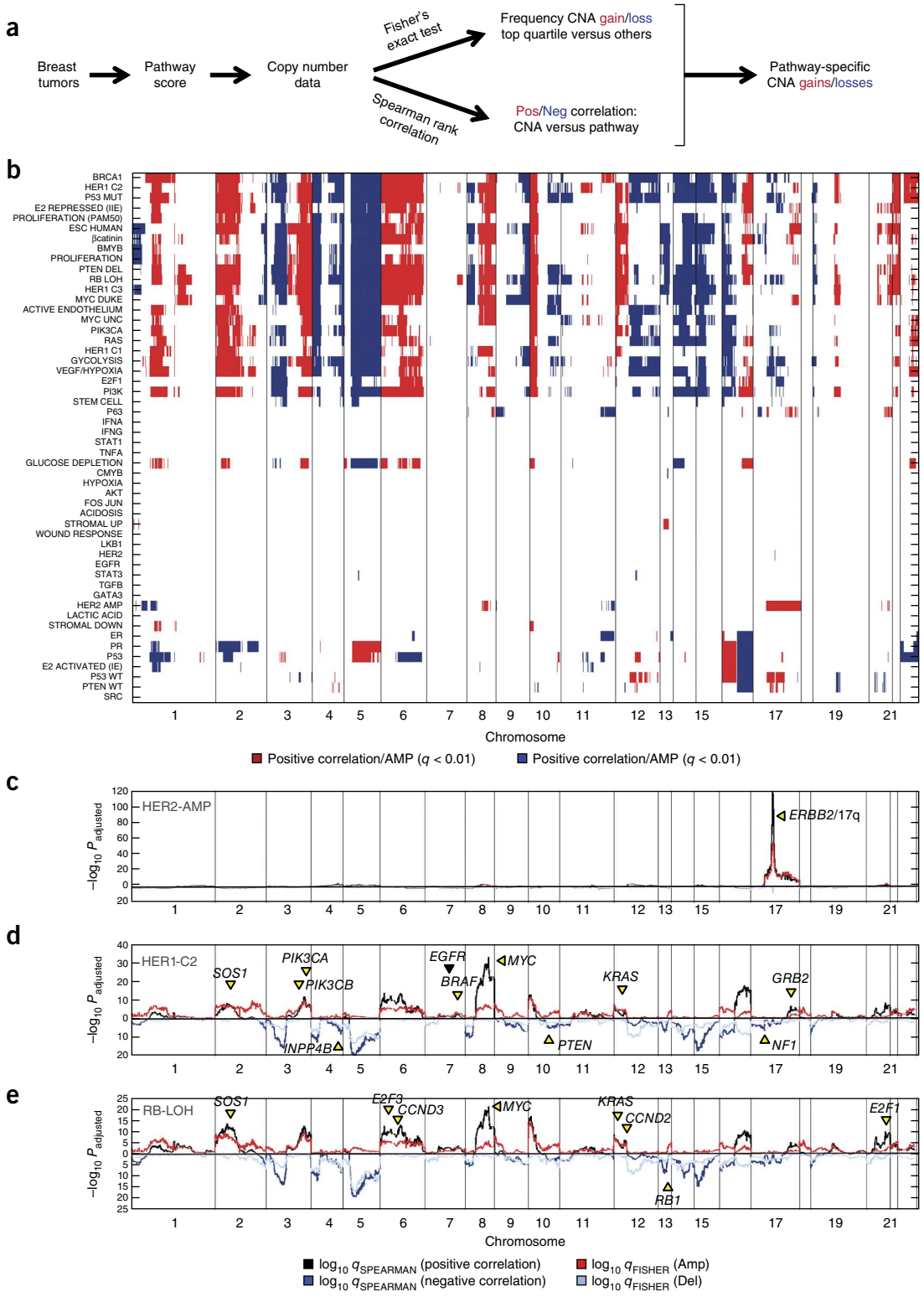


**Figure 2** Identification of genomic pathway–specific CNAs. (**a**) Schematic outlining the strategy used to identify CNAs associated with pathway activity. Gain/loss indicates gains or losses; Pos/Neg indicates positive or negative. (**b**) For each signature, significant copy number gains and losses were calculated. The plot identifies those genes that had a positive Spearman rank correlation and increased amplification frequency ($q < 0.01$) (red) and those that had a negative Spearman rank correlation and an increased frequency of copy number losses in the top-scoring (top quartile) samples with pathway activity ($q < 0.01$) (blue). (**c**–**e**) Spearman rank correlation was used to identify genes positively (black line) or negatively (dark blue) associated with pathway activity, and Fisher's exact test was used to compare the frequency of copy number gains (Amp, red) or losses (Del, light blue) for the HER2-AMP (**c**), HER1-C2 (**d**) and RB-LOH (**e**) signatures. Yellow arrowheads indicate known pathway drivers with $q < 0.01$ for each analysis; the black arrowhead indicates $q < 0.01$ for a single analysis. In each figure, chromosomal boundaries are indicated by vertical black lines.

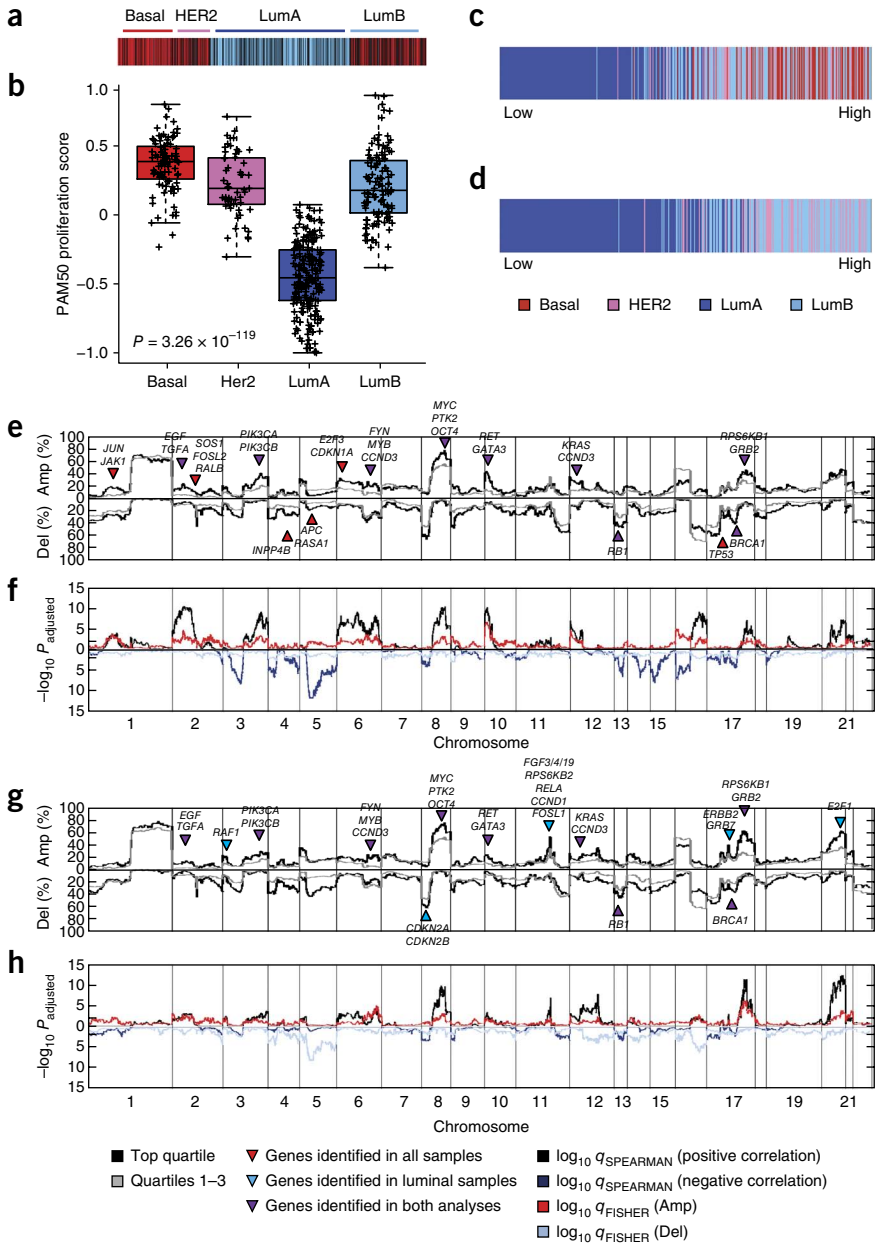


**Figure 3** Identification of DNA CNAs in highly proliferative breast tumors. (**a**,**b**) Distribution of proliferation scores across all tumors (**a**) and by subtype (**b**). (**b**) Box and whisker plots indicate the median score (horizontal line), the interquartile range (IQR, box boundaries) and 1.5 times the IQR (whiskers). 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%) samples. (**d**) Highly proliferative luminal tumors are restricted to LumB (68.0%) and HER2E (32.0%) samples. (**e**) Frequency of CNAs in highly proliferative (black line) and all other (gray line) samples. (**f**) Statistical analyses of CNAs. Indicated are positive correlations (black) and negative correlations (dark blue) by Spearman rank and frequency of amplifications (red) and deletion (light blue) by Fisher's exact test. (**g**) Frequency of CNAs in highly proliferative luminal tumors; the color key used is the same as that in **e**. (**h**) Statistical analyses of CNAs in proliferative luminal tumors; the color key used is the same as that in **f**. Chromosomal boundaries in **e**–**h** are defined by vertical black lines.

score and gene-level DNA segment score (**Supplementary Fig. 3** and **Supplementary Tables 6–57**). Second, we used a Fisher's exact test (Bonferroni corrected) to calculate the frequency of CNA gains (including high-level amplifications and gains) or losses (including LOH and deletions) in samples with high (top quartile) pathway activity compared to all other samples (low activity) (**Supplementary Fig. 4** and **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 (**Fig. 2a**); potential drivers of pathway activity had a positive correlation and a higher amplification frequency in samples with high pathway activity, whereas 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 CNAs (**Fig. 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 had more complex patterns of CNAs 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], as this signature is comprised of genes located at the 17q loci and the *ERBB2*/17q amplification is the dominant driver of this pathway. *ERBB2* was amplified in 84.9% of samples with high (top quartile) pathway activity compared to in 7.3% of low-scoring samples ($q = 1.1 \times 10^{-55}$); likewise, this relationship had a positive Spearman rank correlation ($q = 2.4 \times 10^{-108}$) (**Fig. 2c** and **Supplementary Table 27**). Although several other alterations, including *MYC* amplification ($q = 1.1 \times 10^{-2}$ and $q = 6.3 \times 10^{-3}$), were also associated with this signature, thus identifying a previously known relationship[36], *ERBB2*/17q amplification was the dominant alteration identified, providing a robust positive control for this strategy. As expected, we observed similar results when analyzing the HER2 pathway using the independently developed HER2 (refs. 11,32) signature (**Supplementary Table 26**).

We further validated this strategy by assessing the relationship between CNAs and pathways that are associated with a more complex genomic landscape. Previous studies from our group have suggested that the HER1-C2 (ref. 13) signature measures predominantly 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 the PI3K repressors *INPP4B* and *PTEN* (**Fig. 2d** and **Supplementary Table 24**). We then analyzed CNAs associated with the RB-LOH[16] signature (**Fig. 2e** and **Supplementary Table 47**) and identified associations between it and CNAs of known

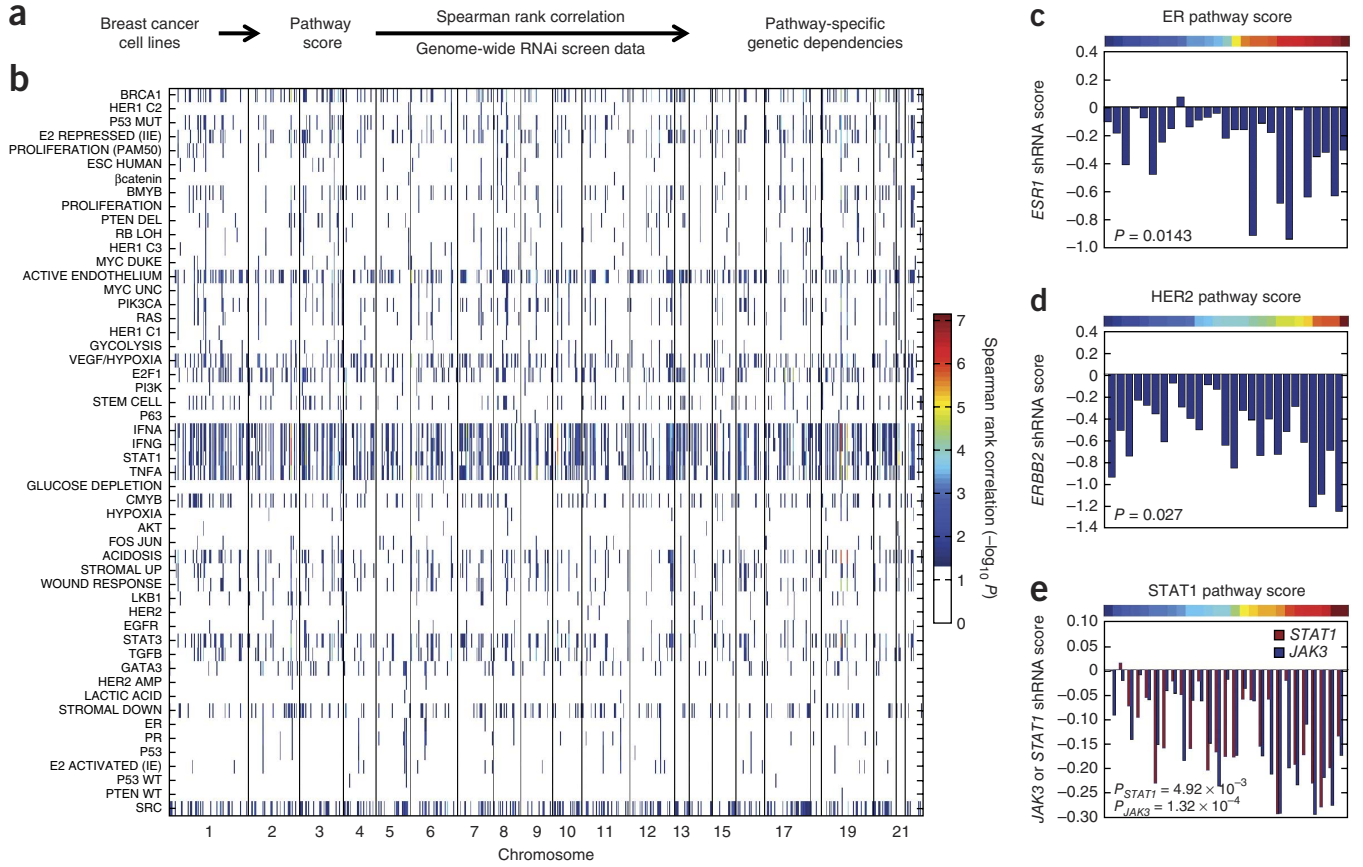


**Figure 4** Identification of genomic pathway–associated essential genes in cell lines. (**a**) Schematic outlining the strategy used to identify pathway-specific genetic dependencies. (**b**) A panel of 27 breast cancer cell lines with both expression data and data from a genome-wide RNAi screen was used to identify pathway-specific genes that are required for cell viability using a negative Spearman rank correlation (with $-\log_{10} P$ values plotted); significant genes ($P < 0.05$) are shown according to chromosome location. Vertical black lines indicate chromosomal boundaries. (**c**–**e**) *ESR1* (**c**), *ERBB2* (**d**) and *STAT1* or *JAK3* (**e**) shRNA levels are inversely associated with the ER, HER2 and STAT1 pathway scores, respectively.

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 CNAs 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 new regulatory components.

### Identification of amplified genes linked to pathway activity

Given the ability of this strategy to identify known CNAs of pathway activity, we next used this approach to identify new drivers of pathway activity. Because highly proliferative luminal tumors have a poor prognosis and poor responses to existing therapies[38,39], we sought to identify amplified genes and/or CNAs 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 altered specifically in highly proliferative luminal tumors while excluding those that are associated with proliferation irrespective of subtype, we performed analyses 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 in which 12 tumor types were studied simultaneously, and the results 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 be considered broadly as two main disease types[40].

Examining the TCGA breast cancer data set using the PAM50 proliferation signature[31], we found that basal-like, LumB and HER2E tumors had the highest proliferation levels (**Fig. 3a**,**b**), with the top quartile (**Fig. 3c**) comprised of patients with basal-like (49.6%), LumB (33.6%) and HER2E (16.8%) tumors, whereas the top quartile of proliferative luminal tumors (**Fig. 3d**) contained patients with LumB (68.0%) and HER2E (32.0%) tumors. 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 statistical strategies discussed previously (**Fig. 3e**,**f** and **Supplementary Table 43**). To identify genes that are specifically amplified in highly proliferative luminal breast cancer, we repeated these analyses using the luminal tumor subset (**Figs. 3g**,**h** and **Supplementary Table 58**). Analyzing both populations of tumors identified three classes of proliferation-associated regions ($q < 0.05$): (i) CNAs associated irrespective of subtype, (ii) CNAs altered in basal-like tumors, and (iii) CNAs altered in highly proliferative luminal tumors. These results allowed us to focus our analyses on those genes within regions that are uniquely altered in highly proliferative luminal tumors by censoring proliferation-associated genes that are altered in basal-like breast cancer (e.g., *TP53* or *INPP4B* loss) or that are 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 nonessential genes in amplified regions that are associated with proliferation in luminal tumors, we next examined data from a genome-wide RNAi screen of multiple breast tumor–derived cell lines[9]. We applied the 52 gene expression signatures to a panel (GSE12777)[41] of breast cancer cell lines (**Supplementary Fig. 5** and **Supplementary Table 59**), 27 of which had mRNA expression data and were also part of an RNAi proliferation screen in which a genome-wide shRNA library (~16,000 genes) had been used to identify essential genes (**Fig. 4a**)[9]. For each signature, we used a negative Spearman rank correlation to identify pathway-specific essential genes (**Fig. 4b** and **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 (refs. 11,32) or STAT1 (ref. 42) signatures as controls (**Fig. 4c–e**) 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 ER-α 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 that are 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 nonessential genes.

### Amplified essential genes linked to luminal tumor proliferation

We next sought to distinguish between essential and nonessential genes within regions amplified specifically 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 Fig. 6a**) and in luminal HER2$^+$ cell lines (**Supplementary Fig. 6b**) in the context of the PAM50 proliferation signature (**Supplementary Table 61**). Comparing the results of these four analyses (**Fig. 5a**) identified 19 genes that were uniquely essential for cell viability in luminal cell lines and that were amplified in highly proliferative luminal tumors (**Fig. 5b**). Two additional genes, *DNAJC5* and *SNX21*, were identified by RNAi analysis but were initially overlooked in the CNA analyses, as they were located at the cusp of two segmented regions; however, because genes overlapping both 5′ and 3′ of these genes were amplified, we included them in further investigations. Of these 21 candidate genes, 12 showed a significant relationship ($P < 0.05$) between DNA copy number levels and mRNA expression in luminal tumors (**Supplementary Fig. 7**). Notably, 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 1,000 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 data set (Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), $n = 1{,}992$) for which both mRNA expression and genomic DNA CNA data were available[3]. Of the 12 genes identified, 9 (*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 Fig. 8**) showed a significant ($P < 0.05$) relationship between CNA status and mRNA expression in luminal breast tumors ($n = 1{,}333$). Notably, eight of the nine genes, the exception being *TRIO*, also showed an increased amplification frequency ($P < 0.05$) in highly proliferative (top quartile) luminal tumors (**Supplementary Fig. 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 35 previously identified significantly mutated genes in human breast cancers reported by TCGA[2]. Using a Fisher's exact test (Bonferroni corrected), we determined that only *TP53* ($q = 7.0 \times 10^{-10}$) and *MAP3K1* ($q = 5.0 \times 10^{-3}$) mutations occurred at significantly different frequencies in highly proliferative

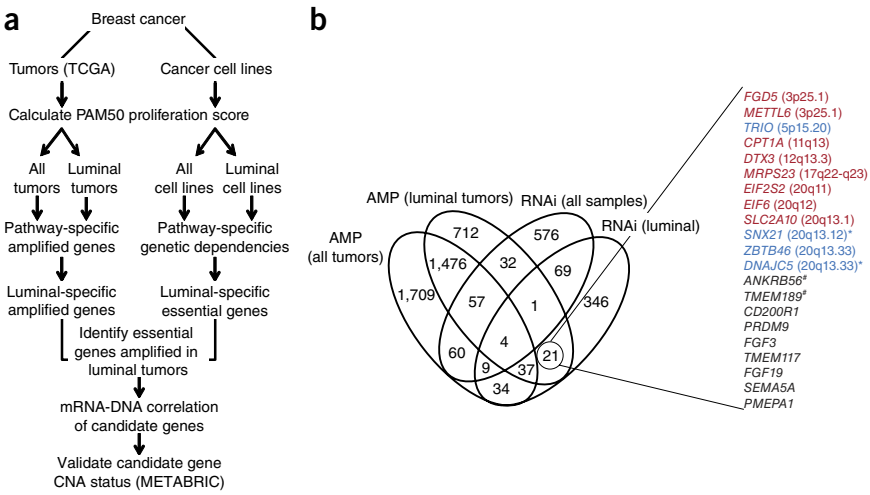


**Figure 5** Identification of essential genes amplified in highly proliferative luminal tumors. (**a**) Schematic outlining the integrated genomic strategy used to identify essential genes amplified in highly proliferative luminal breast tumors. (**b**) Identification of 21 genes in amplified loci that are unique to highly proliferative luminal tumors and are required specifically for luminal cell line proliferation *in vitro*. mRNA expression of the genes in red and blue was significantly associated with CNA status, with the subset highlighted in red being further validated in the METABRIC data set; genes in black did not show a significant mRNA-DNA correlation. Candidate genes demarcated by asterisks are located at the cusp of a CNA segment and were originally excluded but are mentioned here. Genes marked with # were not included on the mRNA expression microarrays, and the correlation between DNA and mRNA expression was not assessed.

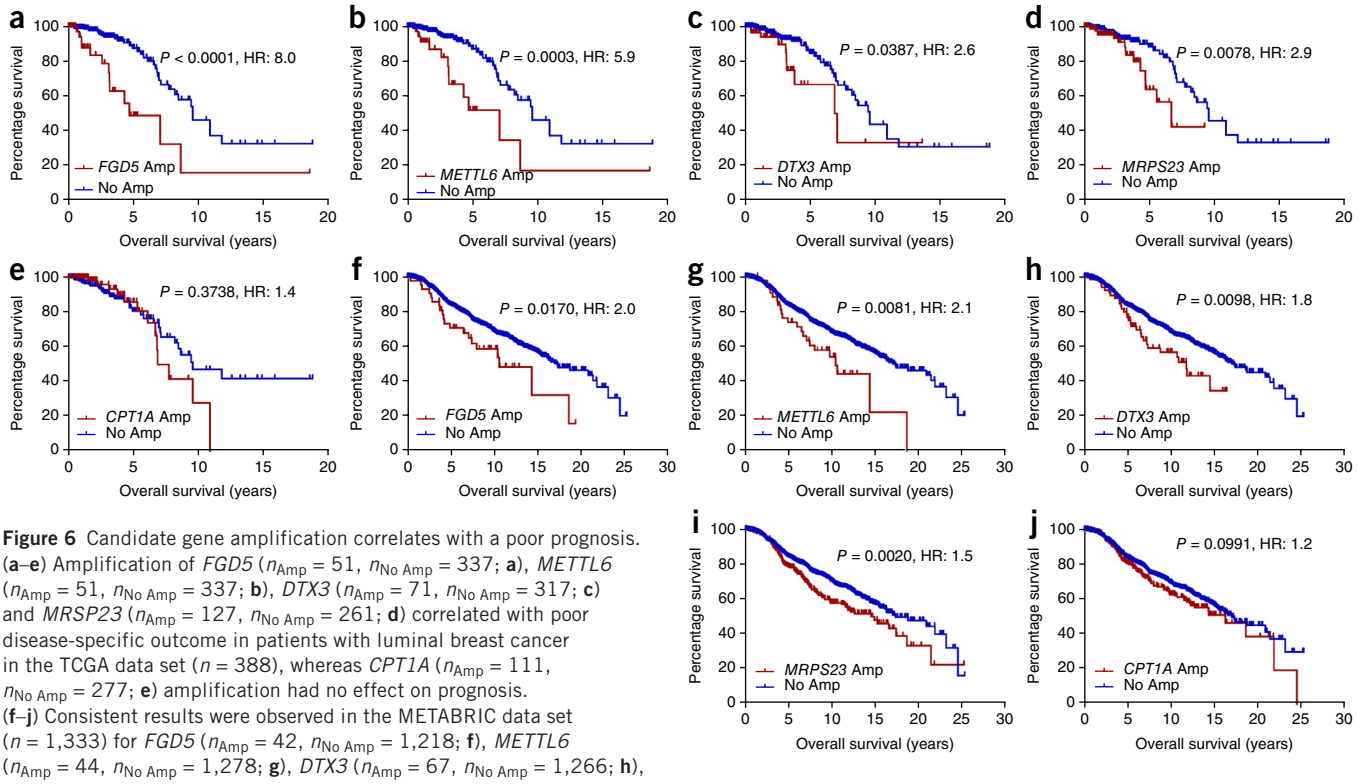


**Figure 6** Candidate gene amplification correlates with a poor prognosis. (**a**–**e**) Amplification of *FGD5* ($n_{Amp}$ = 51, $n_{No\ Amp}$ = 337; **a**), *METTL6* ($n_{Amp}$ = 51, $n_{No\ Amp}$ = 337; **b**), *DTX3* ($n_{Amp}$ = 71, $n_{No\ Amp}$ = 317; **c**) and *MRSP23* ($n_{Amp}$ = 127, $n_{No\ Amp}$ = 261; **d**) correlated with poor disease-specific outcome in patients with luminal breast cancer in the TCGA data set ($n$ = 388), whereas *CPT1A* ($n_{Amp}$ = 111, $n_{No\ Amp}$ = 277; **e**) amplification had no effect on prognosis. (**f**–**j**) Consistent results were observed in the METABRIC data set ($n$ = 1,333) for *FGD5* ($n_{Amp}$ = 42, $n_{No\ Amp}$ = 1,218; **f**), *METTL6* ($n_{Amp}$ = 44, $n_{No\ Amp}$ = 1,278; **g**), *DTX3* ($n_{Amp}$ = 67, $n_{No\ Amp}$ = 1,266; **h**), *MRPS23* ($n_{Amp}$ = 266, $n_{No\ Amp}$ = 1,062; **i**) and *CPT1A* ($n_{Amp}$ = 241, $n_{No\ Amp}$ = 1,029; **j**). Samples in the METABRIC data set missing CNA calls were excluded. For each analysis, the $P$ value was determined by log-rank test, and hazard ratios (HR) are reported.



(top quartile) luminal tumors compared to all other samples; *TP53* mutations occurred more frequently (51.6% compared to 18.6%) and *MAP3K1* (2.1% compared to 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.

We then 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 CNAs 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 anticorrelated with the signature (*DTX3*) and those that showed no correlation (*FGD5*, *METTL6* and *SLC2A10*) between data sets (**Supplementary Fig. 10**). These data suggest that amplification is a key mechanism driving the expression of these genes. However, our data also suggest, not surprisingly, that overall high expression may be the driver for some genes, 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] data sets, we extracted the subset of patients with LumA, LumB or HER2E tumors for which survival data were available (**Supplementary Tables 64** and **65**). We first analyzed data from the TCGA project (**Fig. 6a**–**e**), and despite the relatively short follow-up time (median, 1.7 years), we determined that amplification of *FGD5* ($P < 0.0001$; hazard ratio (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 patients with luminal breast cancer, whereas *CPT1A* amplification had no effect on patient survival ($P = 0.3738$). Extending these analyses to the METABRIC data set (**Figs. 6f**–**j**), which had 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, whereas gain of *CPT1A* had no effect ($P = 0.099$) on the survival of patients with luminal breast cancer. The remaining three genes showed no consistent effect on prognosis (**Supplementary Fig. 11**). Although 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 are essential for cell viability *in vitro*.

We likewise 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, as only one (*TMEM117*) of these genes showed a consistently poor prognosis in the TCGA and METABRIC data sets (**Supplementary Table 66**). We then investigated whether the 12 initial candidate genes were predictive of 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 outperform 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 as likely drivers of proliferation, a highly important prognostic feature.

## DISCUSSION

Numerous studies, including many that have focused on human breast cancer, used large-scale analyses to investigate the genomic landscape of human cancers in order to identify molecular heterogeneity and define new tumor subtypes 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 disease and how to interpret the relevance of identified genomic alterations without relying on inferences of 'known' biology to determine the role that these alterations have in tumorigenesis.

In this study we utilized gene expression signatures of signaling pathways to identify patterns that can distinguish the known subtypes of breast cancer. These signatures were developed largely 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, the use of an experimentally derived pathway signature, as opposed to an 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 as a result of either 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); notably, 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]. Because proliferation is so important, we used 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 amplified uniquely in highly proliferative luminal tumors in the TCGA data set, have a correlation between mRNA expression and DNA copy number and have been shown to be essential for luminal breast cancer cell line viability; we validated 8 of these genes using the independent METABRIC data set. Whereas *FGD5*, *METTL6*, *DTX3* and *MRPS23* amplification was prognostic in luminal tumors, these and many of the other identified genes have been reported previously to regulate tumorigenic characteristics, albeit not necessarily in human breast cancer. For example, *FGD5* has been shown to regulate the proangiogenic function of *VEGF*[43], potentially leading to increased proliferation. *DTX3* purportedly promotes Notch signaling[44,45], whereas *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 altered specifically in subsets of breast cancer, including 3p25 (encompassing *METTL6* and *FGD5*)[2] and 11q13 (*CPT1A*)[3] in luminal breast tumors; however, these studies neither discriminated between essential and nonessential genes within a specific amplicon nor identified the functional consequences of these alterations. In contrast, we have shown that these regions are amplified uniquely in highly proliferative luminal tumors, and we distinguish between amplified genes that are essential for cell proliferation and are thus likely contribute to tumorigenesis and those that are amplified but are not essential. For instance *SRC* (20q12-13), which is co-amplified with *EIF6,* is similarly amplified in a significant ($q < 0.01$) percentage of highly proliferative luminal tumors (**Supplementary Tables 43** and **58**) but was not identified as being essential in highly proliferative luminal breast cancer cell lines in the RNAi screen (**Supplementary Table 60**). Notably, 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 downstream 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, which is 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 major 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 that is 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 downregulation, or drug-mediated inhibition, of *CPT1A* inhibited cancer cell line proliferation, migration and metastasis[53–55], although not in breast cancer cell lines. In addition, 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 cross-platform genomics approach to identify new oncogenes that are essential for cell viability and are amplified in a subset of patients with highly proliferative luminal breast cancer. 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 new therapeutic targets in a subset of human breast cancers for which limited therapeutic opportunities currently exist.

**URLs.** TCGA data portal, https://tcga-data.nci.nih.gov/docs/publications/brca_2012/; GenePattern, http://www.broadinstitute.org/cancer/software/genepattern/; COLT database, http://dpsc.ccbr.utoronto.ca/cancer/index.html; 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/.

## METHODS

Methods and any associated references are available in the online version of the paper.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*


## ACKNOWLEDGMENTS

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 US National Institutes of Health under award number K99-CA166228-01A1 to M.L.G. Additional funding for research reported in this study was provided by the National Cancer Institute of the US National Institutes of Health Breast SPORE program grant P50-CA58223-09A1 and RO1-CA148761-04, as well as grants from the Susan G. Komen for the Cure and the Breast Cancer Research Foundation to C.M.P.


## AUTHOR CONTRIBUTIONS

M.L.G., J.S.P. and C.M.P. conceived and designed the study. M.L.G., G.O.S. and C.F. performed analyses. M.L.G. and C.M.P. wrote the manuscript. All authors have reviewed and approved the final manuscript.

## COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.




1. Perou, C.M. *et al.* Molecular portraits of human breast tumors. *Nature* **406**, 747–752 (2000).
2. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
3. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
4. Wood, L.D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113 (2007).
5. Bild, A.H. *et al.* An integration of complementary strategies for gene-expression analysis to reveal novel therapeutic opportunities for breast cancer. *Breast Cancer Res.* **11**, R55 (2009).
6. Bild, A.H. *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353–357 (2006).
7. Rhodes, D.R. *et al.* Molecular concepts analysis links tumors, pathways, mechanisms, and drugs. *Neoplasia* **9**, 443–454 (2007).
8. Vogelstein, B. & Kinzler, K.W. Cancer genes and the pathways they control. *Nat. Med.* **10**, 789–799 (2004).
9. Marcotte, R. *et al.* Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discov.* **2**, 172–189 (2012).
10. Gatza, M.L. *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.* **13**, R62 (2011).
11. Gatza, M.L. *et al.* A pathway-based classification of human breast cancer. *Proc. Natl. Acad. Sci. USA* **107**, 6994–6999 (2010).
12. Fan, C. *et al.* Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. *BMC Med. Genomics* **4**, 3 (2011).
13. Hoadley, K.A. *et al.* EGFR associated expression profiles vary with breast tumor subtype. *BMC Genomics* **8**, 258 (2007).
14. Troester, M.A. *et al.* Gene expression patterns associated with p53 status in beast cancer. *BMC Cancer* **6**, 276 (2006).
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* **4**, e6693 (2009).
16. Herschkowitz, J.I., He, X., Fan, C. & Perou, C.M. The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas. *Breast Cancer Res.* **10**, R75 (2008).
17. Hu, Z. *et al.* A compact VEGF signature associated with distant metastases and poor outcomes. *BMC Med.* **7**, 9 (2009).
18. Hutti, J.E. *et al.* Oncogenic PI3K mutations lead to NF-κB–dependent cytokine expression following growth factor deprivation. *Cancer Res.* **72**, 3260–3269 (2012).
19. Oh, D.S. *et al.* Estrogen-regulated genes predict survival in hormone receptor–positive breast cancers. *J. Clin. Oncol.* **24**, 1656–1664 (2006).
20. Thorner, A.R. *et al. In vitro* and *in vivo* analysis of B-Myb in basal-like breast cancer. *Oncogene* **28**, 742–751 (2009).
21. Thorner, A.R., Parker, J.S., Hoadley, K.A. & Perou, C.M. Potential tumor suppressor role for the c-Myb oncogene in luminal breast cancer. *PLoS ONE* **5**, e13073 (2010).
22. Troester, M.A. *et al.* Activation of host wound responses in breast cancer microenvironment. *Clin. Cancer Res.* **15**, 7020–7028 (2009).
23. Usary, J. *et al.* Mutation of GATA3 in human breast tumors. *Oncogene* **23**, 7669–7678 (2004).
24. Harrell, J.C. *et al.* Endothelial-like properties of claudin-low breast cancer cells promote tumor vascular permeability and metastasis. *Clin. Exp. Metastasis* **31**, 33–45 (2014).
25. Wong, D.J. *et al.* Module map of stem cell genes guides creation of epithelial cancer stem cells. *Cell Stem Cell* **2**, 333–344 (2008).
26. Ji, H. *et al.* LKB1 modulates lung cancer differentiation and metastasis. *Nature* **448**, 807–810 (2007).
27. Saal, L.H. *et al.* Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proc. Natl. Acad. Sci. USA* **104**, 7564–7569 (2007).
28. Glinsky, G.V., Berezovska, O. & Glinskii, A.B. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *J. Clin. Invest.* **115**, 1503–1521 (2005).
29. Lim, E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat. Med.* **15**, 907–913 (2009).
30. van 't Veer, L.J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
31. Parker, J.S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
32. Chang, J.T. *et al.* SIGNATURE: a workbench for gene expression signature analysis. *BMC Bioinformatics* **12**, 443 (2011).
33. Leone, G. *et al.* Myc requires distinct E2F activities to induce S phase and apoptosis. *Mol. Cell* **8**, 105–113 (2001).
34. Grandis, J.R. *et al.* Requirement of Stat3 but not Stat1 activation for epidermal growth factor receptor–mediated cell growth *in vitro*. *J. Clin. Invest.* **102**, 1385–1392 (1998).
35. Weigman, V.J. *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.* **133**, 865–880 (2012).
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.* **36**, 634–639 (2005).
37. Nevins, J.R. The Rb/E2F pathway and cancer. *Hum. Mol. Genet.* **10**, 699–703 (2001).
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.* **10**, R65 (2008).
39. Perreard, L. *et al.* Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. *Breast Cancer Res.* **8**, R23 (2006).
40. Hoadley, K.A. *et al.* Multi-platform integration of 12 cancer types reveals cell-of-origin classes with distinct molecular signatures. *Cell* **158**, 1–16 (2014).
41. Hoeflich, K.P. *et al. In vivo* antitumor activity of MEK and phosphatidylinositol 3-kinase inhibitors in basal-like breast cancer models. *Clin. Cancer Res.* **15**, 4649–4664 (2009).
42. Rody, A. *et al.* T-cell metagene predicts a favorable prognosis in estrogen receptor–negative and HER2-positive breast cancers. *Breast Cancer Res.* **11**, R15 (2009).
43. Kurogane, Y. *et al.* FGD5 mediates proangiogenic action of vascular endothelial growth factor in human vascular endothelial cells. *Arterioscler. Thromb. Vasc. Biol.* **32**, 988–996 (2012).
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.* **19**, 21–35 (2001).
45. Matsuno, K., Diederich, R.J., Go, M.J., Blaumueller, C.M. & Artavanis-Tsakonas, S. Deltex acts as a positive regulator of Notch signaling through interactions with the Notch ankyrin repeats. *Development* **121**, 2633–2644 (1995).
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* **7**, e32047 (2012).
47. Miluzio, A. *et al.* Impairment of cytoplasmic eIF6 activity restricts lymphomagenesis and tumor progression without affecting normal growth. *Cancer Cell* **19**, 765–775 (2011).
48. Lyng, H. *et al.* Gene expressions and copy numbers associated with metastatic phenotypes of uterine cervical cancer. *BMC Genomics* **7**, 268 (2006).
49. Tan, X.L. *et al.* Genetic variation predicting cisplatin cytotoxicity associated with overall survival in lung cancer patients receiving platinum-based chemotherapy. *Clin. Cancer Res.* **17**, 5801–5811 (2011).
50. Gandin, V. *et al.* Eukaryotic initiation factor 6 is rate-limiting in translation, growth and transformation. *Nature* **455**, 684–688 (2008).
51. Biffo, S. *et al.* Isolation of a novel β4 integrin–binding protein (p27(BBP)) highly expressed in epithelial cells. *J. Biol. Chem.* **272**, 30314–30321 (1997).
52. Shi, Z.Z. *et al.* Genomic alterations with impact on survival in esophageal squamous cell carcinoma identified by array comparative genomic hybridization. *Genes Chromosom. Cancer* **50**, 518–526 (2011).
53. Liu, L., Wang, Y.D., 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* **12**, 154 (2012).
54. Samudio, I. *et al.* Pharmacologic inhibition of fatty acid oxidation sensitizes human leukemia cells to apoptosis induction. *J. Clin. Invest.* **120**, 142–156 (2010).
55. Pacilli, A. *et al.* Carnitine-acyltransferase system inhibition, cancer cell death, and prevention of myc-induced lymphomagenesis. *J. Natl. Cancer Inst.* **105**, 489–498 (2013).

## ONLINE METHODS

**Gene expression data.** Agilent custom 244K whole-genome gene expression microarray data for human breast cancer samples was acquired from the 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[2], (TCGA) data were median centered for each gene. Illumina HT-29 v3 expression data for the METABRIC project ($n$ = 1,992 samples) were 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 were acquired from GEO (GSE12777)[41]. Affymetrix U133+2 data were MAS5.0 normalized using the 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 (476 of which had matched mRNA expression data) and 1,992 METABRIC primary breast tumors using Affymetrix 6.0 SNP arrays as described previously[2,3]. Copy number segmentation and segment calls (i.e., NEUT, AMP, GAIN, HOMD or HETD) were performed using the circular binary segmentation (CBS) algorithm as described previously[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 data set, 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 previously published gene expression signatures was used to examine patterns of pathway activity and/or microenvironmental states (**Supplementary Table 1**). To implement each signature, the methods detailed in the original studies were followed as closely as possible. Of these 52 signatures, 22 signatures[10,11,32] were originally developed using a Bayesian binary regression strategy and are comprised of Affymetrix probe sets with positive and negative regression weights. These signatures were translated to a form that could be applied to non-Affymetrix expression data. 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 data set, the expression data were 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 shown in **Supplementary Table 2**, and the scores for the TCGA data set (**Supplementary Table 3**) and cell line data set are also provided (**Supplementary Table 59**).

**Statistical analyses of signature scores.** To quantify differences in patterns of signature scores across subtypes, ANOVA followed by Tukey's post-test for pairwise comparisons was used (as shown in **Fig. 1b**). To investigate the level of concordance between each of the 52 signatures, the pathway scores calculated for each sample in the TCGA data set (**Supplementary Table 3**) were analyzed. The $R$ values calculated by 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 frequencies 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) pathway activity in LumA, LumB and HER2E ($n$ = 388) samples. The frequencies of mutations associated with each group for each signature are summarized in **Supplementary Tables 60** and **61**.

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

**Analysis of genome-wide RNAi proliferation data.** To identify genes that are required for cell viability in a signature-dependent manner, data from a previously published genome-wide RNAi screen carried out on a panel of 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 genes 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 CNAs of each candidate gene, luminal and HER2E samples from the TCGA ($n$ = 388) and METABRIC ($n$ = 1,333) studies were grouped into those with high (top quartile) and low (all other samples) pathway activity. Samples with copy number gains (including high-level amplifications or gains) or losses (both LOH and homozygous deletions) 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 with tumors classified as LumA, LumB or HER2E and for whom survival data were reported ($n$ = 1,333). For survival analysis of the TCGA data set[2], we extracted patients with tumors classified as LumA, LumB or HER2E and for whom clinical data were available (September 2012). Disease-specific survival was calculated by comparing samples with amplification (including copy number gains and high-level amplification) of a candidate gene against those without. In each data set, 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 and age at diagnosis, a multivariate Cox model was used.

56. Reich, M. *et al.* GenePattern 2.0. *Nat. Genet.* **38**, 500–501 (2006).




doi:10.1038/ng.3073