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:

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

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 $y_{hc} = \begin{bmatrix} y_{hc,r=1,t=0} & \cdots & y_{hc,r=R,t=T} \end{bmatrix}^{transpose}$ 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.

Fast dropout trends and signal/noise weights

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:

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

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)

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

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 $p = 3.6 \times 10^{-8}$, 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.