SUPPLEMENTARY FIGURES

Supplementary Figure 1: Variant filtering decision tree. Variants are filtered according to certain criteria to select cancer relevant variants (Type 2). Variants with an Exome-Sequencing Project (ESP5400) frequency $\geq 0.00009$ or 1000 Genome frequency higher than 0.0005 were removed from the data. Next, variants with a dataset frequency higher than 0.2 (13 in 61 cell lines) were removed. At this step, variants already in the Cosmic database for 2 or more samples were placed in the Type 2 bin directly. Next, we used blat to identify possible false-positive variants. We selected variants that only matched primarily to the variant position +/- 35 bp. The remaining variants that are in regions of segmental duplications and in dbSNP version 135 were also removed.

Supplementary Figure 2: Distribution of variants by annotation. (A) The fraction of variants for a given genomic annotation was plotted as the mean for all 61 samples. Error bars represent the standard deviation of 61 samples. Type 1 and Type 2 variants are plotted separately as open and black bars, respectively. (B) The same plot in panel A was generated for coding variant annotations.

Supplementary Figure 3: The NCI-60 cell lines. The list of cell lines, tissue of origin and MSI status as reported in the COSMIC database.
Supplementary Figure 4: Distribution of variants by type. For each cell line the fraction of variants that are either a single nucleotide variant (SNP) or one to five or more base pair deletions or insertions were plotted as a bar graph for (A) Type 1, (B) Type 2 variants only.

Supplementary Figure 5: Phylogenic conservation of POLε-P286. Multiple sequence alignment of POLε homologs were generated using CLUSTAL X 2.1. P286 is a 100% conserved residue in the exonuclease domain, which is mutated in the HCC2998 cell line.

Supplementary Figure 6: A snapshot of the WES dataset in Variant Analysis by Ingenuity Systems (http://www.ingenuity.com/NCI60_WES). The dataset is freely accessible by users for data query/manipulation.

Supplementary Figure 7: Pattern comparison for cell with wild-type TP53 status. A. Differential activity profiles of RITA (Thiosphene, NSC-652287) and Nutlin-3. Mean-centered graph for drug activities using the CellMiner z-score tool (http://discover.nci.nih.gov/cellminer) (1). B. The figures shows a snapshot of the Excel spreadsheet obtained using the CellMiner pattern comparison tool (http://discover.nci.nih.gov/cellminer) (1).

Supplementary Figure 8: Receiver operating curves (ROC) for cross-validated drug predictors. Cross-validated ROC curves are shown for each cell line grouped by tissue-of-origin. The legend reports the area under the ROC curve (AUC) for each cell line and the number of in-active drugs (n1) and active drugs (n2). In-active drugs for a cell line are defined as having
a normalized log10 GI50 greater than 0.5 and active drugs have a normalized log10 GI50 less than -0.5.

**Supplementary Figure 9: Volcano plots for cancer genes of interest.** For 46 genes of interest (one per page) volcano plots are provided. Blue marks compounds that passed the significance cut-off (p-value < $10^{-4}$) and an imposed cut-off on difference in log GI50 between mutated and wild-type groups (> 0.5). The NSC numbers for these drugs are provided in Supplemental Table 3 for each gene of interest.

**Supplementary Figure 10: Volcano plots for FDA approved cancer drugs of interest.** For 28 drugs of interest (one per page) volcano plots are provided. Blue marks genes that passed the significance cut-off (p-value < $10^{-4}$) and an imposed cut-off on difference in log GI50 between mutated and wild-type groups (> 0.5).

**Supplementary Table 1: Additional data on various NSC compounds.**

**Supplementary Table 2: Top 5 lowest GI50 drugs predicted by the super-learner algorithm for the NCI-60 panel.** For each cell line the 5 drugs with the lowest normalized predicted log GI50 is presented. PredScore is the normalized predicted log GI50 for the cell line based on the model with that cell line removed. The ObsScore is the observed normalized log GI50 for the cell line.
Supplementary Table 3: List of NSC compounds for the significant data points in Figure S10. For each of the 46 genes with volcano plots in Supplement Figure 9 the statistically significant compounds are listed. The compounds are identified by their NSC number. DM is the difference in mean log10 GI50 between the variant subgroup and the non-variant subgroup for that gene.

Supplementary References