Resource

Cell

Predicting Cancer-Specific Vulnerability via Data-Driven Detection of Synthetic Lethality

Livnat Jerby-Arnon,[1,*] Nadja Pfetzer,[2] Yedael Y. Waldman,[1] Lynn McGarry,[2] Daniel James,[2] Emma Shanks,[2] Brinton Seashore-Ludlow,[3] Adam Weinstock,[1] Tamar Geiger,[4] Paul A. Clemons,[3] Eyal Gottlieb,[2] and Eytan Ruppin[1,4,*]

[1]The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 6997801, Israel
[2]Cancer Research UK, The Beatson Institute for Cancer Research, Switchback Road, Glasgow G61 1BD, Scotland, UK
[3]Center for the Science of Therapeutics, Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142, USA
[4]The Sackler School of Medicine, Tel Aviv University, Tel Aviv 6997801, Israel
*Correspondence: livnatje@post.tau.ac.il (L.J.-A.), ruppin@post.tau.ac.il (E.R.)
 http://dx.doi.org/10.1016/j.cell.2014.07.027

SUMMARY

Synthetic lethality occurs when the inhibition of two genes is lethal while the inhibition of each single gene is not. It can be harnessed to selectively treat cancer by identifying inactive genes in a given cancer and targeting their synthetic lethal (SL) partners. We present a data-driven computational pipeline for the genome-wide identification of SL interactions in cancer by analyzing large volumes of cancer genomic data. First, we show that the approach successfully captures known SL partners of tumor suppressors and oncogenes. We then validate SL predictions obtained for the tumor suppressor *VHL*. Next, we construct a genome-wide network of SL interactions in cancer and demonstrate its value in predicting gene essentiality and clinical prognosis. Finally, we identify synthetic lethality arising from gene overactivation and use it to predict drug efficacy. These results form a computational basis for exploiting synthetic lethality to uncover cancer-specific susceptibilities.

INTRODUCTION

Synthetic lethality occurs when the perturbation of two nonessential genes is lethal (Hartwell et al., 1997). This phenomenon offers a unique opportunity to develop selective anticancer drugs that will target a gene whose synthetic lethal (SL) partner is inactive only in the cancer cells (Ashworth et al., 2011; Hartwell et al., 1997). Toward the realization of this potential, screening technologies have been developed to detect SL interactions in model organisms (Byrne et al., 2007; Costanzo et al., 2010; Typas et al., 2008) and in human cell lines (Bassik et al., 2013; Brough et al., 2011; Laufer et al., 2013). However, currently their scope is not sufficiently broad to encompass the large volume of genetic interactions that need to be surveyed across different cancer types. New bioinformatics approaches are hence called for to guide and complement the experimental search for SL interactions in cancer.

Previous computational approaches developed to systematically study genetic interactions have mainly focused on yeast, where there are genome-wide maps of experimentally determined SL interactions (Chipman and Singh, 2009; Kelley and Ideker, 2005; Szappanos et al., 2011; Wong et al., 2004). In cancer, synthetic lethality has been computationally inferred by mapping SL interactions in yeast to their human orthologs (Conde-Pueyo et al., 2009) and by utilizing metabolic models and evolutionary characteristics of metabolic genes (Folger et al., 2011; Frezza et al., 2011; Lu et al., 2013). Here, we analyze the rapidly accumulating cancer genomic data to identify candidate SL interactions via the data mining synthetic lethality identification pipeline (DAISY). We show that genome-wide cancer SL networks can be used to successfully predict gene essentiality, drug response, and clinical prognosis.

RESULTS

The DAISY

DAISY is an approach for statistically inferring SL interactions from cancer genomic data of both cell lines and clinical samples. DAISY applies three statistical inference procedures, each tailored to specific cancer data sets.

The first inference strategy, termed genomic survival of the fittest (SoF), is based on the observation that cancer cells that have lost two SL-paired genes do not survive, they are strongly selected against. Accordingly, as cells harboring SL coinactivation are eliminated from the cell population, SL interactions can be identified by analyzing somatic copy number alterations (SCNA) and somatic mutation data and detecting events of gene coinactivation that occur significantly less than expected. In fact, very similar concepts are already extensively used to analyze the outcomes of small hairpin RNA (shRNA) screens in cell lines, in which essential genes and SL gene pairs are detected by identifying the shRNA probes that have been rapidly eliminated from the cell population (Cheung et al., 2011; Marcotte et al., 2012). More recently, a related concept was implemented to identify synthetic lethality in glioblastoma (Szczurek et al., 2013).

The second inference strategy, shRNA-based functional examination, is based on the notion that the knock down of a
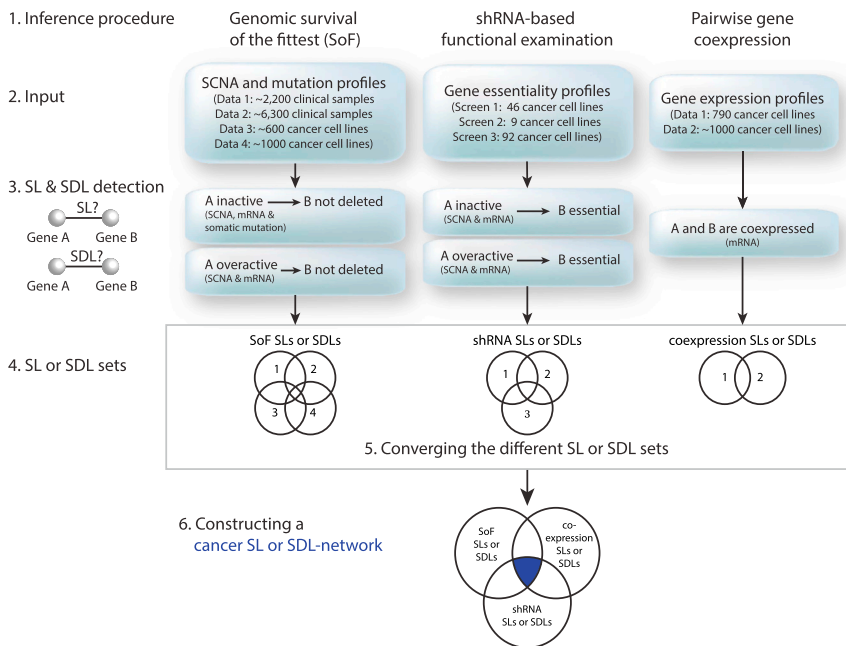
CrossMark

synthetically lethal gene is lethal to cancer cells where its SL partner is inactive. Accordingly, the SL pairs of a given gene can be detected by searching for genes whose underexpression and low copy number induce its essentiality. This can be conducted via an integrative analysis of the results obtained in shRNA essentiality screens and their accompanying SCNA and transcriptomic profiles.

The third procedure, pairwise gene coexpression, is based on the notion that SL pairs tend to participate in closely related biological processes and hence are likely to be coexpressed (Costanzo et al., 2010; Kelley and Ideker, 2005). We show that this trend indeed holds in known SLs that have been experimentally detected in cancer (Figure 2).

Given SCNA, somatic mutation, shRNA, and gene expression data of thousands of cancer samples, DAISY traverses over all gene pairs (~534 million) and examines for every pair if it fulfills each one of the three criteria described above. Gene pairs that fulfill all three criteria in a statistically significant manner are predicted to be SL pairs. Here, we applied DAISY to analyze nine different genome-wide cancer data sets (Barretina et al., 2012; Beroukhim et al., 2010; Cheung et al., 2011; Garnett et al., 2012; Luo et al., 2008; Marcotte et al., 2012; Cancer Genome Atlas Research et al., 2013) (Table S1 available online).

We expanded DAISY to also detect synthetic dosage lethality (Sajesh et al., 2013). While two genes form an SL pair if the inactivation of one gene renders the other essential, two genes form a synthetic dosage lethal (SDL) pair if the overactivity of one of them renders the other gene essential. Importantly, SDL interactions can permit the eradication of cancer cells with overactive oncogenes that are difficult to target directly (such as KRAS), by targeting the SDL partners of such oncogenes. DAISY detects SDL interactions via three inference procedures that are analogous to those outlined

above for detecting SL interactions (Figure 1; Experimental Procedures). More specifically, DAISY defines two genes, A and B, as an SDL pair if their expression is correlated and if the overactivity (amplification and overexpression) of gene A induces the essentiality of gene B. Induced essentiality is detected in two ways: first, according to shRNA screens, by examining if gene B becomes essential when gene A is overactive; second, according to SCNA data, by examining if gene B has a higher SCNA level when gene A is overactive.

Evaluating DAISY Based on Experimentally Detected SL Interactions in Cancer

We first examined DAISY based on SL interactions that have been experimentally tested in cancer. We applied DAISY to identify the SL partners of *PARP1*, the tumor suppressors *VHL* and *MSH2*, and the SDL partners of the oncogene *KRAS*. The predictions were performed for over 7,276 gene pairs that have been experimentally tested in six large scale screens (Bommi-Reddy et al., 2008; Lord et al., 2008; Luo et al., 2009; Martin et al., 2009; Steckel et al., 2012; Turner et al., 2008). For every gene pair, DAISY returns four p values that denote the significance of the SL or SDL interaction between the two genes according to each one of the three inference strategies described in the previous section and according to all three approaches together (Figure 1; Experimental Procedures). We utilized these p values to examine the predictions along an increasing p value threshold and generate receiver operating characteristic (ROC) curves (Extended Experimental Procedures).

The DAISY predictor obtains an overall AUC of 0.779, which shows the concordance between the predicted and observed SL and SDL pairs (empirical p value $<1 \times 10^{-4}$; Figure 2A). To assess which of the inference strategies enables DAISY to correctly predict synthetic lethality, we repeated the predictions when using the p values obtained based on only one inference strategy at a time (Figure 2A). An AUC of 0.683 was obtained by predicting SL interactions based only on the SoF approach. These results are further improved by requiring that the gene pairs will also be coexpressed, reaching to an AUC of 0.770. As shRNA-based functional examination is not predictive on its own (an AUC of 0.477), we modified DAISY to consider the shRNA criterion as a soft constraint (Experimental Procedures).
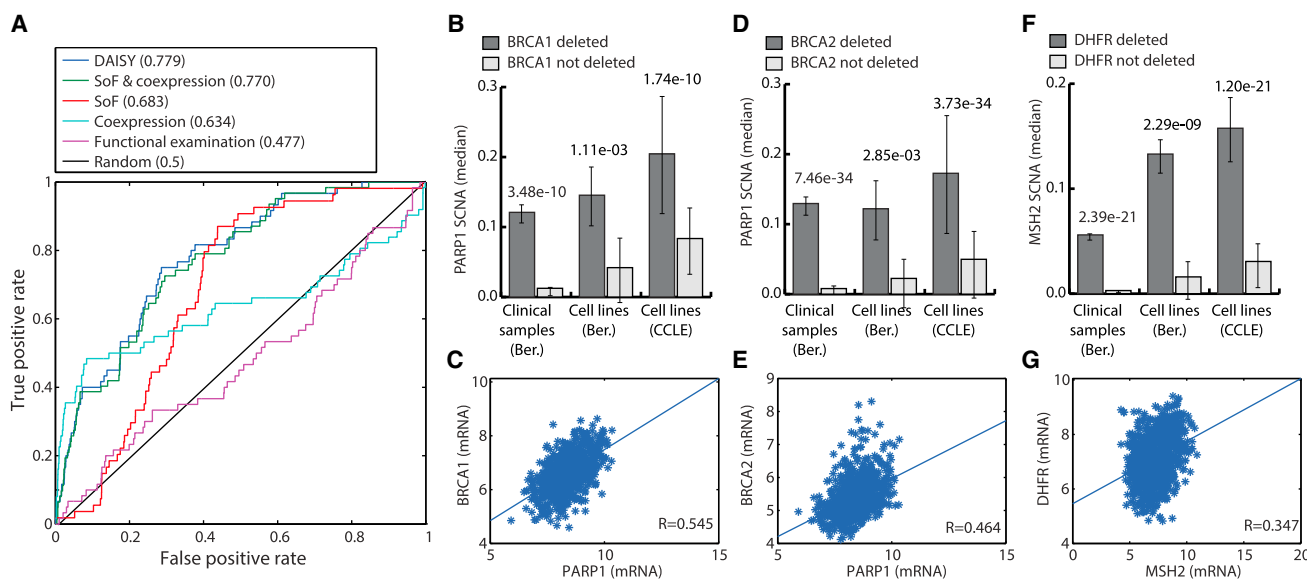
Figure 2. DAISY-Inferred SL and SDL Interactions Match Experimentally Detected Interactions in Cancer

(A) The ROC curves obtained when predicting SL and SDL interactions by applying each of the three inference strategies separately (red, light blue, and purple), the SoF and coexpression approaches together (green), and all three inference strategies together (DAISY, blue). The black line denotes the random ROC-curve. (B–G) The SCNA and expression patterns of experimentally well-established SL pairs: (B and C) *PARP1-BRCA1*, (D and E) *PARP1-BRCA2*, and (F and G) *MSH2-DHFR*, respectively. For each one of these SL pairs the SCNA levels of one gene are significantly higher when its partner is deleted than when its partner is retained, as expected (one-sided Wilcoxon rank sum test) (Barretina et al., 2012; Beroukhim et al., 2010) (B, D, and F), and their expression is significantly correlated (Garnett et al., 2012) (C, E, and G). The error bars in (B), (D), and (F) represent the confidence interval of the SCNA levels across the samples.

Despite the nonpredictability of the shRNA-based functional examination approach in this task, shRNA data are important for the generation of predictive SDL-networks (Supplemental Information; Figure S6). Importantly, DAISY captures well-established and clinically important SL interactions, including the prominent SL interaction between *PARP1* and *BRCA1/BRCA2* and the synthetic lethality between *MSH2* and *DHFR* (Figures 2B–2G).

Experimentally Examining the DAISY-Predicted SL Partners of the Tumor Suppressor VHL

We next turned to experimentally test SL predictions of the tumor suppressor *VHL* that is frequently mutated in cancer, especially in clear cell renal carcinomas (Bommi-Reddy et al., 2008). We applied DAISY to predict the SL partners of *VHL* and identify among these genes those that are essential in renal carcinoma cells (RCC4) exclusively due to the loss of *VHL*, resulting in a set of 44 genes (Extended Experimental Procedures).

We performed a small interfering RNA (siRNA) screen to examine if the predicted genes are preferentially essential in $VHL^{-/-}$ renal carcinoma cells compared with isogenic cells in which pVHL function was restored (Extended Experimental Procedures). Overall, compared to the *VHL*-restored cells, the *VHL*-deficient cells are significantly more sensitive to the knockdown of the predicted *VHL*-SL partners (paired t test p value of 8.25×10^{-4}) (Figure 3A, Table S2). Reassuringly, compared to the *VHL*-restored cells, the *VHL*-deficient cells are not significantly more sensitive to the knockdown of a control set of 30 randomly selected genes (paired t test p value of 0.255). Compared to another screen that searched for the SL partners of *VHL* among

88 kinases (Bommi-Reddy et al., 2008), our screen detected 3.83 times more SL genes (Bernoulli p value of 4.76×10^{-9}; Extended Experimental Procedures).

We then measured the response of the renal cells to nine drugs whose targets were predicted by DAISY to be selectively essential in the *VHL*-deficient renal cells. Of note, these drugs are not currently administered to treat cancer, but are Food and Drug Administration (FDA)-approved to treat other clinical conditions, such as hypertension and depression. We managed to identify effects on cell growth for six out of the nine drugs. As predicted, the *VHL*-deficient cells were significantly more sensitive to each one of these six drugs (higher percentage of inhibition at mideffective concentration) (Figure 3B; Table S2). Reassuringly, this specificity was not observed with the negative control drug Staurosporine, indicating that the selective effect is not due to a general susceptibility of the *VHL*-deficient cells.

Applying DAISY to Construct Genome-wide Networks of SL and SDL Interactions in Cancer

We applied DAISY to identify all gene pairs that are likely to be synthetically lethal in cancer, resulting in an SL network of 2,077 genes and 2,816 SL interactions (Figure 4), and an SDL network of 3,158 genes and 3,635 SDL interactions (Table S3). As each of the nine data sets examined were analyzed separately to identify SL (SDL) pairs, we tested the mutual overlap between the resulting SL (SDL) sets and found it to be significantly higher than expected (Figure S1).

Both networks display scale-free-like characteristics and are enriched with known cancer-associated genes and biological functions (Figures S1 and S2; Table S4). The genes included in
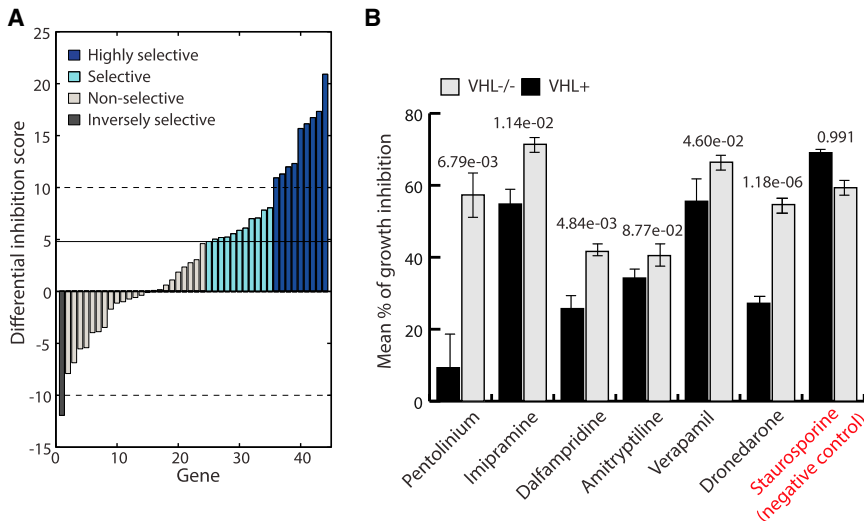
Figure 3. Examining DAISY Predictions of *VHL*-SLs

(A) The differential inhibition score of the predicted SL partners of *VHL*, ordered in ascending order. Highly selective genes (inversely selective) are those with differential inhibition scores >10 (<−10). Selective genes are those with differential inhibition scores ≥ 4.8 (the score of the positive control gene *MYT1*, identified in Bommi-Reddy et al., 2008). Dashed horizontal lines denote the threshold values.

(B) The mean percentage of growth inhibition of *VHL*-deficient (*VHL*^{−/−}) and *VHL*-restored (*VHL*^{+}) cells at the mideffective concentration of each drug. On top are the one-sided t test p values denoting if the inhibition of the *VHL*-deficient cells is significantly higher than the inhibition of the *VHL*-restored cells. The error bars represent the SE across the three replicates of the experiment. See also Table S2.

the networks are significantly overexpressed both in normal tissues and especially in cancers (Wilcoxon rank sum p values <6.29 × 10^{−8}). Interestingly, the network genes are significantly associated with cancer proliferation and less associated with normal proliferation (Waldman et al., 2013). They are highly enriched with human orthologs of mouse essential genes (hypergeometric p values <1 × 10^{−30}) and are evolutionary conserved (Wilcoxon rank sum p values <2.99 × 10^{−17}). Moreover, each one of these properties is further emphasized in genes that have a higher degree in the SL or SDL networks (Supplemental Information; Figure S2).

The SL and SDL pairs are highly enriched with genes that interact in the protein-protein interaction (PPI) network (hypergeometric p values <4.02 × 10^{−7}). Testifying to their importance, genes included in the SL or SDL networks have a higher degree in the PPI network compared to other genes, especially if their degree in the SL or SDL network is high (Wilcoxon rank sum p values <5.79 × 10^{−22}; Figure S2D). Examining the genomic location of the SL and SDL pairs we find that while SL pairs tend to reside on different chromosomes, or at a large distance from each other on the same chromosome, the SDL gene pairs show the opposite behavior. The latter trend is observed also when identifying SDL interactions without considering the SoF approach. Discarding SDL gene pairs that reside close to each other depreciates the predictive signal of the network (Supplemental Information; Figure S3).

As a direct experimental validation of the predicted SL and SDL interactions is yet impossible on a genome scale, we tested the interactions by examining their utility in three fundamental prediction assignments, the prediction of gene essentiality, clinical prognosis, and drug efficacy. In all tasks, the networks are utilized to generate cancer-specific predictions given a genomic characterization of a specific cancer cell line or clinical sample.

SL-Based Prediction of Gene Essentiality in Cancer Cell Lines

Predicting gene essentiality based on the SL network is cell-line-specific. Indeed, examining the results of shRNA screens, the majority of genes are essential in very few cancer cell lines

(Supplemental Information; Figure S4A). As we examined the predictions based on the results obtained in shRNA gene knockdown screens, we constructed an SL network without any shRNA data to avoid potential circularity. Based on this SL network and the genomic profiles of the cell lines, we predicted a gene as essential in a given cell line if one or more of its SL partners is inactive in that cell line (see Supplemental Information for further details, analyses, and results).

Overall, we predicted gene essentiality in 129 different cancer cell lines and examined the predictions based on the results of two large-scale gene essentiality screens (Cheung et al., 2011; Marcotte et al., 2012). Per cell line the predicted essential genes are significantly enriched with genes that were found experimentally to be essential in the pertaining cell line (empirical p value < 2.52 × 10^{−4}; Supplemental Information; Figure 5A; Table S5). Furthermore, the higher the number of predicted inactive SL partners a gene has the more essential it is according to the experimental data (Figures 5B and 5C). Of note, the SL network succeeds more in predicting gene essentiality in cell lines with a higher number of gene deletions (Supplemental Information; Figures S4B and S4C; Table S5). Indeed, in such cell lines it is more likely that gene essentiality arises due to synthetic lethality. Finally, we predicted gene essentiality based on gene pairs that are human orthologs of yeast SLs (Conde-Pueyo et al., 2009). This, however, leads to markedly inferior performance, testifying to the value of the DAISY-inferred SLs (Supplemental Information; Figures S4D and S4E; Table S5).

We improved the unsupervised SL-based gene essentiality predictions described above by considering additional features that describe the state of a specific gene in a given cell line according to the SL network (e.g., the average SCNA level of its SL partners). Using these features, we trained neural network models on gene essentiality data (Extended Experimental Procedures). The performances of these supervised prediction models on unseen test sets resulted in ROC curves with AUCs of 0.755 and 0.854 for the Marcotte et al. (2012) and Achilles (Cheung et al., 2011) data, respectively (Figures 5D and 5E). For comparison, we considered the nine cell lines that were tested in both screens and utilized the shRNA scores obtained in one screen
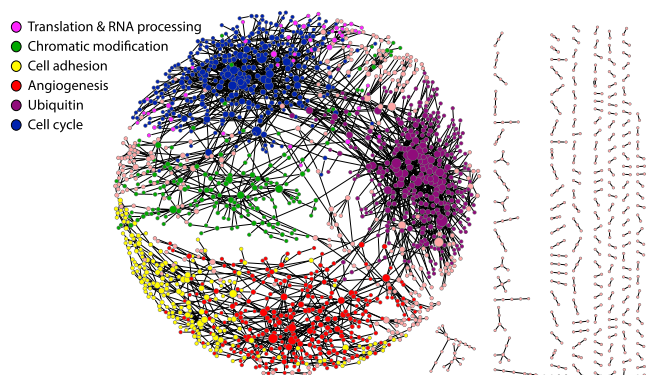
Figure 4. The SL Network

Each node represents a gene, and each edge represents an inferred SL interaction. Genes that are included in one of the six major clusters of the network are colored according to the main biological process that their cluster is enriched for. Node size is proportional to the number of SL pairs a gene has. See also Figure S2 and Tables S3 and S4.

Legend:
- Translation & RNA processing
- Chromatic modification
- Cell adhesion
- Angiogenesis
- Ubiquitin
- Cell cycle

to predict gene essentiality according to the other screen (Extended Experimental Procedures). Using the Achilles screen to predict gene essentiality as reported in the Marcotte screen, or vice versa, results in inferior prediction performance, with AUCs of 0.663 and 0.706, respectively.

To further examine the SL-based gene essentiality predictions, we conducted a whole genome siRNA screen in the breast cancer cell line BT549 under normoxia and hypoxia (Extended Experimental Procedures; Table S6). We defined a refined set of essential genes, composed of genes that are essential in BT549 according to our siRNA screen under both conditions and according to the shRNA screen of Marcotte et al. (2012). Indeed, the performance of the SL-based predictor (that was not trained on gene essentiality data of BT549) is further improved when tested on this refined set of essential genes, obtaining an AUC of 0.951 (Figures 5F and S4F–S4K; Supplemental Information).

Counderexpression of SL Pairs Is a Marker of Better Prognosis in Breast Cancer

To examine the SL network in a clinical setting, we analyzed gene expression and 15 year survival data in a cohort of 1,586 breast cancer patients (Curtis et al., 2012). We postulated that counderexpression of two SL-paired genes would increase tumor vulnerability and result in better prognosis. To test this hypothesis, we classified the patients according to each SL pair into two groups: patients whose tumors counderexpressed the two SL-paired genes (SL$^-$ group) and patients whose tumors expressed at least one of these genes (SL$^+$ group). For each SL pair, we computed a signed Kaplan-Meier (KM) score (Extended Experimental Procedures). The higher the signed KM score is, the better the prognosis of the SL$^-$ group is compared to the SL$^+$ group. Indeed, the signed KM score of the SL pairs is significantly higher than those of randomly selected gene pairs (one-sided Wilcoxon rank sum p value of 3.09×10^{-59}). To examine if this result arises from the mere essentiality of genes in the SL network rather than the interaction between them, we repeated the analysis with

randomly selected gene pairs involving genes from the SL network that are not connected by SL interactions. Reassuringly, the SL pairs have significantly higher signed KM scores also compared to these random SL network gene pairs (one-sided Wilcoxon rank sum p value of 2.00×10^{-9}). Highly significant KM plots were obtained based on 271 SL pairs (log rank and Cox regression p values <0.05, following multiple hypotheses testing correction) (Figure 6A; Table S7).

Next, we classified the patients according to all the SL pairs in the network together. For each sample, we computed a global SL score that denotes the number of SL pairs it counderexpressed. As predicted, samples that counderexpressed a high number of SL pairs had a significantly better prognosis compared to those that counderexpressed a low number of SL pairs (log rank p value of 1.482×10^{-7}; Figures 6B and 6C). Again, we examined if this result is due to the mere essentiality of the SL network genes rather than due to the specific SL interactions; we repeated this analysis using 10,000 topology preserving randomized networks consisting of the breast cancer essential genes (Marcotte et al., 2012) (Extended Experimental Procedures). Reassuringly, none of these random networks managed to predict patient survival as significantly as the SL network.

Because breast cancer is a highly heterogeneous disease, we examined whether higher global SL scores are associated with improved prognosis in specific and more homogenous groups of patients—all consisting of the same subtype, grade, or genomic instability level (Bilal et al., 2013). This is indeed the case for all groups except one—grade 1 patients. The global SL scores provide the most significant separation in the grade 2 normal-like subtype and moderate genomic instability groups (log rank p values of 8.64×10^{-5}, 1.01×10^{-3}, and 1.25×10^{-4}, respectively). As expected, the global SL score is significantly negatively correlated with the tumor grade and genomic instability level (Spearman correlation coefficients of -0.407 and -0.267, p values of 2.58×10^{-62} and 2.43×10^{-27}, respectively) and highly associated with the tumor subtype (ANOVA p value of 4.25×10^{-102}; Figure S5). Normal-like tumors have the highest global SL scores, while basal tumors have the lowest scores (Figure S5E). Notably, the prognostic value of the global SL score is significant even when accounting for the tumor grade, subtype, or genomic instability level (Cox p values of 7.18×10^{-4}, 3.12×10^{-7}, and 4.37×10^{-8}, respectively). Lastly, the prognostic value of the global SL scores is superior to that obtained by using genomic instability levels (Figures S5I and S5J).

Harnessing SDL Interactions to Predict Drug Efficacy

We utilized the SDL network to predict the response of various cancer cell lines to anticancer drugs. As these drugs mainly target oncogenes, we used the SDL network to predict their efficacy rather than the SL network, whose performance is indeed inferior in this task (Supplemental Information). Based on the SDL network and the genomic profiles of the cancer cell lines, we predicted for each drug which cell lines are sensitive and which are resistant to its administration (Extended Experimental Procedures). More specifically, if one of the drug targets had more than one overexpressed SDL partner in a given cell line,
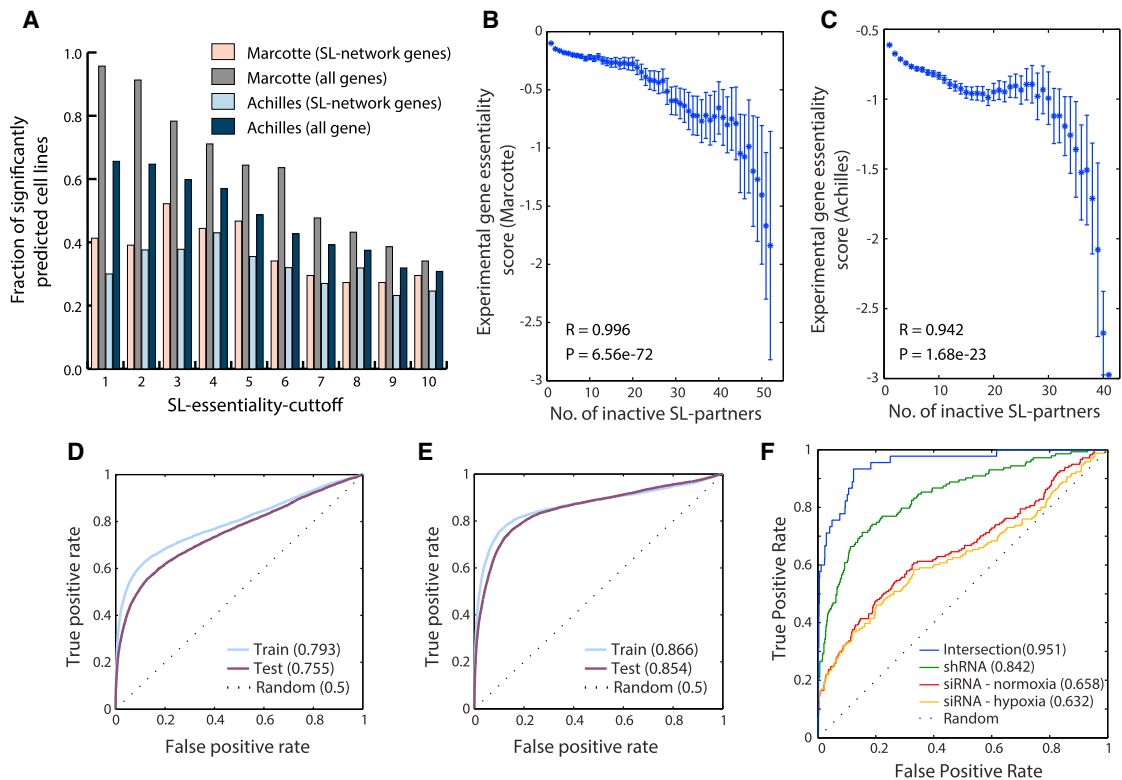
Figure 5. Predicting Cell-Specific Gene Essentiality Based on the SL Network

(A) The fraction of cell lines for which the unsupervised SL network predictor successfully differentiated between essential and nonessential genes (y axis). This is plotted for ten different SLessentiality$_{cutoff}$ values (x axis), denoting the minimal number of inactive SL partners a predicted essential gene has (Extended Experimental Procedures). Results are shown when considering only SL network genes (pink and light blue) or all genes (gray and blue) as the background random model and testing the predictions based on the Marcotte et al. (2012) (pink and gray) or Achilles (Cheung et al., 2011) (blue and light blue) screen. All the results are statistically significant (empirical p value <3.40 × 10^{-3}).

(B and C) The experimental essentiality scores (median and confidence interval) of genes across different cancer cell lines as a function of the number of SL partners they have lost, according to (B) the Marcotte and (C) the Achilles screens (lower experimental gene essentiality scores denote higher essentiality).

(D and E) The ROC curves obtained when predicting gene essentiality across the (D) Marcotte and (E) Achilles cancer cell lines via the supervised SL-based predictors.

(F) The ROC curves obtained by predicting gene essentiality in BT549 via the supervised SL-based predictor and testing the predictions with the genes that were found essential in all screens (blue), in the Marcotte shRNA screen (green), or in the siRNA screen we conducted under normoxia (red) and under hypoxia (yellow). The predictor was trained based on the gene essentiality reported in Marcotte et al. (2012), excluding the BT549 cell line data that was used exclusively for testing. See also Figure S4 and Tables S5 and S6.

the cell line was predicted to be sensitive to the drug administration (Supplemental Information).

To test this approach, we utilized two data sets of drug efficacies that were measured in a panel of cancer cell lines: (1) the Cancer Genome Project (CGP) data (Garnett et al., 2012), and (2) the Cancer Therapeutics Response Portal (CTRP) data (Basu et al., 2013). The SDL network enabled to predict the response of 593 cancer cell lines to 23 drugs and of 241 cancer cell lines to 33 additional drugs when utilizing the CGP and CTRP data sets to test the predictions, respectively. Overall, drugs are significantly more effective in the predicted sensitive cell lines than in the predicted resistant cell lines (empirical p values <5.34 × 10^{-4}; Figures 7A and 7B; Table S8). Considering only the predictions that were obtained for drugs with a sufficiently high number of SDL interactions increases the fraction of drugs that are significantly predicted (Figure 7C). As predicted, the effi-

cacies of drugs increase with the number of overexpressed SDL partners that their targets have in a given cell line (Figure 7D). Exceptions to this trend may be explained by noting that drug efficacy is determined only partially by the essentiality of the drug targets, and additional factors, like the drug membrane permeability, may affect drug efficacies. For comparison, we predicted drug response by applying two other well established approaches: (1) based on the mutation and copy-number status of the drug target(s), and (2) based on the genomic instability index of the cancer cells. The SDL network generates significant predictions for more than twice as many drugs compared to these competing predictors (Supplemental Information).

Focusing on the drugs that were most accurately predicted by using the SDL-network, we found that each one of the SDL interactions involving the targets of these drugs enables, on its own, to accurately predict the response to the pertaining drug
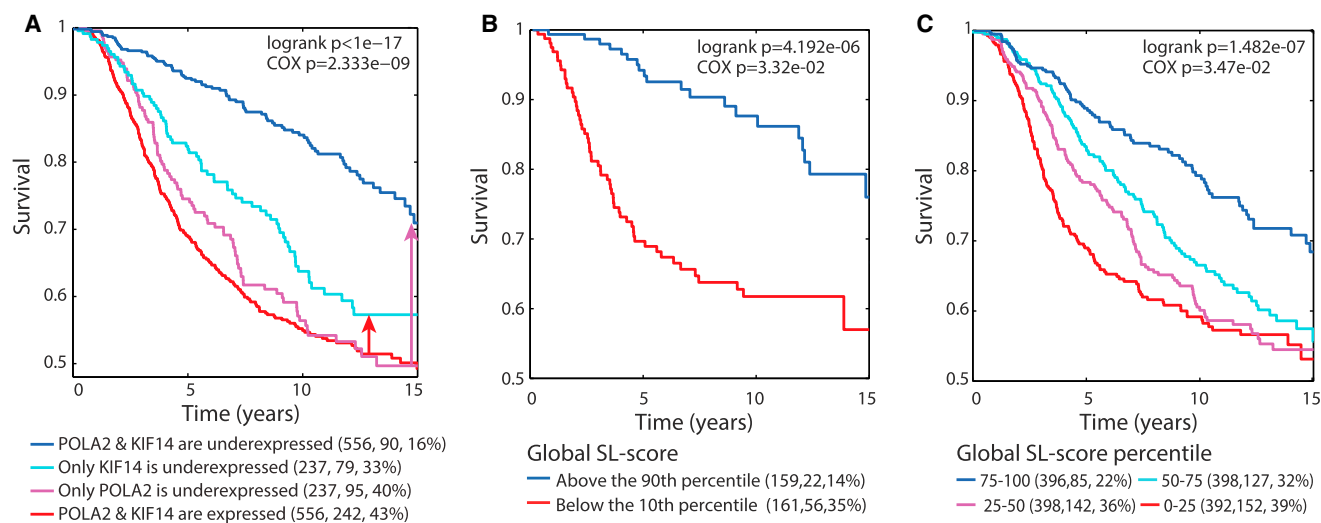
Figure 6. Predicting Clinical Prognosis Based on the SL Network

In parenthesis next to the name of each group are the number of patients and the number and percentage of deaths in that group.

(A) The KM plot obtained when dividing the breast cancer samples according to the expression of *POLA2* and *KIF14* (the most predictive SL pair in terms of breast cancer survival). The red and purple arrows point to the estimated effect of *KIF14* underexpression, in the context of *POLA2* expression and underexpression, respectively.

(B) KM plots depicting the survival of samples that counderexpressed a high number of SL pairs (global SL score above the 90th percentile, in blue) and of samples that counderexpressed a low number of SL pairs (global SL score below the 10th percentile, in red).

(C) The KM plots depict the survival of breast cancer patients uniformly divided into four groups according to their global SL score. As predicted, higher global SL scores are characterized with better 15 year survival.

See also Figure S5 and Table S7.

(Figure 7E; Extended Experimental Procedures). Among these interactions is the predicted SDL interaction between *EGFR* and *IGFBP3*, whose overexpression should accordingly induce sensitivity to drugs targeting *EGFR*. Reassuringly, it has been shown that *IGFBP3* is underexpressed in Gefitinib-resistant cells, and the addition of recombinant *IGFBP3* restored the ability of Gefitinib to inhibit cell growth (Guix et al., 2008). Another interesting example is the predicted SDL interaction between *PARP1* and *MDC1*. The latter contains two *BRCA1* C-terminal motifs and also regulates *BRCA1* localization and phosphorylation in DNA damage checkpoint control (Lou et al., 2003). Indeed, *BRCA1/BRCA2* are known to be synthetically lethal with *PARP1* (Lord et al., 2008).

In a manner analogous to that described earlier for predicting gene essentiality, we utilized the SDL network to build supervised neural network predictors of drug efficacies in cancer cell lines (Extended Experimental Procedures). Using only 53 features, we predicted drug efficacies with Spearman correlation coefficients of 0.721 and 0.547 and p values $<1 \times 10^{-350}$ for the CGP and CTRP data, respectively (Figures 7F–7I). We further examined our SDL-based predictors by analyzing results of a large pharmacological screen carried out recently by the same team as CTRP. In this study, the efficacies of ~500 compounds were measured across >850 cancer cell lines (P.A.C., personal communication). One hundred and twenty six of the tested compounds have at least one target in the SDL network, enabling to predict the response to their administration. Based the SDL network and the genomic profiles of these cell lines (Barretina et al., 2012), we predicted

the efficacies of these drugs by using the unsupervised and supervised predictors (trained on the CTRP data). The SDL-based predictors obtained significant predictions (p value < 0.05) of drug efficacy for 83 (65.87%) and 70 (55.6%) drugs, when applying the unsupervised or supervised approach, respectively.

DISCUSSION

DAISY is a genome scale, data-driven, approach for the identification of cancer SL and SDL interactions. As shown, DAISY successfully captures the results obtained in key large scale experimental studies exploring SLs in cancer, identifies valid SL interactions, and enables to predict gene essentiality, drug efficacy, and clinical prognosis in cancer.

DAISY presents a complementary effort to genetic and chemical screens, narrowing down the number of gene pairs that need to be examined experimentally to detect SL and SDL interactions in cancer. Based on the ROC curve presented in Figure 2A, an experimental screen for discovering SL interactions could be designed to check the SL pairs predicted by DAISY such that 5%, 25%, 50%, or 70% of all the SL interactions that are out there will be detected by examining only 0.25%, 4%, 14%, or 24% of all possible gene pairs, respectively. Hence, testing only the most confident predictions will enable to find up to 20 times more SL pairs than expected by random. Likewise, by applying DAISY to design an siRNA screen for detecting the SL interactions of *VHL* we identified almost four times as many SL interactions compared to a screen that was

Figure 7. The SDL Network Predicts the Efficacy of Anticancer Drugs in Cancer Cell Lines

(A and B) The prediction signal obtained when predicting the response of cancer cell lines to different drugs and testing the predictions based on (A) the CGP data and (B) the CTRP data. Drugs that are significantly predicted via the unsupervised SDL-based predictor are colored in red.

(C) The fraction of significantly predicted drugs when considering only drugs whose targets have an above-threshold number of SDL interactions (unsupervised).

(D) The IC_{50} (left) and area-under-dose-curve (right) of drugs decrease in cell lines where their target(s) have an increasing number of overexpressed SDL partners (lower values denote higher efficacy).

(E) A subnetwork of the SDL network that enables to significantly predict the sensitivity to ten anticancer drugs. Each node denotes a gene: a drug target (purple) or an SDL partner of a drug target (pink). Each edge represents an SDL interaction. Self-loops denote that the gene is an SDL-partner of itself (see Supplemental Information for further discussion and results concerning such SDL interactions). The edge color denotes the predictive power of the SDL interaction (blue, and red denote a p value lower than 1×10^{-3}, and 0.05, respectively). The significance of the prediction based on the entire set of SDL interactions is written next to the name of the drug, in parenthesis; predictions that were tested based on the CGP and CTRP data are written in black and blue, respectively.

(F–I) The drug efficacy predictions obtained by the supervised SDL-based predictors. (F) The predicted versus experimental IC_{50} log values of 41 drugs measured across 414 cancer cell lines (CGP data). (G) The predicted versus experimental area-under-dose-curve of 50 drugs measured across 241 cancer cell lines (CTRP data). (H–I) For each cancer cell line we computed the Spearman correlation between the measured and predicted efficacies of different drugs in it. The histograms show the distribution of these correlation coefficients across the different cancer cell lines in (H) the CGP and (I) the CTRP data. The blue lines mark the median correlation coefficient.

See also Figure S6 and Table S8.

designed by applying biological reasoning. In light of these results DAISY could facilitate a more rapid and rational discovery of SL interactions in cancer by guiding focused experimental screens.

Nonetheless, DAISY has several limitations one needs to account for. First, it is restricted to the identification of SL interactions in cancer, as it is based on unique cancer-specific data that captures the genomic instability of cancer cells (e.g., SCNA). As such DAISY cannot be tested by applying it to identify SL interactions in model microorganisms as yeast. Second, DAISY identifies SL interactions based on large scale genomic data and shRNA screens, which are at times noisy and inaccurate (Bhinder et al., 2014). Third, as DAISY is based on the identification of gene inactivation, additional mechanisms of gene inactivation, such as epigenetic and posttranscriptional regulation, should be accounted for in the future. Fourth, the genomic location of genes may result in false-negative and false-positive predictions of SL and SDL interactions, respectively (see Supplemental Information for further analysis). Last, the ability of the SL network to accurately predict gene essentiality in vivo remains to be determined.

We have shown that SL and SDL interactions have a marked cumulative effect (Figures 5B, 5C, and 7D). Thus, a gene can form a useful drug target due to the (possibly partial) inactivation or overactivation of several of its SL or SDL partners, respectively. SL-based treatment can therefore be especially effective in targeting genetically unstable tumors that harbor many gene deletions and amplifications. Furthermore, a drug may be able to kill a broad array of genomically heterogeneous cells, each sensitive to the drug due to the inactivity (overactivity) of a different subset of the SL (SDL) partners of the drug targets. Targeting a gene with many inactive SL and/or overactive SDL partners may hence counteract the development of treatment resistance, especially if the SL/SDL partners reside on different chromosomes or in distant genomic locations. Moreover, SL-based treatment can induce the reactivation of a tumor suppressor or the inactivation of an oncogene by targeting its SL or SDL pair, respectively.

Four main translational challenges could potentially be tackled by utilizing SL and SDL networks: (1) ranking existing treatments for a given patient based on the genomic characteristics of the tumor, as initially shown here in cell lines; (2) repurposing approved drugs that are currently used to treat other diseases to treat cancer, as shown here for treating a *VHL*-deficient cancer; (3) systematically identifying new drug targets; and (4) predicting cancer prognosis, as shown here for breast cancer. Taken together, SL and SDL network-based analysis combined with personalized genomics can provide an important future tool for assessing response to treatment and for developing more selective and effective personalized therapeutics.

EXPERIMENTAL PROCEDURES

Description of DAISY

DAISY identifies candidate SL and SDL interactions by applying three separate statistical inference procedures. Each procedure has its own input and outputs a set of candidate SL or SDL pairs. Gene pairs that are identified as candidate

SL or SDL pairs by all three procedures are identified by DAISY as SL or SDL pairs, respectively. The three inference procedures are described below (comments in parenthesis refer to changes made to identify SDL pairs):

(1) The genomic SoF procedure analyzes a set of input data sets denoted as $SoF_{data\ sets}$. Each data set includes SCNA profiles of cancer samples and optionally their mRNA and somatic mutation profiles. For every pair of genes, denoted as A and B, and every data set S in $SoF_{data\ sets}$, a Wilcoxon rank sum test is conducted to examine if gene B has a significantly higher SCNA level in samples in which gene A is inactive (overactive) than in the rest of the samples. The output consists of gene pairs that, according to at least one of the data sets in SoF data sets, pass the test described above in a statistically significant manner (a Wilcoxon rank sum p value <0.05 following Bonferroni correction for multiple hypotheses testing).

(2) The shRNA-based functional examination procedure analyzes a set of data sets denoted as $shRNA_{data\ sets}$. Each data set includes the results obtained in a gene essentiality (shRNA) screen together with the SCNA and gene expression profiles of the cancer cell lines examined in that screen. For every pair of genes, denoted as A and B, and every data sets S in $shRNA_{data\ sets}$, a Wilcoxon rank sum test is conducted to examine if gene B has significantly lower shRNA scores in samples in which gene A is inactive (overactive) than in the rest of the samples (the lower the shRNA score is, the more essential the gene is). The output consists of gene pairs that, according to at least one of the data sets in $shRNA_{data\ sets}$, pass the test described above in a statistically significant manner (a Wilcoxon rank sum p value <0.05).

(3) The pairwise gene coexpression procedure is given a set of transcriptomic data sets of cancer samples and returns gene pair whose expression, in at least one of the data sets, is significantly positively correlated (a Spearman correlation coefficient $\geq R_{min}$ and a p value < 0.05 following Bonferroni correction for multiple hypotheses testing).

The candidate SL or SDL pairs that are identified in the first and third procedures are obtained with highly stringent statistical cutoffs, a p value <0.05 following Bonferroni correction. The data obtained in shRNA screens has a low statistical power and is hence utilized (in the second procedure) only to further refine the already highly statistically significant SL and SDL sets identified in the first and third procedures.

The first and second procedures are based on the detection of gene inactivation and overactivation in the samples analyzed. A gene is defined as inactive in a sample if it is underexpressed and its SCNA is below −0.3 or if it is mutated with a deleterious mutation. The latter refers to nonsense and frame-shift mutations. Likewise, a gene is defined as overactive in a sample if it is overexpressed and its SCNA is above 0.3. Of note, the SCNA parameters (−0.3 and 0.3) used here are more stringent cutoffs compared to those used in the literature to define gene amplification and deletion (Beroukhim et al., 2010). A gene is defined as underexpressed in a given sample if its expression level is below the 10th percentile of its expression levels across all samples in the data set, and similarly, as overexpressed if its expression level is above its 90th percentile. In the third procedure we set R_{min} to 0.5.

To find the candidate pairs and construct the SL and SDL networks, we applied DAISY with the data sets listed in Table S1 and traversed over all ~535 million gene pairings. To do so efficiently, DAISY was implemented and run on the HTcondor architecture, which enables parallel computing (Thain et al., 2005).

Network Availability and Visualization

Interactive maps of the networks are accessible through http://www.cs.tau.ac.il/~livnatje/SL_network.zip and can be explored using the freely available Cytoscape software (Cline et al., 2007). The maps include different gene properties and annotations, as well as alternative views that dissect the network hubs or genes with specific characteristics. We clustered the SL and SDL networks by applying the Girvan-Newman fast greedy algorithm as implemented by the GLay Cytoscape plug-in (Morris et al., 2011; Su et al., 2010) and performed gene annotation enrichment analysis for every network and every network cluster via DAVID (Huang et al., 2009).

REFERENCES

Ashworth, A., Lord, C.J., and Reis-Filho, J.S. (2011). Genetic interactions in cancer progression and treatment. Cell *145*, 30–38.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, 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.

Bassik, M.C., Kampmann, M., Lebbink, R.J., Wang, S., Hein, M.Y., Poser, I., Weibezahn, J., Horlbeck, M.A., Chen, S., Mann, M., et al. (2013). A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility. Cell *152*, 909–922.

Basu, A., Bodycombe, N.E., Cheah, J.H., Price, E.V., Liu, K., Schaefer, G.I., Ebright, R.Y., Stewart, M.L., Ito, D., Wang, S., et al. (2013). An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell *154*, 1151–1161.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899–905.

Bhinder, B., Shum, D., and Djaballah, H. (2014). Comparative analysis of RNAi screening technologies at genome-scale reveals an inherent processing inefficiency of the plasmid-based shRNA hairpin. Comb. Chem. High Throughput Screen. *17*, 98–113.

Bilal, E., Dutkowski, J., Guinney, J., Jang, I.S., Logsdon, B.A., Pandey, G., Sauerwine, B.A., Shimoni, Y., Moen Vollan, H.K., Mecham, B.H., et al. (2013). Improving breast cancer survival analysis through competition-based multidimensional modeling. PLoS Comput. Biol. *9*, e1003047.

Bommi-Reddy, A., Almeciga, I., Sawyer, J., Geisen, C., Li, W., Harlow, E., Kaelin, W.G., Jr., and Grueneberg, D.A. (2008). Kinase requirements in human cells: III. Altered kinase requirements in VHL-/- cancer cells detected in a pilot synthetic lethal screen. Proc. Natl. Acad. Sci. USA *105*, 16484–16489.

Brough, R., Frankum, J.R., Costa-Cabral, S., Lord, C.J., and Ashworth, A. (2011). Searching for synthetic lethality in cancer. Curr. Opin. Genet. Dev. *21*, 34–41.

Byrne, A.B., Weirauch, M.T., Wong, V., Koeva, M., Dixon, S.J., Stuart, J.M., and Roy, P.J. (2007). A global analysis of genetic interactions in Caenorhabditis elegans. J. Biol. *6*, 8.

Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.

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.

Chipman, K.C., and Singh, A.K. (2009). Predicting genetic interactions with random walks on biological networks. BMC Bioinformatics *10*, 17.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., et al. (2007). Integration of biological networks and gene expression data using Cytoscape. Nat. Protoc. *2*, 2366–2382.

Conde-Pueyo, N., Munteanu, A., Solé, R.V., and Rodríguez-Caso, C. (2009). Human synthetic lethal inference as potential anti-cancer target gene detection. BMC Syst. Biol. *3*, 116.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., et al. (2010). The genetic landscape of a cell. Science *327*, 425–431.

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.

Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E., and Shlomi, T. (2011). Predicting selective drug targets in cancer through metabolic networks. Mol. Syst. Biol. *7*, 501.

Frezza, C., Zheng, L., Folger, O., Rajagopalan, K.N., MacKenzie, E.D., Jerby, L., Micaroni, M., Chaneton, B., Adam, J., Hedley, A., et al. (2011). Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. Nature *477*, 225–228.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature *483*, 570–575.

Guix, M., Faber, A.C., Wang, S.E., Olivares, M.G., Song, Y., Qu, S., Rinehart, C., Seidel, B., Yee, D., Arteaga, C.L., and Engelman, J.A. (2008). Acquired resistance to EGFR tyrosine kinase inhibitors in cancer cells is mediated by loss of IGF-binding proteins. J. Clin. Invest. *118*, 2609–2619.

Hartwell, L.H., Szankasi, P., Roberts, C.J., Murray, A.W., and Friend, S.H. (1997). Integrating genetic approaches into the discovery of anticancer drugs. Science *278*, 1064–1068.

Huang, W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.

Kelley, R., and Ideker, T. (2005). Systematic interpretation of genetic interactions using protein networks. Nat. Biotechnol. *23*, 561–566.

Laufer, C., Fischer, B., Billmann, M., Huber, W., and Boutros, M. (2013). Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. Nat. Methods *10*, 427–431.

Lord, C.J., McDonald, S., Swift, S., Turner, N.C., and Ashworth, A. (2008). A high-throughput RNA interference screen for DNA repair determinants of PARP inhibitor sensitivity. DNA Repair (Amst.) 7, 2010–2019.

Lou, Z., Chini, C.C.S., Minter-Dykhouse, K., and Chen, J. (2003). Mediator of DNA damage checkpoint protein 1 regulates BRCA1 localization and phosphorylation in DNA damage checkpoint control. J. Biol. Chem. 278, 13599–13602.

Lu, X., Kensche, P.R., Huynen, M.A., and Notebaart, R.A. (2013). Genome evolution predicts genetic interactions in protein complexes and reveals cancer drug targets. Nat. Commun. 4, 2124.

Luo, B., Cheung, H.W., Subramanian, A., Sharifnia, T., Okamoto, M., Yang, X., Hinkle, G., Boehm, J.S., Beroukhim, R., Weir, B.A., et al. (2008). Highly parallel identification of essential genes in cancer cells. Proc. Natl. Acad. Sci. USA 105, 20380–20385.

Luo, J., Emanuele, M.J., Li, D., Creighton, C.J., Schlabach, M.R., Westbrook, T.F., Wong, K.-K., and Elledge, S.J. (2009). A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. Cell 137, 835–848.

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

Martin, S.A., McCarthy, A., Barber, L.J., Burgess, D.J., Parry, S., Lord, C.J., and Ashworth, A. (2009). Methotrexate induces oxidative DNA damage and is selectively lethal to tumour cells with defects in the DNA mismatch repair gene MSH2. EMBO Mol. Med. 1, 323–337.

Morris, J.H., Apeltsin, L., Newman, A.M., Baumbach, J., Wittkop, T., Su, G., Bader, G.D., and Ferrin, T.E. (2011). clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics 12, 436.

Sajesh, B.V., Guppy, B.J., and McManus, K.J. (2013). Synthetic genetic targeting of genome instability in cancer. Cancers 5, 739–761.

Steckel, M., Molina-Arcas, M., Weigelt, B., Marani, M., Warne, P.H., Kuznetsov, H., Kelly, G., Saunders, B., Howell, M., Downward, J., and Hancock, D.C. (2012). Determination of synthetic lethal interactions in KRAS oncogene-dependent cancer cells reveals novel therapeutic targeting strategies. Cell Res. 22, 1227–1245.

Su, G., Kuchinsky, A., Morris, J.H., States, D.J., and Meng, F. (2010). GLay: community structure analysis of biological networks. Bioinformatics 26, 3135–3137.

Szappanos, B., Kovács, K., Szamecz, B., Honti, F., Costanzo, M., Baryshnikova, A., Gelius-Dietrich, G., Lercher, M.J., Jelasity, M., Myers, C.L., et al. (2011). An integrated approach to characterize genetic interaction networks in yeast metabolism. Nat. Genet. 43, 656–662.

Szczurek, E., Misra, N., and Vingron, M. (2013). Synthetic sickness or lethality points at candidate combination therapy targets in glioblastoma. Int. J. Cancer 133, 2123–2132.

Thain, D., Tannenbaum, T., and Livny, M. (2005). Distributed computing in practice: the Condor experience. Concurr. Comp-Pract. E. 17, 323–356.

Turner, N.C., Lord, C.J., Iorns, E., Brough, R., Swift, S., Elliott, R., Rayter, S., Tutt, A.N., and Ashworth, A. (2008). A synthetic lethal siRNA screen identifying genes mediating sensitivity to a PARP inhibitor. EMBO J. 27, 1368–1377.

Typas, A., Nichols, R.J., Siegele, D.A., Shales, M., Collins, S.R., Lim, B., Braberg, H., Yamamoto, N., Takeuchi, R., Wanner, B.L., et al. (2008). High-throughput, quantitative analyses of genetic interactions in E. coli. Nat. Methods 5, 781–787.

Waldman, Y.Y., Geiger, T., and Ruppin, E. (2013). A genome-wide systematic analysis reveals different and predictive proliferation expression signatures of cancerous vs. non-cancerous cells. PLoS Genet. 9, e1003806.

Wong, S.L., Zhang, L.V., Tong, A.H.Y., Li, Z., Goldberg, D.S., King, O.D., Lesage, G., Vidal, M., Andrews, B., Bussey, H., et al. (2004). Combining biological networks to predict genetic interactions. Proc. Natl. Acad. Sci. USA 101, 15682–15687.