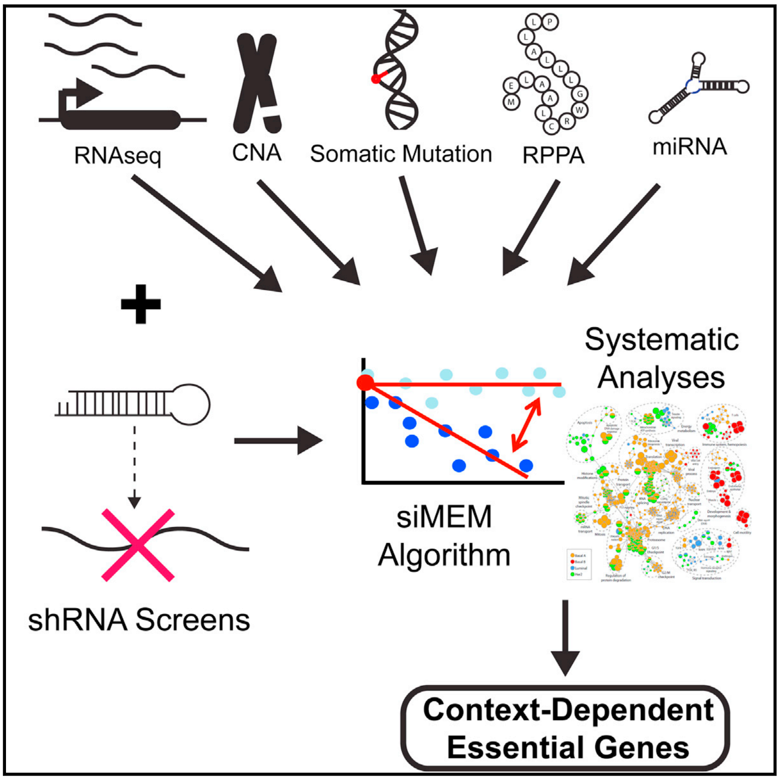
Cell

Functional Genomic Landscape of Human Breast Cancer Drivers, Vulnerabilities, and Resistance

Graphical Abstract



Authors

Richard Marcotte, Azin Sayad,
Kevin R. Brown, ..., Dana Pe'er,
Jason Moffat, Benjamin G. Neel

Correspondence

benjamin.neel@nyumc.org

In Brief

Pooled shRNA screens of a large panel of breast cancer cell lines, coupled with an improved analytical tool, siMEM, and integration with genomic and proteomic data, identify general and context-dependent essential genes in breast cancer. This study constitutes the largest functional characterization of breast cancer to date.

Highlights

- We screened 77 breast cancer lines using a genome-wide pooled shRNA library
- We developed an algorithm (siMEM) to improve identification of context-dependent genes
- Integrating screen results with genomic data reveals potential "drivers"
- *BRD4* is essential for luminal cancer, and mutant *PIK3CA* confers BET-I resistance

Accession Numbers

GSE73526
GSE74702

Marcotte et al., 2016, Cell *164*, 293–309
January 14, 2016 ©2016 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.11.062

Functional Genomic Landscape of Human Breast Cancer Drivers, Vulnerabilities, and Resistance

Richard Marcotte,[1,9,12] Azin Sayad,[1,9] Kevin R. Brown,[2] Felix Sanchez-Garcia,[4] Jüri Reimand,[2,10] Maliha Haider,[1]
Carl Virtanen,[1] James E. Bradner,[5,6] Gary D. Bader,[2] Gordon B. Mills,[7] Dana Pe'er,[4] Jason Moffat,[2,3]
and Benjamin G. Neel[1,8,11,*]

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, ON M5G 1L7, Canada
[2]The Donnelly Centre
[3]Department of Molecular Genetics
University of Toronto, ON M5S 3E1, Canada
[4]Columbia University, New York, NY 10027, USA
[5]Department of Medical Oncology, Dana-Farber Cancer Institute
[6]Department of Medicine
Harvard Medical School, Boston, MA 02215, USA
[7]Department of Systems Biology, Sheikh Khalifa Al Nahyan Ben Zayed Institute for Personalized Cancer Therapy,
University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[8]Laura and Isaac Perlmutter Cancer Centre, NYU-Langone Medical Center, NY 10016, USA
[9]Co-first author
[10]Present address: Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada
[11]Present address: Laura and Isaac Perlmutter Cancer Centre, NYU-Langone Medical Center, NY 10016, USA
[12]Present address: National Research Council, Royalmount Avenue, Montreal, QC H4P 2R2, Canada
*Correspondence: benjamin.neel@nyumc.org

http://dx.doi.org/10.1016/j.cell.2015.11.062

SUMMARY

Large-scale genomic studies have identified multiple somatic aberrations in breast cancer, including copy number alterations and point mutations. Still, identifying causal variants and emergent vulnerabilities that arise as a consequence of genetic alterations remain major challenges. We performed whole-genome small hairpin RNA (shRNA) "dropout screens" on 77 breast cancer cell lines. Using a hierarchical linear regression algorithm to score our screen results and integrate them with accompanying detailed genetic and proteomic information, we identify vulnerabilities in breast cancer, including candidate "drivers," and reveal general functional genomic properties of cancer cells. Comparisons of gene essentiality with drug sensitivity data suggest potential resistance mechanisms, effects of existing anti-cancer drugs, and opportunities for combination therapy. Finally, we demonstrate the utility of this large dataset by identifying BRD4 as a potential target in luminal breast cancer and *PIK3CA* mutations as a resistance determinant for BET-inhibitors.

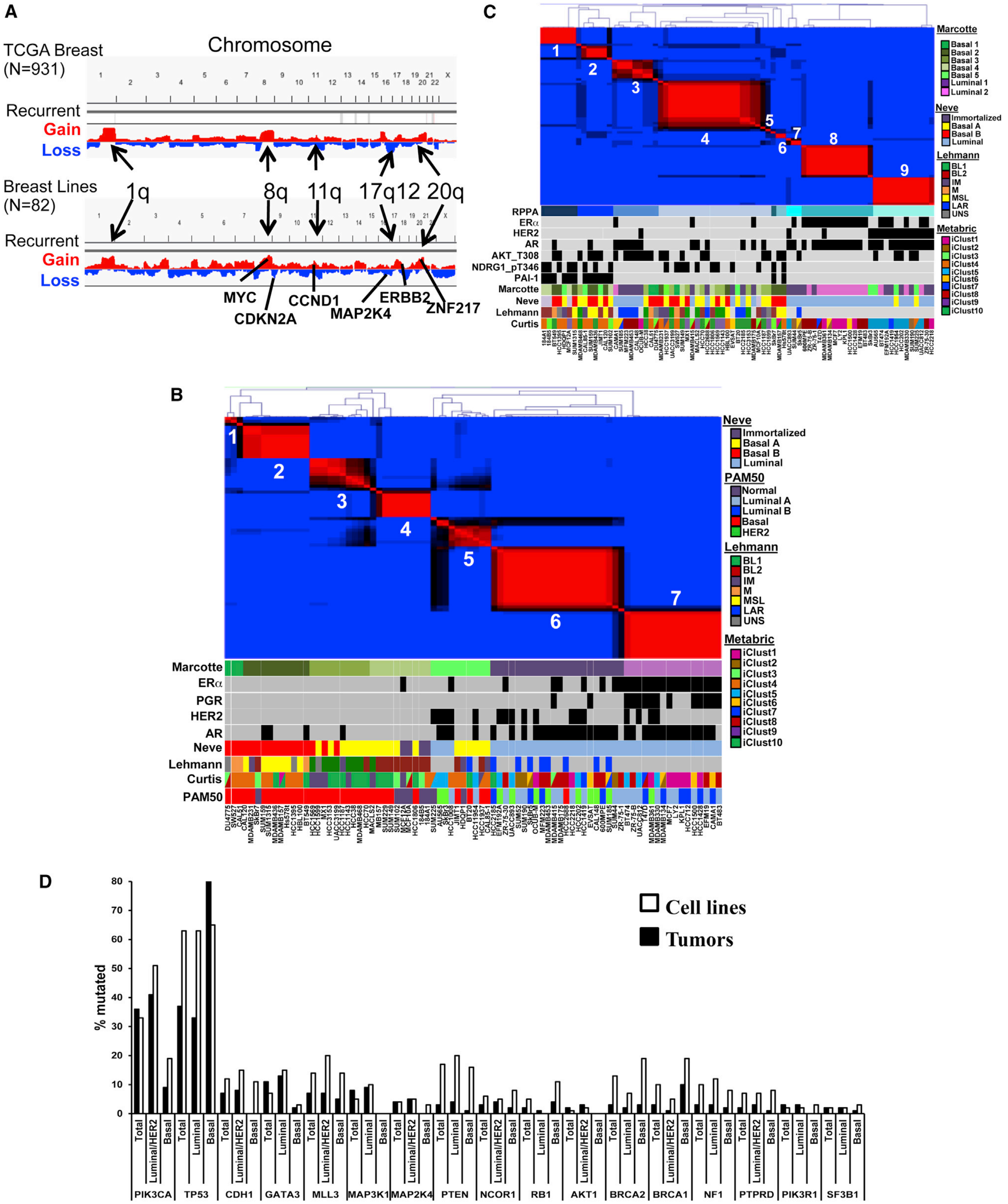
INTRODUCTION

Breast cancer is the second leading cause of cancer death in women. Better detection and therapy have led to >85% 5-year survival, yet half of affected women die from their disease. This outcome reflects incomplete understanding of the molecular alterations, heterogeneity, and determinants of drug response in breast tumors. Genetic and epigenetic abnormalities in breast cancer have been defined, but identifying causal defects and exploiting them for target discovery remain challenging.

"Breast cancer" actually comprises molecular subtypes that predict prognosis and drug response. Early profiling studies identified "intrinsic subtypes": luminal A and B, basal-like (basal), HER2$^+$ and normal-like (Perou et al., 2000; Sørlie et al., 2001). These were joined by a "claudin-low" subtype that, like basal breast cancer, is typically estrogen receptor-negative (ER$^-$), progesterone receptor-negative (PR$^-$), and HER2-negative (HER2$^-$) (Hennessy et al., 2009; Prat et al., 2010). Basal and luminal B tumors have the worst prognosis; claudin-low tumors have intermediate outcome (Prat et al., 2010). Clinically, intrinsic subtypes can be defined by the "PAM50" classifier (Parker et al., 2009).

These molecular subtypes complement, but do not fully overlap, pathologic classification by ER, PR, and HER2 status (Parker et al., 2009). Luminal tumors are typically ER$^+$/PR$^+$, and basal tumors are usually "triple negative" (ER$^-$, PR$^-$, HER2$^-$). Breast cancer cell lines generally fall into four subtypes: basal A or B, HER2$^+$, and luminal (Neve et al., 2006; Prat et al., 2010). Basal A lines resemble "basal" tumors; basal B lines are enriched for claudin-low genes.

Recent large-scale RNA and proteomic profiling studies have further divided luminal and "triple negative" breast cancer (TNBC) into at least ten subtypes (Curtis et al., 2012; Lehmann et al., 2011; Cancer Genome Atlas Network, 2012), and next-generation sequencing (NGS) has identified multiple aberrations in breast tumors (Banerji et al., 2012; Ellis et al., 2012; Shah et al.,



(legend on next page)

2012; Stephens et al., 2012; Cancer Genome Atlas Network, 2012). Whether breast cancer lines represent these new categories and have mutational profiles like tumors remains unresolved.

Moreover, genomics often cannot distinguish "passenger" mutations from "drivers" that promote tumorigenesis and might be therapeutic targets. Highly recurrent defects (e.g., HER2 amplification) point to drivers and some have led to "targeted therapies" (e.g., Trastuzumab). Many other abnormalities, some clearly oncogenic, occur at low frequency, and some drivers are difficult to target (e.g., *MYC*, *RAS*). However, the collateral genotoxic, proteotoxic, and metabolic stresses caused by the abnormal tumor genome can cause "emergent dependencies," potentially providing alternate therapeutic options.

Functional genomics, partnered with genomic data, can identify targets coupled to biomarkers (Zender et al., 2008). Pooled shRNA libraries enable genome-wide "drop-out" screens, which can identify cancer drivers and context-dependent events. Several groups have performed shRNA screens (Cheung et al., 2011; Marcotte et al., 2012), but most surveyed relatively few cell lines of the same cancer type and none represented the diversity of neoplasms such as breast cancer. Here, we report the results of genome-wide shRNA screens of >75 breast cancer lines with genomic, transcriptomic, and proteomic annotation. Employing an improved statistical framework (siMEM), we provide an integrated map of subtype- and context-dependent essentiality in breast cancer cells.

RESULTS

Breast Cancer Lines Are Reasonable Models

We performed genomic and proteomic analysis on 78 breast cancer and four immortalized mammary cell lines (Table S1A). Copy number abnormalities (CNAs) were similar (r = 0.7) in lines and breast tumors, with all major CNAs represented (Figures 1A and S1A). RNA sequencing (RNA-seq) and non-negative matrix factorization (NMF) yielded seven clusters (Figures 1B and S1B). Compared with the Neve classification (Neve et al., 2006), we found four basal, two luminal/HER2^{-}, and one mixed cluster(s). The extra basal clusters mainly sub-divided the basal A and B subtypes (Figures 1B and S1C) and resembled the additional subgroups seen in an extensive survey of TNBC (Lehmann et al., 2011). Most luminal/HER2 cell lines fell into Clusters 6 and 7, which were distinguished by *ERBB2* and *ESR1* expression, respectively. The NMF clusters also related to specific METABRIC "iClusters" (Curtis et al., 2012). Every iCluster was present in the panel, although iClusters 2 and 7 each were represented by less than five lines (Figure S1C). Lines defined as "basal" by PAM50 generally fell into our basal clusters and those of Lehmann (Lehmann et al., 2011), but PAM50-derived signatures did not place luminal/HER2 lines into subgroups similar to those seen by NMF or the Curtis classification.

The top 50% variable proteins by reverse-phase protein array (RPPA) formed nine clusters by NMF (Figures 1C and S1D). With few exceptions, RPPA-(R) and RNA clusters differed markedly. Most (13/18) HER2^{+} lines fell into R-Cluster 9. R-Cluster 8 consisted mainly of expression-derived Cluster 7 lines and was driven by ERα, GATA3, and BCL2. Two small R-clusters were enriched for luminal/HER2 lines: R-Cluster 3 was mainly ER^{-}/AR^{+} and featured high p-AKT (pT308 and pS473) and p-AMPKα (pT172). R-Cluster 7 (three lines) was distinguished by high G6PD, p-4EBP, and reactivity to a VHL antibody that cross-reacts with Epiplakin. The other R-clusters were enriched for basal lines. R-Cluster-1 contained three of the four "normal breast" lines and was driven by NDRG1, MYC, TAZ, and p-YAP. R-Cluster 2 also had high NDRG1, MYC, TAZ, and p-YAP, as well as high PAI-1 and phospho- and total EGFR (Table S1B). R-Cluster-4, the largest, was a default basal cluster.

Exome sequencing of genes mutated in $\geq$3% of breast tumors in COSMIC and TCGA (Table S1C) showed that all frequent somatic mutations in breast cancer were found in our cell line panel. *TP53* and *PIK3CA* mutations (23% and 26%, respectively, in tumors) were seen in 63% and 33% of lines, respectively. *TP53* is mutated more often in TNBC/basal tumors (80% versus 26%) (Cancer Genome Atlas Network, 2012), but its mutation frequency was similar in basal and luminal/HER2 lines. For most genes, mutation rate and distribution were comparable in tumors and lines (Figure 1D).

We also profiled microRNAs (miRNAs) by NanoString. ERα is the major determinant of miRNA levels in breast tumors (Dvinge et al., 2013; Riaz et al., 2013). Similarly, unsupervised clustering revealed three miRNA groups in cell lines, two basal and one luminal (Figure S1E). Overall, we conclude that a sufficiently large cell line panel represents the genomic and proteomic landscape of breast tumors and provides a reasonable template for identifying context-dependent essential genes.

Improved Prediction of Gene Essentiality

To identify genes required for proliferation/survival ("essentials"), we used pooled lentiviral shRNA dropout screens (Marcotte et al., 2012). Nearly all (77/82) lines gave satisfactory data (Table S1A). Using our earlier metric, zGARP, we scored 402 genes as essential in at least 50% of lines (Table S2A). These included most (261/297 and 218/291, respectively) genes defined earlier as "general essential" or "core essential" in ovarian, pancreas, and selected breast cancer lines (Hart et al., 2014; Marcotte et al., 2012). Not surprisingly, genes annotated as having "housekeeping" roles (e.g., translation, splicing, proteasome, cell cycle) were prominent general essentials (Table S2B).

Figure 1. Genomic/Proteomic Characterization

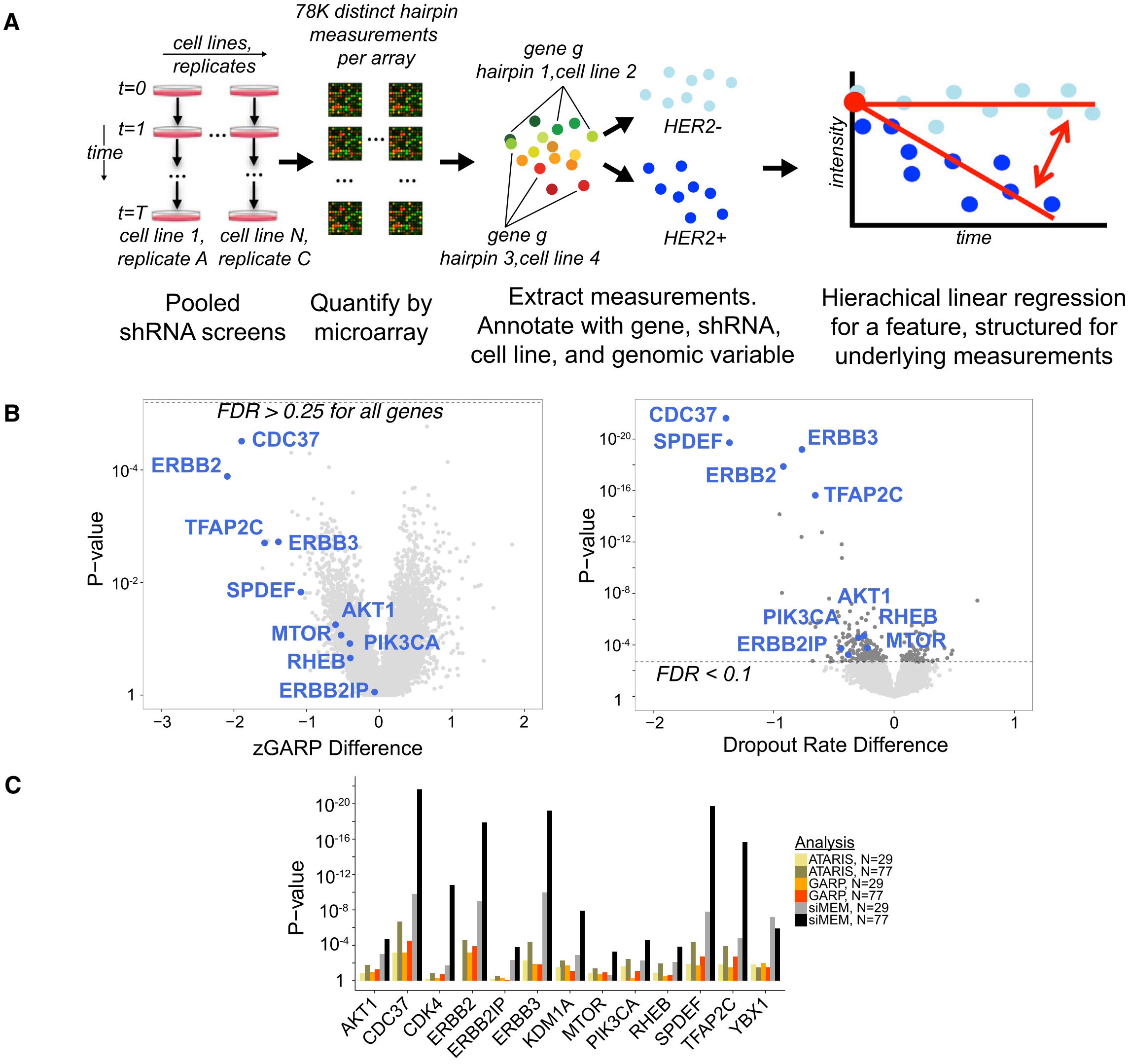
(A) CNA profiles of breast tumors (top) from TCGA and cell lines (bottom).

(B) NMF clustering of RNA-seq data for breast cancer lines. *ESR1* (ER), *ERBB2* (HER2), *PGR* (PR), and *AR* (AR) expression are represented by black squares. Lines were assigned to published subtypes (colored boxes).

(C) NMF clustering of RPPA data.

(D) Frequency of indicated mutations in cell lines and tumors, grouped into basal and luminal/HER2 subtypes. Tumor data are from COSMIC.

See also Figure S1 and Table S1.



D

Achilles dataset (N=102)		ATARIS			siMEM		
analysis	**gene**	**p-value**	**FDR**	**rank**	**p-value**	**FDR**	**rank**
BRAF mutant vs. normal	*BRAF*	$2x10^{-5}$	0.143	1	$3x10^{-14}$	$3.1x10^{-10}$	1
KRAS mutant vs. normal	*KRAS*	$2x10^{-5}$	0.157	1	$7.3x10^{-8}$	$2x10^{-4}$	1
	BCL2L1	Not Significant			$1.4x10^{-5}$	0.025	6
PIK3CA mutant vs. normal	*PIK3CA*	$2x10^{-5}$	0.053	1	$5x10^{-9}$	$5.1x10^{-5}$	3
More essential with increasing expression *(N=83 samples with matching CCLE expression)*	*HNF1B*	$2x10^{-5}$	0.075	1	$3.2x10^{-9}$	$2.2x10^{-6}$	3
	PAX8	$2x10^{-5}$	0.075	2	$2.93x10^{-6}$	$6.97x10^{-4}$	6
	E2F3	$4x10^{-5}$	0.075	3	$8.01x10^{-5}$	0.01	11
	SOX10	$6x10^{-5}$	0.075	5	$1.81x10^{-11}$	$2.29x10^{-8}$	1
	BCL2L1	$1.4x10^{-4}$	0.096	9	$1.72x10^{-10}$	$1.74x10^{-7}$	2
	MYB	$3.4x10^{-4}$	0.213	12	0.003	0.111	70
	KRAS	Not Significant			$7.16x10^{-8}$	$3.62x10^{-5}$	4
	HDAC4	Not Significant			$1.4x10^{-4}$	0.016	13
	MAP3K3	Not Significant			$5.2x10^{-4}$	0.037	23
	MYC	Not Significant			$7.2x10^{-4}$	0.047	24
	CCNE1	Not Significant			$8.7x10^{-4}$	0.052	31
	JAK3	Not Significant			$9.4x10^{-4}$	0.054	32

(legend on next page)

By contrast, neither zGARP, nor other algorithms (ATARIS [Shao et al., 2013], RIGER [Barbie et al., 2009], RSA [König et al., 2007]), identified known subtype-specific essential genes from our large dataset. Such methods summarize replicate shRNA measurements into single "hairpin" or "gene" scores, which are compared between subtypes by t tests or similar statistics. This approach leads to loss of information about measurement variance, limiting statistical power to detect biological differences.

Hierarchical ("mixed-effect") linear models allow systematic measurement effects, such as hairpin differences or heterogeneous genetic contexts, to be specified and used in significance calculations. Such a model could increase sensitivity for detecting biological differences in screens by avoiding information loss, while limiting false positives. We therefore developed the small interfering RNA (siRNA)/shRNA mixed-effect model (siMEM), which considers the level of each shRNA to be a regression function of its initial abundance, baseline trend in abundance over time, and difference in abundance trend between samples sharing a common feature (Figures 2A, S2A, and S2B; Supplemental Experimental Procedures).

Using siMEM and previous metrics, we sought genes selectively required in HER2$^+$ lines (n = 17). Reassuringly, siMEM-detected known HER2$^+$-associated essentials ("known positives"), such as *ERBB2*, its dimerization partner *ERBB3*, PI3K/mTOR pathway members (*PIK3CA*, *AKT1/2*, *RHEB*, *MTOR*), *CDC37* (encodes an ERBB2 co-chaperone), and two transcription factors (*TFAP2C*, *YBX1)* in the HER2 (ERBB2) pathway. Almost none of these survived false discovery rate (FDR) correction using GARP or ATARIS (Figure 2B; Table S2C). Only siMEM predicted "known positives" from the data in our earlier screen (Marcotte et al., 2012) and it greatly improved their prediction rankings and p values (Figures 2C and S2C). When classes (normal/HER2$^+$) were shuffled randomly for each gene, siMEM p values were close to the expected uniform distribution (Figure S2D). Regression structures that ignored systematic measurement effects produced many (incorrectly) significant p values (Figures S2E and S2F). By contrast, siMEM produced the best fit and ranking of known positives (Figures S2B, S2G, and S2H). Finally, we applied siMEM and ATARIS to the "Achilles" dataset (Cheung et al., 2011): siMEM was better at predicting *BRAF*, *KRAS*, or *PIK3CA* essentiality in cognate mutant cells and in finding genes more essential with increased expression, which are enriched for drivers (Figure 2D; also see below).

Breast- and Subtype-Specific Essential Genes

We focus here on gene essentiality relative to the Neve classification, which most closely resembles clinical subtypes, but Tables S3A–S3G provide essentiality data for each subtype in Figures 1B and 1C. Comparing basal with luminal/HER2 cell lines, we found 975 and 985 subtype-specific essentials (FDR < 0.1), respectively (Figure 3A; Tables S3F and S3G). The top luminal/HER2-essentials were *FOXA1*, a pioneer factor for ERα (Lupien et al., 2008), *SPDEF*, which promotes luminal differentiation and survival of ERα^+ cells (Buchwalter et al., 2013), *CDK4* and *CCND1*, which form a complex targeted by Palbociclib in ER$^+$ breast cancer (Dhillon, 2015), and *TFAP2C*, which directs *ERBB2* expression (Bosher et al., 1995). Other "expected" luminal/HER2-essential genes included PI3K/mTOR pathway components *(PIK3CA*, *PDPK1*, *AKT1/2*, *RHEB*, *MTOR*) and ER-interacting proteins/co-activators (*KMT2D*, *EP300*, *GATA3*, *KDM1A*, *DNM1L*, *NCOA*2).

The top basal-selective essentials, *PSMB3* and *PSMA6*, encode proteasome subunits (Table S3F), a dependency seen earlier (Petrocca et al., 2013). The next most essential basal-specific gene was *ATP6V1B2*, which encodes a component of the vacuolar ATPase required for lysosomal acidification that is the target of Bafilomycin A1 (BafA1). Notably, basal lines were 5-fold more sensitive and basal A lines were 7-fold more sensitive to BafA1 than luminal/HER2 lines (Figures S3A and S3B). Other genes reputedly more important in basal breast cancer scored as "basal-essential," including *PLK1*, *EGFR*, *FZD7*, *SLC7A11*, *CTNNB1*, *LRP5*, *FZD8*, and *TWIST2* (Jamdade et al., 2015; Maire et al., 2013; Timmerman et al., 2013), but we also saw other potential vulnerabilities (Table S3F).

We selected several subtype-specific genes for orthogonal testing with siRNAs. Multiple basal-specific, luminal-specific, and HER2-specific genes validated and demonstrated the predicted subtype preference (Figure 3B; Table S3L). Overall, the validation rate was ~70%, with most siRNAs showing >80% knockdown (Figure 3C; data not shown).

The genomics of basal breast cancer and high-grade serous ovarian cancer (HGSC) are very similar (Cancer Genome Atlas Research Network, 2011; Cancer Genome Atlas Network, 2012). Remarkably, in a pairwise comparison with luminal-specific (this screen) or HGSC- or pancreatic cancer-specific essentials (Marcotte et al., 2012), only 20 essential genes differed between basal breast cancer and HGSC. By contrast, thousands of differences were seen in all other comparisons (Figure S3C).

We analyzed subtype-specific essential gene sets for preferred pathways and protein-protein interactions (PPIs) (Figures 3D and 3E; Tables S3H–S3K). As expected, HER2-specific essential pathways included EGF, PI3K, and mTOR signaling. Other functions important in this subtype included regulation of eIF2, aerobic ATP synthesis/TCA cycle, chromatin-modifying enzymes, "response to gamma radiation" (including *YAP1*, *ATR*, and *ATM*), as well as an EP300/BRCA1 PPI sub-network (Table S3J). EP300 is a BRCA1 co-activator (Pao et al., 2000), and BRCA1 is phosphorylated via the PI3K/AKT pathway, which also is required in HER2$^+$ lines (Figure 2C; Tables S2C and S3B).

Figure 2. siMEM Overview

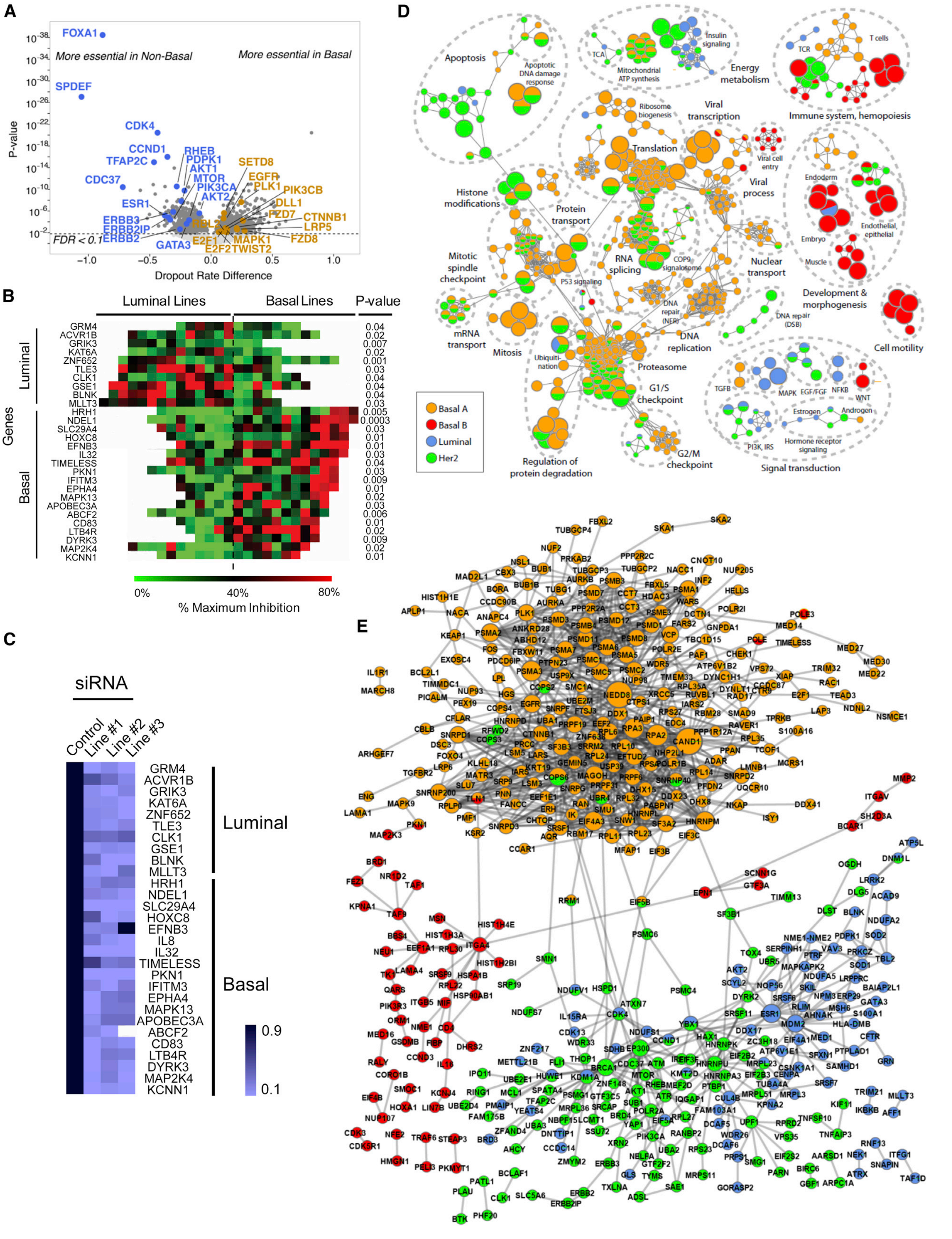
(A) Experimental scheme. Samples were hybridized to microarrays and dropout was quantified. Hierarchical linear regression summarizes data as a combination of initial measurement intensity, baseline trend, and difference in essentiality associated with changes in a genomic covariate (light blue versus dark blue).

(B) Volcano plot of zGARP (left) and siMEM (right) essentiality differences associated with HER2$^+$ lines. Dotted lines show FDR cut-off.

(C) siMEM produces the best p values for known positives.

(D) *BRAF*, *PIK3CA*, or *KRAS* mutant versus normal and expression versus essentiality analyses of the Achilles dataset (n = 102).

See also Figure S2, Table S2, and Supplemental Experimental Procedures.



(legend on next page)

Notably, ATM is essential for HER2$^+$ tumors (Stagni et al., 2015) and it also phosphorylates BRCA1 (Cortez et al., 1999; Gatei et al., 2000). Preferential sensitivity to loss of DNA damage sensors might explain the observed synergy of chemotherapy and Trastuzumab.

Top enriched pathways and PPIs for basal A lines were dominated by genes for splicing, the proteasome and mitosis (Figures 3D and 3E; Table S3H). Other required functions included the COP9 signalosome (CSN) and a PPI sub-network defined by CAND1/NEDD8 (Figure 3E). CSN and CAND1/NEDD8 regulate SKP1/CUL1/F-box (SCF) complexes (Flick and Kaiser, 2013). While the core SKP1/CUL1 complex showed no subtype specificity, several F-box genes were selectively essential in basal A lines, including *FBXW11*/β-*TrCP2* (Table S3B). *FBXL6* and *FBXO15* were more essential in basal B or luminal/HER2 lines, respectively (Table S3B; data not shown). Hence, F-box proteins might impart subtype-specific functions to SKP1/CUL1.

Lack of functional annotation (<50% of genes annotated) resulted in a relative paucity of basal B and luminal nodes when compared to basal A- and HER2- nodes (>65% of genes annotated, Figures S3D–S3F). Nevertheless, essential pathways and PPI networks for luminal lines included epithelial development, MDM2, PI3K, and hormone receptor (*ESR1*) signaling (Figures 3D and 3E). The latter two are targets of known drugs for luminal breast cancer. Less expected "luminal-enriched" pathways/PPIs included redox-related (SOD1, SOD2, ENOX1) and mitochondrial (e.g., electron transport chain, mitochondrial ribosome) proteins. By contrast, basal B-essentials were enriched for genes related to polarity (*PARD3*, *PAR3D*), cell-cell junctions and adhesion (*CDH2*, *CLDN1*, *CLDN4*, *ITGA4*, *ITGAV*, *ITGB5*), embryonic development, organ morphogenesis, fatty acid metabolism, and T cell immunity (Figures 3D and 3E). Some of these genes, such as *SOX9* (Guo et al., 2012a), *KLF4* (Yu et al., 2011), and *ALOX5AP* (Kim et al., 2005), have reported roles in breast cancer, although not specifically in basal B tumors.

cis- and *trans*-Essential Interactions with Common CNAs

There are hundreds of CNAs in breast cancer (Curtis et al., 2012; Cancer Genome Atlas Network, 2012), yet for most, the key driver gene(s) is unclear. METABRIC defines 30 regions of copy number gain and 15 deletions (Curtis et al., 2012). ISAR, being more sensitive for small amplicons, identifies 83 recurrent CNAs (Sanchez-Garcia et al., 2014). We predicted significant (FDR < 0.2) *cis*-essential genes (more essential in amplicon$^+$ lines) for 9/83 ISAR regions. Four corresponded to genes in a METABRIC amplicon (Figure S4A; Table S4A): *EGFR* (ISAR(I)-34/METABRIC(M)-10), *CCND1* (I-52/M-21), *ERBB2* (I-70/M-35), and *TFAP2C* (I-81/M-42). The others were unique to ISAR-defined regions (Table S4B): *CTSS* (I-6), *ESR1* (I-30), *RALGAPA1* (I-62), *FOXA1* (I-63), and *BCL2* (I-76).

Even for known drivers (or for deletions), targeting the key gene can be difficult. "*trans*-Essential" genes can suggest alternative strategies. Combining all METABRIC regions, we identified 2,560 *trans*-essentials, an average of 58 per CNA (range 0–285; Figures 4A and S4A; Table S4A). Only 61 (~3%) *trans*-essentials showed significantly increased or decreased expression in sensitive lines (Figure S4B and Supplemental Experimental Procedures); hence, most would not be found by gene expression surveys. Expected *trans*-essentials were seen for the *CCND1* (*CDK4*, *USP18*) (Guo et al., 2012b) and *ERBB2* (*ERBB3*, *CDC37*, *PIK3CA*) amplicons and for *CDKN2A* deletions (*CCND1*, *CDK6*) (Figures 2B and S4A; Table S4A). It can be difficult to know if a *trans*-essential is "expected" for deletions, especially if the cognate tumor suppressor is undefined. Even so, we saw intriguing associations with "druggable" targets for region 27, containing *RB1* (more sensitive to *MAP2K2* depletion), region 11 (more sensitive to *TLK2*, *BRD4*, or *ACVR1B* depletion), and region 40 (more sensitive to *PTK6* or *MAP2K4* depletion) (Table S4A).

METABRIC region 14 includes *MYC*, which is generally deemed "undruggable." Notably, *MYC* was the most essential gene in region 14-amplified lines (Figure 4A), but was not differentially essential by FDR, probably because of its requirement in most tumor cells (Dang, 2012). Pathway analysis of the 91 region 14 *trans*-essentials (FDR < 0.2; Table S4A) revealed genes for mitosis, DNA replication, and RNA metabolism (Table S4C), all known MYC functions (Dang 2012). MYC transcriptional targets (Figure 4B) and genes encoding MYC-interacting proteins (Table S4D) also were strongly enriched: 46% of *MYC trans*-essential genes were MYC transcriptional targets/interactors. We tested two *MYC trans*-essentials potentially amenable to drug discovery; indeed, amplified lines were preferentially sensitive to *MINK1* or *USP5* depletion (Figure 4C). We also validated *YAP1* and *BRCA1* as *trans*-essential for METABRIC regions 35 (contains *ERBB2*), and 36 (putative driver: *ZNF652*), respectively (Figures S4C and S4D).

HELIOS integrates CNA, expression, mutation, and essentiality into a single score that predicts *cis*-essential genes (Sanchez-Garcia et al., 2014). The initial HELIOS report, using data from our earlier screen, identified and validated ten potential drivers. Using our expanded dataset, the HELIOS score increased for most known drivers and previously validated genes (Figure 4D; Table S4E**)**. We also tested two new predictions and found that amplicon$^+$ lines were more sensitive to siRNA-mediated depletion (Figure 4E).

Figure 3. Subtype-Specific Essential Genes

(A) Volcano plot of basal-specific and luminal/HER2-specific essentials.

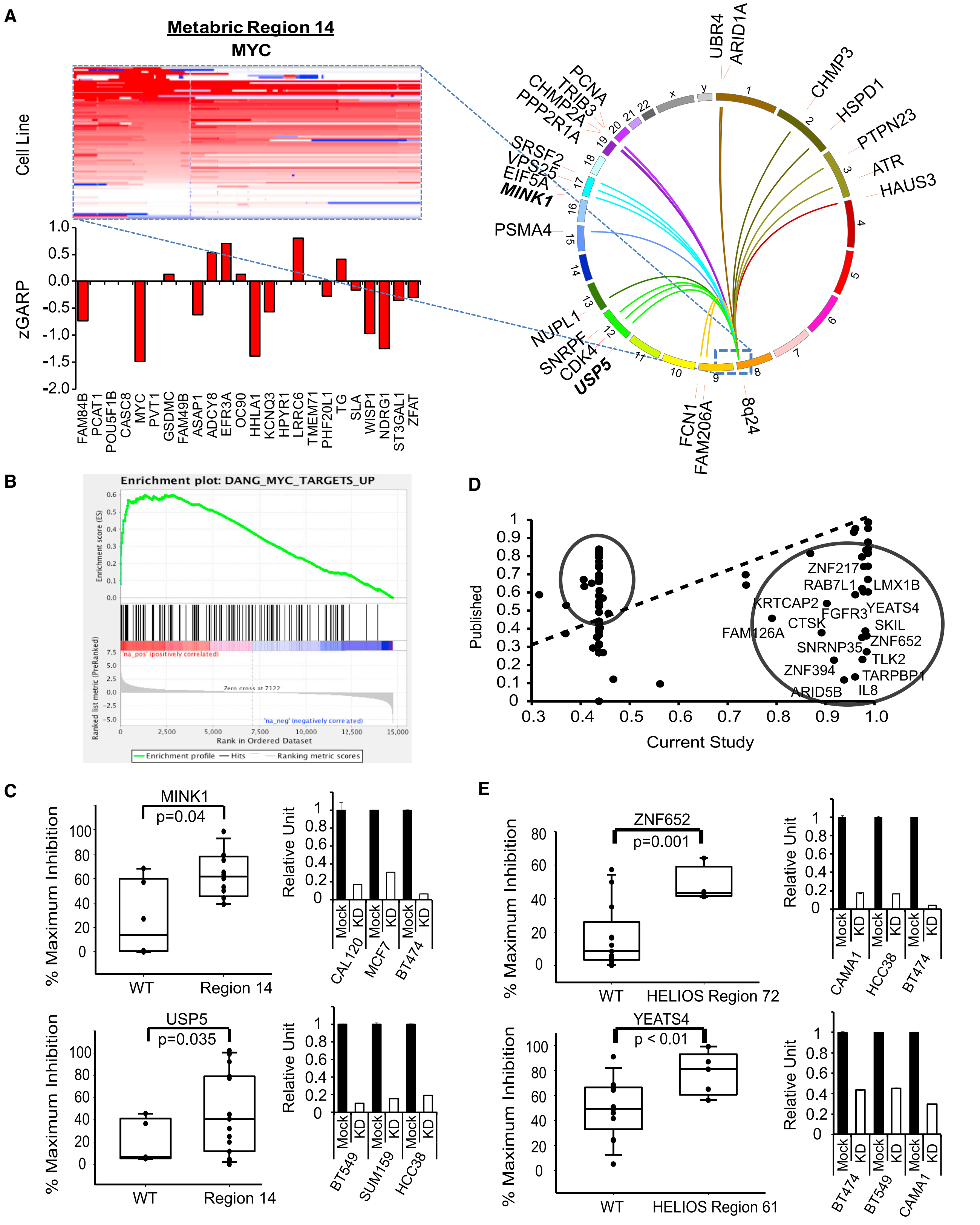
(B) Heatmap shows % proliferation-inhibition, compared to general essential *RPL9* (100% inhibition), after pooled siRNA treatment (p values: one-sided t test).

(C) Knockdown efficiency (by qRT-PCR) of siRNAs for genes in (B).

(D) Subtype-specific pathways. Each node represents a process; functionally similar nodes are grouped and labeled by enriched function. Nodes are colored according to the subtype in which the process is enriched; processes enriched in more than one subtype have multiple colors. Red, basal B; orange, basal A; green, HER2$^+$; blue, luminal.

(E) PPI networks for subtype-specific genes. Nodes represent genes and are multi-colored if present in multiple subtypes; edges represent interactions.

See also Figure S3 and Table S3.



(legend on next page)

Functional Genomic Clustering Reveals Groups Not Captured by Expression Profiling

Using NMF clustering, we grouped lines based on shared dependencies (“functional genomic clustering”) (Marcotte et al., 2012). Six “functional clusters” (fClusters) were observed, two containing lines designated as basal by expression profiling (fCluster-4 and fCluster-5), two luminal/HER2 clusters (fCluster-2 and fCluster-6), and two (fCluster-1 and fCluster-3) comprising a mix of basal and luminal/HER2 lines (Figure 5A). Thus, as we saw earlier (Marcotte et al., 2012), “basal” and “luminal/HER2” lines have distinct patterns of gene dependency. Yet, while there was little additional separation in our earlier study, with our expanded panel, HER2 (mainly fCluster-2) and ER^+ (fCluster-6) lines largely segregated into distinct fClusters. Genes determining the ER^+ (fCluster-6), $HER2^+$ (fCluster-2), and basal (fCluster-4) clusters (Table S5A) overlapped substantially with luminal-, HER2A-, or basal- essential genes, respectively (Figure 3). fCluster-1 was enriched for genes curated as H3K27-trimethylated, neuroactive peptides, or as involved in cytokine-cytokine interactions. fCluster-3 was enriched for annotations for cell cycle (G1/S and mitosis), DNA replication, and immune system genes, whereas fCluster-5 was enriched for genes involved in the immune system, lipid metabolism, and NGF signaling (Table S5A).

Drug Sensitivity and Gene Essentiality

We also compared gene essentiality and sensitivity data for 90 drugs tested against 84 breast cancer lines (Daemen et al., 2013), most of which (69) were included in our panel. Using siMEM, we identified genes whose essentiality correlated with sensitivity to mTOR/PI3K/ERBB2/AKT or EGFR/MEK/ERK inhibitors. Hierarchical clustering revealed distinct positive (red) and negative (blue) correlation clusters associated with drug sensitivity (Figure 5B; Supplemental Experimental Procedures). Reassuringly, genes for PI3K/AKT pathway components were required in lines sensitive to the cognate inhibitors. Sensitivity also correlated with essentiality of the luminal markers *ESR1*, *FOXA1*, and *GATA3*, consistent with the known sensitivity of luminal tumors to these agents. Likewise, EGFR/MEK/ERK inhibitor response correlated with sensitivity to *EGFR*, *GRB2*, *SOS1*, *MAPK1*, *MAPK3*, or *MAP2K1* depletion. Interestingly, response to EGFR/MEK/ERK inhibitors correlated with dependence on the NF-kB pathway: *RELA*, *REL*, and *NKAP* were more essential in such cells. These results comport with reports of NF-kB activation in response to EGFR, RAS, RAF, or MEK activation (Pan and Lin, 2013) and suggest that NF-kB inhibitors might be effective in basal breast cancer.

Drug sensitivity/essentiality comparisons also identified negative regulatory/tumor suppressor pathways. For example, *PTEN* was more essential in lines that were insensitive to mTOR/PI3K/ERBB2/AKT or EGFR/MEK/ERK inhibitors, consistent with the effects of PTEN deletion/inactivation (Worby and Dixon, 2014). Likewise, *MDM2* and *TP53* essentiality were associated with sensitivity or resistance to Nutlin-3A treatment, respectively.

Unsupervised analysis of the whole gene essentiality/drug sensitivity dataset revealed five clusters. Most drugs with a similar mechanism of action fell into the same cluster, and pathway analysis confirmed that essentiality clusters were enriched for genes implicated in the pathways targeted by their respective agents (Figure S5A; Tables S5B and S5C). Unanticipated clusters also emerged. For example, sensitivity to 11 drugs, which included alkylating agents, topoisomerase inhibitors, and cell cycle/cell cycle checkpoint inhibitors, correlated with essentiality of genes “associated with the H3K27me3 mark” (e.g., *PRDM13*, *NKX2-5*, *HOXC8*, *PAX7*, *HES2*) and for “neuropeptides and neurotransmitter signaling” (Figures S5B, box 2, S5C, and S5D). Notably, we had validated one of these genes, *HOXC8*, in our siRNA assays (Figures 3B and 3C).

Screen/drug sensitivity data might suggest drug combinations to kill resistant cells and/or negative regulators associated with drug resistance. For example, drugs targeting the PI3K/mTOR pathway (Cluster-1) strongly anti-correlated with *BCL2L1* essentiality (i.e., cell lines resistant to PI3K/mTOR inhibitors required *BCL2L1*). Interestingly, drug combinations targeting the PI3K/mTOR pathway and BCL-X_L are reported for several malignancies (Muranen et al., 2012; Rahmani et al., 2013). Another known combination predicted by our data is EGFR plus HDAC inhibitors (Zhang et al., 2015). Suggested combinations awaiting validation include RAF/MEK and CDK4 inhibitors, EGFR inhibitors with Cluster-5 drugs, BET-Is with Cluster-4 drugs, especially epirubicin and vinorelbine, PLK1 inhibitors with Nutlin-3A or PI3K/AKT inhibitors or Nutlin-3A with Cluster-5 drugs (Table S5B).

We also used DGIdb to identify essential genes that are potentially “druggable” (Griffith et al., 2013). Genes for kinases, phosphatases, and histone modifying enzymes were the most frequently essential, although other categories were represented (Figure 5C; Table S5D). Inhibitors exist for only a small fraction of most potential targets, especially the histone modifiers; a larger percentage of essential kinases had a known inhibitor (Figures 5C, 5D, and S5E).

Figure 4. *cis*- and *trans*-Essential Genes for CNAs

(A) Heatmap showing 8q24 amplification (METABRIC-14, containing *MYC*) in cell lines. Red, amplification; blue, deletion. Bar graph shows average zGARP score for genes in the amplified region in $amplicon^+$ lines. CIRCOS plot depicts top 20 significant genes (by siMEM) in $amplicon^+$ versus $amplicon^-$ cells.

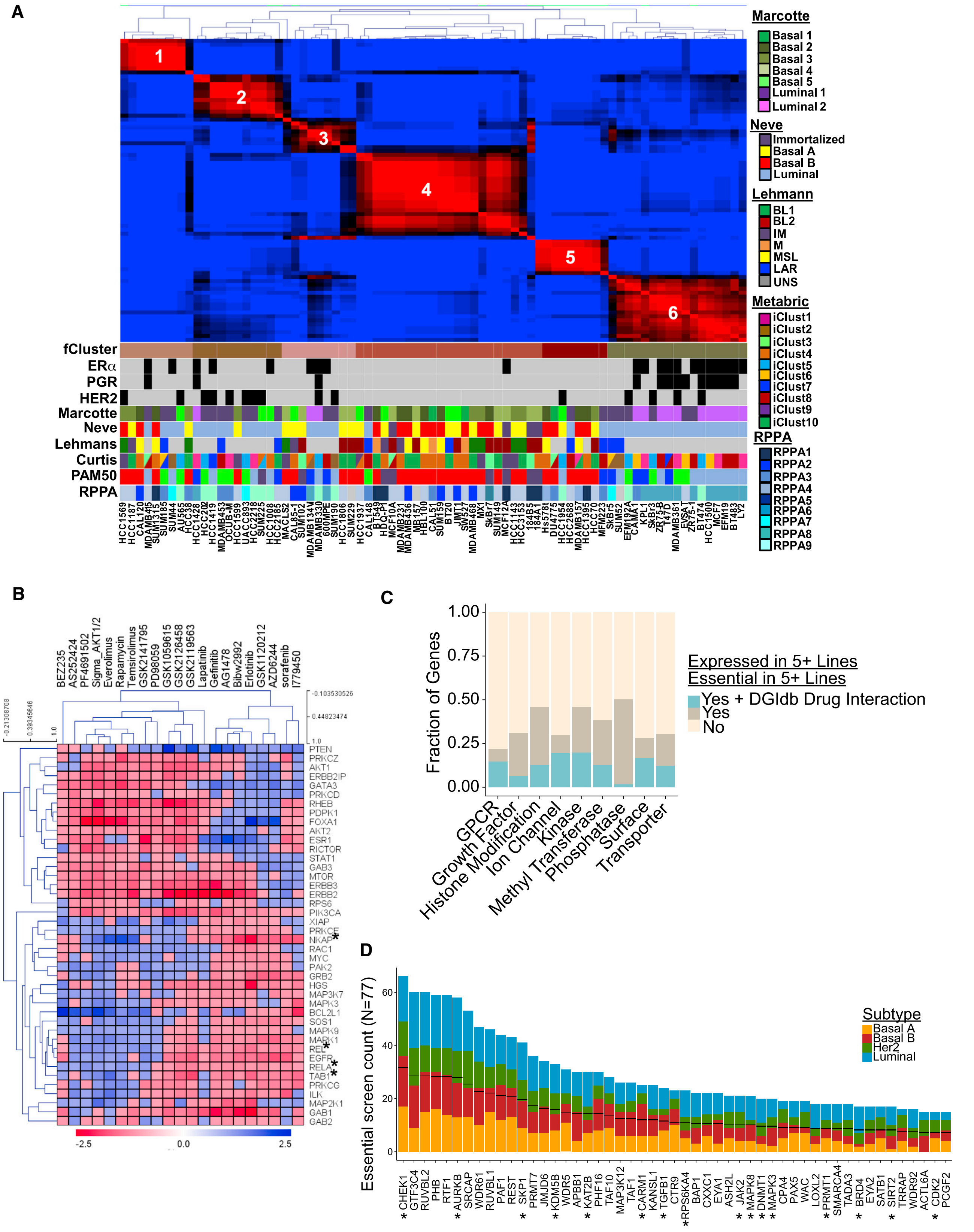
(B) GSEA of *trans*-essential genes for MYC targets ($FDR < 0.0001$).

(C) Validation of 8q24 *trans*-essential genes with siRNAs. y axis, % maximum inhibition; bar graphs, knockdown efficiency (by qRT-PCR) of siRNAs.

(D) Correlation between published HELIOS scores (y axis) (Sanchez-Garcia et al., 2014) and new scores (x axis) obtained using our screen data. Circled genes deviate from earlier score and represent potential new amplified drivers.

(E) Validation of HELIOS genes with siRNAs. y axis, % maximum inhibition; bar graphs, knockdown efficiency of siRNAs. P values were calculated by one-sided t test.

See also Figure S4 and Table S4.



(legend on next page)

Additional Functional Genomic Properties of Cancer Cells

For most genes, essentiality decreased as expression increased (Figure 6A, right); such genes are enriched for housekeeping functions (Table S6A). A smaller set of genes became more essential with increased expression (Figure 6A, left): 16 of the 20 top-ranked genes in this group are known drivers in breast or other cancers (Table S6B). We suspected that other genes whose essentiality increased with increased expression might be drivers and tested several using siRNAs (Figures 6B and S6A). Indeed, 11/20 (55%) were more essential in lines with increased expression ($R > 0.3$). Genes more essential with increased expression showed lower expression overall than genes whose essentiality lessened with increased expression (Figure S6B). The former were more variably expressed, although, consistent with the behavior of known oncogenes (e.g., *ESR1*, *ERBB2*).

"CYCLOPS" (Nijhawan et al., 2012) and "GO" (Solimini et al., 2012) genes show increased essentiality upon heterozygous deletion of their cognate genomic regions. We identified 224 genes (FDR < 0.2) that were more essential with copy number loss (Figure 6C; Table S6C); their essentiality also correlated strongly with decrease in their expression (Figure 6D; Spearman $\rho = 0.74$). These genes overlapped significantly with CYCLOPS and GO genes, only five showed homozygous deletion in any line, and their protein products were enriched for housekeeping functions (Figure S6C; Table S6C; Supplemental Experimental Procedures). Thus, our data validate the CYCLOPS/GO concept and provide many other candidate members of this class of genes.

PIK3CA Mutations Drive Resistance to BET-I

BRD4, encoding a BET bromodomain-containing co-activator (Shi and Vakoc, 2014), was preferentially essential in luminal/HER2 lines (Figure 7A; Table S3G). Moreover, luminal/HER2 lines were more sensitive to *BRD4* depletion by siRNAs (Figures 7B and S7B), and expression of shRNA-resistant *BRD4* cDNA abrogated inhibition by *BRD4* shRNA (Figure S7C).

We tested the BET domain inhibitor (BET-I) JQ1 on a subset of our lines, expecting greater sensitivity in luminal/HER2 cells. Cell line GI50s ranged from low nM (<100) to μM (>2.5), with lines that showed high JQ1 sensitivity undergoing apoptosis, while resistant lines had slower cell-cycle progression (Figures S7D–S7F). However, many luminal/HER2 lines sensitive to *BRD4* knockdown were JQ1-resistant. By contrast, most basal lines that were sensitive to *BRD4* knockdown were JQ1-sensitive (Figure 7C; data not shown). In contrast to previous studies (Shi and Vakoc, 2014), JQ1 sensitivity did not reflect impaired *MYC* expression: sensitive and resistant cell lines displayed similar decreases in *MYC* mRNA (Figure S7G), and exogenous MYC did not convert JQ1-sensitive lines to JQ1-resistance (Figures S7H and S7I).

Instead, integrative analysis revealed a strong correlation between JQ1 resistance and *PIK3CA* mutation (Figure 7C). Overexpression of wild-type or mutant *PIK3CA* conferred JQ1 resistance on JQ1-sensitive SkBR3 cells (Figure 7D), establishing a causal relationship between PI3K and resistance. Moreover, A66, a PIK3CA-specific inhibitor, but not TGX-221 (PIK3CB-specific), increased the JQ1 sensitivity of resistant cells, as did the mTOR inhibitors rapamycin or Torin (Figures 7E and 7F). The one basal line (SUM159) sensitive to *BRD4* depletion but JQ1-resistant also has a *PIK3CA* mutation, and PIK3CA inhibitor treatment sensitized these cells to JQ1 (Figure S7J). Finally, combining JQ1 and Everolimus enhanced their respective anti-tumor effects (Figure 7G). In concert, these data indicate that BRD4 has bromodomain (BrD)-dependent and BrD-independent effects in breast cancer cells and establishes *PIK3CA* mutations as a BET-I resistance mechanism.

DISCUSSION

Most dropout screens analyze relatively few lines of any single cancer histotype. By contrast, we provide gene essentiality data for a large set of breast cancer lines with genomic, proteomic, and drug response annotation, and an analytic tool, siMEM, that more precisely measures differential essentiality. Our results identify and provide initial validation of synthetic lethal relationships with expression subtypes and CNAs, yield insight into essential pathways that correlate with anti-cancer drug response, and reveal general features of functional genomic screens. Illustrating the utility of combining genomic/functional genomic data, we identify and validate *BRD4* as a luminal/HER2-selective essential gene, uncover BET-independent requirements for BRD4 in luminal/HER2 cells, and reveal *PIK3CA* mutations as a potential resistance mechanism to BET-Is in vitro and in vivo.

The breadth of our screen has several advantages. Many have argued that breast cancer lines only partly reflect tumor heterogeneity (Hollestelle et al., 2010; Kao et al., 2009; Neve et al., 2006). But there are at least ten breast cancer subtypes (Curtis et al., 2012; Lehmann et al., 2011; Cancer Genome Atlas Network, 2012); only a large panel could possibly represent such heterogeneity (Figures 1 and S1). Our screen identified nearly all known breast cancer drivers linked to the appropriate biomarker (Figures 3A, 4A, and S4A). The increased power of our dataset also revises the identification of putative targets of

Figure 5. Screen Refines Classification and Pathway Identification

(A) NMF clustering of screen results (zGARP). *ESR1*, *ERBB2*, and *PGR* expression are shown by black squares. Colored boxes indicate major published subcategories.

(B) Unsupervised analysis of essential genes implicated in PI3K/mTOR or EGFR/MEK/ERK pathways. Heatmap shows association of essentiality for each gene (this study) with sensitivity to drugs targeting these pathways (Daemen et al., 2013). The asterisk (*) indicates genes belonging to the NF-κB pathway.

(C) Fraction of essential genes overlapping with reported "druggable" gene categories or gene-drug interactions (DGIdb).

(D) Top-ranked histone-modifying enzymes deemed essential in our screen, by breast cancer subtype. *Reported gene-drug interaction in DGIdb. Black lines represent 50% of lines in which the gene is essential.

See also Figure S5 and Table S5.

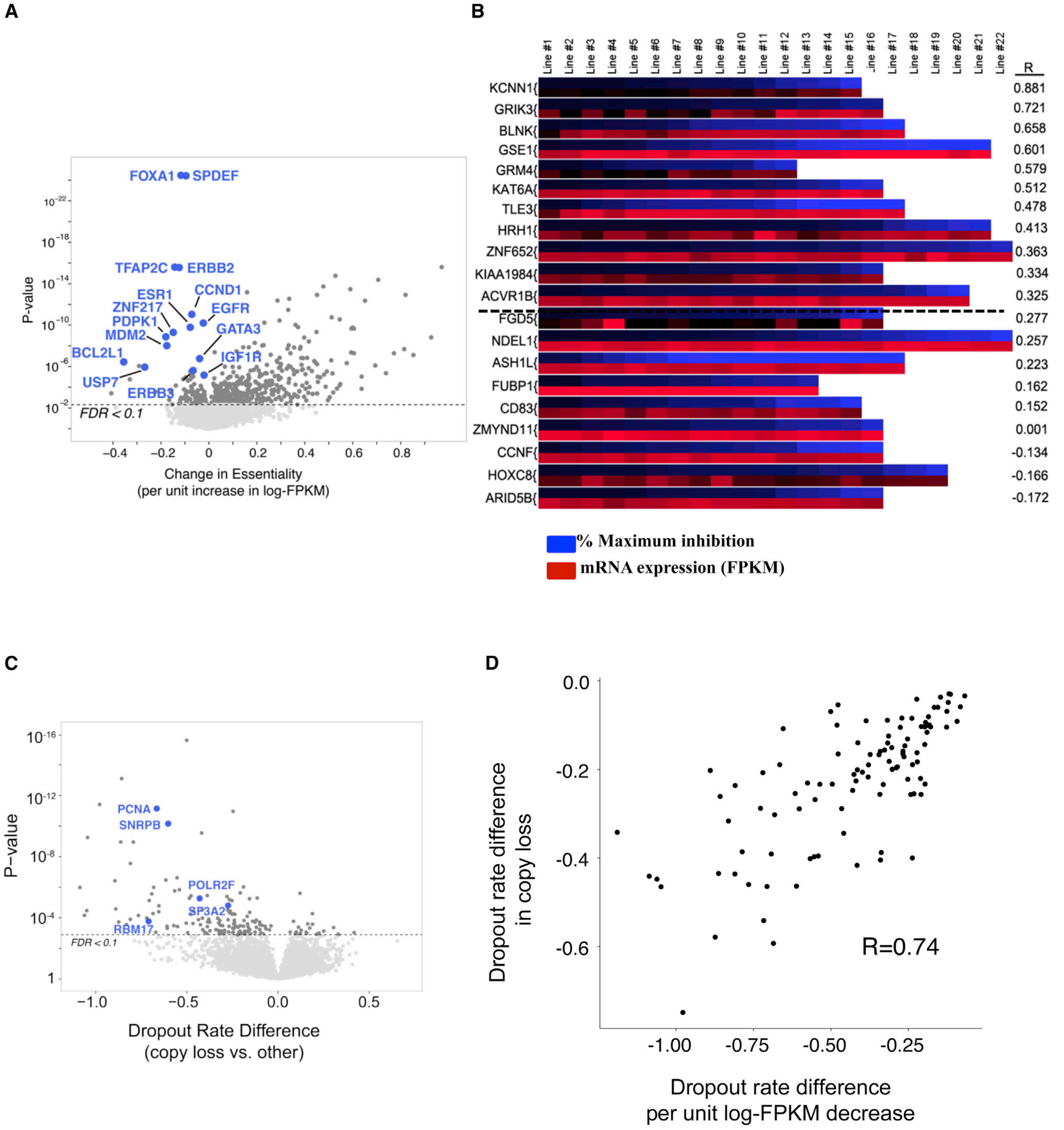


Figure 6. Additional Features of shRNA Screens

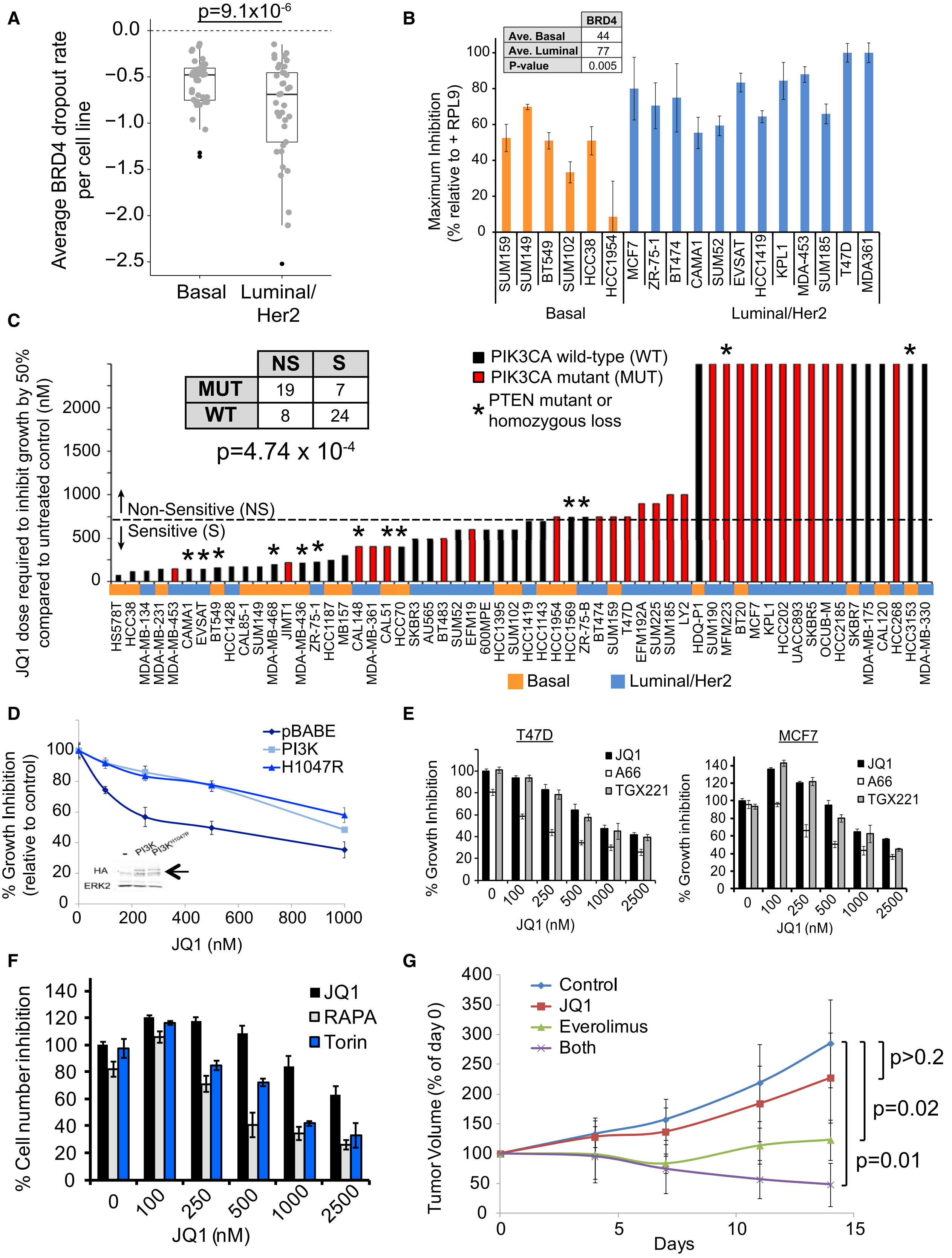
(A) Volcano plot of relationship between essentiality and gene expression. x axis, change in dropout rate per unit increase in expression log-FPKM; y axis, p value.

(B) Heatmap showing % inhibition of proliferation following knockdown by siRNA in cell lines. For each gene, the upper row (blue) represents maximum growth inhibition, while the lower row (red) represents mRNA levels of the same gene in each line. R, Pearson correlation.

(C) Vulnerabilities associated with genomic loss (CYCLOPS genes).

(D) Strong agreement (Spearman $\rho = 0.74$, p value $< 2.2 \times 10^{-16}$) between genes more essential with heterozygous loss (FDR < 0.25) and genes whose essentiality changes significantly with expression (FDR < 0.25).

See also Figure S6 and Table S6.



(legend on next page)

some breast cancer amplicons and strengthens the identification of others by HELIOS (Figure 4D; Table S4E). Thus, if enough cell lines are tested, they provide valid surrogates for probing core cancer cell properties, such as proliferation/survival.

Conventional algorithms for sh/siRNA screens generate hairpin-level and/or gene-level scores that summarize multiple measurements and fail to identify known differential essential genes. By contrast, siMEM greatly improves detection of essentials associated with CNAs, gene expression, somatic mutations, or cancer subtype without increasing the false positive rate. ‘‘Hits’’ suggested by siMEM have a high validation rate ($\sim$60%–70%) (Figures 3B and 6B), and an analogous approach can be applied to any pooled screen (e.g., CRISPR/Cas9 screens).

Our screen identified ‘‘general’’ and ‘‘context-specific’’ essentials. As expected, general essentials are enriched for housekeeping functions, yet some show a gradient of essentiality tied to specific genetic parameters. For example, specific splicing factors (data not shown, but see Hsu et al., 2015) and proteasome genes are preferentially required in basal lines (Table S3F). A splicing inhibitor is in clinical trials (E7107; NCT00459823), and several proteasome inhibitors are approved drugs (Dou and Zonder, 2014) and could be repurposed for breast cancer therapy.

Our data provide strong confirmation of earlier work suggesting distinct subtype-specific vulnerabilities. The pivotal roles of hormone receptors in luminal breast cancer, of ERBB2 signaling in HER2$^+$ disease, and of EGFR and WNT signaling in basal breast cancer are confirmed by our screen hits (Figures 3A, 3D, and 3E; Tables S3F and S3G). We also identify several ‘‘druggable’’ targets, including *EFNB3*/*EPHA4*, *MAP2K4*, *MAPK13*, and *IL32*, for basal breast cancer, the most lethal form of the disease. How these genes promote basal breast cancer is unclear. EFNB3/EPHA4 are a ligand/receptor pair that promotes neuronal proliferation and survival (Furne et al., 2009; Takemoto et al., 2002). MAP2K4 phosphorylates and activates MAPK13 (O’Callaghan et al., 2014); MAPK13 and IL32 are linked to IL-1 signaling (Netea et al., 2005; Yousif et al., 2013), which also is basal-specific in our screen. Basal A cells are preferentially susceptible to *CAND1*-*NEDD8* depletion. A NEDD8 inhibitor, MLN4924, is in phase 1 trials (NCT00677170, NCT01862328); TNBC patients might benefit from this agent.

Basal B cell lines are claudin-low-like, represent a unique TNBC subset, and have EMT-like, cancer stem cell-like, and mammary stem cell-like gene signatures (Lim et al., 2010), like those seen in chemotherapy-resistant cells (Creighton et al., 2009). Basal B lines also showed unique essentialities: basal B-essentials are enriched for motility, immune-related, developmental and neuronal, and cell junction and adhesion genes, several of which validate in siRNA experiments (Figure 3B). We also find marked functional similarity between basal breast cancer and HGSC (Figure S3C). Our results and the shared genomics of these tumors (Cancer Genome Atlas Research Network, 2011; Cancer Genome Atlas Network, 2012) argue for similar treatment strategies and drug discovery efforts.

Consistent with earlier work (Davoli et al., 2013; Solimini et al., 2012), our results suggest that for many amplicons, multiple genes contribute to increased fitness. For some amplicons, no clear *cis*-essential gene was identified. Failure to identify such genes might be technical (e.g., insufficient amplicon$^+$ lines). More likely, these amplicons select for multiple weak drivers, miRNAs/long non-coding RNAs (lncRNAs), or genes dispensable for proliferation/survival, but mediating other cancer hallmarks. For other amplicons, the key gene(s) cannot be targeted directly, nor can deleted tumor suppressor genes be restored. ‘‘*trans*-Essentials’’ provide insight into pathways perturbed by CNAs and can suggest more tractable drug targets. For example, METABRIC region 14, containing *MYC*, confers dependency on a MYC-regulated functional network. Two genes in this network, *MINK1* and *USP5*, are potential drug targets and validate by siRNA. Potentially druggable *trans*-associations also exist for common deletions: e.g., *RB1*-deleted lines are more sensitive to *MAP2K2* depletion, whereas *CDKN2A*-deleted lines rely more on *KAT6B*, *ADRBK1*, *SYK*, and *DNMT3A.*

As expected, genes encoding targets of known anti-cancer drugs are more essential in lines sensitive to those agents. But other genes, without known or obvious connections to the target pathway, also show essentiality strongly correlated with specific drug sensitivity. Also, gene essentiality can anti-correlate with drug sensitivity. Such genes might mediate therapy resistance and suggest potential combination strategies.

BRD4 was implicated in cancer by studies of NUT midline carcinoma, which often harbors a *BRD4-NUT* translocation (French et al., 2003). Subsequently, BRD4 emerged as a potential target for many other neoplasms (Shi and Vakoc, 2014). We identified *BRD4* as more essential in luminal/HER2 lines (Figures 7A and 7B; Table S3G). In hematologic malignancies, BET-I sensitivity

Figure 7. *BRD4* Is Luminal-Essential, and *PIK3CA* Mutations Cause BET-I Resistance

(A) Box plot showing *BRD4* dropout in each line, by subtype.

(B) *BRD4* siRNAs confirm pooled screen results. Averages are maximum percent inhibition ($p = 0.005$, Student’s t test).

(C) Effect of JQ1 on breast cancer lines. Table (inset) shows number of lines, grouped by JQ1 sensitivity (NS, non-sensitive; S, sensitive) and *PIK3CA* status (mut, mutated; WT, wild-type). Red shading shows lines with *PIK3CA* mutations. Mutant lines were more likely to be JQ1-resistant ($p < 4.7 \times 10^{-4}$, chi-square test). Sensitive lines have GI50 < 750 nM. *Lines with *PTEN* mutation/homozygous deletion.

(D) WT or mutant *PIK3CA* (H1047R) renders JQ1-sensitive SkBr3 line resistant to JQ1. Inset: immunoblot showing expression of PIK3CA-p110α. Arrow indicates the specific band.

(E) JQ1 cooperates with PIK3CA (A66; 1 μM), but not with PIK3CB (TGX; 1), inhibitors to decrease MCF7 and T47D proliferation. ‘‘0’’ JQ1 represents A66 or TGX alone.

(F) JQ1 cooperates with mTOR inhibitors (rapamycin; 0.5 nM, Torin; 50 nM) to decrease MCF7 proliferation. ‘‘0’’ represents rapamycin or Torin alone.

(G) JQ1 and Everolimus cooperatively inhibit xenograft growth. MCF7 cells (2×10^6) were injected into mammary fat pads of athymic nude mice bearing a slow release estrogen pellet. When tumors were 5 × 5 mm ($\sim$21 days), mice were grouped into: (1) control, (2) JQ1 (50 mg/kg/day intraperitoneally [IP]), (3) Everolimus (5 mg/kg/day by gavage), and (4) JQ1 + Everolimus daily. Tumors were measured with calipers every 3–4 days. P value, one-sided Student’s t test.

See also Figure S7.

correlates with *MYC* downregulation and is antagonized by exogenous *MYC* expression (Shi and Vakoc, 2014). Very recently, mouse basal-like breast tumors caused by *MYC* overexpression and mutant *PIK3CA* were found to be sensitive to combined BET/PI3K inhibition, as was a human basal line, SUM159 (Stratikopoulos et al., 2015). However, we saw no correlation between JQ1 sensitivity and basal *MYC* levels or the ability of JQ1 to inhibit *MYC* expression. Nor does forced *MYC* expression alter JQ1 sensitivity (Figures S7G–S7I).

Instead, using our genomic data, we found that *PIK3CA* mutations are biomarkers of BET-I resistance. Moreover, they are functional biomarkers, as treating cell lines or xenografts with a BET-I/mTOR inhibitor combination improves efficacy (Figures 7F and 7G). Our results have clear clinical implications, as Everolimus is approved for ER$^+$ breast cancer, and BET-Is are in clinical trials. *PIK3CA* mutations are most frequent in luminal tumors, so such patients would likely benefit most from BET-I/mTOR-I combinations. But our results and those of Stratikopoulos et al. (2015) also suggest a role for BET-Is as single agents in basal tumors. Surprisingly, and for unclear reasons, in basal lines, *PTEN* mutation/homozygous deletion predicts BET-I sensitivity (Figure 7C; data not shown).

Finally, as breast cancer lines can be JQ1-insensitive, but *BRD4*-dependent, BRD4 must have (a) BRD-independent function(s). Although the detailed mechanism is unclear, mutant *PIK3CA* confers JQ1-resistance, so PI3K pathway activation can selectively abrogate BRD-dependent, but not BRD-independent functions of BRD4. Thus, our integrated functional genomic approach not only can suggest new treatment strategies for breast tumor subtypes, but also reveals new features of breast cancer biology.

EXPERIMENTAL PROCEDURES

For additional details and computational methods, see Supplemental Experimental Procedures.

Cell Lines

Cell lines were from the American Type Culture Collection (ATCC), Asterands, Deutsche Sammlung von Mikroorganismen und Zellkultrenen GmbH (DSMZ), or were available in-house (Table S1).

Genomics/Proteomics

SNP-Arrays

Genomic DNA was amplified with the Illumina Infinium Genotyping kit, hybridized to Human Omni-Quad Beadchips, and analyzed on an iScan (Illumina). Data were quantified in GenomeStudio Version 2010.2 (Illumina) using Omni-Quad Multiuse_H manifest (April 2011 release), containing data from Genome-Build 37, Hg19.

RNA-Seq

RNA was reverse transcribed using the Illumina TruSeq Stranded mRNA kit. Libraries were sized (Agilent Bioanalyzer), normalized, and pooled (six each), and loaded onto an Illumina cBot. Paired-end sequencing (50 cycles) was performed on an Illumina HiSeq 2000.

Targeted Sequencing

DNA for 126 genes (1.264 Mbp) mutated at $\geq$3% frequency in breast or ovarian carcinoma was captured using Agilent SureSelect XT, loaded onto the cBot, and subjected to paired-end sequencing (100 cycles).

miRNA

miRNA expression was assessed by using the nCounter Human V2 miRNA Assay Kit (Cat# GXA-MIR2-48) and a NanoString counter.

RPPA

RPPAs were generated and analyzed as described (Tibes et al., 2006). For all lines, fresh media was added at 80%–90% confluency, and cells were harvested 16 hr later.

shRNA/siRNA Experiments

Pooled screens with the TRC-II library were performed as described (Marcotte et al., 2012). HCC712, ZR-75-30, MDA-MB-175VII, UACC812, and UACC3199 failed quality control. For validation, cells (1,000–3,000) seeded in 96-well plates for 24 hr were transfected with Dharmacon SMARTPOOL siRNAs (10 nM) using Lipofectamine RNAimax (Life Technologies). After 7 days, cells were stained with Alamar blue (Life Technologies), which measures redox activity and is as a surrogate for cell number. Percent maximum inhibition, corrected for transfection efficiency, was determined using siRNAs for the general essential *RPL9*.

Xenografts

MCF7 cells (5×10^{6}) were mixed 1:1 with growth factor-reduced Matrigel (BD Biosciences) and injected into mammary fat pads of athymic nude mice (Charles River). When tumors were 5 × 5 mm, mice were separated into control and drug-treated groups. JQ1 was synthesized (Filippakopoulos et al., 2010). Everolimus was purchased from Selleckchem.

RNA-seq and screen data are deposited in Gene Expression Omnibus (GEO: GSE73526 and GEO: GSE74702). Genomics and proteomics data are available at http://neellab.github.io/bfg/. All code is available upon request from A.S. and siMEM code will be posted at http://neellab.github.io/simem/.

All animal studies were approved by the University Health Network Animal Care Committee, under Animal Use Protocol (#1239).

ACCESSION NUMBERS

The accession numbers for the RNA-seq and screen data reported in this paper are Gene Expression Omnibus (GEO): GSE73526 and GSE74702.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.11.062.

AUTHOR CONTRIBUTIONS

R.M., A.S., J.M., and B.G.N. designed the study. R.M. and M.H. performed experiments. G.B.M. performed RPPAs. A.S. designed/implemented siMEM with input from J.M. and K.R.B. K.R.B., C.V., R.M., and A.S. performed genomic and statistical analyses. J.R. and G.D.B. performed pathway enrichment and PPI analyses. F.S.G. and D.P. implemented HELIOS. J.B. provided JQ1. R.M., A.S., and B.G.N. wrote the paper with help from all authors.

ACKNOWLEDGMENTS

This work was supported by the Canadian Foundation for Innovation (J.M.) and the Ontario Research Fund (B.G.N. and J.M.). B.G.N. was a Canada Research Chair, Tier 1, and work in his laboratory was supported in part by the Princess Margaret Cancer Foundation. R.M. was partly funded by a postdoctoral fellowship from the Canadian Breast Cancer Foundation. This work was also supported by NIH grants R37 CA49132 (B.G.N.), P41 GM103504, R01 GM070743, U41 HG006623 (G.D.B.), and PO1 CA099031 (G.B.M.), a Komen SAC and Promise grant (G.B.M.), and the M.D. Anderson CCSG functional proteomics core (CA16672).

Received: July 27, 2015
Revised: October 9, 2015
Accepted: November 23, 2015
Published: January 14, 2016

REFERENCES

Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K.K., Carter, S.L., Frederick, A.M., Lawrence, M.S., Sivachenko, A.Y., Sougnez, C., Zou, L., et al. (2012). Sequence analysis of mutations and translocations across breast cancer subtypes. Nature *486*, 405–409.

Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature *462*, 108–112.

Bosher, J.M., Williams, T., and Hurst, H.C. (1995). The developmentally regulated transcription factor AP-2 is involved in c-erbB-2 overexpression in human mammary carcinoma. Proc. Natl. Acad. Sci. USA *92*, 744–747.

Buchwalter, G., Hickey, M.M., Cromer, A., Selfors, L.M., Gunawardane, R.N., Frishman, J., Jeselsohn, R., Lim, E., Chi, D., Fu, X., et al. (2013). PDEF promotes luminal differentiation and acts as a survival factor for ER-positive breast cancer cells. Cancer Cell *23*, 753–767.

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70.

Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609–615.

Cheung, H.W., Cowley, G.S., Weir, B.A., Boehm, J.S., Rusin, S., Scott, J.A., East, A., Ali, L.D., Lizotte, P.H., Wong, T.C., et al. (2011). Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proc. Natl. Acad. Sci. USA *108*, 12372–12377.

Cortez, D., Wang, Y., Qin, J., and Elledge, S.J. (1999). Requirement of ATM-dependent phosphorylation of brca1 in the DNA damage response to double-strand breaks. Science *286*, 1162–1166.

Creighton, C.J., Li, X., Landis, M., Dixon, J.M., Neumeister, V.M., Sjolund, A., Rimm, D.L., Wong, H., Rodriguez, A., Herschkowitz, J.I., et al. (2009). Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. Proc. Natl. Acad. Sci. USA *106*, 13820–13825.

Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al.; METABRIC Group (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346–352.

Daemen, A., Griffith, O.L., Heiser, L.M., Wang, N.J., Enache, O.M., Sanborn, Z., Pepin, F., Durinck, S., Korkola, J.E., Griffith, M., et al. (2013). Modeling precision treatment of breast cancer. Genome Biol. *14*, R110.

Dang, C.V. (2012). MYC on the path to cancer. Cell *149*, 22–35.

Davoli, T., Xu, A.W., Mengwasser, K.E., Sack, L.M., Yoon, J.C., Park, P.J., and Elledge, S.J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell *155*, 948–962.

Dhillon, S. (2015). Palbociclib: first global approval. Drugs *75*, 543–551.

Dou, Q.P., and Zonder, J.A. (2014). Overview of proteasome inhibitor-based anti-cancer therapies: perspective on bortezomib and second generation proteasome inhibitors versus future generation inhibitors of ubiquitin-proteasome system. Curr. Cancer Drug Targets *14*, 517–536.

Dvinge, H., Git, A., Gräf, S., Salmon-Divon, M., Curtis, C., Sottoriva, A., Zhao, Y., Hirst, M., Armisen, J., Miska, E.A., et al. (2013). The shaping and functional consequences of the microRNA landscape in breast cancer. Nature *497*, 378–382.

Ellis, M.J., Ding, L., Shen, D., Luo, J., Suman, V.J., Wallis, J.W., Van Tine, B.A., Hoog, J., Goiffon, R.J., Goldstein, T.C., et al. (2012). Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature *486*, 353–360.

Filippakopoulos, P., Qi, J., Picaud, S., Shen, Y., Smith, W.B., Fedorov, O., Morse, E.M., Keates, T., Hickman, T.T., Felletar, I., et al. (2010). Selective inhibition of BET bromodomains. Nature *468*, 1067–1073.

Flick, K., and Kaiser, P. (2013). Set them free: F-box protein exchange by Cand1. Cell Res. *23*, 870–871.

French, C.A., Miyoshi, I., Kubonishi, I., Grier, H.E., Perez-Atayde, A.R., and Fletcher, J.A. (2003). BRD4-NUT fusion oncogene: a novel mechanism in aggressive carcinoma. Cancer Res. *63*, 304–307.

Furne, C., Ricard, J., Cabrera, J.R., Pays, L., Bethea, J.R., Mehlen, P., and Liebl, D.J. (2009). EphrinB3 is an anti-apoptotic ligand that inhibits the dependence receptor functions of EphA4 receptors during adult neurogenesis. Biochim. Biophys. Acta *1793*, 231–238.

Gatei, M., Scott, S.P., Filippovitch, I., Soronika, N., Lavin, M.F., Weber, B., and Khanna, K.K. (2000). Role for ATM in DNA damage-induced phosphorylation of BRCA1. Cancer Res. *60*, 3299–3304.

Griffith, M., Griffith, O.L., Coffman, A.C., Weible, J.V., McMichael, J.F., Spies, N.C., Koval, J., Das, I., Callaway, M.B., Eldred, J.M., et al. (2013). DGIdb: mining the druggable genome. Nat. Methods *10*, 1209–1210.

Guo, W., Keckesova, Z., Donaher, J.L., Shibue, T., Tischler, V., Reinhardt, F., Itzkovitz, S., Noske, A., Zürrer-Härdi, U., Bell, G., et al. (2012a). Slug and Sox9 cooperatively determine the mammary stem cell state. Cell *148*, 1015–1028.

Guo, Y., Chinyengetere, F., Dolinko, A.V., Lopez-Aguiar, A., Lu, Y., Galimberti, F., Ma, T., Feng, Q., Sekula, D., Freemantle, S.J., et al. (2012b). Evidence for the ubiquitin protease UBP43 as an antineoplastic target. Mol. Cancer Ther. *11*, 1968–1977.

Hart, T., Brown, K.R., Sircoulomb, F., Rottapel, R., and Moffat, J. (2014). Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. Mol. Syst. Biol. *10*, 733.

Hennessy, B.T., Gonzalez-Angulo, A.M., Stemke-Hale, K., Gilcrease, M.Z., Krishnamurthy, S., Lee, J.S., Fridlyand, J., Sahin, A., Agarwal, R., Joy, C., et al. (2009). Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics. Cancer Res. *69*, 4116–4124.

Hollestelle, A., Nagel, J.H., Smid, M., Lam, S., Elstrodt, F., Wasielewski, M., Ng, S.S., French, P.J., Peeters, J.K., Rozendaal, M.J., et al. (2010). Distinct gene mutation profiles among luminal-type and basal-type breast cancer cell lines. Breast Cancer Res. Treat. *121*, 53–64.

Hsu, T.Y., Simon, L.M., Neill, N.J., Marcotte, R., Sayad, A., Bland, C.S., Echeverria, G.V., Sun, T., Kurley, S.J., Tyagi, S., et al. (2015). The spliceosome is a therapeutic vulnerability in MYC-driven cancer. Nature *525*, 384–388.

Jamdade, V.S., Sethi, N., Mundhe, N.A., Kumar, P., Lahkar, M., and Sinha, N. (2015). Therapeutic targets of triple-negative breast cancer: a review. Br. J. Pharmacol. *172*, 4228–4237.

Kao, J., Salari, K., Bocanegra, M., Choi, Y.L., Girard, L., Gandhi, J., Kwei, K.A., Hernandez-Boussard, T., Wang, P., Gazdar, A.F., et al. (2009). Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. PLoS ONE *4*, e6146.

Kim, J.H., Hubbard, N.E., Ziboh, V., and Erickson, K.L. (2005). Attenuation of breast tumor cell growth by conjugated linoleic acid via inhibition of 5-lipoxygenase activating protein. Biochim. Biophys. Acta *1736*, 244–250.

König, R., Chiang, C.Y., Tu, B.P., Yan, S.F., DeJesus, P.D., Romero, A., Bergauer, T., Orth, A., Krueger, U., Zhou, Y., and Chanda, S.K. (2007). A probability-based approach for the analysis of large-scale RNAi screens. Nat. Methods *4*, 847–849.

Lehmann, B.D., Bauer, J.A., Chen, X., Sanders, M.E., Chakravarthy, A.B., Shyr, Y., and Pietenpol, J.A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J. Clin. Invest. *121*, 2750–2767.

Lim, E., Wu, D., Pal, B., Bouras, T., Asselin-Labat, M.L., Vaillant, F., Yagita, H., Lindeman, G.J., Smyth, G.K., and Visvader, J.E. (2010). Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. Breast Cancer Res. *12*, R21.

Lupien, M., Eeckhoute, J., Meyer, C.A., Wang, Q., Zhang, Y., Li, W., Carroll, J.S., Liu, X.S., and Brown, M. (2008). FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell *132*, 958–970.

Maire, V., Némati, F., Richardson, M., Vincent-Salomon, A., Tesson, B., Rigaill, G., Gravier, E., Marty-Prouvost, B., De Koning, L., Lang, G., et al. (2013). Polo-like kinase 1: a potential therapeutic option in combination with conventional chemotherapy for the management of patients with triple-negative breast cancer. Cancer Res. *73*, 813–823.

Marcotte, R., Brown, K.R., Suarez, F., Sayad, A., Karamboulas, K., Krzyzanowski, P.M., Sircoulomb, F., Medrano, M., Fedyshyn, Y., Koh, J.L., et al. (2012). Essential gene profiles in breast, pancreatic, and ovarian cancer cells. Cancer Discov. *2*, 172–189.

Muranen, T., Selfors, L.M., Worster, D.T., Iwanicki, M.P., Song, L., Morales, F.C., Gao, S., Mills, G.B., and Brugge, J.S. (2012). Inhibition of PI3K/mTOR leads to adaptive resistance in matrix-attached cancer cells. Cancer Cell *21*, 227–239.

Netea, M.G., Azam, T., Ferwerda, G., Girardin, S.E., Walsh, M., Park, J.S., Abraham, E., Kim, J.M., Yoon, D.Y., Dinarello, C.A., and Kim, S.H. (2005). IL-32 synergizes with nucleotide oligomerization domain (NOD) 1 and NOD2 ligands for IL-1beta and IL-6 production through a caspase 1-dependent mechanism. Proc. Natl. Acad. Sci. USA *102*, 16309–16314.

Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.P., Tong, F., et al. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell *10*, 515–527.

Nijhawan, D., Zack, T.I., Ren, Y., Strickland, M.R., Lamothe, R., Schumacher, S.E., Tsherniak, A., Besche, H.C., Rosenbluh, J., Shehata, S., et al. (2012). Cancer vulnerabilities unveiled by genomic loss. Cell *150*, 842–854.

O'Callaghan, C., Fanning, L.J., and Barry, O.P. (2014). p38δ MAPK: Emerging Roles of a Neglected Isoform. Int. J. Cell Biol. *2014*, 272689.

Pan, D., and Lin, X. (2013). Epithelial growth factor receptor-activated nuclear factor kappaB signaling and its role in epithelial growth factor receptor-associated tumors. Cancer J. *19*, 461–467.

Pao, G.M., Janknecht, R., Ruffner, H., Hunter, T., and Verma, I.M. (2000). CBP/p300 interact with and function as transcriptional coactivators of BRCA1. Proc. Natl. Acad. Sci. USA *97*, 1020–1025.

Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. *27*, 1160–1167.

Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747–752.

Petrocca, F., Altschuler, G., Tan, S.M., Mendillo, M.L., Yan, H., Jerry, D.J., Kung, A.L., Hide, W., Ince, T.A., and Lieberman, J. (2013). A genome-wide siRNA screen identifies proteasome addiction as a vulnerability of basal-like triple-negative breast cancer cells. Cancer Cell *24*, 182–196.

Prat, A., Parker, J.S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J.I., He, X., and Perou, C.M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res. *12*, R68.

Rahmani, M., Aust, M.M., Attkisson, E., Williams, D.C., Jr., Ferreira-Gonzalez, A., and Grant, S. (2013). Dual inhibition of Bcl-2 and Bcl-xL strikingly enhances PI3K inhibition-induced apoptosis in human myeloid leukemia cells through a GSK3- and Bim-dependent mechanism. Cancer Res. *73*, 1340–1351.

Riaz, M., van Jaarsveld, M.T., Hollestelle, A., Prager-van der Smissen, W.J., Heine, A.A., Boersma, A.W., Liu, J., Helmijr, J., Ozturk, B., Smid, M., et al. (2013). miRNA expression profiling of 51 human breast cancer cell lines reveals subtype and driver mutation-specific miRNAs. Breast Cancer Res. *15*, R33.

Sanchez-Garcia, F., Villagrasa, P., Matsui, J., Kotliar, D., Castro, V., Akavia, U.D., Chen, B.J., Saucedo-Cuevas, L., Rodriguez Barrueco, R., Llobet-Navas, D., et al. (2014). Integration of genomic data enables selective discovery of breast cancer drivers. Cell *159*, 1461–1475.

Shah, S.P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., et al. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature *486*, 395–399.

Shao, D.D., Tsherniak, A., Gopal, S., Weir, B.A., Tamayo, P., Stransky, N., Schumacher, S.E., Zack, T.I., Beroukhim, R., Garraway, L.A., et al. (2013). ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens. Genome Res. *23*, 665–678.

Shi, J., and Vakoc, C.R. (2014). The mechanisms behind the therapeutic activity of BET bromodomain inhibition. Mol. Cell *54*, 728–736.

Solimini, N.L., Xu, Q., Mermel, C.H., Liang, A.C., Schlabach, M.R., Luo, J., Burrows, A.E., Anselmo, A.N., Bredemeyer, A.L., Li, M.Z., et al. (2012). Recurrent hemizygous deletions in cancers may optimize proliferative potential. Science *337*, 104–109.

Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA *98*, 10869–10874.

Stagni, V., Manni, I., Oropallo, V., Mottolese, M., Di Benedetto, A., Piaggio, G., Falcioni, R., Giaccari, D., Di Carlo, S., Sperati, F., et al. (2015). ATM kinase sustains HER2 tumorigenicity in breast cancer. Nat. Commun. *6*, 6886.

Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R., et al.; Oslo Breast Cancer Consortium (OSBREAC) (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature *486*, 400–404.

Stratikopoulos, E.E., Dendy, M., Szabolcs, M., Khaykin, A.J., Lefebvre, C., Zhou, M.M., and Parsons, R. (2015). Kinase and BET Inhibitors Together Clamp Inhibition of PI3K Signaling and Overcome Resistance to Therapy. Cancer Cell *27*, 837–851.

Takemoto, M., Fukuda, T., Sonoda, R., Murakami, F., Tanaka, H., and Yamamoto, N. (2002). Ephrin-B3-EphA4 interactions regulate the growth of specific thalamocortical axon populations in vitro. Eur. J. Neurosci. *16*, 1168–1172.

Tibes, R., Qiu, Y., Lu, Y., Hennessy, B., Andreeff, M., Mills, G.B., and Kornblau, S.M. (2006). Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Mol. Cancer Ther. *5*, 2512–2521.

Timmerman, L.A., Holton, T., Yuneva, M., Louie, R.J., Padró, M., Daemen, A., Hu, M., Chan, D.A., Ethier, S.P., van 't Veer, L.J., et al. (2013). Glutamine sensitivity analysis identifies the xCT antiporter as a common triple-negative breast tumor therapeutic target. Cancer Cell *24*, 450–465.

Worby, C.A., and Dixon, J.E. (2014). Pten. Annu. Rev. Biochem. *83*, 641–669.

Yousif, N.G., Al-Amran, F.G., Hadi, N., Lee, J., and Adrienne, J. (2013). Expression of IL-32 modulates NF-κB and p38 MAP kinase pathways in human esophageal cancer. Cytokine *61*, 223–227.

Yu, F., Li, J., Chen, H., Fu, J., Ray, S., Huang, S., Zheng, H., and Ai, W. (2011). Kruppel-like factor 4 (KLF4) is required for maintenance of breast cancer stem cells and for cell migration and invasion. Oncogene *30*, 2161–2172.

Zender, L., Xue, W., Zuber, J., Semighini, C.P., Krasnitz, A., Ma, B., Zender, P., Kubicka, S., Luk, J.M., Schirmacher, P., et al. (2008). An oncogenomics-based in vivo RNAi screen identifies tumor suppressors in liver cancer. Cell *135*, 852–864.

Zhang, L., Zhang, Y., Mehta, A., Boufraqech, M., Davis, S., Wang, J., Tian, Z., Yu, Z., Boxer, M.B., Kiefer, J.A., et al. (2015). Dual inhibition of HDAC and EGFR signaling with CUDC-101 induces potent suppression of tumor growth and metastasis in anaplastic thyroid cancer. Oncotarget *6*, 9073–9085.

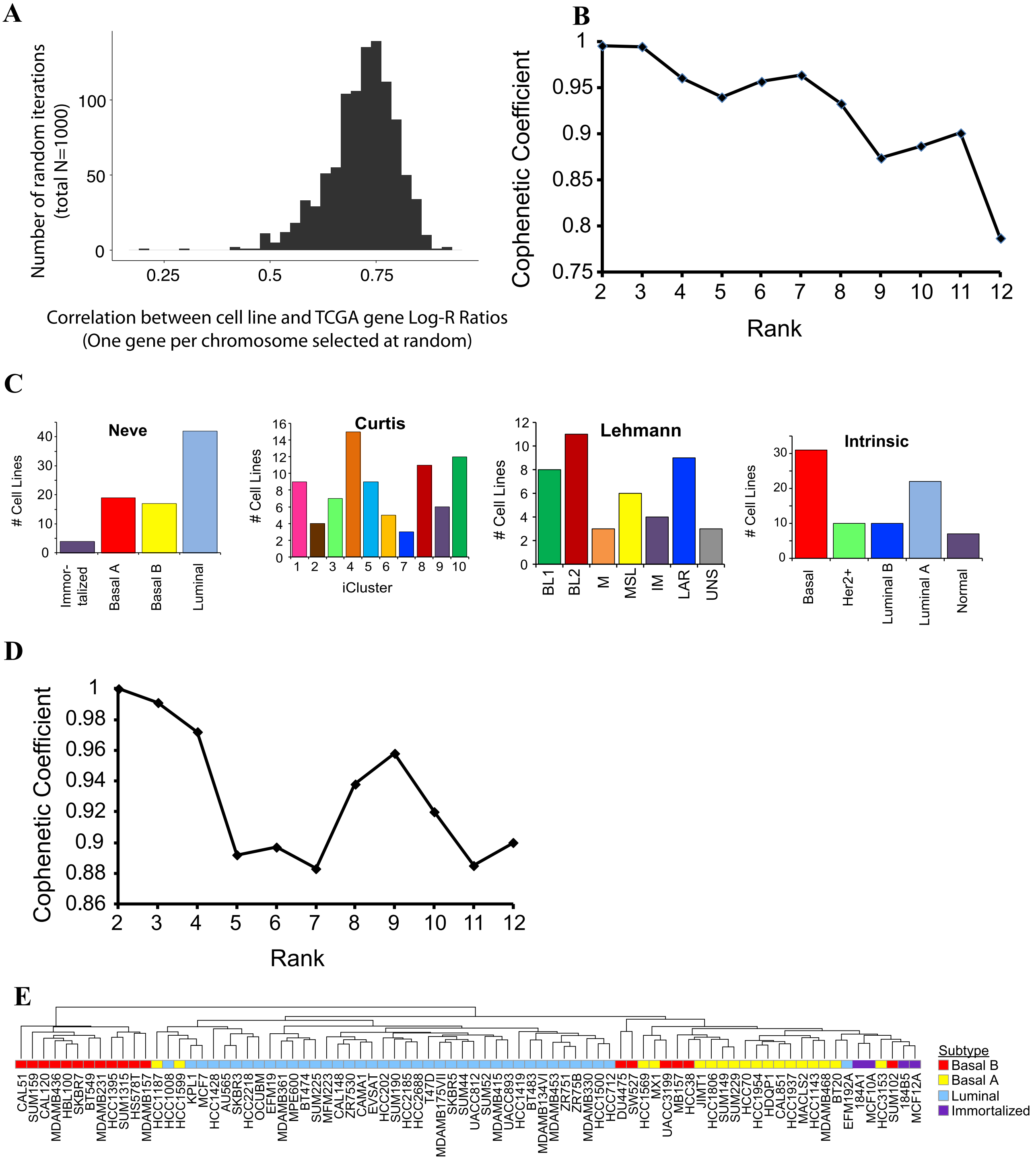


Figure S1. Breast Cancer Cell Lines Are Reasonable Surrogates for Breast Tumors, Related to Figure 1

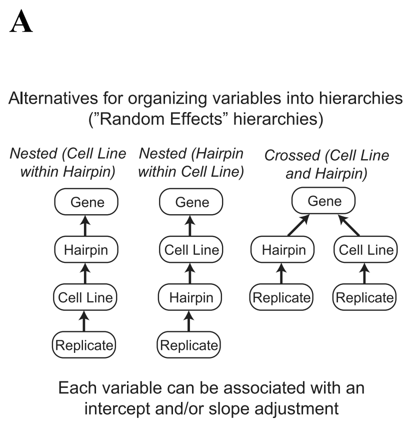
(A) Correlation of the individual gene Log-R ratios between our cell line panel and TCGA samples.

(B) Cophenetic correlation of the RNaseq NMF clustering.

(C) Breakdown of cell lines according to their PAM50, Neve, Lehmans, and Curtis classifications.

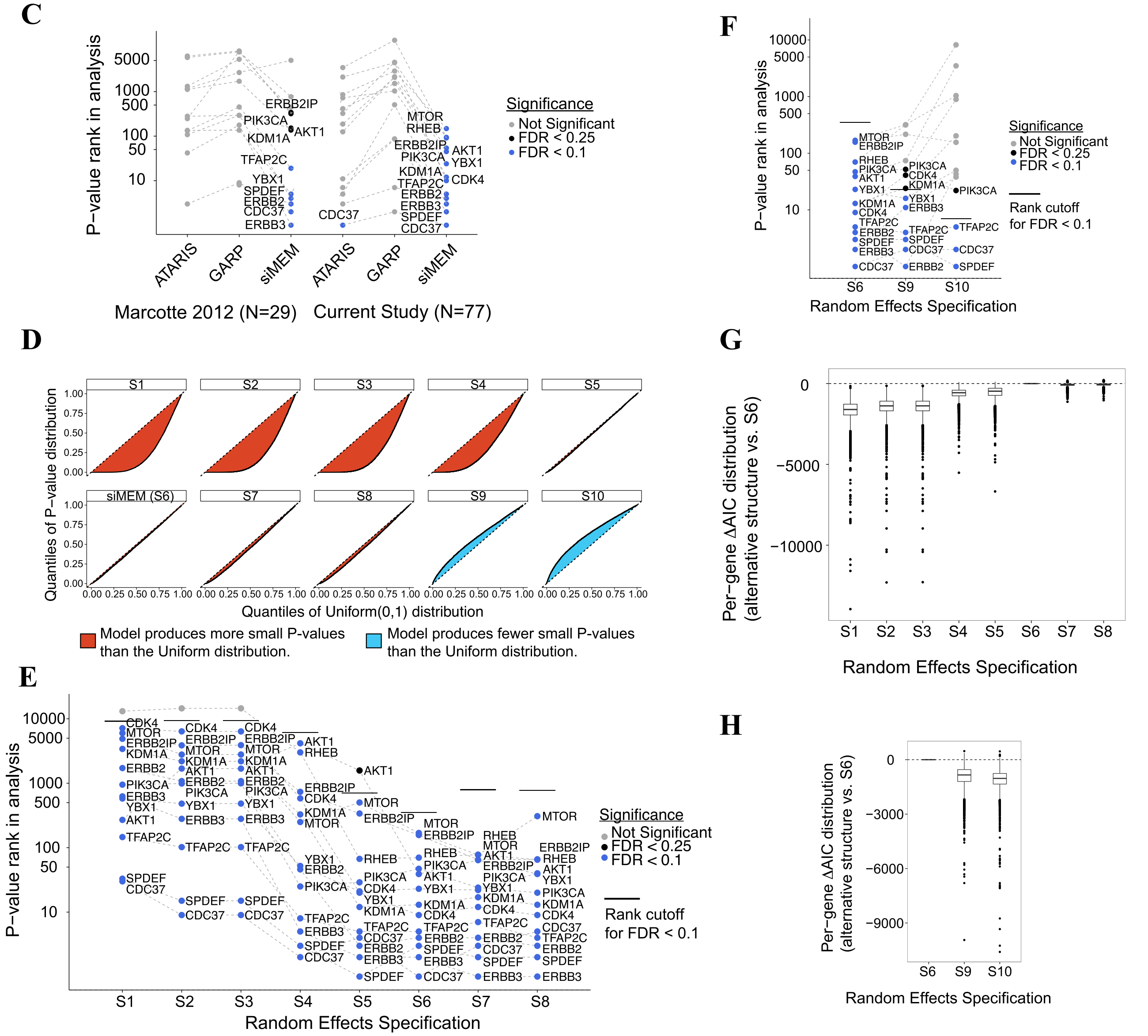
(D) Cophenetic correlation for the RPPA NMF clustering.

(E) Hierarchical clustering of miRNA expression data.


A
Alternatives for organizing variables into hierarchies
("Random Effects" hierarchies)
Nested (Cell Line within Hairpin)
Nested (Hairpin within Cell Line)
Crossed (Cell Line and Hairpin)
Gene
Hairpin
Cell Line
Replicate
Each variable can be associated with an intercept and/or slope adjustment

B

Structure Name	Relation Between Variables	Variables Included in Structure	Detailed Description	Compared to siMEM (S6)
S1	Single variable	Hairpin	Hairpin (random intercept)	Simpler
S2	Single variable	Hairpin	Hairpin (random slope)	Simpler
S3	Single variable	Hairpin	Hairpin (random intercept+slope)	Simpler
S4	Nested variables	Cell Line, Hairpin	Nested: Cell Line (random intercept), nested in Hairpin (random intercept+slope)	Simpler
S5	Nested variables	Cell Line, Hairpin	Nested: Cell Line (random slope), nested in Hairpin (random intercept+slope)	Simpler
S6 (siMEM)	Nested variables	Cell Line, Hairpin	Nested: Cell Line (random intercept+slope), nested in Hairpin (random intercept+slope)	Same
S7	Nested variables	Cell Line, Hairpin, Replicate	Nested: Replicate (random interept), nested in Cell Line (random intercept+slope), nested in Hairpin (random intercept+slope)	More complex
S8	Nested variables	Cell Line, Hairpin, Replicate	Nested: Replicate (random slope), nested in Cell Line (random intercept+slope), nested in Hairpin (random intercept+slope)	More complex
S9	Nested variables	Cell Line, Hairpin	Nested: Hairpin (random intercept+slope), nested in Cell Line (random intercept+slope)	Different – nested variables
S10	Crossed variables	Cell Line, Hairpin	Crossed Cell Line(random intercept+slope) and Hairpin (random intercept+slope)	Different – crossed variables



(legend on next page)

Figure S2. Systematic Assessment of Alternate Model Structures, Related to Figure 2

Testing of model fit, prediction of known positives, and prediction with randomized classes for the HER2+-associated analysis presented in Figure 2.

(A) Alternative model random effect hierarchies can be used to relate hairpin and cell line variables: nested (“cell line in hairpin”), nested (“hairpin in cell line”) or crossed (“hairpin x cell line”). Each of these approaches is plausible a priori, and associated with several model variants.

(B) Each variable (cell line, hairpin, replicate) can be represented by an intercept and/or slope adjustment (random effect), giving rise to numerous potential model structures, with a representative ten shown. S1-S5 are simpler variants of siMEM structure S6, whereas S7 and S8 are more complex. S9 and S10 use different variable hierarchies.

(C) Compared with ATARIS and GARP, only siMEM correctly assigns significance to known positives, both in the Marcotte 2012 (n = 29) and current (n = 77) screens, and also shifts positives to the top of the prediction list. ATARIS does not produce scores or rankings for nearly half the genes, thereby lowering its rankings of scored genes relative to GARP.

(D) Quantile-quantile plots comparing model p value and Uniform distributions from randomized class analysis. When predicting differences between randomly assigned classes, the distribution of p values produced by S1-S4 is skewed heavily toward small p values (below diagonal), whereas S5 and S6 (siMEM) produce distributions closest to the Uniform p value distribution. S9 and S10 produce p value distributions skewed toward large p values (above diagonal), and appear to be excessively conservative in their predictions. If samples are assigned randomly to classes without biological significance, regression analyses that fail to account for systematic measurement effects produce many significant predictions, whereas siMEM and other methods produce close to the expected random distribution of p values.

(E) Spurious predictions are greatly reduced, while ranking of known positives improves, between S1 and S5, thereafter remaining stable (S6-S8). siMEM structure S6 produces the best ranking, combined with the fewest number of significance predictions.

(F) S9 and S10 also predict fewer than ten significant genes each, miss most known positives, and produce worse p value rankings for positives, regardless of significance.

(G) Negative ΔAIC distributions show greatly improved model fits for S6 relative to simpler alternatives (S1-S5), while indicating no further improvement upon added model complexity (S7-S8).

(H) Alternative nested (S9) and crossed (S10) model structures produce substantially worse model fits.

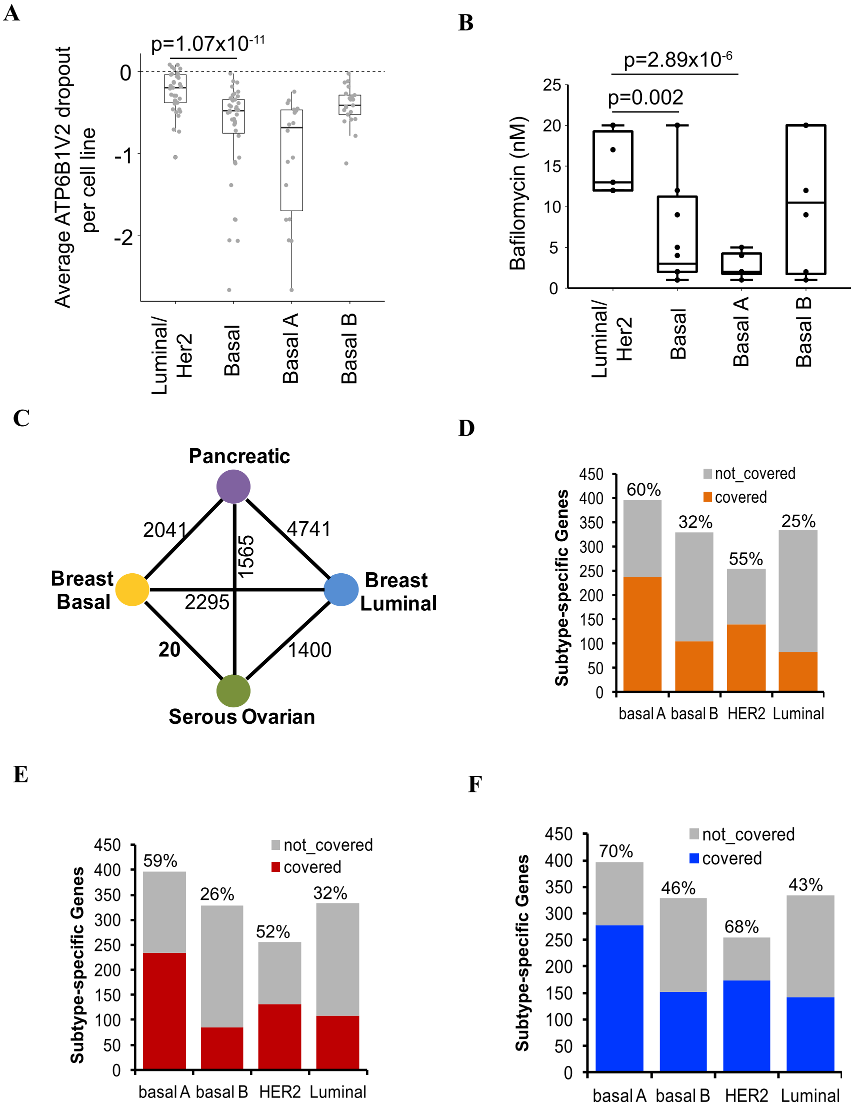


Figure S3. Subtype- and Tissue-Specific Essential Genes, Related to Figure 3

(A) Drop-out rate of *ATP6V1B2* in the 77 cell lines, broken down by subtype.

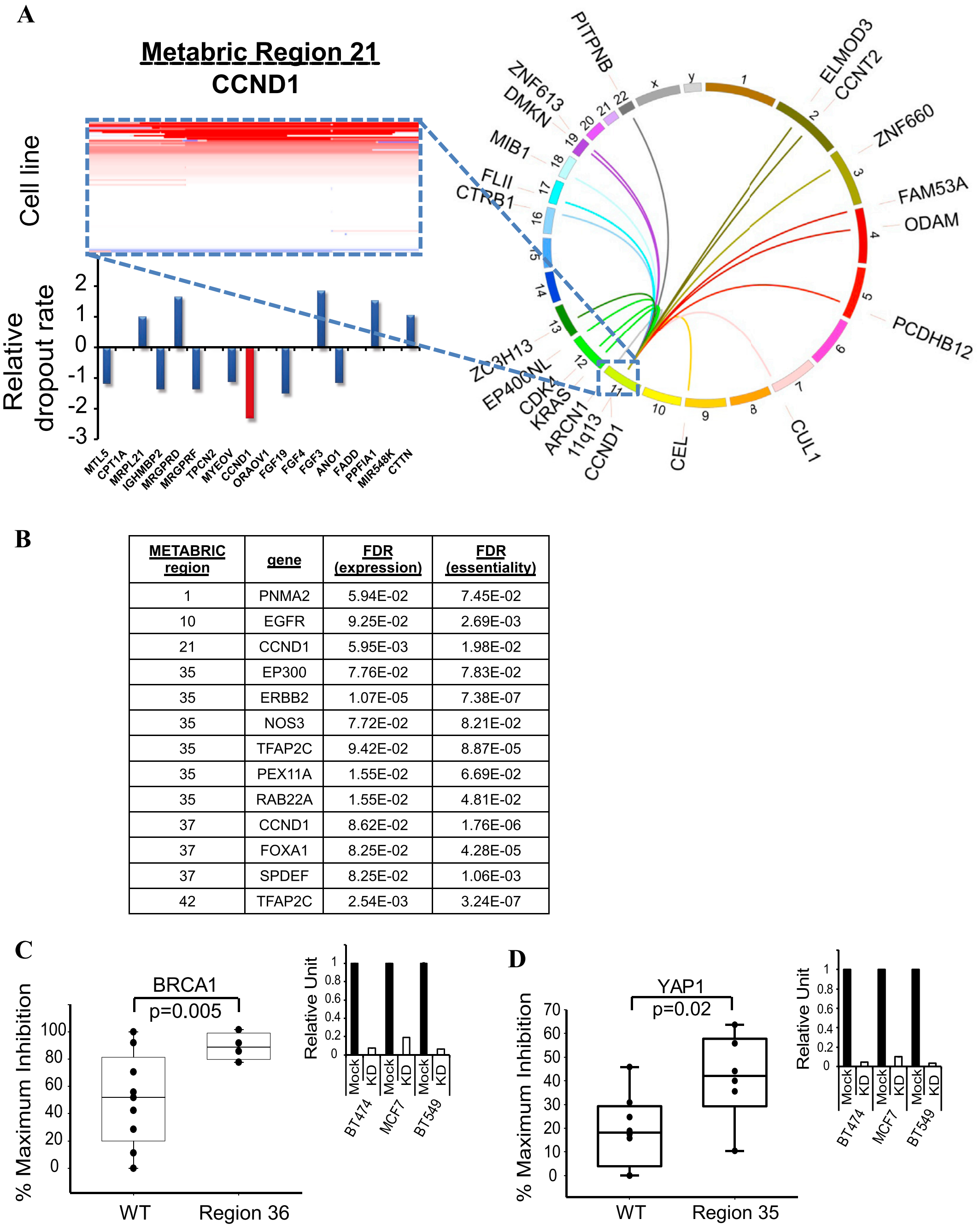
(B) Basal cell lines (n = 12), especially basal A lines, are more sensitive to Bafilomycin (10nM) than luminal cell lines (n = 8). p values are calculated by 1-sided Student's t test.

(C) Overlap of breast luminal-, breast basal-, pancreatic- and ovarian-specific genes, as determined by siMEM. Note that ovarian-specific genes are highly enriched for basal-specific genes.

(D) Number of genes that were annotated with a known function in Figure 3D.

(E) Number of genes that were annotated with a known function in Figure 3E.

(F) Total number of annotated genes from the pathway and PPI analyses.



METABRIC region	gene	FDR (expression)	FDR (essentiality)
1	PNMA2	5.94E-02	7.45E-02
10	EGFR	9.25E-02	2.69E-03
21	CCND1	5.95E-03	1.98E-02
35	EP300	7.76E-02	7.83E-02
35	ERBB2	1.07E-05	7.38E-07
35	NOS3	7.72E-02	8.21E-02
35	TFAP2C	9.42E-02	8.87E-05
35	PEX11A	1.55E-02	6.69E-02
35	RAB22A	1.55E-02	4.81E-02
37	CCND1	8.62E-02	1.76E-06
37	FOXA1	8.25E-02	4.28E-05
37	SPDEF	8.25E-02	1.06E-03
42	TFAP2C	2.54E-03	3.24E-07

Figure S4. Expression and Validation of *trans*-Essential Genes, Related to Figure 4

(A) IGV-generated heat map showing the amplification status of the 11q13 region (METABRIC region 21, which contains *CCND1*) in our panel of breast cancer lines. Red = amplification, blue = deletion. The bar graph below the heat map shows the average zGARP score for each gene in the amplified region in lines containing the amplified region. The CIRCOS plot depicts the top 20 statistically most essential genes, as determined by siMEM in 11q13-amplified versus non-amplified lines.

(B) List of genes that scored as *trans*-essential in the context of the indicated METABRIC region, and that were associated with increased expression in lines with containing that region.

(C) *BRCA1* is *trans*-essential for region 36. y axis; % maximum inhibition; p values were calculated by one-sided Student's t test.

(D) *YAP1* is *trans*-essential for region 35. y axis; % maximum inhibition; p values were calculated by one-sided Student's t test.

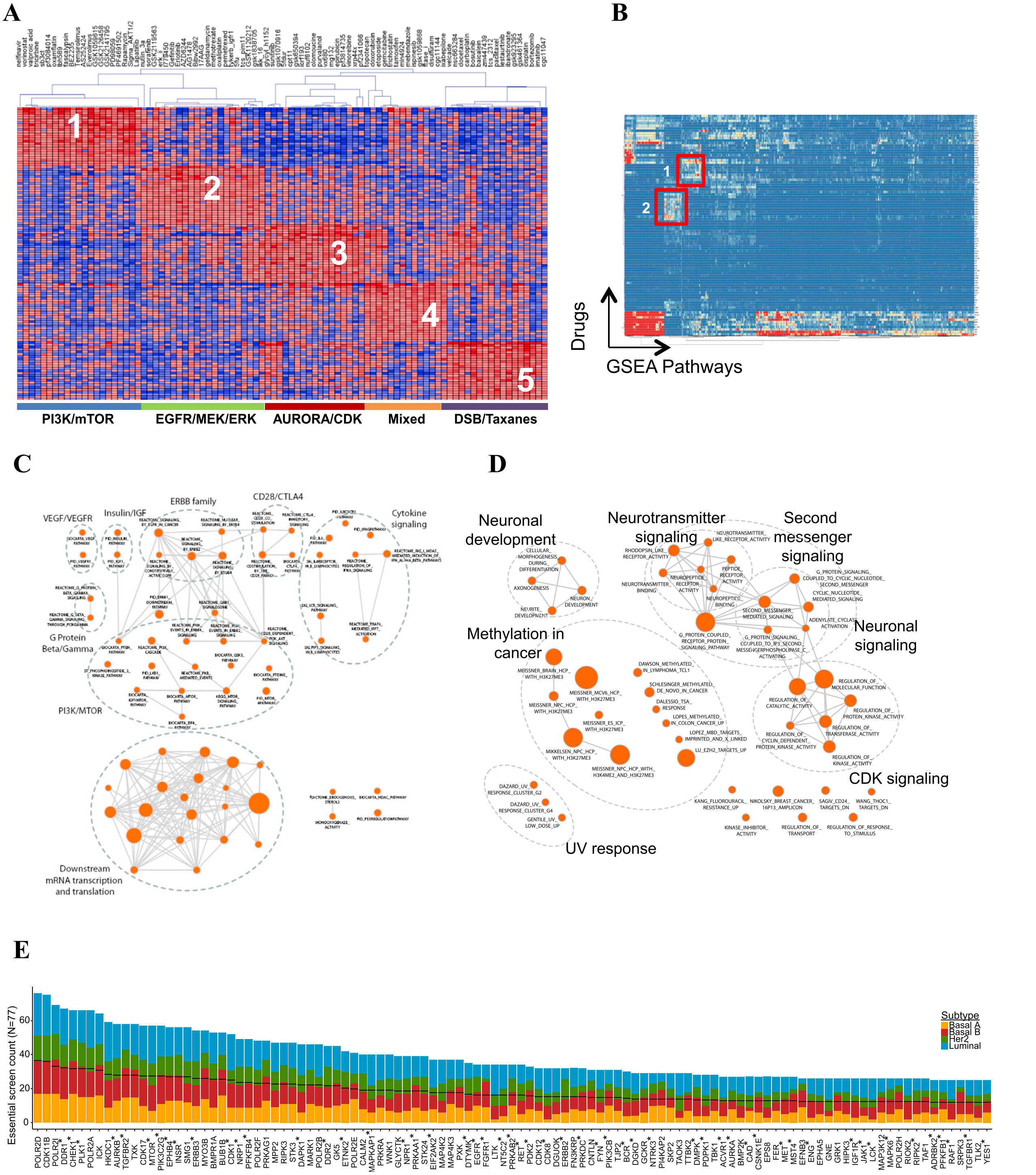


Figure S5. Integration of Screen Results with Drug Response, Related to Figure 5

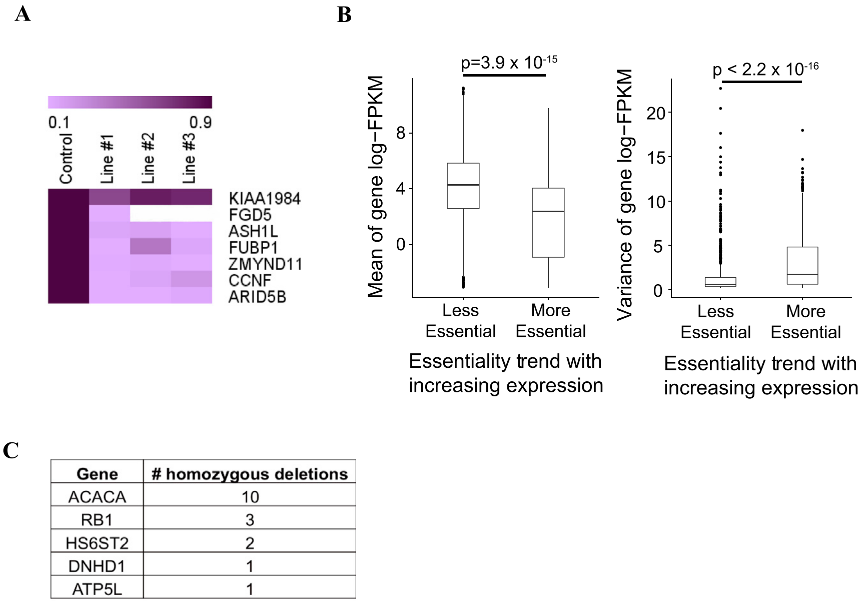
(A) NMF clustering of gene essentiality associated with published drug sensitivity data (Daemen et al., 2013).

(B) Heatmap of pathway enrichment for gene essentiality associated with each drug treatment. (1) indicates PI3K signaling-related pathways. (2) indicates neuronal-related pathways.

(C) Pathway enrichment analyses for essential genes associated with response to topotecan.

(D) Pathway enrichment analysis of essential genes associated with GSK2126458 sensitivity. Network image was generated by using Cytoscape.

(E) Top kinases essential according to our screen broken down by subtypes. *Reported gene-drug interaction in DGIdb. For each gene, the black bar represents 50% of the lines in which the gene is essential.



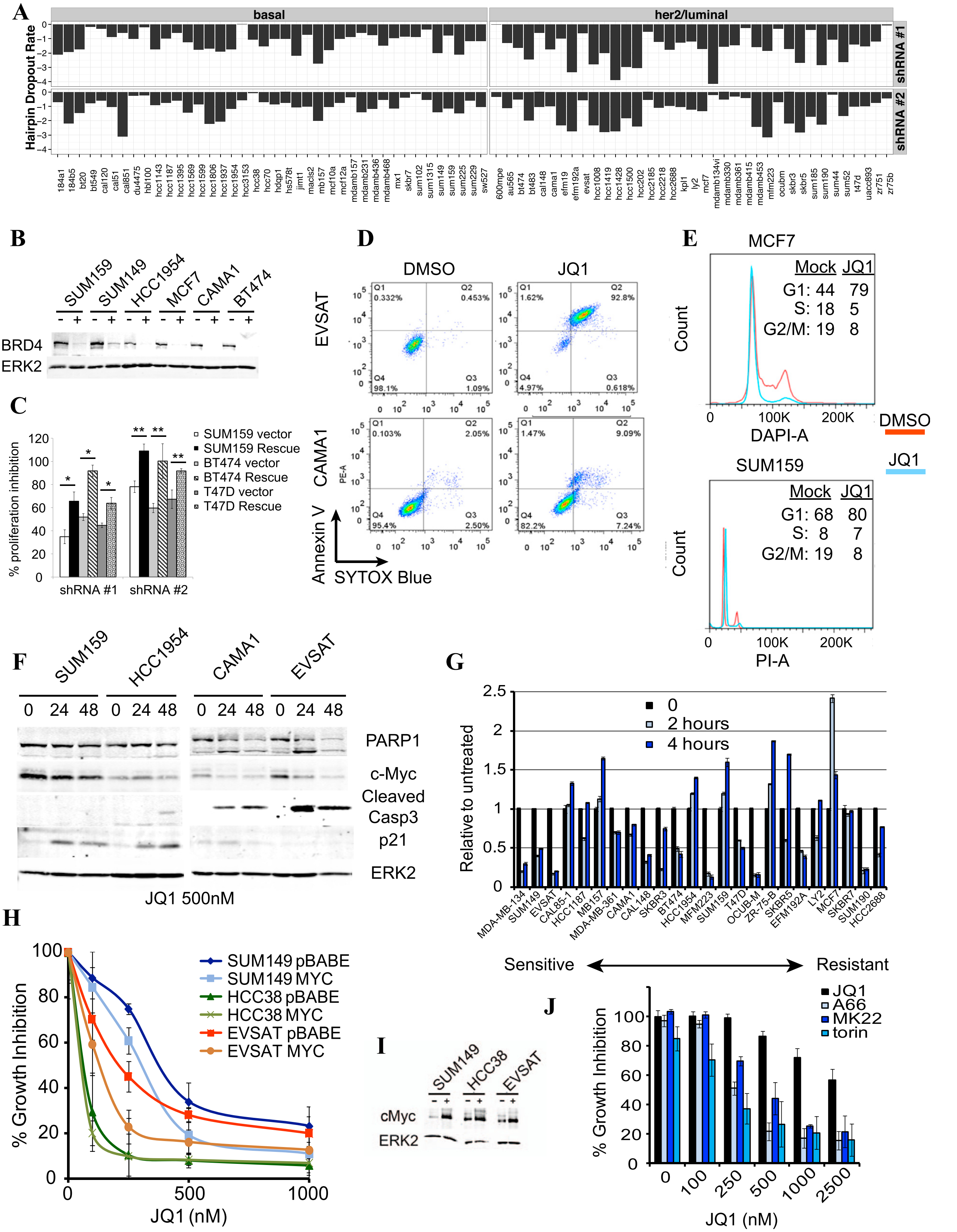
Gene	# homozygous deletions
ACACA	10
RB1	3
HS6ST2	2
DNHD1	1
ATP5L	1

Figure S6. Integration of Screen Results with Expression and Copy Number Status, Related to Figure 6

(A) Heat map depicting knockdown efficiency (by q-RT-PCR) of siRNAs targeting the subtype-specific genes in Figure 6B in representative cell lines. For others, see Figure 3C.

(B) Genes whose essentiality increases with increasing expression tend to have significantly lower and more variable expression patterns than genes whose essentiality decreases with increasing expression.

(C) CYCLOPS and STOP genes that show homozygous deletion in lines in the panel.



(legend on next page)

Figure S7. BRD4 Regulates Proliferation and Survival of Breast Cancer Lines in a MYC-Independent Manner, Related to Figure 7

(A) Dropout rates of the two best *BRD4* shRNAs across 77 breast cancer cell lines. Note that rate is significantly higher in Luminal/HER2 lines (right side of bar plot) and that the drop-out rate is highly correlated between the two shRNAs (i.e., lines sensitive to shRNA #1 are highly susceptible to shRNA #2). Also note that some basal lines are sensitive to *BRD4* depletion.

(B) Immunoblots showing BRD4 depletion following siRNA knockdowns in representative cell lines.

(C) *BRD4* shRNA-resistant cDNA rescues proliferation inhibition caused by *BRD4* depletion by the two shRNAs shown in (A). $^*p < 0.003$, $^{**}p < 0.02$ by Student's t test.

(D) Annexin V/Sytox Blue flow cytometry of JQ1-sensitive cells (500 nM JQ1 for 48 hr). Note that these cells undergo cell death.

(E) Cell-cycle profile of JQ1-resistant cell lines, assessed by DAPI or PI staining, as indicated. Note the G1 arrest in treated cells.

(F) Immunoblot confirms the flow cytometry data. Notice that p21, a cell-cycle inhibitor, is increased in JQ1-resistant cells and not in sensitive cells, whereas the JQ1-sensitive (but not the JQ1-resistant) line shows cleaved caspase-3.

(G) qRT-PCR for *MYC* mRNA following JQ1 treatment (500 nM).

(H) Percent proliferation inhibition of JQ1-sensitive cell lines expressing exogenous *MYC* cDNA.

(I) Immunoblot of exogenous MYC.

(J) JQ1 cooperates with p110α (A66; 1uM), AKT (MK22; 500 nM), and mTOR inhibitors (Torin; 50 nM) to decrease SUM159 cell proliferation.

Cell

Supplemental Information

Functional Genomic Landscape of Human Breast Cancer Drivers, Vulnerabilities, and Resistance

Richard Marcotte, Azin Sayad, Kevin R. Brown, Felix Sanchez-Garcia, Jüri Reimand, Maliha Haider,Carl Virtanen, James E. Bradner, Gary D. Bader, Gordon B. Mills, Dana Pe’er, Jason Moffat and Benjamin G. Neel

TABLE OF CONTENTS

siMEM

Model overview

The dropout behavior of a typical gene in our screen was measured in triplicate by 5 shRNAs (hairpins) at 3 time-points across the 77 cell lines that survived quality control (for a total of 3465 measurements per gene). The siMEM (si/shRNA Mixed-Effect Model) hierarchical linear model represents each short time course as a line with a specific intercept and slope. It adjusts for hairpin- and cell line-induced systematic measurement effects by representing assay measurements as the sum of several components:

$$\begin{aligned}
&\textit{overall intercept}\\
&+\textit{average line slope across all hairpins}\\
&+\textit{difference in slope associated with genomic variable}\\
&+\textit{hairpin intercept \& slope adjustments (}1\textit{ per hairpin)}\\
&+\textit{hairpin in cell line intercept \& slope adjustments (}1\textit{ per hairpin / cell line combination)}\\
&+\textit{random error components (}1\textit{ per hairpin / cell line combination)}
\end{aligned}$$

Several random effect components are associated with the measurements; values for these components are derived during the process of model optimization. A hairpin-specific intercept and slope adjustment is associated with all measurements of a given hairpin. An additional cell line intercept and slope adjustment is associated with all measurements of that hairpin in a specific cell line. The overall combination of intercept, slope and random effects defines a line that approximates the measurements for a hairpin in a specific cell line. The random error component summarizes any remaining differences between this model and measurements for a specific hairpin and cell line.

The siMEM model estimates the magnitude and error associated with each component in the equation above (e.g., the difference in slope associated with a “genomic variable,” such as

HER2+ status, subtype, mutation status). We then test whether the estimated magnitude of this slope difference is consistent with the null hypothesis that the true slope difference is zero (described in detail in later sections). A small p-value indicates that the observed magnitude is very unlikely, allowing us to reject the null hypothesis.

When a genomic variable is categorical, the associated difference in slope quantifies the differential essentiality (abbreviated DE) between two classes. For a continuous variable (e.g., expression log-FPKM), DE is the average up/down slope difference associated with each unit increase of the genomic variable. For example, as HER2 log-FPKM increases across cell lines, the average slope for hairpins targeting HER2 will tend to decrease, indicating increasing essentiality.

siMEM model specification

More formally, the hierarchical linear model is defined as:

$$y_{hc} = X_{hc}\beta + Z_h b_h + Z_{h,c} b_{hc} + \varepsilon_{hc}$$

with *h* indexing the hairpins targeting a given gene, *c* indicating a given cell line, *r=1,…,R* indicating replicates, and *t=0,…,T* representing time-points. The column vector $y_{hc} = \left[\begin{array}{ccc} y_{hc,r=1,t=0} & \cdots & y_{hc,r=R,t=T} \end{array} \right]^{transpose}$ contains *h*- and *c*- associated measurements across all replicates *r=1,…,R* and time-points *t=0,…,T*. For our data, y_{hc} is a 9-element column vector containing the 3 replicate x 3 time-point measurements that generate the dropout trend associated with *h* and *c*.

X_{hc} is the *(T+1)* x *R* row, *3* column fixed-effect design matrix, containing *1* in all rows of its first column, time values in the second column, and either all *1s* or all *0s* in the third column (depending on whether *c* is, or is not, associated with the condition, respectively, e.g., HER2+/HER2-). If the genomic variable is continuous (e.g., cell line HER2 log-FPKM values), that value is repeated for each row of the third column instead of *0/1*. In practice, continuous genomic values are median-centered across cell lines prior to modeling.

$\beta = \begin{bmatrix} \beta_0 & \beta_1 & \beta_D \end{bmatrix}^{transpose}$ is the fixed-effect coefficient column vector. The coefficients summarize average measurement intensity at time $t=0$ (β_0), linear slope (β_1) for the baseline condition, and slope difference associated with a genomic variable (β_D). β_D coefficient estimates (magnitude, error, p-value) are the most relevant, as they summarize DE magnitude and significance. The volcano plots presented in this paper display the magnitude and p-value of β_D for each gene in our assay (along the x and y axes, respectively).

Model random-effect components are critical to generating adequate error and p-value estimates for β_D. Models with very different random effects structures (Fig. S2A, S2B, discussed further below) produce very similar estimates for coefficient magnitudes. However, coefficient errors and p-values can differ greatly (see Fig. S2E and S2F).

The model includes random effect regressor matrices Z_h (hairpin-specific effect) and $Z_{h,c}$ (for hairpin-specific cell line effects). b_h is a two (2-) element column vector, containing hairpin *h* intercept and trend adjustments relative to the overall average intercept and trend. b_h is drawn from a bivariate Gaussian distribution ($b_h \sim N(0,\Sigma)$); this distribution represents hairpin-specific adjustments for the set of all hairpins targeting the gene, with the assay hairpins considered random instances from the set. The "0" is shorthand for $\begin{bmatrix} 0 & 0 \end{bmatrix}^{transpose}$, indicating that the b_h adjustments, as a group, average to the overall gene intercept and slope. Σ is a 2 x 2 variance-covariance matrix that contains the variance of hairpin intercept and slope adjustments as two diagonal entries, and the covariance between hairpin intercept and slope adjustments in both off-diagonal entries.

Similarly, b_{hc} is the intercept/slope adjustment specific to h and c, and is also drawn from a bivariate Gaussian ($b_{hc} \sim N(0,\Sigma_h)$) summarizing the distribution of individual cell line intercepts/slopes around the hairpin h intercept/slope. A separate 2 x 2 Σ_h matrix is estimated for each h. This conditional structure gives rise to the "cell line nested in hairpin" nomenclature (Fig. S2A, Fig. S2B). The modeling of b_h (or b_{hc}) as a random instance from the set of all such possible intercept and slope adjustments gives rise to the "random effect" terminology.

Finally, ε_{hc} is a random error term associated with each h- and c-specific line, with h x c error terms estimated for each gene. Once overall intercept/slope, genomic variable slope difference, b_h and b_{hc} have been summed, ε_{hc} accounts for any remaining differences between the model and measurements. The calculation of each ε_{hc} assumes that the measurements associated with h and c have no systematic error (or variance) trends (e.g., errors linked to measurement intensity or time-point). In other words, the sizes of the error bars around the line are assumed to be constant, regardless of measurement intensity or time-point.

Pooled screen data deviate severely from this assumption, as seen by plotting the dropout measurements for any gene that substantially impacts proliferation. Replicate measurement error bars widen systematically as intensity decreases and as time increases (in other words, the measurements are heteroscedastic). As detailed in later sections, we mitigate this problem by using precision, or inverse-variance, weights.

Although we refer to hairpins in the above model specification, siMEM is equally applicable to other similar types of screens (e.g., CRISPR/Cas9 screens). The model is agnostic to the specifics of the biological entity being measured, as long as several such entities map to each gene, and each produces multiple measurements in cell lines or samples.

Finally, the siMEM model can be simplified to analyze individual hairpin DEs. Considering a single hairpin obviates the multiple hairpin adjustment; hence, only cell line adjustments are included in the model.

Single time-point model

The model also can be simplified to enable analysis of end-point measurements, such as the Achilles (Cheung et al., 2011) dataset (omitting the universal reference samples) or a single time-point subset of our measurements. The fixed-effect coefficients then simplify to $\beta = \begin{bmatrix} \beta_0 & \beta_D \end{bmatrix}^{transpose}$, with β_0 being the mean in the baseline condition and β_D the difference in means associated with the genomic variable. If the genomic variable is continuous, β_D is the slope of the line through the "measurement intensity vs. genomic variable" scatterplot, and β_0 is the intercept of that line when the genomic variable is 0.

The random effects structure also is simpler. Variable nesting structure does not change (cell line nested in hairpin), but the random effects are now univariate: $b_h \sim N(0,\sigma^2)$ and $b_{hc} \sim N(0,\sigma_h^2)$, with σ^2 and σ_h^2 representing variances of the Gaussians.

P-values

The siMEM model produces estimates of the magnitudes, errors and *t*-statistics (*t*-statistic = magnitude/error) for each fixed-effect coefficient (β_0, β_1, β_D). These are used to estimate the probability of observing the magnitude of β_0, β_1 or β_D, given the null hypothesis that the real magnitude is 0. P-values are obtained by comparing *t*-statistics to a *t* distribution, with the denominator degrees of freedom estimated using the “inner-outer” (or “between-within”) heuristic (Pinheiro and Bates, 2000). When comparing alternative model structures, Gaussian-based p-values are used. All p-values are two-sided, and are adjusted using the False Discovery Rate (FDR) method of Benjamini & Hochberg (Benjamini and Hochberg, 1995).

Regularization of random effects using weakly informative priors

Estimation of some random effect parameters can be computationally difficult when the number of random effect groups is small: for example, when five hairpins target a gene. In some cases, likely positive parameter values, such as variances, are estimated as 0. This issue can be addressed by imposing a weakly informative prior on random effect parameter estimates (Chung et al., 2015; Chung et al., 2013). These priors ensure that parameter estimates are always positive, yielding slightly more conservative error estimates and model predictions.

Following Chung *et al.*’s default distribution choice, we applied Wishart priors to the 2 x 2 Σ matrices summarizing intercept and slope adjustments, and Gamma priors for σ^2 variance parameters summarizing slope or intercept adjustments. These priors are implemented in Chung *et al.*’s accompanying *blme* R package, and applied to our models. We performed a large number of model fits for a variety of analyses (e.g.: HER2+ vs. HER2-, luminal vs. basal, essentiality with changing expression, etc…) with or without the priors, to confirm that prediction results are very similar in magnitude and significance.

Measurement weights

Measurement variance trends and precision weights

As do other high-throughput measurement assays, our data and those from Project Achilles show prominent and systematic measurement variance trends. In our case, replicate measurement variance increases as mean replicate measurement intensity decreases and as time increases. The overall shape of the mean-variance relationship is highly platform-dependent. Accounting for systematic variance trends is more consequential to model prediction performance than underlying distributional assumptions (Law et al., 2014). As a recent example, Law *et al.* used a linear model approach assuming Gaussian distributions, but taking into account systematic variance trends, to model RNAseq differential expression, and demonstrated prediction performance as good if not better than the most popular published models based on Negative Binomial count-specific distributions. This finding is consistent with mixed-effect model simulation results (Jacqmin-Gadda et al., 2007), which show that data with unequal variances substantially reduce parameter confidence interval coverage from the nominal 95%. However, even severe deviations from Gaussian distributional assumptions led to little reduction in confidence interval coverage. In short, linear model prediction performance tends to be robust to deviations from Gaussian distributional assumptions, but not to the presence of systematic measurement variance trends.

To address this issue, we use precision, or inverse-variance, weights for measurements. The small number of replicates at each time-point results in imprecise variance estimates when triplicate measurements from each hairpin are considered in isolation. As this issue occurs frequently in high throughput assays, an established solution (see Law *et al.* for a recent example) is to model replicate measurement variance as a smooth function of the mean measurement intensity. Thus, hairpins with similar mean intensities are assumed to have similar variances.

We estimate a separate measurement mean-variance function for each cell line and time-point pair. This function is obtained by applying local regression to the scatter plot of replicate means vs. variances using the R *locfit* (Loader, 2013) package. Replicate hairpin measurements are then assigned a precision weight equal to the inverse of their smoothed variance. To avoid extremely large weights, smoothed variance is set to a minimum of 0.01. These weights, and the associated measurements, are then used to perform weighted regression. Although, by default, a separate function is estimated for each cell line and time-point combination, the *siMEM* R package allows user-defined sample groupings, thus allowing flexibility for different replication designs.

<u>Fast dropout trends and signal/noise weights</u>

Previously, we highlighted the issue of "fast dropout" hairpins, particularly among those targeting general essential genes (Marcotte et al., 2012). In such cases, the trend for a hairpin tends to sharply decrease between the first and second time-points, and flattens between the second and third. Such plots are non-linear, even in the log-scale.

We mitigated this issue in a data-driven manner, using biological control features available on the Gene Modulation Array Platform (Ketela et al., 2011) used to evaluate our pooled screens. The GMAP platform probes a large number of human and mouse RNAi Consortium hairpins (Moffat et al., 2006; Root et al., 2006). Because our pooled screens were performed on human cells, measurements for the mouse hairpin pool provide a large number of potential negative controls, allowing us to quantify the probability that a particular human measurement is signal or noise, given its intensity. We used Bayes' rule to estimate the signal/noise probability as a function of measurement intensity, specifically:

$$Pr(S=1\,|\,x)=\frac{Pr(x\,|\,S=1)\,Pr(S=1)}{Pr(x\,|\,S=1)\,Pr(S=1)+Pr(x\,|\,S=0)\,Pr(S=0)}$$

with x being measurement intensity, $S=1$ and $S=0$ representing signal and noise states, respectively, and assumed to be equally probable *a priori* ($Pr(S=1)$ and $Pr(S=0)$ set to 0.5). We used arrays from the initial time-points (T0) of our assays, before substantial dropout occurs, thus ensuring that the signal and noise distributions were not confounded by decreases in human measurements occurring at later time-points. Mouse and human measurements were first averaged among T0 replicates of each cell line, before being merged across cell lines. Thus, a single signal/noise vs. intensity function was estimated for all cell lines. This function was sigmoidal, with high (>10) measurement intensities assigned probabilities ~1, whereas low (<7) intensities had probabilities ~0.2 (see data file accompanying the siMEM R package).

Next, we weighted measurements from later time-points in each *h*- and *c*-specific dropout time-course according to the signal/noise probability of measurements at the previous time-point. For example, if measurements at T1 had a signal/noise probability of 0.2 (mean intensity of ~7 or lower), T2 measurements were assigned a weight of 0.2. T0 measurements were assigned weights of 1. For fast dropout hairpins, T1 measurements tend to be low, and T2 measurements are correspondingly assigned much less weight in the model fitting. This approach helps to mitigate the systematic non-linearity observed with fast dropout hairpins.

Because the signal/noise function is calculated using measurements available on the GMAP platform, this weighting is study-specific. However, when considered as a heuristic to mitigate “fast dropout” trend non-linearity, the approach is applicable to other short time-course dropout studies with a user-defined sigmoidal or other function that can be used to penalize later, low-intensity time-points. In exploratory analyses to gauge the relative importance of precision and signal/noise weights, inclusion of precision weights alone produced model improvements an order of magnitude greater than signal/noise weights alone.

Hairpin- and cell line-specific weights

Assigning weights to individual measurements also enables hairpin- and/or cell line-specific weighting. The associated measurements are assigned a weight proportional to the total weight assigned to all cell lines or hairpins in the gene-level analysis. We use hairpin weights to filter hairpins whose initial (T0) measurements are close to, or below the noise threshold for, the platform. For our screens, we assign a weight of 0 to any hairpin whose mean T0 measurement intensity across all screens is < 8.5 (log2 scale). This cutoff was selected based on the signal/noise function described above. For example, eight hairpins target *HER2* in our dataset, but only four of these are used in the analysis after the low T0 filter. This approach avoids flat trends resulting from measurements starting at T0 and continuing (at later time-points) within the noise range of the measurement platform. After applying this filter, approximately 9,000 hairpins are excluded from analysis. For analyses using the Achilles dataset, we assign a 0 weight to all hairpins with a mean measurement intensity below 5 (log2 scale) in the universal sample replicates.

Another potential application of hairpin weights is to incorporate measures of on-target hairpin activity, such as the ATARIS C-score (Shao et al., 2013). Measurements associated with each hairpin can be weighed according to the hairpin C-score, with higher C-scores indicating greater likelihood of on-target activity. In several analyses incorporating ATARIS C-scores, both for our data and the Achilles dataset, we noted further improvement in predictions beyond those presented in **Results**. However, approximately half the genes in our assay do not have assigned

C-scores (as a result of not having any ATARIS solutions). Of the remainder, more than a thousand genes have two or more ATARIS solutions, with one C-score per solution. Further work is necessary to address these issues, so we have not incorporated C-scores into the analyses presented here.

Nevertheless, our approach correctly identifies many known breast cancer vulnerabilities, and predicts novel ones that subsequently can be confirmed by validation experiments (see **Results**). Our analysis suggests that the direct analysis of assay measurements, rather than measurement-derived summary scores, is most consequential for improving prediction performance, with hairpin on-target activity weights providing potentially important, but not prediction-critical, information.

Rescaling and combining weights

In a model excluding all measurement weights described above, each measurement has a weight of 1, and the sum of weights applied to the measurements is equal to the number of measurements. Increasing the total weight applied to the measurements also results in smaller p-values. For example, assigning a weight of 10 to each measurement produces predictions with much smaller p-values than the same measurements analyzed with a weight of 1. Consequently, significance predictions can be inflated if weighting strategies greatly increase the total weight of the measurements. This potential problem is particularly relevant for precision weights where, in most instances, variances associated with measurements at "high" intensities (10 or above on a log2 scale) are far smaller than 1, resulting in correspondingly larger weights. Additionally, the bulk of measurements associated with any gene are high. Applied as is, the total precision weights for the measurements are much greater than the number of measurements, which sometimes can result in a dramatic increase in the predicted significance. To counter this problem, we rescaled each precision weight using a constant, so that the sum of all precision weights for a gene was equal to the number of measurements (once zero-weighted measurements were excluded).

When multiple weights (precision, signal/noise, hairpin or cell line) were associated with the same measurement, they were multiplied (after rescaling) to obtain a combined weight, and again rescaled to sum to the total number of measurements. This produced the final weight applied to each measurement in the analysis. All analyses of our data used precision and signal/noise weights. Analyses of the Achilles dataset use precision weights. Hairpin binary 0/1 weights were also used, but only to omit measurements for hairpins with low T0 (our study) or universal sample (Achilles) intensities.

Relative Dropout Rate

In general, genes that are more essential tend to be associated with larger differences between conditions. In other words, the magnitudes of β_I and β_D are correlated. Ranking significant analysis predictions by the magnitude of β_D will thus tend to favor generally essential genes, even if the magnitude of β_D is small relative to β_I. To mitigate this issue, we formulated a complementary measure of effect size that considers the magnitude of the difference (β_D) relative to the magnitude of the baseline trend (β_I)

$$Relative\ \ Dropout\ \ Rate = sign(\beta_D)\frac{max(|\beta_1|,\ |\beta_1+\beta_D|)}{min(|\beta_1|,\ |\beta_1+\beta_D|)+median(\beta_1)}$$

The median value of the genome-wide distribution of β_1, which is reliably modestly negative, is added to the denominator to moderate unusually large ratios. The Relative Dropout Rate is restricted to categorical analyses.

Performance assessment

Alternative structures for model random effects

To evaluate the impact of our model design on prediction performance, we considered a range of alternative model random effect structures (Fig. S2B). We distinguish between model structures that are "simpler" or "more complex" variants of the siMEM structure (S6, Fig. S2B) and those that are "different." A model is simpler if it can be transformed to S6 by adding a random intercept or slope for a variable, or by adding a variable to the nesting structure (cell line). More complex models can be transformed to S6 by removing a variable (replicate) from the variable nesting structure. We consider a simpler model with comparable prediction performance to be preferable.

Models S9 and S10 use a different nested or crossed approach to relate hairpin and cell line variables (Fig. S2A-B). Model S9 (hairpin nested in cell line) assumes that hairpin adjustments depend on the cell line. This structure can be a good design choice if cell line characteristics are of primary importance for modeling measurements, while hairpin characteristics are secondary. A biological example would be cell lines that are generally more susceptible to shRNA-mediated knockdown, regardless of hairpin details. Model S10 assumes that the hairpin and cell line adjustments are independent of each other, and that each contributes separately to explaining measurements.

By contrast, siMEM structure S6 can work best if the observed cell line trend mostly depends on the specific hairpin (e.g., if a hairpin is ineffective). In that case, the dropout trend will be flat, regardless of the cell line in which measurements are made. Alternately, a potent on-target hairpin will tend to have larger dropout trend. Thus, the measurements from any cell line would be explained primarily by an overall hairpin intercept/trend, with a secondary cell line adjustment included to reduce differences between the overall hairpin trend and cell line-specific measurements.

We evaluated model performance by three criteria: model fit, prediction of known positives, and prediction in a random class analysis.

Alternative model fits

Akaike's Information Criterion, or AIC (Akaike, 1976), quantifies how well a model represents measurements. An AIC value is produced for each gene-specific model in an analysis. We assessed alternative model fits by using ~15,000 separate sets of measurements arising from the same assay and sharing underlying characteristics. The difference in per-gene AIC values ($\Delta AIC=AIC_{S6}-AIC_{alternative}$) indicates whether the alternative model is better (positive difference) or worse (negative) than S6. A ΔAIC of -10 or lower is strong evidence in favor of S6. As illustrated by the HER2+ analysis, S6 greatly outperformed simpler or different alternatives (Fig. S2G, S2H). The more complex alternatives S7 and S8 have ΔAIC distributions centered at 0, indicating no overall improvement resulting from additional model complexity (Fig. S2G). Ten alternative models were fit for each gene-specific set of measurements; in almost all cases, S6 was the simplest model that produces the best AIC values.

Although the HER2+ analysis is discussed in detail as a representative case, ΔAIC distributions were very similar in other analyses and using other data, including our previously published set of 72 breast, pancreatic and ovarian cancer screens (Marcotte et al., 2012). This remained true when the classes are biologically meaningful (e.g., subtype/tissue essentials) or when class assignments were randomized per gene (discussed below). We also performed an analogous assessment of alternatives for the model used to analyze the Achilles data. In all cases, the "cell line nested in hairpin" structure was much better as assessed by AIC (data not shown).

Prediction of known positives

We examined HER2+-dependent DE predictions because of their obvious biological and clinical relevance and because the subject has been extensively studied, providing us with a substantial number of literature-backed genes with which to test our predictions (Table S2C). Furthermore, while the large differences between luminal and basal breast subtypes (or tissues) make predictions easier, the differences are (relatively) less pronounced for classifications such as HER2+ vs. HER2-. For example, we predict about 2,000 differences (at FDR < 0.1) between breast basal and luminal lines, and comparable numbers for pairwise tissue comparisons (except basal vs. ovarian, Fig. S3C), but only a few hundred HER2+-specific vulnerabilities in breast cancer (Table S3B).

As seen in Fig S2E, the overall number of predictions drops 50-fold between structures S1 and S6, before increasing for S7 and S8. There was a concordant improvement in ranking for HER2+-associated genes from S1 to S5, with S5 to S8 producing comparable rankings. In short, up to a certain point, additional model structures eliminated many spurious predictions, while known positives rose to the top of the p-value rankings. The rankings were comparable from S5 to S8, but S6 produced the fewest overall predictions. These trends mirror the previously discussed improvement in AIC (Fig. S2G, S2H), indicating that the best model structure according to AIC also produces the best DE predictions in a biologically meaningful analysis.

Structures S9 and S10 performed worse, each predicting few significant genes, and failing to predict most of the known positives. Know positives also had worse p-value rankings with these models, regardless of significance (Fig. S2F). As discussed below, the small number of significant predictions with these model structures might be due to their overly conservative predictions.

Predictions using data with randomly assigned classes

Finally, we evaluated the prediction performance of alternative models (Fig. S2B) when cell lines were randomly assigned to two classes. Randomization was separate for each gene. In the example below, cell lines were classified in the same numbers as the HER2+/- classes (62/77 in one class, 15/77 in another). These results are representative of additional analyses performed with different class ratios. To mitigate the potential confounding effects of subtypes, the random class assignment was performed separately for cell lines of each Neve subtype (basal A, basal B, HER2+, luminal) before being combined. Thus, a similar proportion of cell lines from each subtype were randomly assigned to each class.

The randomized data were analyzed using each model, and the resulting p-values were compared to the Uniform(0,1) distribution using quantile-quantile Plots (Fig. S2D). A line below the diagonal indicates a p-value distribution skewed towards small values, whereas a line above the diagonal indicates enrichment for larger values.

Models S1-S4 produced a substantial enrichment for small p-values (Fig. S2D), consistent with the lower AIC values (Fig. S2G) and the large number of predictions in the HER2+ analysis (Fig. S2E). Models S7 and S8 were closer to the Uniform, but performed no better than S6 on this analysis, and might thus be unnecessarily complex. By contrast, models S9 and S10 produced a dearth of small p-values. Their predictions might be excessively conservative, again consistent with the worse model fits (Fig. S2H) and prediction of known positives (Fig. S2F). The results from models S5 and S6 were closest to the Uniform distribution (Fig. S2D). However, considered in conjunction with its better model fits (Fig S2G) and prediction of known positives (Fig S2E), S6 represents the best combination of model structure, complexity and prediction performance.

A similar analysis, applied to our previously published set of 72 screens (Marcotte et al. 2012), yielded comparable results. Finally, an analogous comparison of end-point model alternatives, using the Achilles data with randomized classes, showed that the "cell line nested in hairpin" structure, with random intercepts for each variable, produced the best results (data not shown).

Comparison to Parallel Mixed Model

Recently, Ramo et al. (Ramo et al., 2014) published Parallel Mixed Model (PMM), a hierarchical linear modeling algorithm to quantify kinome-wide siRNA screens assessing the impact of different pathogens on cells. Their approach has some similarity to siMEM, most prominently the application of hierarchical models to si/shRNA data, and allowing weights for different siRNAs according to quality measures of on-target effect. However, key differences in model assumptions and structure make PMM inapplicable to the genome-scale shRNA screens referenced in this manuscript.

Although the data modeled by PMM contains multiple siRNAs targeting each gene, each siRNA is measured once per screen, and the model does not account for systematic effects due to different siRNAs. Furthermore, all screens modeled by PMM are performed in the same cell line. The model does not account for screens performed across highly genetically heterogeneous cell lines, as is the case for our data or that of project Achilles. As we have shown, modeling these systematic reagent (si/shRNA) and cell line effects is key to making credible predictions in published genome-scale screens, and siMEM adjusts for both these factors (Figure 2, S2).

Furthermore, the PMM model assumes that each pathogen induces a global difference in cell essentiality profile. The observed difference in each gene's essentiality is modeled as a combination of the global pathogen-associated essentiality difference and a gene-specific essentiality difference. The pathogen variable modeled by PMM is analogous to a genomic variable, such as HER2 status or subtype, modeled in our context. The PMM structure that estimates a global pathogen (or genomic variable) effect is suited to situations where we expect to see thousands of differences between two classes of screens, for example when predicting thousands of significant differences between two cancer types. However, this modeling assumption may not be well suited to the vast majority of class comparisons examined in this manuscript, or those of interest to researchers, which involve at most a few dozen or hundred significant differences, and where the vast majority of genes in the genome are reasonably expected to have similar essentiality between comparison classes. Comparisons in which we expect to see many differences between classes are very much the exception to the rule. Finally, PMM does not model measurement heteroscedasticity, and is restricted to single time-point experimental designs.

In conclusion, although PMM's model structure and assumptions are not well-suited to the genome-scale screens referenced in this manuscript, it does provide an example of the general strategy of quantifying loss of function screens using hierarchical linear models.

R implementation and computational details

The *blme* (Dorie, 2014) v1.0.1 and *lme4* (Bates et al., 2014a; Bates et al., 2014b) v1.0.5 packages were used to fit all described linear mixed-effect models. Mean-variance function estimation was implemented by using local polynomial regression fits from the *locfit* 1.5-9.1 package. To reduce analysis time, *doMC* 1.3.1 was used to parallelize computations on a user-specified number of processor cores. Given the complex structure of our assay and pooled screen data in general, a Bioconductor (Gentleman et al., 2004) *ExpressionSet* structure was used to consolidate and link measurements, hairpin/gene annotations, and cell-line/replicate/time-point annotations. To facilitate community use, the *siMEM* R package used to generate many of our analyses is available, along with detailed instructions and sample workflows, from A.S.. Unless otherwise noted, all plots were generated using the R *ggplot2* (Wickham, 2009) v0.9.3.

Additional Methods

Screen data processing and normalization

Pooled screens were performed in triplicate, and infected cells were allowed to proliferate under standard growth conditions. Timepoints were taken for gDNA isolation and subsequent hybridizations depended on the population doubling; typically Passage 0 (P0), P2-3, and P5-6 were used to determine dropout (see (Marcotte et al., 2012).)

The T0 measurements for the EFM19, HCC1954, HCC38 screens were omitted for technical reasons. T0 measurements, regardless of cell line, represent the initial abundance of shRNAs before cell line-specific selection effects, leading to highly correlated T0 measurements across cell lines. Our analyses showed a median correlation of 0.92 between pairs of T0 arrays from different cell lines, compared to correlations of 0.94-0.97 for replicate arrays within a cell line, a median correlation of 0.79 between T1 arrays of different cell lines and median correlation of 0.68 between T2 arrays from different cell lines. Based on this similarity, we used to T0 measurements of the MCF7 screen to provide T0 measurements for the HCC1954 and HCC38 screens, and T0 measurements from the SW527 screen to provide initial measurements for the EFM19 screen.

As in our earlier screens, triplicate arrays for each time-point of each screen were normalized separately by using Cyclic Loess (Dudoit et al., 2002) to mitigate technical artifacts. In the course of our subsequent analyses, we observed that summarizing screen data using linear models sometimes produced highly skewed predictions (visible as extremely lop-sided volcano plots). This problem was coupled with a global shift of the β_D distribution mode away from 0. Although the shift was modest for hairpin-level analyses, its impact was amplified at the gene level, because each gene is targeted by multiple shifted hairpins. Consequently, many genes targeted by these hairpins would be deemed erroneously significant.

In our previous analyses, replicates were normalized *within* a time-point, without considering potential distortions *across* the time-points of a short time-series. Given the ubiquity of measurement artifacts in high-throughput assays, there is no guarantee that a theoretically flat hairpin trend will produce assay measurements showing a flat time-course. Although we previously made GARP scores comparable across cell lines using Z-normalization, a different approach is needed to mitigate this issue for measurements.

For these reasons, we performed an additional Quantile Normalization (Bolstad et al., 2003), including all arrays for a given time-point, irrespective of cell line. Performing this additional normalization within each time-point diminished the issue of global shifts across time-points, and centered the mode of the β_D distribution at 0, in the process removing erroneously significant predictions. We also Quantile-Normalized Achilles replicate-level array data before analysis.

Update of gene annotations for 78K screen

In order to update gene ID and symbol annotations, genes were first matched to the latest available list of Entrez gene IDs using the Bioconductor *AnnotationDbi* package (Pages et al., 2014). Genes with existing IDs had their symbols and descriptions updated. Genes without matching IDs were matched using Refseq IDs and canonical symbols. If a match was found, associated information (Entrez gene Id, symbol, description) was updated. Genes that did not match using these criteria were manually examined using the NCBI gene website and matched if possible. Remaining genes, typically no longer existing, were removed from the dataset. The updated annotations contained 77,156 hairpins mapped to 15,709 genes.

SNP arrays and copy number analysis

Genomic DNA (750 ng) from each line and control normal female DNA (Biochain Lot # B502039) were amplified by using the Illumina Infinium Genotyping multi-use kit. Amplified DNA was fragmented, precipitated, and one third was hybridized to Human Omni-Quad Beadchips, incubated at 48°C for 18 hrs, washed and stained as per the manufacturer's protocol, and analyzed on an iScan (Illumina). Data files were quantified in GenomeStudio Version 2010.2 (Illumina) using Omni-Quad Multiuse_H manifest (Released April 2011), containing data from GenomeBuild 37, Hg19. All samples passed staining, extension, target removal, hybidization (independent controls) stringency metrics, non-polymorphic control, and non-specific binding (sample-dependent) controls.

SNP array data were segmented by Circular Binary Segmentation, or CBS (Olshen et al., 2004), using the Bioconductor *DNAcopy* package (Seshan and Olshen, 2014), with 10,000 permutations, alpha 0.001, and undoing of segment splits less than 1.5 standard deviations apart. CBS segments were mapped to genes using the Bioconductor *CNTools* package (Zhang, 2014), with the same gene start-end coordinates used to map the RNAseq reads. Gene-level copy gains and losses were defined by Log-R Ratio (LRR) cutoffs of +/- 0.2 respectively. We performed a per-gene LRR comparison for cell lines profiled in-house and by the Cancer Cell Line Encyclopedia (CCLE (Barretina et al., 2012)) using Affymetrix SNP arrays. This analysis showed that gene-level LRR values were highly linearly correlated, with in-house LRRs equal to approximately 0.37 times CCLE LRRs. Thus, our gain/loss cutoffs of +/- 0.2 are comparable to CCLE cutoffs of +/- 0.5.

Correlation of cell line and tumor CNA profiles

TCGA breast level 3 segmented copy number data were obtained for 1,021 tumor samples and mapped to genes, as described above. For cell lines and tumors, total LRR for each gene was then obtained by summing gene LRRs across samples. Although it is tempting to quantify similarity of tumor and cell line CNA profiles by using a Pearson correlation incorporating all genes, the strong association between LRR values for genomically proximal genes invalidates the required data independence assumptions, as can be seen by the very obvious paths and curve patterns on the tumor vs. cell line LRR scatterplot. Instead, we used a sampling approach, randomly selecting one gene from each chromosome and correlating the resulting 22 tumor/cell line pairs of LRR values. This exercise was repeated 1,000 times, yielding the strongly positive distribution of correlation coefficients with a peak around 0.7 (Fig S1A).

RNAseq

RNA (1 ug) from each sample was reverse transcribed into cDNA by using the Illumina TruSeq Stranded mRNA kit. Libraries were sized on an Agilent Bioanalyzer, and their concentrations were validated by qPCR. Six different libraries were normalized to 10nM and pooled, 13pM of pooled libraries were loaded onto an Illumina cBot for cluster generation, and the flow cell was subjected to 50-cycles of paired-end sequencing on an Illumina HiSeq 2000. Genomic alignment was performed with STAR (v2.3.0) (Dobin et al., 2013), using default parameters, except that –out SAMstrandField was set to intronMotif. The median number of reads/sample was 45M (min. 18M, max 160M). Reads (average 47M/cell line) were aligned to the NCBI Build 37 reference human genome, using Gencode V19 transcript models. The median percentage of aligned reads was 97% (min 93%, max 98%). Gene expression levels were estimated with Cufflinks (v.2.2.1) (Trapnell et al. 2010), using default parameters and the Gencode V19 GTF file. All resulting cufflinks output files were merged using a bespoke script written in R (v.3.0.3).

Targeted sequencing

DNA for 126 genes (1.264Mbp) mutated ≥3% frequency in breast or ovarian carcinoma was captured using Agilent SureSelect XT. For target capture, 750ng of a library generated from DNA (3ug) from each sample was hybridized for 24hrs (Agilent Custom Design 059771). Enriched libraries were sized, and concentrations were validated as above. Libraries from 41 and 42 cell lines, respectively, were normalized to 10nM and pooled, and 9 nmoles of each pool was loaded onto an Illumina cBot for cluster generation, and subjected to 100 paired-end sequencing cycles on an Illumina HighSeq 2000. FASTQ files were generated using Illumina CASAVA (v1.8.2) software. Sample quality was assessed by using the FASTQC v. 0.10.1 software package. Reads were aligned to the hg19 Human reference genome using BWA-MEM (v0.7.7), with an average read-depth of 430/site. Alignment quality was assessed using BAMQC (v2014-030-21), followed by marking of duplicates (Picard v0.1.19), indel realignment, base quality score recalibration and variant calling using HaplotypeCaller (GATK v3.0.0, dbSNP v138). Variants were filtered to a minimum depth of 10 and a quality by depth (QD) of 2. All variants were annotated by using Annovar (v2013-08) with its default set of databases, with inclusion of the COSMIC (v68) and Clinvar (v2013-11-05) databases. These files were converted into HTML for ease of viewing and analysis. To find variants of interest, we created custom scripts in PERL that filter all annotated variant files for changes that affect coding regions. Variants were filtered to include only those found to have matches in COSMIC or Clinvar (designated "pathogenic") and to have a minor allele frequency of 0.2.

miRNA analysis

Expression of miRNAs was assessed by using the nCounter® Human V2 miRNA Assay Kit (Cat# GXA-MIR2-48). Assays (200 ng total RNA) followed the standard protocol, which enables multiplexed direct digital counting of miRNAs. Sample preparation involved multiplexed annealing of specific tags (miRtags) to target miRNAs, ligation, and enzymatic purification to remove unligated tags. For hybridizations, 5 μL of each miRNA multiplex assay were mixed with 20 μL NanoString nCounter reporter probe mix and 5 μL capture probe mix (30 μL total volume), and then incubated at 65°C for 18-24 hrs. Post-hybridization samples were run on the nCounter analysis system, images were processed and barcode counts were tabulated in comma separated value (CSV) format.

Data were received in three batches, and normalized using the positive control method and the six positive controls provided in the kit. Exploratory clustering of the data revealed prominent batch effects, which were corrected using ComBat (Johnson et al., 2007). Subsequent clustering revealed no visible batch-effects.

Cell line subtyping

Intrinsic (PAM50)

Three signatures for centroid-based classification of breast cancer into intrinsic subtypes (Hu et al., 2006; Parker et al., 2009; Sorlie et al., 2003) were obtained from Supplementary Materials published by Wiegelt (Weigelt et al., 2010). Expression of each gene in the classifier was median-centered across cell lines prior to classification. For each of the three signatures, Pearson correlation was used to match each cell line to an intrinsic subtype, defined as the subtype with the highest associated correlation coefficient. If all subtypes had a correlation of less than 0.1, the cell line was not classified. A majority vote among the three classifiers was used to assign a consensus intrinsic subtype to each cell line. In the few cases where each signature predicted a different subtype, the PAM50 classification was used.

Neve (luminal/basal A/basal B) subtypes

Neve *et al.* derived signatures that classify breast cancer cell lines into luminal, basal A and basal B subtypes (Neve et al., 2006). These signatures consist of 305 unique Affymetrix U133plus2 probe sets mapping to 240 unique genes. To classify our cell lines using these signatures, we initially extracted expression values for 230 genes overlapping with the signature, and, following the Neve methodology, we subjected the expression data to hierarchical clustering by using average linkage and the Pearson correlation distance metric. Although this approach clearly identified luminal and basal lines, it failed to cleanly subdivide the basal cluster into basal A and B classes.

Instead, we found that three-component NMF (Lee and Seung, 1999; Lee and Seung, 2001) clustering of the top 10% (or the top 5% or 20%) of genes with highest expression variance clearly separated cell lines of known subtype into luminal, basal A and basal B clusters. Therefore, we used NMF clustering to assign the remaining cell lines. For the subtype analyses presented in Figure 3D-3E, and given the distinct underlying biology, high HER2 expression (see below) was used to further distinguish "HER2+" cell lines among the luminal group. Note that "HER2+" was not part of the original Neve classification.

Receptor high/low expression status

We used the R *mixtools* package (Benaglia et al., 2009) to fit two-component Gaussian mixture (not mixed-effect) models to classify *ERBB2* (HER2), *ESR1* (ER), *PGR* (PR), and *AR* (AR) expression into high and low classes. *AR*, *ESR1* and *PGR* were not expressed above the noise level (FPKM > 0.1) in a substantial fraction of cases. These values are clearly non-Gaussian, and a large number of cell lines assigned the same (log-) FPKM value would lead to distorted Gaussian model fits. We therefore defined cell lines with a noise-level expression value as having a low receptor status, and omitted these samples from the mixture model fitting for that receptor.

Assignment of receptor (HER2/ESR1/PGR) status

After determining receptor high/low expression, samples with high HER2 were assigned to the HER2+ subtype. Of the remaining samples, those with high *ESR1* or *PGR* were assigned to the ER subtype. The remaining samples were classified as triple negative (TNBC).

Claudin-low subtyping

Following the classification approach of Prat (Prat et al., 2010), expression data were extracted for the 1920 "intrinsic" gene list published by Parker (Parker et al., 2009). In total, 1677 genes matching the intrinsic list by symbol were included. This list was filtered further to remove non-expressed genes, and the result was hierarchically clustered by Pearson correlation. The clusters were examined to identify the sub-tree containing previously identified claudin-low cell lines. Other cell lines in the same sub-tree were then defined as claudin-low.

Lehmann TNBC classification

The Lehmann TNBC subtype (Lehmann et al., 2011) was assigned by using the TNBCType web server (Chen et al., 2012) on the 44 cell lines identified as TNBC by the aforementioned three-receptor classification.

Curtis integrative subyping

The integrative subtype signatures (Curtis et al., 2012) comprise 10 class-specific centroids, each with values for 715 expression and 39 copy number Illumina array probes. These probes were mapped to genes using the accompanying annotations, resulting in 607 unique genes for expression, and 39 for copy number. We extracted the corresponding per-gene expression (log-FPKM) and copy number (LRR) values from our data. Because two data types were included in the same centroid, gene-specific expression or copy number data were median-centered and rescaled using the standard deviation across samples. In cases where multiple Illumina probes for the same gene were included in the published signature, per-gene values from our data were duplicated, so that each Illumina probe for the same gene was assigned the same values across our cell lines. Cell line values were then compared to the published centroids by using Pearson correlation, and the integrative cluster was defined as the centroid yielding the highest correlation coefficient. Copy number LRR data were not available for 3 of our cell lines (HCC1395, SUM229, ZR7530). As copy number probes accounted for only 5% of all Integrative cluster probes, we assigned subtypes to these lines using only the expression portion of the signature.

Subtype DE analyses

For each of the subtypes described above, all cell lines were dichotomized to one specific class (e.g., luminal) or another, and siMEM analyses were performed. For the Lehmann TNBC subtypes, siMEM analyses were restricted to the set of 44 TNBC cell lines. Genes were removed from these analyses if they were only expressed above noise levels (defined as FPKM > 0.1) in < 5 cell lines. Our aim was to remove genes showing substantial dropout differences despite not being expressed, which strongly suggests off-target effects. This filtering does not apply to the expression vs. essentiality analysis (detailed below). Expression filtering also was not performed for analyses where the overall number of predictions is of primary interest, such as the pairwise tissue comparison overview (Fig S3C).

Expression vs. essentiality analysis

Genes expressed above noise levels (FPKM > 0.1) in less than 20% (15/77) cell lines were excluded from this analysis, as were those whose expression varies little across cell lines (expression standard deviation < 0.5). Per-gene expression log-FPKM values were median-centered prior to siMEM analysis.

Copy gain- and loss-associated DE analyses

We first dichotomized the per-gene copy number results into gain (or loss) and other classes. A gene was analyzed provided a minimum of 3 cell lines fell into the gain (or loss) category. Each gene was then analyzed using siMEM to determine whether its essentiality is significantly associated with copy status.

Comparing copy loss DE predictions to CYCLOPS and STOP/GO genes

To determine whether our predictions agreed with previously identified CYCLOPS genes (Nijhawan et al., 2012), we obtained the list of 6,084 genes examined in the original report, and matched them to genes in our copy loss vs. essentiality analysis. This comparison resulted in a 4,293 gene overlap. Following the CYCLOPS analysis, we used a more permissive FDR < 0.25 significance threshold, and required that a gene become more essential in samples with copy loss. From the overlapping gene list, we predicted 114 significant genes. Forty-nine (49) are predicted CYCLOPS genes, with 11 of these genes predicted as significant in both analyses (Fisher's Exact Test p = 3.6×10^{-8}, odds ratio=11.6; 95% CI 5.2-24).

We also obtained the published list of STOP/GO genes (Solimini et al., 2012). Matching these genes, identified by symbol, to our data resulted in 1,058 STOP and 682 GO genes. Using a cutoff of FDR < 0.25 to identify significant genes, we found that 23/1,058 STOP genes were significantly more essential in copy loss lines, whereas 62/682 GO genes satisfied the same criterion, resulting in a GO/STOP odds ratio of 4.5 (GO/STOP = (62/620) / (23/1035)).

We applied a sampling with replacement bootstrap approach (Efron and Tibshirani, 1994) to determine the significance and 95% confidence interval for the odds ratio. We separately sampled with replacement from the 1,058 STOP genes and 682 GO genes, tabulated the number of significant genes in each sample, and calculated the resulting odds ratio. This process was repeated 100,000 times, producing a corresponding number of bootstrap odds ratios. The logarithms of these ratios were calculated, and the resulting distribution of log-ratios was verified to be symmetric and centered at *log(4.5)*. To determine whether the observed GO/STOP ratio is significantly > 1 (i.e., log-ratio > 0), the number of bootstrap log-ratios smaller or equal to 0 was counted, resulting in a bootstrap p-value < 0.00001. The 95% confidence interval for the observed ratio of 4.5 was obtained by using the 2.5th and 97.5th percentiles of the bootstrap log-ratio distribution, and converted to the exponential to obtain the equivalent interval for the ratio.

METABRIC and ISAR region trans-DE analyses

Breast CNA regions associated with expression changes *in trans* were identified from the METABRIC dataset (Curtis et al., 2012). Using genomic coordinates provided for these regions, genes were assigned to each by testing for at least partial coordinate overlap. The LRR values of each gene of a region were then averaged to obtain a single LRR per region (per cell line). The region-specific LRR value was discretized, with cutoffs of +/- 0.2 indicating gains and losses, respectively. Intermediate values were considered copy-neutral.

We then performed DE analyses for each METABRIC region, examining essentiality changes associated with copy gains and losses (each vs. copy-neutral). A minimum of three cell lines with gains or losses was required for each analysis. The above analysis was repeated for the 83 regions of focal gain identified by the ISAR algorithm (Sanchez-Garcia et al., 2014). Plots illustrating the top METABRIC region DE predictions were produced using CIRCOS software v0.67 (Krzywinski et al., 2009).

Testing expression changes for trans-DE genes

Our goal was to test the extent to which expression and essentiality changes co-occur in gain vs. normal, and separately, in the copy loss vs. normal, trans-DE analyses. For each of the two analyses, we extracted the list of differentially essential genes (FDR < 0.1) for every Curtis region, and checked differential expression between copy gain and normal (or copy loss and normal) using a Wilcoxon RankSum test. Once this test was performed for all genes from all regions, p-values were FDR-adjusted, and the number of genes with expression FDR < 0.1 were counted.

From a total of 1,450 genes differentially essential in the METABRIC copy gain vs. normal analysis (for any region), 32 genes with siMEM FDR < 0.1 also showed differential expression FDR < 0.1. To assess the overlap significance, we determined how many genes met the FDR < 0.1 threshold for each region-specific analysis. To this end, we randomly picked an identical number of genes from that analysis, and tested whether those genes were differentially expressed at FDR < 0.1 using the Wilcoxon test. This process was repeated 1,000 times. The observation of 32 genes was statistically significant (permutation p=0.003; mean expected by chance 14.5). However, this resulted in only ~2-fold enrichment above random background, and only accounted for ~2% of all DE genes. Thus, regardless of statistical significance, co-occurring changes in expression and essentiality arose only in a small fraction of trans-DE genes.

For copy loss vs. normal analysis, 1,108 genes were found to be differentially essential, with 29 genes both differentially essential and expressed (permutation p < 0.001, mean expected by chance 11.1).

Tissue-specific DE

Our previously published ovarian (N=15) and pancreatic (N=28) cancer screens were used in conjunction with the complete set of breast screens in the current study to perform all pairwise DE analyses between breast luminal, breast basal, ovarian and pancreatic lines. As previously noted, for these analyses, results were not filtered to remove mostly non-expressed genes. We choose to include these in our totals because differentially essential, but non-expressed, hairpins, though "off target," are still targeting some gene in the genome. Therefore, including these hairpins in the analysis increases power.

Comparisons to drug sensitivity data

We obtained the per cell line *–*$log_{10}IC50$ values for 90 drugs previously profiled on breast cancer lines (Daemen et al., 2013). For each drug, the negative *–*$log_{10}IC50$ values for cell lines also profiled in the present study were split into quartiles, with cell lines in the first and fourth representing drug-resistant and -sensitive lines, respectively. Cell lines with *–*$log_{10}IC50$ values in the second and third quartiles were excluded from the analysis of each drug. In a few cases, identical *–*$log_{10}IC50$ values are assigned to >25% of cell lines, and all lines with identical values were included in the analysis.

Sensitive vs. resistant DE analyses were performed for each drug, followed by GSEA as described above. GSEA results were parsed to obtain the list of all significant pathways for each of the 90 analyses. To group and explore pathway similarity between drugs, pathway –$log_{10}(FDR)$ significance values were hierarchically clustered using Ward's method and correlation distance metric. The 50% of pathways with lowest *–*$log_{10}(FDR)$ variances across the drug analyses were removed prior to clustering.

Subtype-specific pathway and network analyses

Enriched pathways were computed with g:Profiler (Reimand et al., 2011) for the subtype-specific analyses (Fig. 3D), using biological processes from Gene Ontology and pathways from

KEGG and Reactome. For the protein-protein interaction (PPI) analysis (Fig. 3E), the human PPI network was retrieved from BioGRID version 3.2.114 (Chatr-Aryamontri et al., 2015), and filtered to extract physical PPIs. Results were visualized using Cytoscape with the Enrichment Map plugin (Merico et al., 2010). Node size corresponds to the number of interactions (node degree).

GSEA analysis and enrichment map visualization

Prior to gene set analysis, all genes in DE analyses were ranked using the equation:

$$score_{gene} = -sign(magnitude_{gene}) * \log_{10}(P - value_{gene})$$

This score highly ranks genes whose essentiality increases significantly in the condition of interest, while those with decreasing essentiality occupy the lowest ranks. GSEA (Mootha et al., 2003; Subramanian et al., 2005) command-line software (v2.2) was used in pre-ranked analysis mode, with 1000 permutations, exclusion of small (<15) and large (>500) gene sets, and a weighted scoring scheme. Gene sets were from v5.0 of MSigDB (Subramanian et al., 2005). All gene sets from the Chemical and Genetic Perturbations, Canonical Pathways and GO MSigDB categories were included in our analysis.

Enrichment map visualizations for GSEA analysis results were generated using Cytoscape (Shannon et al., 2003) v3.2 with the Enrichment Map (Merico et al., 2010) plugin, using default filters for GSEA analysis results: p-value < 0.005, FDR < 0.05, edges are shown between gene set nodes if the two gene sets have an overlap metric of 0.5 or greater.

DGIdb

The Drug Gene Interaction Database (DGIdb) (Griffith et al., 2013) was used to define lists of "druggable" targets in the GPCR, Growth Factor, Histone Modification, Hormone Activity, Ion Channel, Kinase, Methyl Transferase, Phospholipase, Surface and Transporter categories. For the Surface category, a bespoke Java program (available from K.R.B. upon request) was used to query Ensembl and extract the number of transmembrane and extra-cellular domains for each gene. Surface genes were those with at least one of each domain. For genes in each druggable category, the subset with a DGIdb-annotated drug interaction was also extracted.

Immunoblots

Transfected cells were lysed in RIPA buffer (10 mM Na phosphate [pH 7.0], 150 mM NaCl, 1.0% NP-40, 0.1% SDS, 1.0% Na deoxycholate, 10 mM NaF, 2 mM EDTA, supplemented with a protease inhibitor cocktail), and incubated on ice for 20 minutes. Lysates were clarified by centrifugation for 15 minutes at maximum speed (14,000 rpm) at 4°C in a tabletop centrifuge (Eppendorf 5424 R), resolved by SDS-PAGE, and transferred onto PVDF membranes. The following antibodies were used for blotting, all at concentrations recommended by their manufacturer: BRD4 (Bethyl), ERK2 (Santa Cruz), cleaved Capsapse-3 (Cell Signaling), p21 (Cell Signaling), HA (Covance), MYC (Cell Signaling), and PARP1 (Cell Signaling). Infrared fluorescent-conjugated secondary antibodies (at their manufacturer-recommended concentrations), and the Odyssey infrared imaging system (LI-COR biotechnology, NE) were used for detection.

Flow cytometry

For Annexin V/SYTOX blue experiments, cells were resuspended in 1X Annexin V binding buffer (BD), supplemented with 2% serum, Annexin V-PE (1/300), and SYTOX blue. Cells were incubated for 20 minutes in the dark, and then analyzed on an LSR II Flow Cytometer (Becton-Dickson, Mountain View, CA). Data were analyzed with FlowJo software (TreeStar, Ashland, OR). For cell cycle analysis, 1x10^6 cells were fixed for 1 hour at 4°C with 70% ethanol, and washed once with ice-cold 1xPBS. Cell pellets were digested with RNase A (0.5mg/ml) for one hour, after which 10ul of a 1mg/ml PI solution were added to the cell suspension. Stained cells were analyzed by flow cytometry, as above.

qRT-PCR

Cells (30-50,000) were seeded on 24-well plates, and 24 hours later, were transfected with Dharmacon SMARTPOOL siRNAs (10nM) using Lipofectamine RNAimax (Life Technologies). Media were changed the following day, and cells were allowed to proliferate for 24 hours before lysis in RLT buffer (Qiagen mRNeasy kit). RNA was isolated following the manufacturer's instructions, quantified by Nanodrop, reverse-transcribed by using the Superscript First-Strand synthesis kit (Life Technologies), and quantified by using SYBR green (Life Technologies) on a CFX96 (Bio-Rad).

REFERENCES

Akaike, H. (1976). An information criterion (AIC). Math Sci *14*, 5-9.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D., *et al.* (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603-607.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014a). lme4: Linear mixed-effects models using Eigen and S4.

Bates, D., Melchler, M., Bolker, B., and Walker, S. (2014b). Fitting Linear Mixed-Effects Models using lme4. In ArXiv e-prints, pp. 5823.

Benaglia, T., Chauveau, D., Hunter, D., and Young, D. (2009). mixtools: An r package for analyzing finite mixture models. Journal of Statistical Software *32*, 1-29.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological), 289-300.

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185-193.

Chen, X., Li, J., Gray, W.H., Lehmann, B.D., Bauer, J.A., Shyr, Y., and Pietenpol, J.A. (2012). TNBCtype: A Subtyping Tool for Triple-Negative Breast Cancer. Cancer informatics *11*, 147-156.

Cheung, H.W., Cowley, G.S., Weir, B.A., Boehm, J.S., Rusin, S., Scott, J.A., East, A., Ali, L.D., Lizotte, P.H., Wong, T.C*., et al.* (2011). Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proc Natl Acad Sci U S A *108*, 12372-12377.

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., and Dorie, V. (2015). Weakly informative prior for point estimation of covariance matrices in hierarchical models. Journal of Educational and Behavioral Statistics *40*, 136-157.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., and Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. Psychometrika *78*, 685-709.

Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y*., et al.* (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346-352.

Daemen, A., Griffith, O.L., Heiser, L.M., Wang, N.J., Enache, O.M., Sanborn, Z., Pepin, F., Durinck, S., Korkola, J.E., Griffith, M*., et al.* (2013). Modeling precision treatment of breast cancer. Genome Biol *14*, R110.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Dorie, V. (2014). blme: Bayesian Linear Mixed-Effects Models.

Dudoit, S., Yang, Y.H., Callow, M.J., and Speed, T.P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica sinica *12*, 111-140.

Efron, B., and Tibshirani, R.J. (1994). An introduction to the bootstrap (CRC press).

Gentleman, R.C., Carey, V.J., Bates, D.M., and others (2004). Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology *5*, R80.

Griffith, M., Griffith, O.L., Coffman, A.C., Weible, J.V., McMichael, J.F., Spies, N.C., Koval, J., Das, I., Callaway, M.B., Eldred, J.M*., et al.* (2013). DGIdb: mining the druggable genome. Nature methods *10*, 1209-1210.

Hu, Z., Fan, C., Oh, D.S., Marron, J.S., He, X., Qaqish, B.F., Livasy, C., Carey, L.A., Reynolds, E., Dressler, L*., et al.* (2006). The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics *7*, 96.

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J.-M., and Thiébaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. Computational Statistics & Data Analysis *51*, 5142-5154.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118-127.

Ketela, T., Heisler, L.E., Brown, K.R., Ammar, R., Kasimer, D., Surendra, A., Ericson, E., Blakely, K., Karamboulas, D., Smith, A.M*., et al.* (2011). A comprehensive platform for highly multiplexed mammalian functional genetic screens. BMC Genomics *12*, 213.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res *19*, 1639-1645.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol *15*, R29.

Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788-791.

Lee, D.D., and Seung, H.S. (2001). Algorithms for non-negative matrix factorization. Paper presented at: Advances in neural information processing systems.

Lehmann, B.D., Bauer, J.A., Chen, X., Sanders, M.E., Chakravarthy, A.B., Shyr, Y., and Pietenpol, J.A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. The Journal of clinical investigation *121*, 2750-2767.

Loader, C. (2013). locfit: Local Regression, Likelihood and Density Estimation.

Marcotte, R., Brown, K.R., Suarez, F., Sayad, A., Karamboulas, K., Krzyzanowski, P.M., Sircoulomb, F., Medrano, M., Fedyshyn, Y., Koh, J.L*., et al.* (2012). Essential gene profiles in breast, pancreatic, and ovarian cancer cells. Cancer Discov *2*, 172-189.

Merico, D., Isserlin, R., Stueker, O., Emili, A., and Bader, G.D. (2010). Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One *5*, e13984.

Moffat, J., Grueneberg, D.A., Yang, X., Kim, S.Y., Kloepfer, A.M., Hinkle, G., Piqani, B., Eisenhaure, T.M., Luo, B., Grenier, J.K*., et al.* (2006). A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. Cell *124*, 1283-1298.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E*., et al.* (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature genetics *34*, 267-273.

Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.P., Tong, F*., et al.* (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell *10*, 515-527.

Nijhawan, D., Zack, T.I., Ren, Y., Strickland, M.R., Lamothe, R., Schumacher, S.E., Tsherniak, A., Besche, H.C., Rosenbluh, J., Shehata, S*., et al.* (2012). Cancer vulnerabilities unveiled by genomic loss. Cell *150*, 842-854.

Olshen, A.B., Venkatraman, E.S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics *5*, 557-572.

Pages, H., Carlson, M., Falcon, S., and Li, N. (2014). AnnotationDbi: Annotation Database Interface.

Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z*., et al.* (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology : official journal of the American Society of Clinical Oncology *27*, 1160-1167.

Pinheiro, J., and Bates, D. (2000). Mixed-effects models in S and S-PLUS Springer. New York.

Prat, A., Parker, J.S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J.I., He, X., and Perou, C.M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res *12*, R68.

Ramo, P., Drewek, A., Arrieumerlou, C., Beerenwinkel, N., Ben-Tekaya, H., Cardel, B., Casanova, A., Conde-Alvarez, R., Cossart, P., Csucs, G*., et al.* (2014). Simultaneous analysis of large-scale RNAi screens for pathogen entry. BMC Genomics *15*, 1162.

Reimand, J., Arak, T., and Vilo, J. (2011). g:Profiler--a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res *39*, W307-315.

Root, D.E., Hacohen, N., Hahn, W.C., Lander, E.S., and Sabatini, D.M. (2006). Genome-scale loss-of-function screening with a lentiviral RNAi library. Nature methods *3*, 715-719.

Sanchez-Garcia, F., Villagrasa, P., Matsui, J., Kotliar, D., Castro, V., Akavia, U.D., Chen, B.J., Saucedo-Cuevas, L., Rodriguez Barrueco, R., Llobet-Navas, D*., et al.* (2014). Integration of genomic data enables selective discovery of breast cancer drivers. Cell *159*, 1461-1475.

Seshan, V.E., and Olshen, A. (2014). DNAcopy: DNA copy number data analysis.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.

Shao, D.D., Tsherniak, A., Gopal, S., Weir, B.A., Tamayo, P., Stransky, N., Schumacher, S.E., Zack, T.I., Beroukhim, R., Garraway, L.A*., et al.* (2013). ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens. Genome Res *23*, 665-678.

Solimini, N.L., Xu, Q., Mermel, C.H., Liang, A.C., Schlabach, M.R., Luo, J., Burrows, A.E., Anselmo, A.N., Bredemeyer, A.L., Li, M.Z*., et al.* (2012). Recurrent hemizygous deletions in cancers may optimize proliferative potential. Science *337*, 104-109.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S*., et al.* (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A *100*, 8418-8423.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S*., et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Weigelt, B., Mackay, A., A'Hern, R., Natrajan, R., Tan, D.S., Dowsett, M., Ashworth, A., and Reis-Filho, J.S. (2010). Breast cancer molecular profiling with single sample predictors: a retrospective analysis. The Lancet Oncology *11*, 339-349.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis (Springer New York).

Zhang, J. (2014). CNTools: Convert segment data into a region by sample matrix to allow for other high level computational analyses.