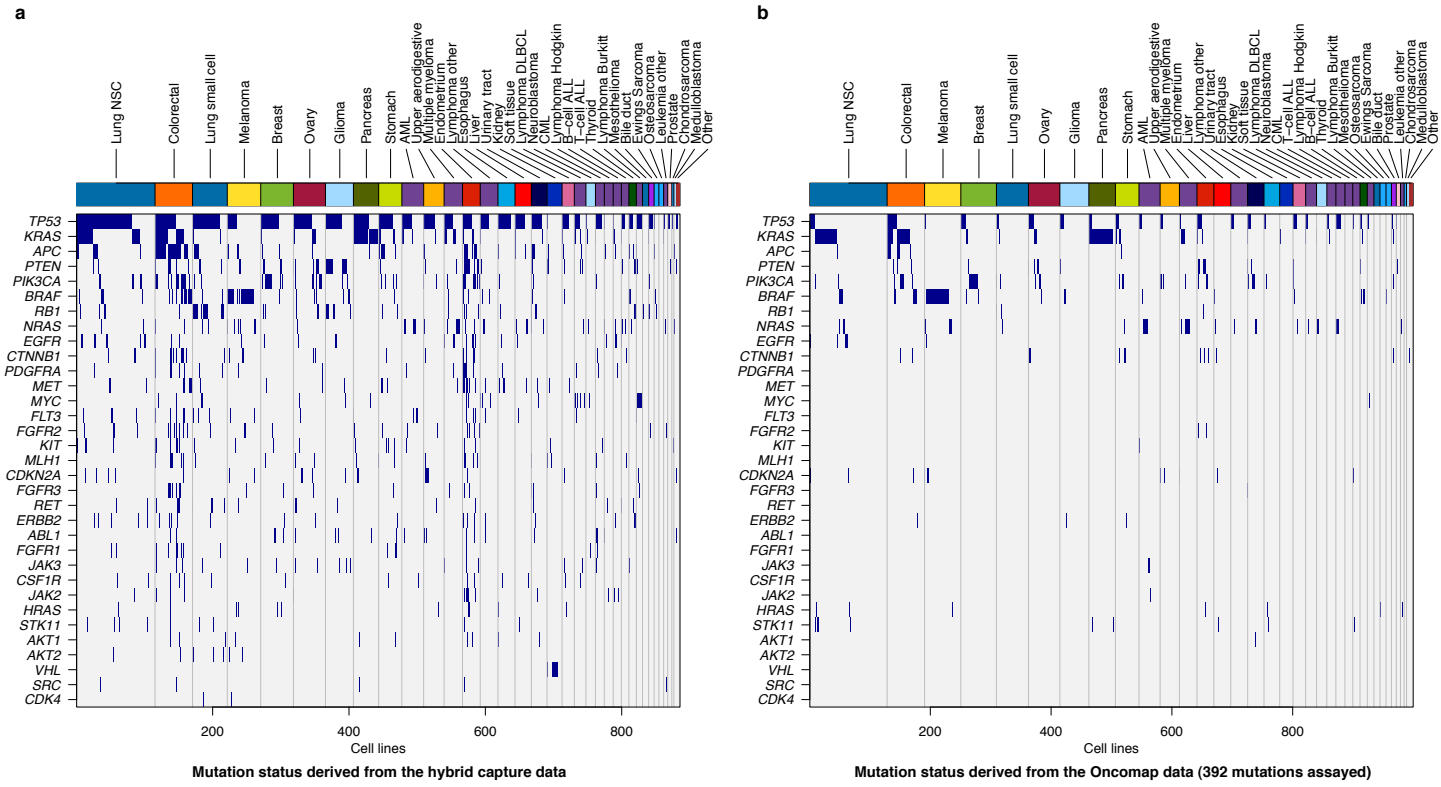
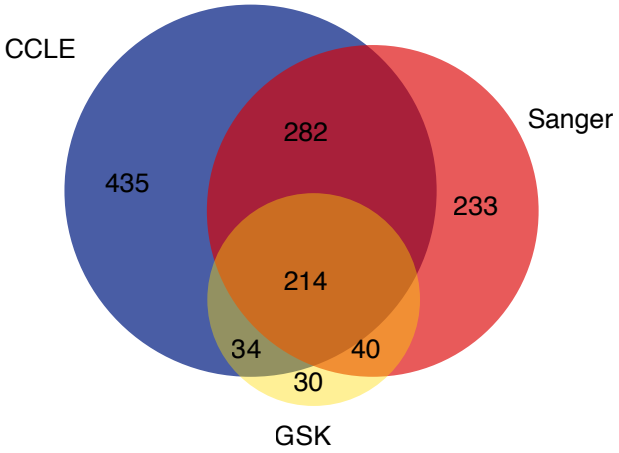
SUPPLEMENTARY INFORMATION

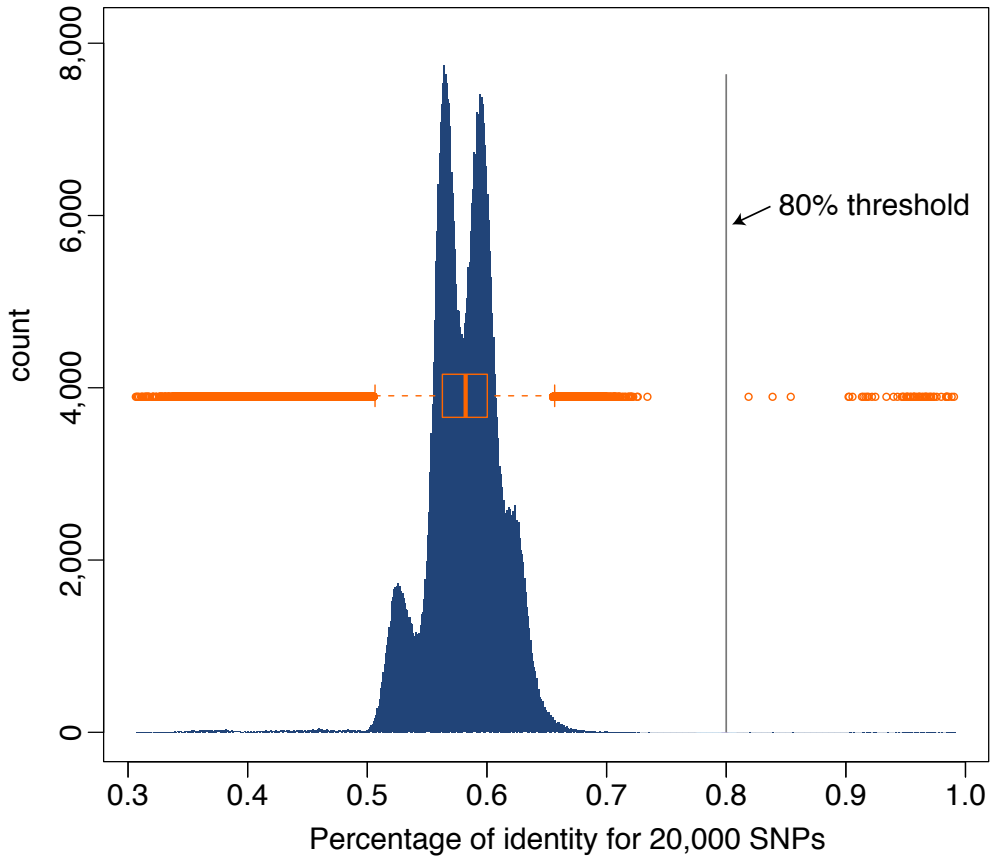
doi:10.1038/nature11003



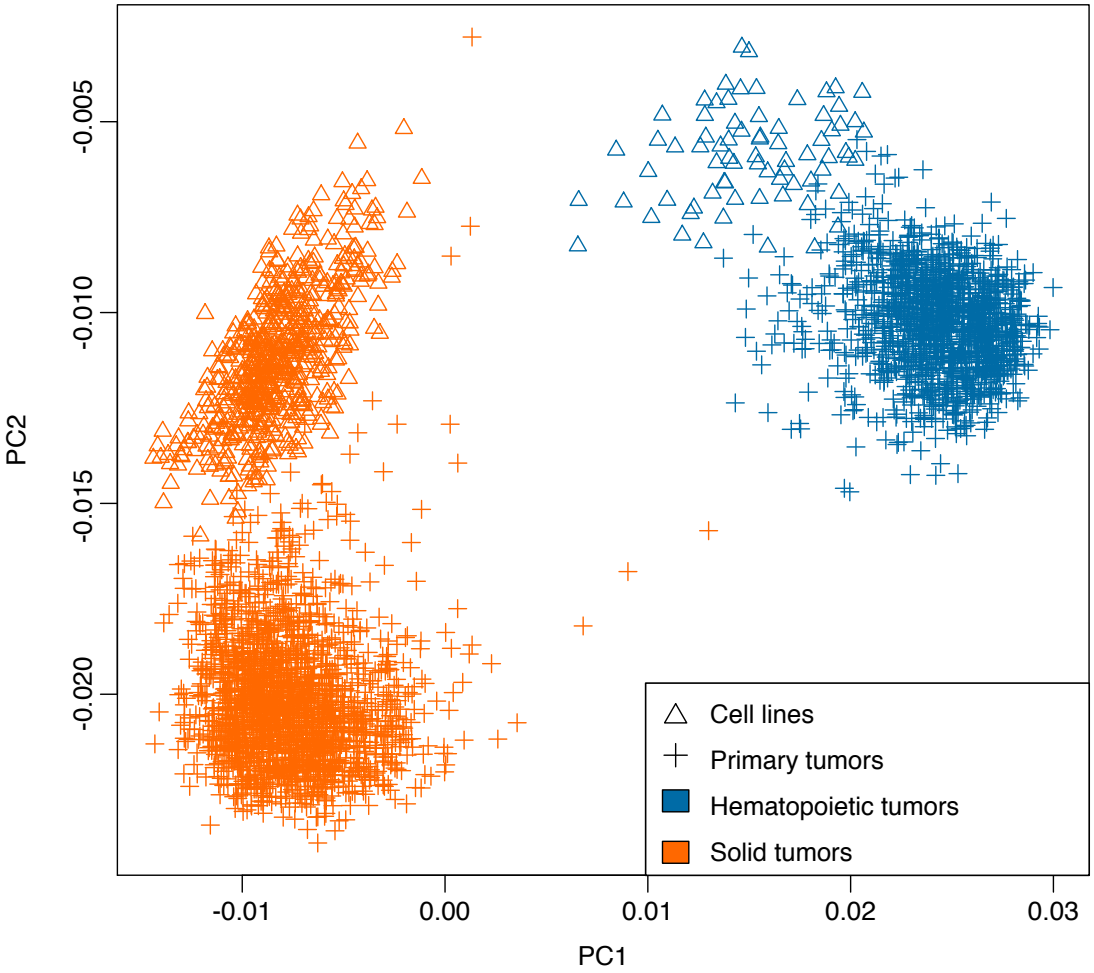
Supplementary Figure 1: Matrix-representation of point mutation data derived from hybrid capture/sequencing (**a**) or OncoMap analysis (**b**) across all CCLE cell lines. Blue bars indicate detection of at least one mutation in the relevant gene. Cell lines are sorted by lineage and cancer gene mutation frequency. The genes are sorted by mutation frequency in the hybrid capture data.



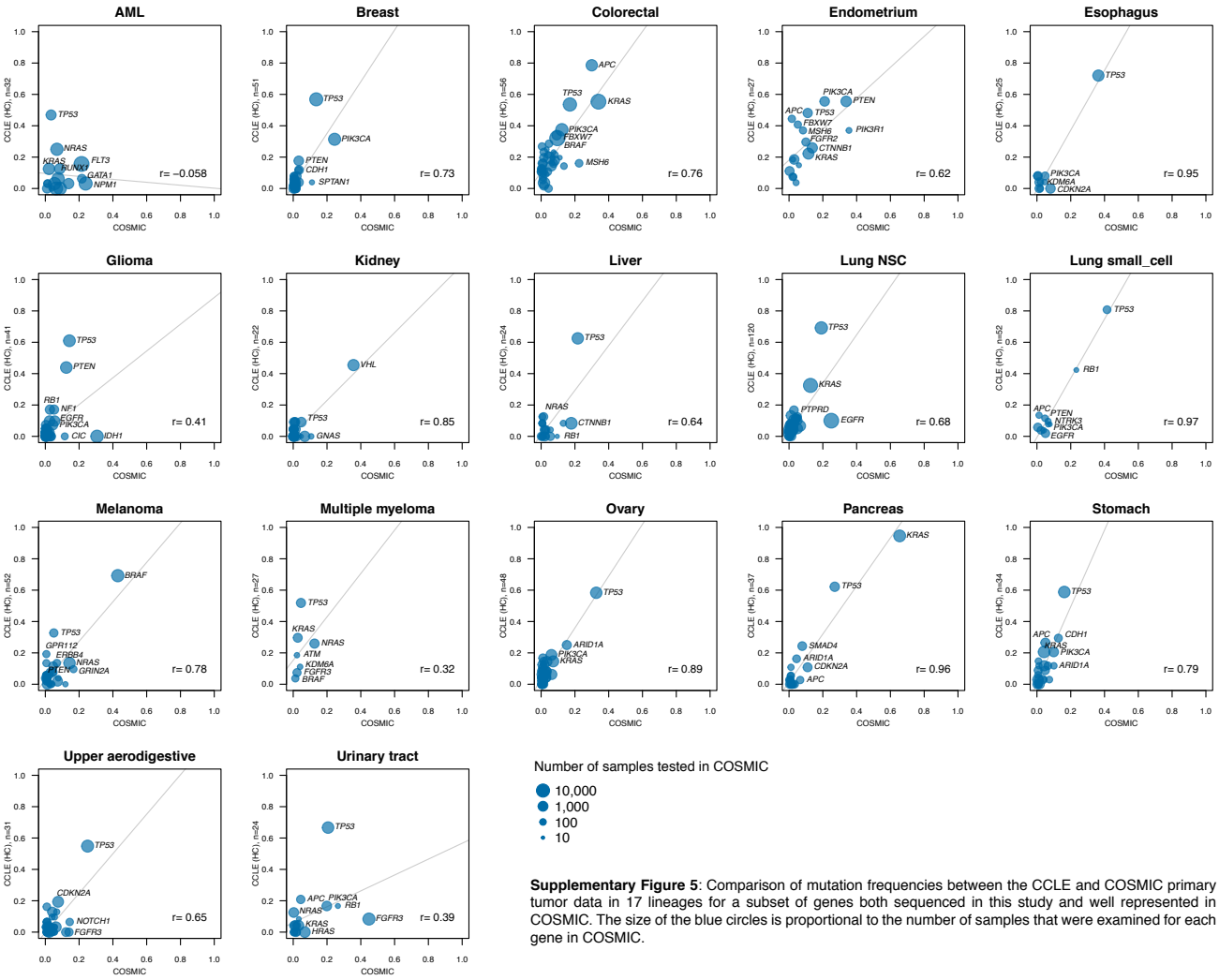
Supplementary Figure 2: Venn diagram of overlap between the cell lines in the CCLE, the Sanger Institute's Cancer Cell Line Project (http://www.sanger.ac.uk/genetics/CGP/CellLines/) and the GlaxoSmithKline (GSK) Cancer Cell Line Genomic Profiling dataset (https://cabig.nci.nih.gov/caArray_GSKdata/).



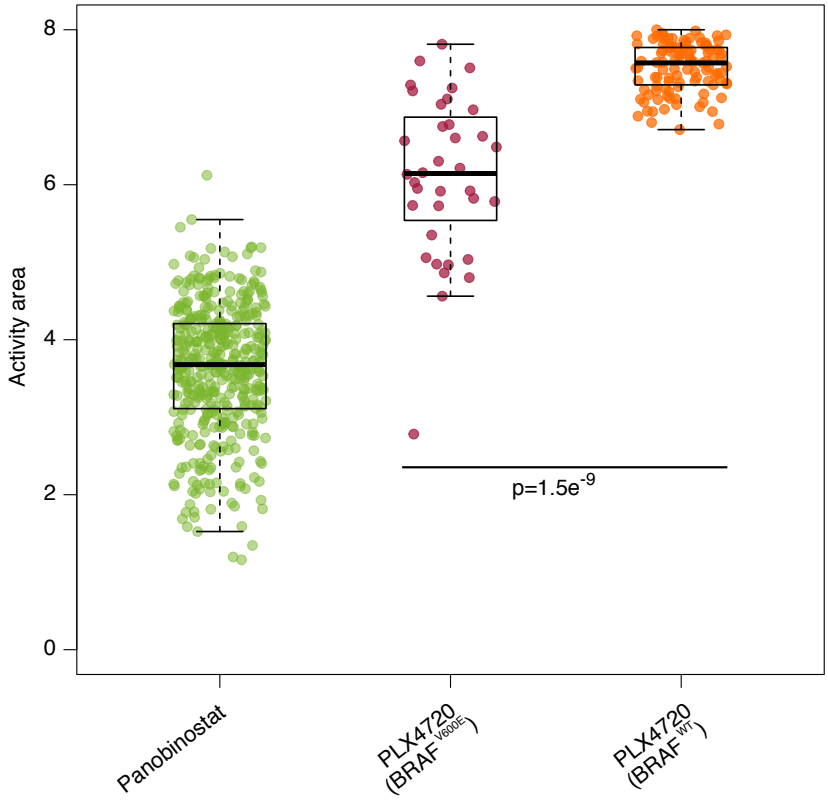
Supplementary Figure 3: Distribution of the percentages of SNP identity across cell lines. The boxplot (orange) corresponds to the same values shown as a histogram in blue. Values above 80% identity denote cell line pairs presumed to derive from the same individual.



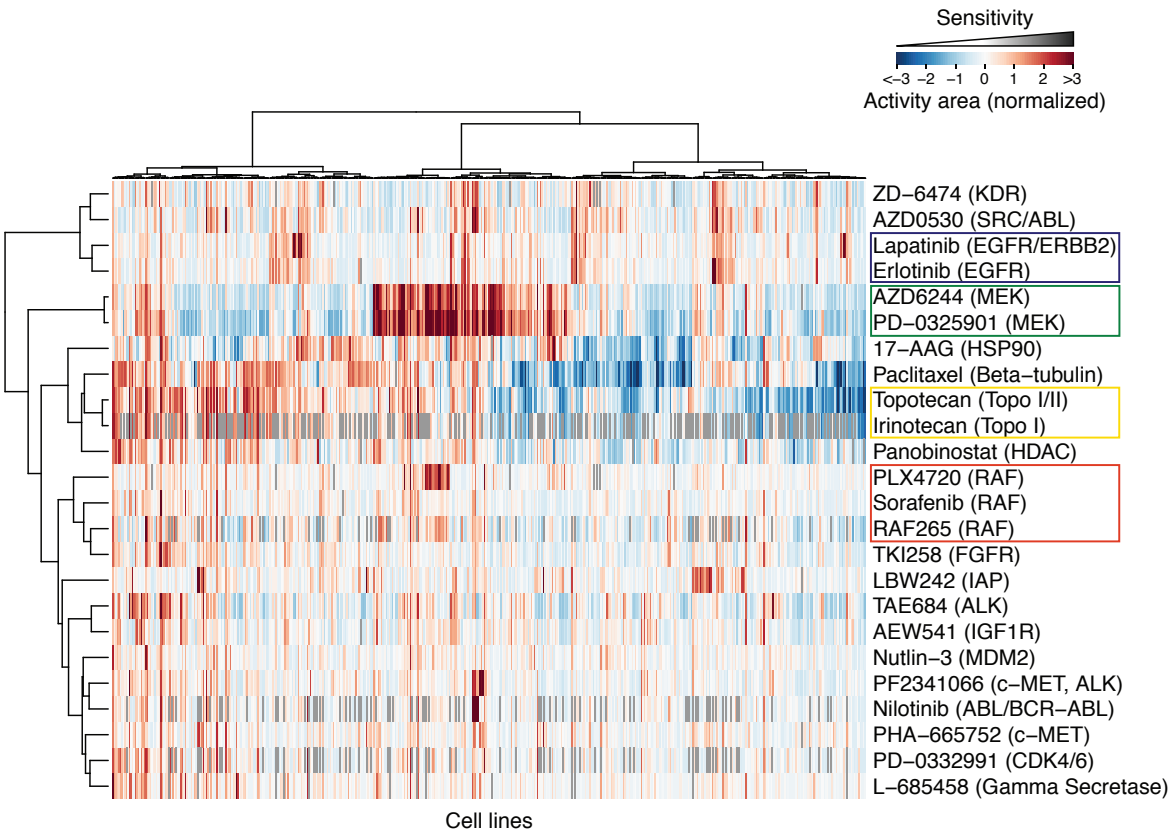
Supplementary Figure 4: Principal component analysis of expression data from cell lines and primary tumors, for the 1,000 most varying genes. Those derived from solid tumors are shown in orange, and those corresponding to hematopoietic lineages are shown in blue. This figure indicates that there was no cross-contamination between hematopoietic and solid tumor-derived cell lines. Expression data was also used to ascertain the original lineage of certain cell lines.



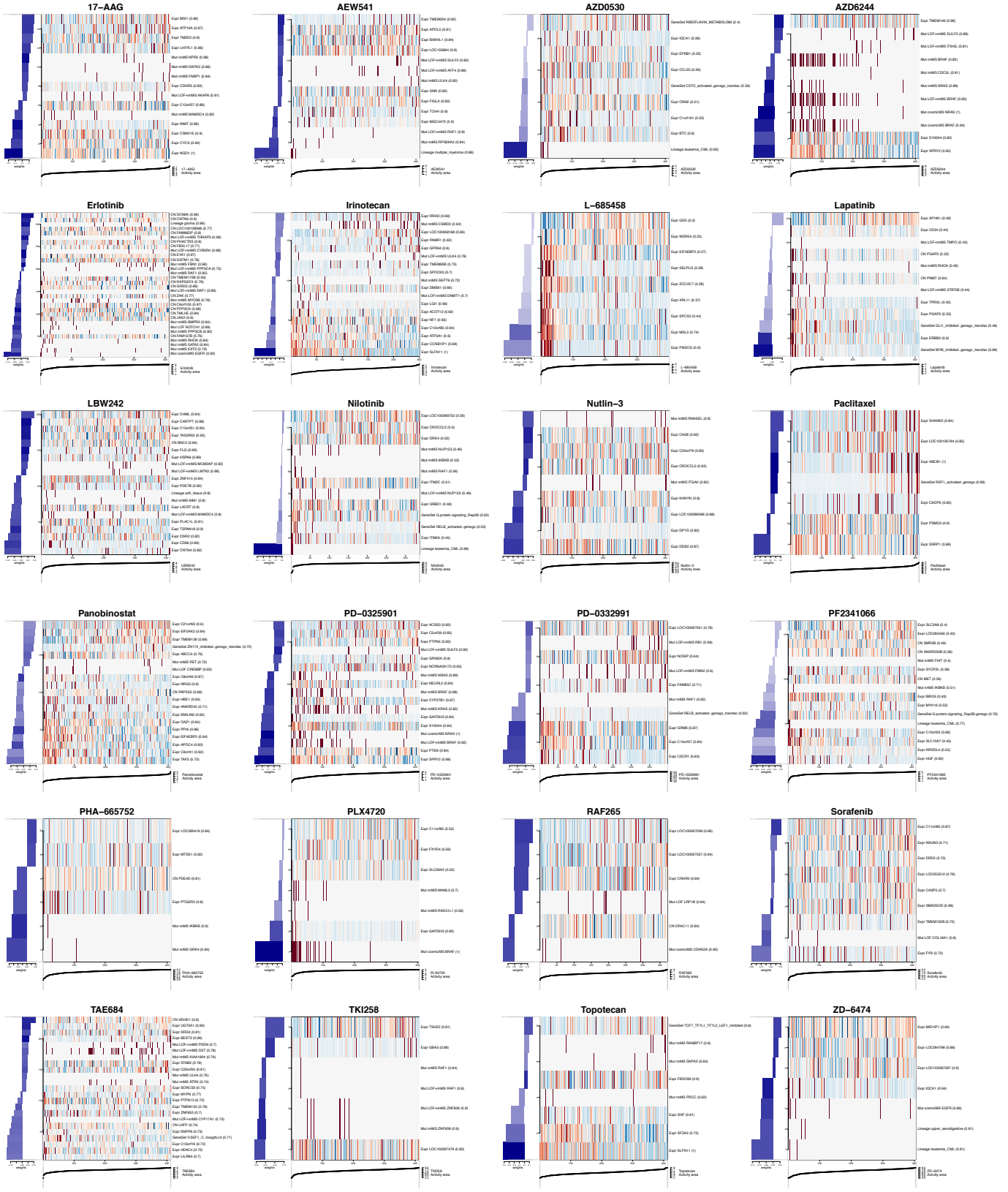
Supplementary Figure 5: Comparison of mutation frequencies between the CCLE and COSMIC primary tumor data in 17 lineages for a subset of genes both sequenced in this study and well represented in COSMIC. The size of the blue circles is proportional to the number of samples that were examined for each gene in COSMIC.



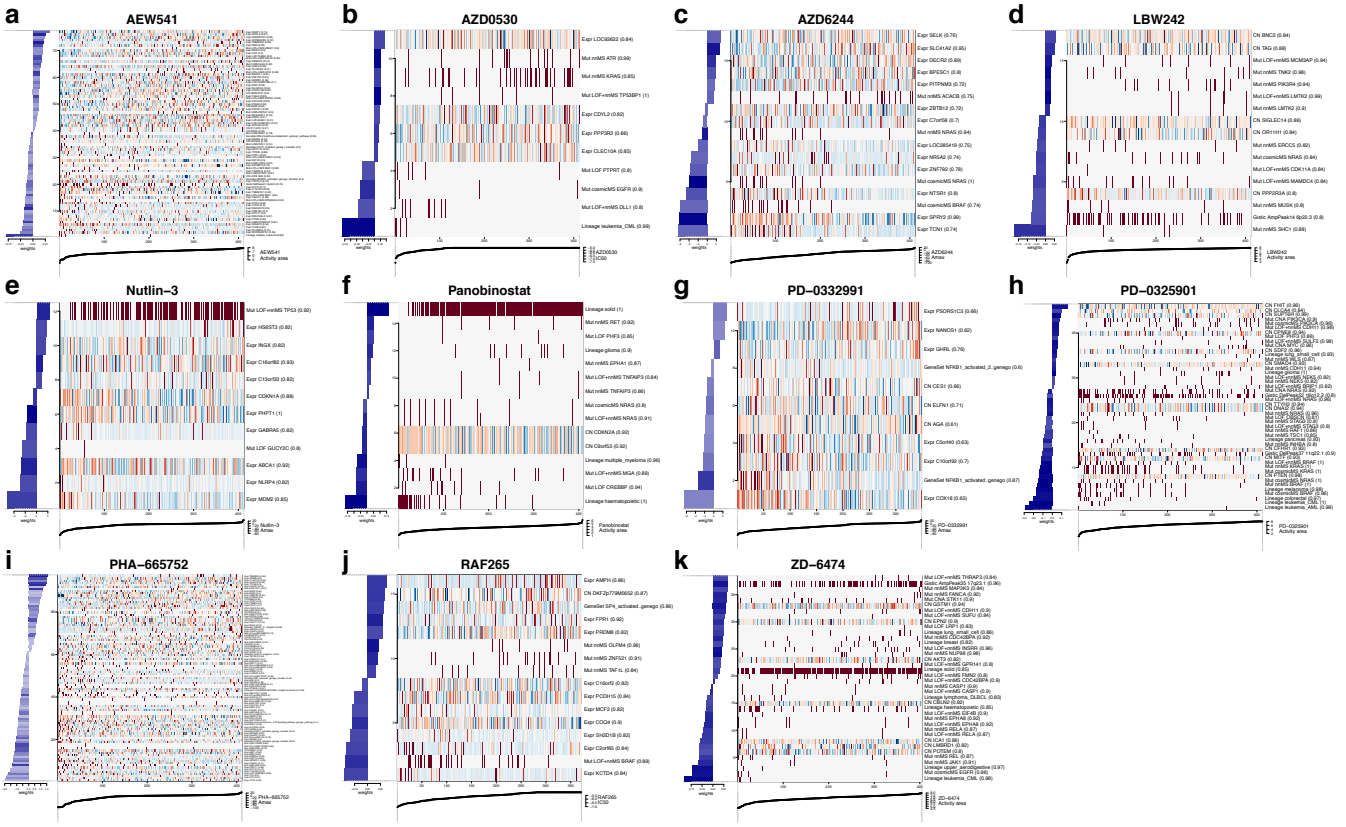
Supplementary Figure 6: Drug responses for Panobinostat (green) and PLX4720 (orange/purple), as measured by the activity area.



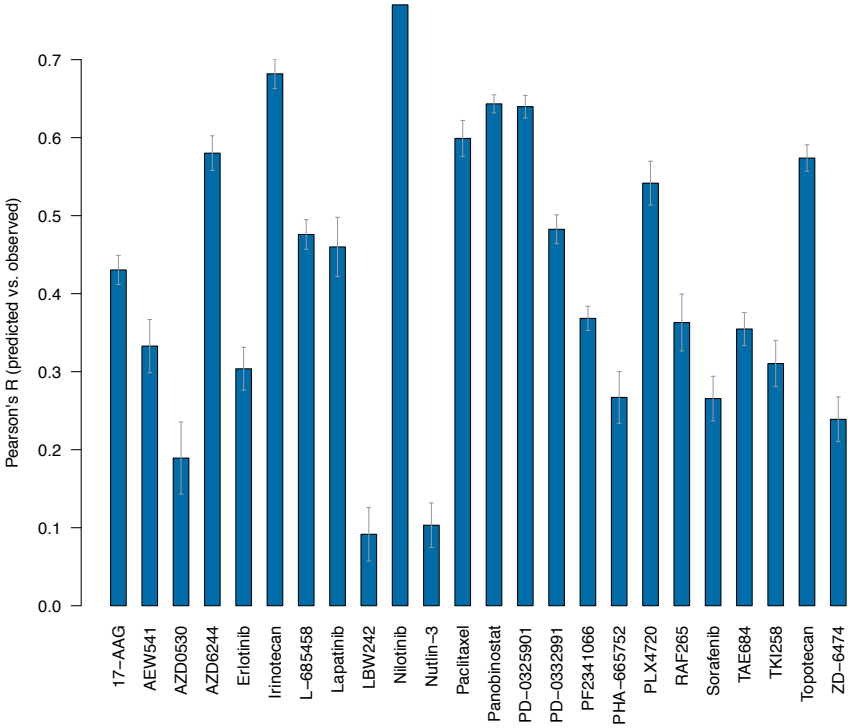
Supplementary Figure 7: Clustered response profiles across all cell lines for all the tested compounds, based on the activity area between the response curve and the zero dose effect level, up to the top tested concentration of 8 μM. Each row was median-normalized. Colored boxes denote compounds with similar targets that cluster together.



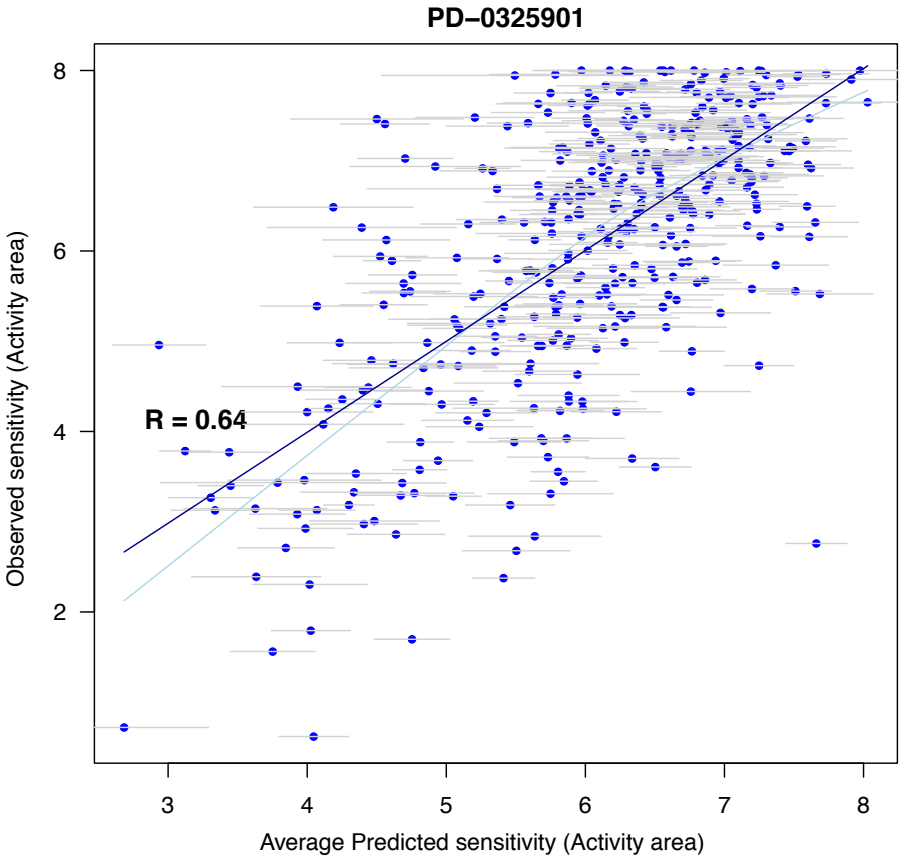
Supplementary Figure 8: Elastic net-based predictive modeling results for 24 compounds using all features as input and activity area as the sensitivity metric. In this figure (and some panels in supplementary figure 9), *8 – activity area* was used as the vector of response values, so that low values correspond to sensitive cell lines and are located towards the left of the heatmap, to keep consistent with other sensitivity metrics (IC_{50}, A_{max}). The frequency of each feature in the models is indicated in parenthesis next to the feature name and the shades of blue in the barplot next to the heatmap are proportional to that value. The cutoff for the bootstrapping frequency was 80% unless it would yield to a model with less than 5 features.



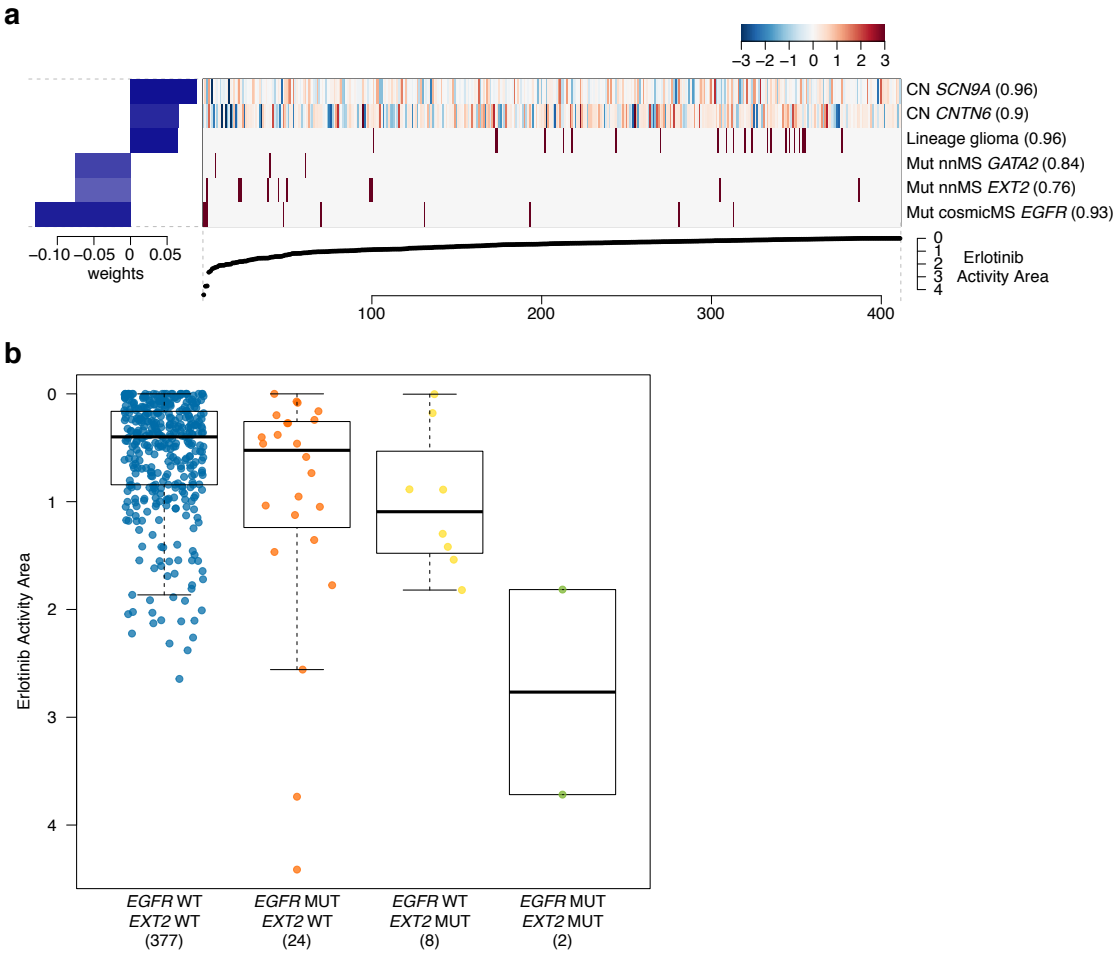
Supplementary Figure 9: Elastic net models built using a reduced feature set and/or utilizing A_{max} or IC_{50} as the sensitivity metric and/or with a lower bootstrapping cutoff so as to show more features. Panel d, f, h and k were built using copy number, mutation, and lineage data only, all others were built using all features, *i.e.* including expression data. Panels c, e, g and i were built using A_{max} as a sensitivity metric; panels b and j were built using IC_{50} as a sensitivity metric; all others



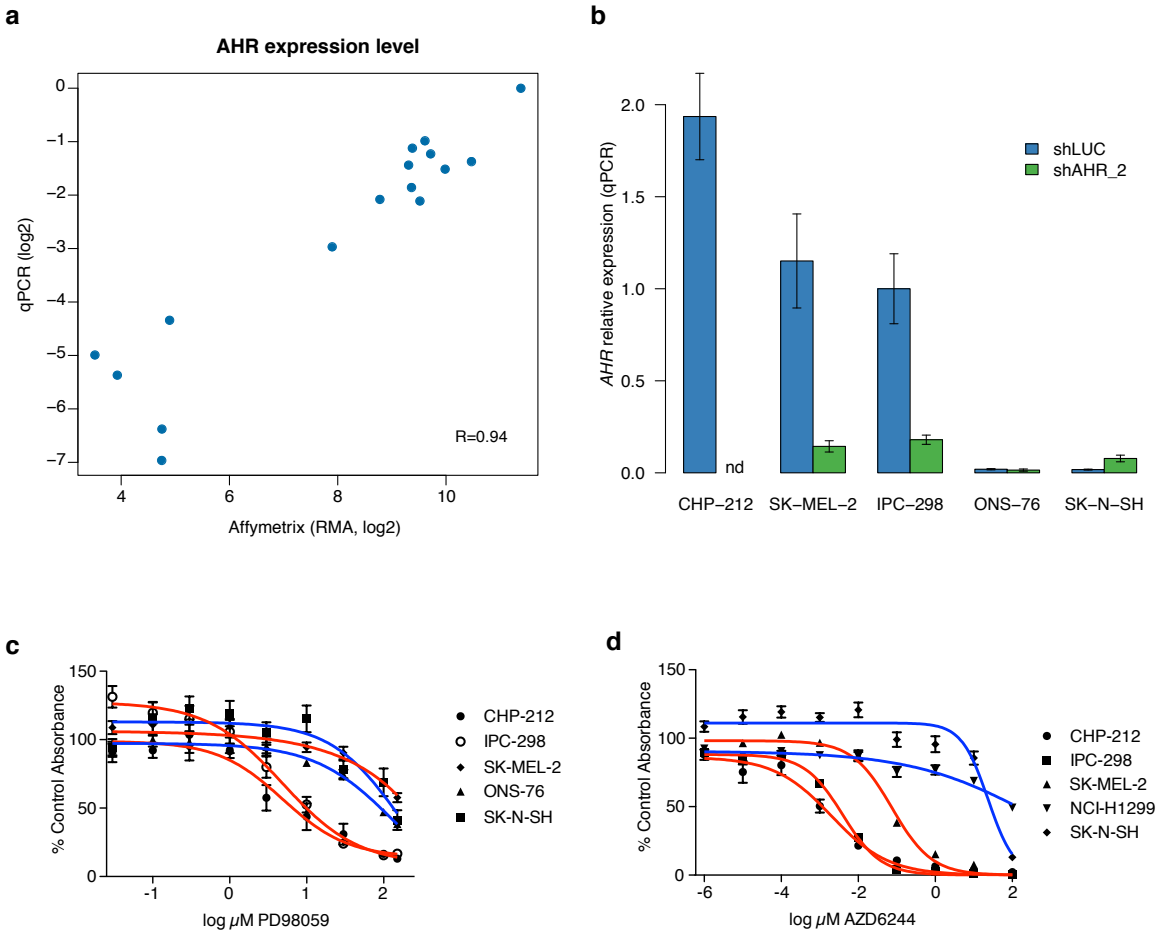
Supplementary Figure 10: Performance of the predictive models in Supplementary Figure 8 as measured by Pearson's correlation coefficients between observed and predicted sensitivity values (mean of 10 iterations of 10-fold cross-validations).



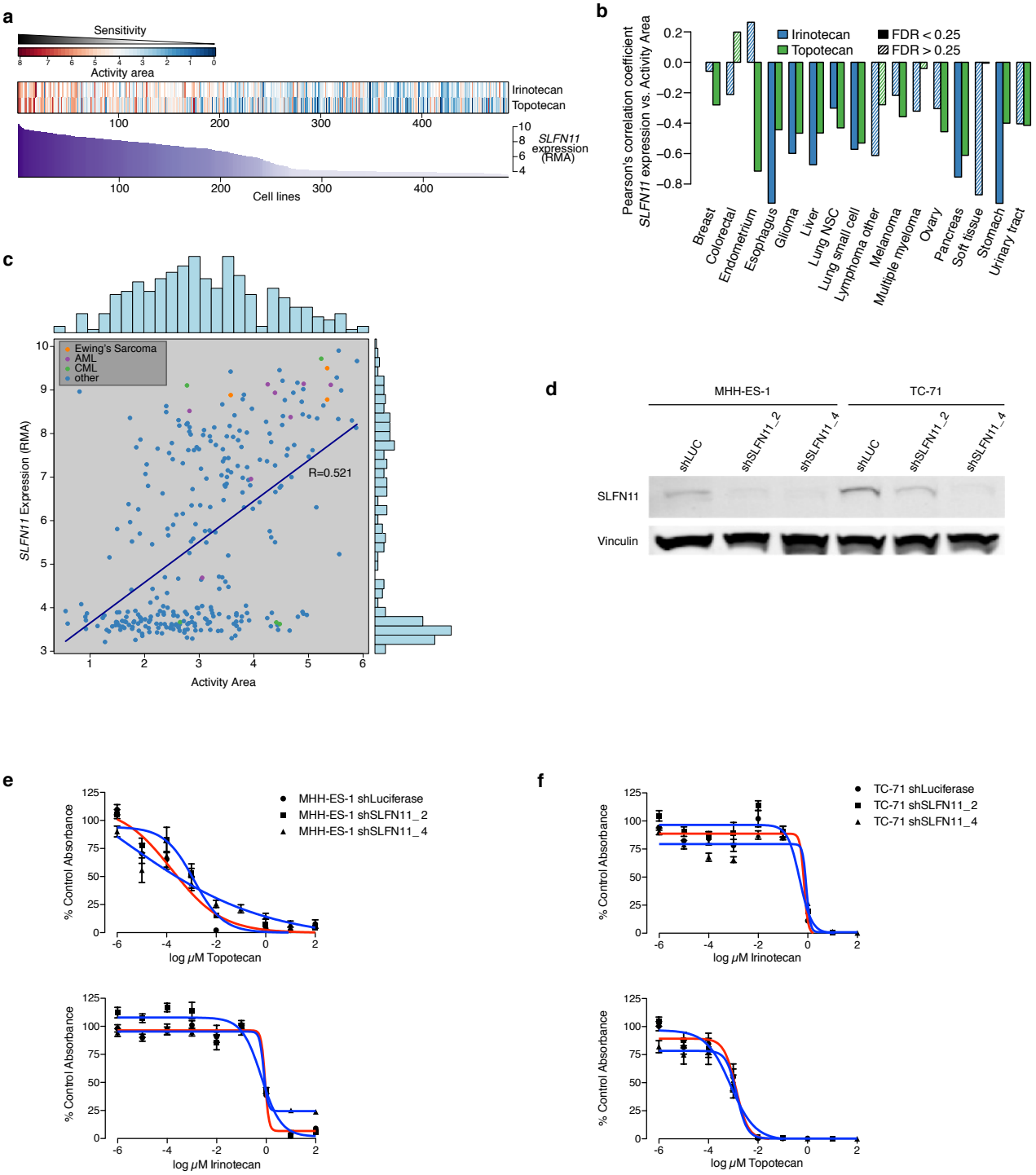
Supplementary Figure 11: Predicted sensitivity values for PD-0325901 (Activity area), versus observed sensitivities. Horizontal grey bars represent the standard deviation of the predicted values across 10 iterations of cross-validation.



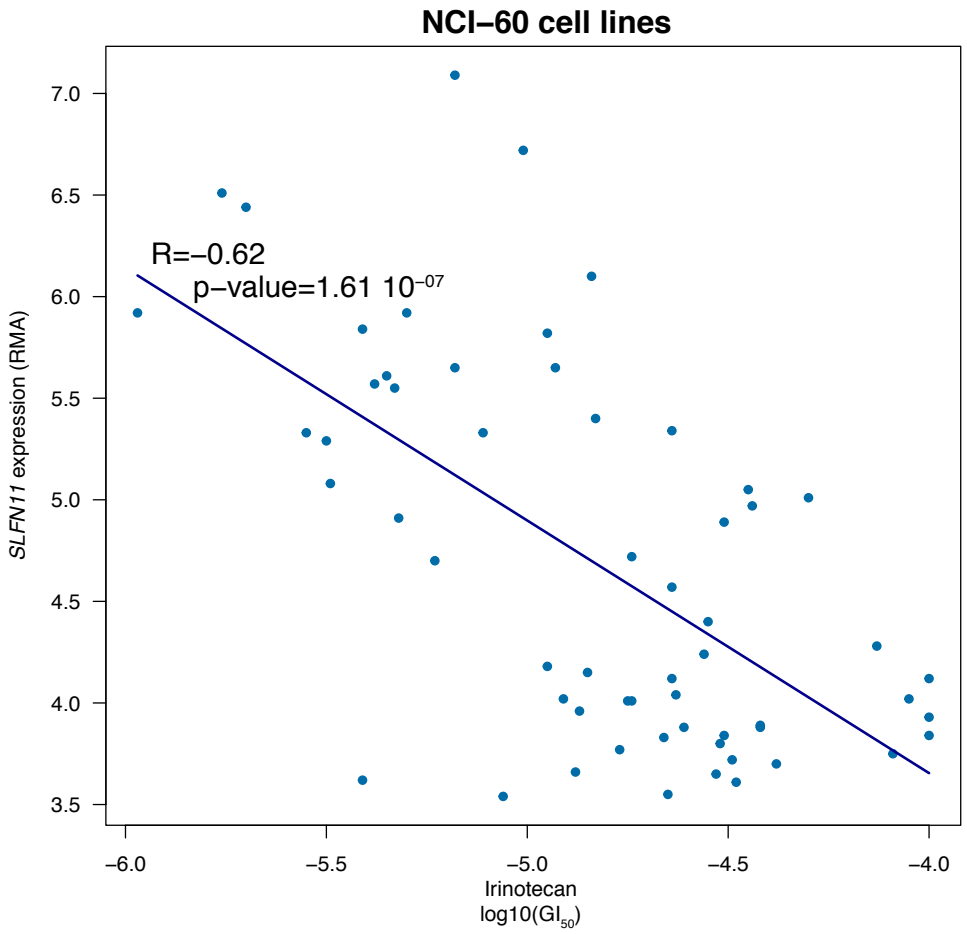
Supplementary Figure 12: *EXT2* variations may enrich for sensitivity to EGFR inhibitors. **a**. Elastic net-based predictive modeling results for Erlotinib using DNA copy number, mutations and lineages as input, and activity area as the sensitivity metric. *EGFR* mutations at or near recurrent COSMIC sites and *EXT2* missenses are the top two predictive features. (The top 3 predictive features for sensitivity and top 3 predictive features for insensitivity are shown. Bootstrapping cutoff: 75%) **b.** Drug response for Erlotinib (activity area) segregated by mutation status for *EGFR* and *EXT2* (all non-polymorphic missenses, supplemented by missenses reported by the COSMIC CGP project).



Supplementary Figure 13: *AHR* expression measurements correlates with sensitivity to MEK inhibitors in selected *NRAS*-mutant cell lines. **a**. Correlation between *AHR* expression measured by microarray analysis (X axis) or quantitative RT-PCR (Y axis). **b**. Mean *AHR* expression in five cell lines after shRNA knockdown using lentiviral constructs targeting luciferase control (shLUC; blue) or *AHR* (shAHR_2; green). Expression levels are relative to IPC-298 cells expressing shLUC. nd: not detected. (Measurement of *AHR* knockdown in the CHP-212 cell line was not possible due to the very low number of surviving cells). Error bars: standard deviation between replicates (n=3). **c, d**. Pharmacologic growth inhibition curves for selected *NRAS*-mutant cell lines using the MEK inhibitor PD-98059 (b) or AZD-6244 (c). Cell lines with high (red) or low (blue) *AHR* expression are indicated. Error bars: standard deviation between replicates (n=6).



Supplementary Figure 14: **a**. Heatmap of irinotecan and topotecan sensitivity ranked according to *SLFN11* expression. **b.** Correlation between SLFN11 mRNA expression and topoisomerase inhibitor sensitivity (activity area) across 16 lineages with sufficient representation (>10 cell lines). **c**. Scatter plot of *SLFN11* expression versus irinotecan sensitivity. Ewing's sarcoma, AML and CML cell lines are depicted as colored dots. **d**. Anti-SLFN11 and anti-vinculin control immunoblots of lysates from MHH-ES-1 and TC-71 cells expressing shRNA against *SLFN11* (shSLFN11_2 and shSLFN11_4) or luciferase (shLUC) as control. **e-f**. Pharmacologic growth inhibition curves for MHH-ES-1 (d) and TC-71 (e) cells expressing shRNA against SLFN11 (shSLFN11_2 and shSLFN11_4, blue lines) or luciferase (shLUC, red lines) as control. Error bars: standard deviation between replicates (n=6).



Supplementary Figure 15: Correlation between SLFN11 expression levels across NCI-60 cell lines (Affymetrix U-133 arrays normalized with RMA) and sensitivity to irinotecan (GI50). Data were obtained using the CellMiner database[41]. As evidenced by the significant inverse correlation between irinotecan GI50 and SLFN11 expression (R = -0.62), SLFN11 remains predictive of the sensitivity to irinotecan in independent datasets.

SUPPLEMENTARY TABLES

Supplementary Table 1 - CCLE cell lines and associated data.
The "CCLE name" consists of alpha-numeric characters from the Primary name and the site of the primary tumor from which the cell line was derived. The "Notes" field describes cell lines that appear to derive from the same individual based on SNP fingerprint matching. "Expression arrays", "SNP arrays", "OncoMap", "Hybrid capture/sequencing" and "Drug sensitivity profiling" refer to the existence of these data to date for the specified cell line.

Supplementary Table 2 - OncoMap assays and associated mutations.
Details of the 456 mass spectrometric assays that were used to assess the presence 392 mutations in 33 genes are shown. The genomic positions refer to hg18 genomic coordinates.

Supplementary Table 3 - Pearson's correlation coefficients for the comparison of DNA copy number data between cell lines and primary tumors.

Supplementary Table 4 - Pearson's correlation coefficients for the comparison of gene expression data between cell lines and primary tumors.

Supplementary Table 5 - Pearson's correlation coefficients for the comparison of mutation frequencies between cell lines and primary tumors.

Supplementary Table 6 - Compounds used for pharmacologic profiling
(source: Thomson Reuters Integrity database).

Supplementary Table 7 - Known predictors of drug response identified by regression-based predictive modeling.
ActArea: Activity area, the area between the pharmacologic dose-response curve and the zero inhibition level. Amax: inhibition observed at the maximal drug concentration. IC50: concentration where the fitted dose response curve first crosses -50%. All models were run using all genomic features as input, including expression data. The bootstrapping threshold refers to the threshold applied to the frequency at which predictive features are present in the models after bootstrapping.

Supplementary Table 8 - Features predicting sensitivity (using the activity area) to the MEK inhibitors PD-0325901 and AZD6244.
These features were elicited by the elastic net analysis using all CCLE genomic and expression features as input data.

Supplementary Table 9 - Features predicting sensitivity (using the activity area) to the MEK inhibitor AZD6244.
These features were obtained using the naïve Bayes classifier and all CCLE genetic and expression features as input data.

Supplementary Table 10 - MSigDB pathways enriched in the list of features predicting PD-0325901 sensitivity.
Targets of the SOX10 transcription factor appear as significantly enriched.

Supplementary Table 11 - Drug sensitivity data for 504 cell lines (481 unique) and 24 anticancer drugs.

Supplementary Table 12 - List of 1651 genes selected for hybrid capture/massively parallel sequencing.