

# The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity

Jordi Barretina<sup>1,2,3†\*</sup>, Giordano Caponigro<sup>4\*</sup>, Nicolas Stransky<sup>1\*</sup>, Kavitha Venkatesan<sup>4\*</sup>, Adam A. Margolin<sup>1†\*</sup>, Sungjoon Kim<sup>5</sup>, Christopher J. Wilson<sup>4</sup>, Joseph Lehár<sup>4</sup>, Gregory V. Kryukov<sup>1</sup>, Dmitriy Sonkin<sup>4</sup>, Anupama Reddy<sup>4</sup>, Manway Liu<sup>4</sup>, Lauren Murray<sup>1</sup>, Michael F. Berger<sup>1†</sup>, John E. Monahan<sup>4</sup>, Paula Morais<sup>1</sup>, Jodi Meltzer<sup>4</sup>, Adam Korejwa<sup>1</sup>, Judit Jané-Valbuena<sup>1,2</sup>, Felipa A. Mapa<sup>4</sup>, Joseph Thibault<sup>5</sup>, Eva Bric-Furlong<sup>4</sup>, Pichai Raman<sup>4</sup>, Aaron Shipway<sup>5</sup>, Ingo H. Engels<sup>5</sup>, Jill Cheng<sup>6</sup>, Guoying K. Yu<sup>6</sup>, Jianjun Yu<sup>6</sup>, Peter Aspasia Jr<sup>4</sup>, Melanie de Silva<sup>4</sup>, Kalpana Jagtap<sup>4</sup>, Michael D. Jones<sup>4</sup>, Li Wang<sup>4</sup>, Charles Hatton<sup>3</sup>, Emanuele Palessandro<sup>3</sup>, Supriya Gupta<sup>1</sup>, Scott Mahan<sup>1</sup>, Carrie Sougnez<sup>2</sup>, Robert C. Onofrio<sup>1</sup>, Ted Liefeld<sup>1</sup>, Laura MacConaill<sup>3</sup>, Wendy Winckler<sup>1</sup>, Michael Reich<sup>1</sup>, Nanxin Li<sup>5</sup>, Jill P. Mesirov<sup>1</sup>, Stacey B. Gabriel<sup>1</sup>, Gad Getz<sup>1</sup>, Kristin Ardlie<sup>1</sup>, Vivien Chan<sup>6</sup>, Vic E. Myer<sup>4</sup>, Barbara L. Weber<sup>4</sup>, Jeff Porter<sup>4</sup>, Markus Warmuth<sup>4</sup>, Peter Finan<sup>4</sup>, Jennifer L. Harris<sup>5</sup>, Matthew Meyerson<sup>1,2,3</sup>, Todd R. Golub<sup>1,3,7,8</sup>, Michael P. Morrissey<sup>4\*</sup>, William R. Sellers<sup>4\*</sup>, Robert Schlegel<sup>4\*</sup> & Levi A. Garraway<sup>1,2,3\*</sup>

The systematic translation of cancer genomic data into knowledge of tumour biology and therapeutic possibilities remains challenging. Such efforts should be greatly aided by robust preclinical model systems that reflect the genomic diversity of human cancers and for which detailed genetic and pharmacological annotation is available<sup>1</sup>.

Here we describe the Cancer Cell Line Encyclopedia (CCLE): a compilation of gene expression, chromosomal copy number and massively parallel sequencing data from 947 human cancer cell lines.

When coupled with pharmacological profiles for 24 anticancer drugs across 479 of the cell lines, this collection allowed identification of genetic, lineage, and gene-expression-based predictors of drug sensitivity. In addition to known predictors, we found that plasma cell lineage correlated with sensitivity to IGF1 receptor inhibitors; *AHR* expression was associated with MEK inhibitor efficacy in *NRAS*-mutant lines; and *SLFN11* expression predicted sensitivity to topoisomerase inhibitors. Together, our results indicate that large, annotated cell-line collections may help to enable preclinical stratification schemata for anticancer agents. The generation of genetic predictions of drug response in the preclinical setting and their incorporation into cancer clinical trial design could speed the emergence of ‘personalized’ therapeutic regimens<sup>2</sup>.

Human cancer cell lines represent a mainstay of tumour biology and drug discovery through facile experimental manipulation, global and detailed mechanistic studies, and various high-throughput applications. Numerous studies have used cell-line panels annotated with both genetic and pharmacological data, either within a tumour lineage<sup>3–5</sup> or across multiple cancer types<sup>6–12</sup>. Although affirming the promise of systematic cell line studies, many previous efforts were limited in their depth of genetic characterization and pharmacological interrogation.

To address these challenges, we generated a large-scale genomic data set for 947 human cancer cell lines, together with pharmacological profiling of 24 compounds across ~500 of these lines. The resulting collection, which we termed the Cancer Cell Line Encyclopedia (CCLE), encompasses 36 tumour types (Fig. 1a and Supplementary Table 1; see also <http://www.broadinstitute.org/ccle>). All cell lines were characterized by several genomic technology platforms. The mutational status of >1,600 genes was determined by targeted massively parallel sequencing, followed by removal of variants likely to be germline events (Supplementary Methods). Moreover, 392 recurrent mutations affecting 33

known cancer genes were assessed by mass spectrometric genotyping<sup>13</sup> (Supplementary Table 2 and Supplementary Fig. 1). DNA copy number was measured using high-density single nucleotide polymorphism arrays (Affymetrix SNP 6.0; Supplementary Methods). Finally, messenger RNA expression levels were obtained for each of the lines using Affymetrix U133 plus 2.0 arrays. These data were also used to confirm cell line identities (Supplementary Methods and Supplementary Figs 2–4).

We next measured the genomic similarities by lineage between CCLE lines and primary tumours from Tumorscape<sup>14</sup>, expO, MILE and COSMIC data sets (Fig. 1b–d and Supplementary Methods). For most lineages, a strong positive correlation was observed in both chromosomal copy number and gene expression patterns (median correlation coefficients of 0.77, range = 0.52–0.94,  $P < 10^{-15}$ , for copy number, and 0.60, range = 0.29–0.77,  $P < 10^{-15}$ , for expression, respectively; Fig. 1b, c and Supplementary Tables 3 and 4), as has been described previously<sup>3–5,15</sup>. A positive correlation was also observed for point mutation frequencies (median correlation coefficient = 0.71, range = −0.06–0.97,  $P < 10^{-2}$  for all but 3 lineages; Supplementary Fig. 5), even when *TP53* was removed from the data set (median correlation coefficient = 0.64, range = −0.31–0.97,  $P < 10^{-2}$  for all but 3 lineages; Fig. 1d and Supplementary Table 5). Thus, with relatively few exceptions (Supplementary Information), the CCLE may provide representative genetic proxies for primary tumours in many cancer types.

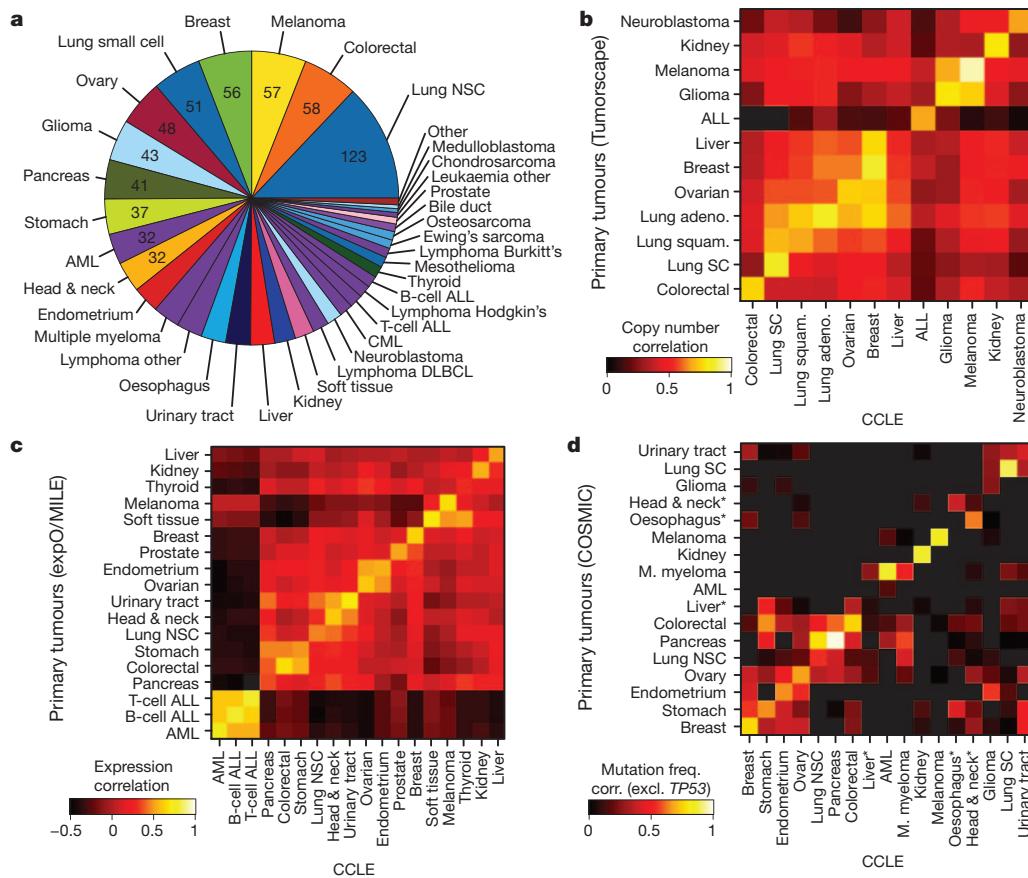
Given the pressing clinical need for robust molecular correlates of anticancer drug response, we incorporated a systematic framework to ascertain molecular correlates of pharmacological sensitivity *in vitro*. First, 8-point dose–response curves for 24 compounds (targeted and cytotoxic agents) across 479 cell lines were generated (Supplementary Tables 1 and 6, and Supplementary Methods).

These curves were represented by a logistical sigmoidal function with a maximal effect level ( $A_{max}$ ), the concentration at half-maximal activity of the compound ( $EC_{50}$ ), a Hill coefficient representing the sigmoidal transition, and the concentration at which the drug response reached an absolute inhibition of 50% ( $IC_{50}$ ).

Broadly active compounds, exemplified by the HDAC inhibitor LBH589 (panobinostat), showed a roughly even distribution of  $A_{max}$  and  $EC_{50}$  values across most cell lines (Fig. 2a). In contrast, the RAF inhibitor PLX4720 had a more selective profile:  $A_{max}$  or  $EC_{50}$  values for most cell lines could be categorized as ‘sensitive’ or ‘insensitive’ to

<sup>1</sup>The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. <sup>2</sup>Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>3</sup>Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>4</sup>Novartis Institutes for Biomedical Research, Cambridge, Massachusetts 02139, USA. <sup>5</sup>Genomics Institute of the Novartis Research Foundation, San Diego, California 92121, USA. <sup>6</sup>Novartis Institutes for Biomedical Research, Emeryville, California 94608, USA. <sup>7</sup>Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. <sup>8</sup>Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA. <sup>†</sup>Present addresses: Novartis Institutes for Biomedical Research, Cambridge, Massachusetts 02139, USA (J.B.); Sage Bionetworks, 1100 Fairview Ave. N., Seattle, Washington 98109, USA (A.M.); Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA (M.F.B.).

\*These authors contributed equally to this work.



**Figure 1 | The Cancer Cell Line Encyclopedia.** **a**, Distribution of cancer types in the CCLE by lineage. **b**, Comparison of DNA copy-number profiles (GISTIC G-scores) between cell lines and primary tumours. The diagonal of the heat map shows the Pearson correlation between corresponding tumour types. Because cell lines and tumours are separate data sets, the correlation matrix is asymmetric: the top left showing how well the tumour features correlate with the average of the cell lines in a lineage, and the bottom right showing the converse. **c**, Comparison of mRNA expression profiles between cell lines and primary tumours. For each tumour type, the log fold change of the 5,000 most variable genes is calculated between that tumour type and all others. Pearson correlations between tumour type fold changes from primary tumours and cell lines are shown as a heat map. **d**, Comparison of point mutation frequencies between cell lines and primary tumours in COSMIC (v56), restricted to genes that are well represented in both sample sets but excluding *TP53*, which is highly prevalent in most tumour types. Pairwise Pearson correlations are shown as a heat map. Asterisk indicates that the correlations of oesophageal, liver, and head and neck cancer mutation frequencies are restored when including *TP53*.

PLX4720, with sensitive lines enriched for the *BRAF*<sup>V600E</sup> mutation (Fig. 2a). To capture simultaneously the efficacy and potency of a drug, we designated an ‘activity area’ (Fig. 2b and Supplementary Fig. 6). The 24 compounds profiled showed wide variations in activity area, and those with similar mechanisms of action clustered together (Supplementary Fig. 7).

Genomic correlates of drug sensitivity may be extracted by predictive models using machine learning techniques<sup>6,10</sup>. We therefore assembled all CCLE genomic data types into a matrix wherein each feature was converted to a *z*-score across all lines (Supplementary Methods). Next, we adapted a categorical modelling approach that used a naive Bayes classification and discrete sensitivity calls, or an elastic net regression analysis<sup>16</sup> for continuous sensitivity measurements. Both approaches were applied to all compounds and genomic data with or without gene expression features (Supplementary Methods). Prediction performance was determined using tenfold cross-validation, and the elastic net features were bootstrapped to retain only those that were consistent across runs (Supplementary Methods).

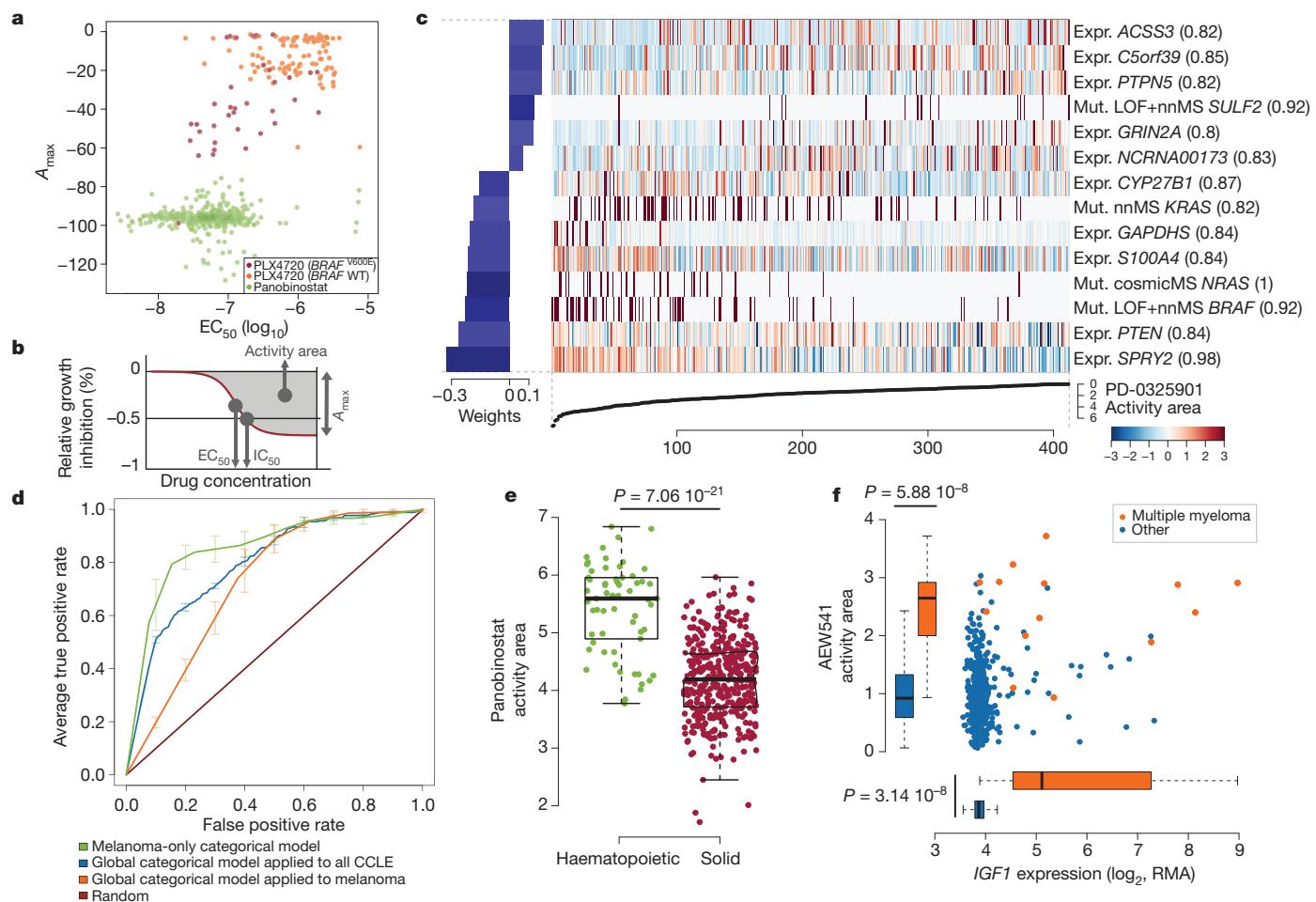
Out of >50,000 input features, the regression-based analysis identified multiple known features as top predictors of sensitivity to several agents (Supplementary Table 7 and Supplementary Figs 8 and 9), with robust cross-validated performance (Supplementary Fig. 10 and 11). For example, activating mutations in *BRAF* and *NRAS* were among the top four predictors of sensitivity in models generated for the MEK inhibitor PD-0325901 (ref. 10) (Fig. 2c). Additional predictive features for MEK inhibition included expression of *PTEN*, *PTPN5* and *SPRY2* (which encodes a regulator of MAPK output). *KRAS* mutations were also identified, albeit with a lower predictive value (Fig. 2c, Supplementary Tables 8 and 9 and Supplementary Fig. 8).

Other top predictors included *EGFR* mutations and *ERBB2* amplification/overexpression for erlotinib<sup>8</sup> and lapatinib<sup>17</sup>, respectively;

*BRAF*<sup>V600E</sup> for RAF inhibitors (PLX4720 (ref. 18) and RAF265); *HGF* expression and *MET* amplification for the MET/ALK inhibitor PF-2341066 (ref. 19); and *MDM2* overexpression for Nutlin-3 (ref. 20) sensitivity. Variants affecting the *EXT2* gene, which encodes a glycosyltransferase involved in heparin sulphate biosynthesis, were significantly correlated with erlotinib effects (Supplementary Fig. 12). This observation is intriguing in light of a report linking heparin sulphate with erlotinib sensitivity<sup>21</sup>. In addition, *NQO1* expression was identified as the top predictive feature for sensitivity to the Hsp90 inhibitor 17-AAG, a quinone moiety metabolized by NAD(P)H:quinone oxidoreductase (*NQO1*). *NQO1* produces a high-potency intermediate (17-AAGH2)<sup>22</sup>, and has previously been identified as a potential biomarker for Hsp90 inhibitors<sup>23</sup>.

Because some genetic/molecular alterations occur commonly in specific tumour types, lineage may become a confounding factor in predictive analyses. Indeed, a classifier built using the entire cell-line data set performed suboptimally when applied exclusively to melanoma-derived cell lines (Fig. 2d), whereas a model built with only melanoma cell lines performed better (Fig. 2d). Predictive features in the melanoma-only model showed a strong overexpression of genes regulated by the transcription factors MITF and SOX10 (Supplementary Table 10), which may also help predict RAF inhibitor drug sensitivity in melanoma cell lines.

Nonetheless, lineage emerged as the predominant predictive feature for several compounds. For example, elastic net studies of the HDAC inhibitor panobinostat identified haematological lineages as predictors of sensitivity (Fig. 2e and Supplementary Fig. 9). Interestingly, most clinical responses to panobinostat and related compounds (for example, vorinostat and romidepsin) have been observed in haematological cancers. Similarly, most multiple myeloma cell lines (12 of 14 lines tested) exhibited enhanced sensitivity to the IGF1 receptor inhibitor



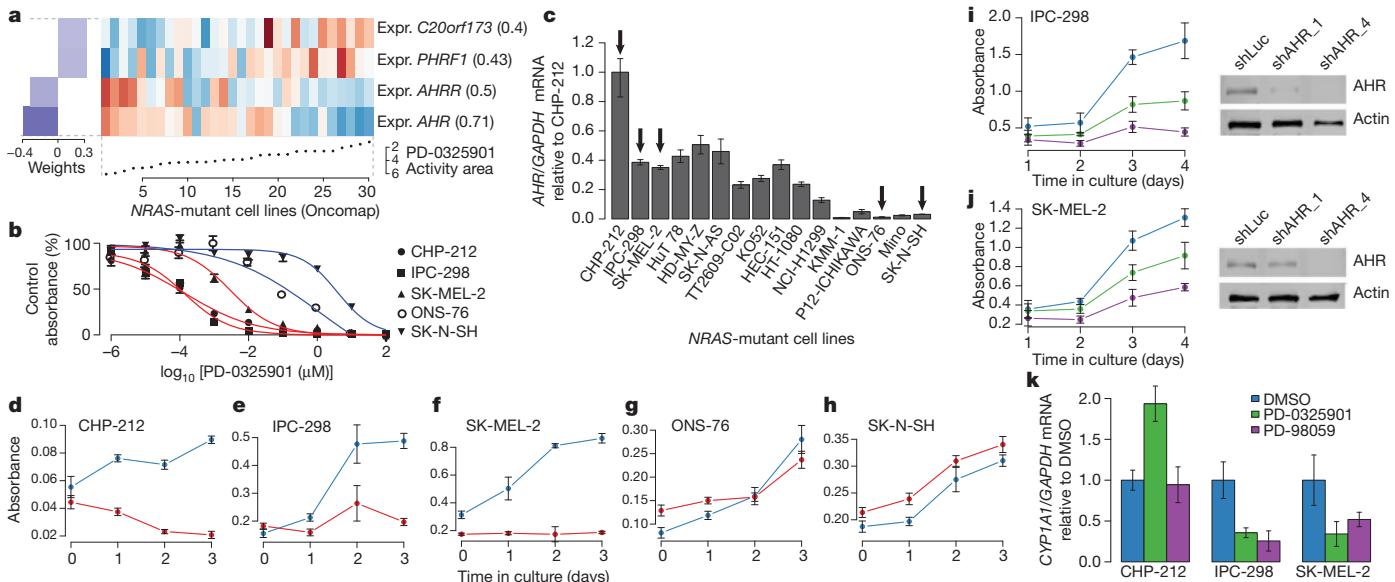
**Figure 2 | Predictive modelling of pharmacological sensitivity using CCLE genomic data.** **a, b.** Drug responses for panobinostat (green) and PLX4720 (orange/purple) represented by the high-concentration effect level ( $A_{\max}$ ) and transitional concentration ( $EC_{50}$ ) for a sigmoidal fit to the response curve. **(b).** **c.** Elastic net regression modelling of genomic features that predict sensitivity to PD-0325901. The bottom curve indicates drug response, measured as the area over the dose–response curve (activity area), for each cell line. The central heat map shows the CCLE features in the model (continuous z-score for expression and copy number, dark red for discrete mutation calls), across all cell lines (x axis). Bar plot (left): weight of the top predictive features for sensitivity (bottom) or insensitivity (top). Parentheses indicate features present in >80% of models after bootstrapping. LOF, loss of function mutation; nnMS, non-neutral missense mutation (Supplementary Methods).

AEW541 (Fig. 2f and Supplementary Figs 8 and 9) and showed high *IGF1* expression (Fig. 2f). Interestingly, elevated *IGF1R* expression also correlated with AEW541 sensitivity (Supplementary Fig. 9). The CCLE results indicate that multiple myeloma may be a promising indication for clinical trials of *IGF1* receptor inhibitors<sup>24</sup> and that these drugs may have enhanced efficacy in cancers with high *IGF1* or *IGF1R* expression.

Whereas *BRAF* and *NRAS* mutations are known single-gene predictors of sensitivity to MEK inhibitors, several ‘sensitive’ cell lines lacked mutations in these genes, whereas other lines harbouring these mutations were nonetheless ‘insensitive’ (Fig. 2c). The elastic net regression model derived from the subset of cell lines with validated *NRAS* mutations identified elevated expression of the *AHR* gene (which encodes the aryl hydrocarbon receptor) as strongly correlated with sensitivity to the MEK inhibitor PD-0325901 (Fig. 3a). This finding was interesting in light of previous studies indicating that a related MEK inhibitor (PD-98059) may also function as a direct *AHR* antagonist<sup>25</sup>. We therefore hypothesized that the enhanced sensitivity of some *NRAS*-mutant cell lines to MEK inhibitors might relate to a coexistent dependence on *AHR* function.

**d**, Specificity and sensitivity (receiver operating characteristic curves) of cross-validated categorical models predicting the response to a MEK inhibitor, PD-0325901 (activity area). Mean true positive rate and standard deviation ( $n = 5$ ) are shown when models are built using all lines (global categorical model, in blue and orange), or within only melanoma lines (green). **e**, Activity area values for panobinostat between cell lines derived from haematopoietic ( $n = 61$ ) and solid tumours ( $n = 387$ ). The middle bar, median; box, inter-quartile range; bars extend to 1.5× the inter-quartile range. **f**, Distribution of activity area values for AEW541 relative to *IGF1* mRNA expression. Orange dots, multiple myeloma cell lines ( $n = 14$ ); blue dots, cell lines from other tumour types ( $n = 434$ ). Box-and-whisker plots show the activity area or mRNA expression distributions relative to each cell line type (line, median; box, inter-quartile range), with bars extending to 1.5× the inter-quartile range.

To test this hypothesis, we first confirmed the correlation between *AHR* expression and sensitivity to MEK inhibitors in a subset of *NRAS*-mutant cell lines (Fig. 3b and Supplementary Fig. 13). Next, we performed short hairpin RNA (shRNA) knockdown of *AHR* in cell lines with high or low *AHR* expression (Fig. 3c). Silencing of *AHR* suppressed the growth of three *NRAS*-mutant cell lines with elevated *AHR* expression (Fig. 3d–f), but had no effect on the growth of two lines with low *AHR* expression (Fig. 3g, h). The growth inhibitory effect was confirmed with two additional shRNAs, where evidence for dose dependence was also apparent (Fig. 3i, j). We also tested the hypothesis that allosteric MEK inhibitors may suppress *AHR* function by measuring the effect of PD-0325901 and PD-98059 on endogenous *CYP1A1* mRNA, a transcriptional target of *AHR* in some contexts. Both compounds reduced *CYP1A1* levels in *NRAS*-mutant melanoma cells (IPC-298 and SK-MEL-2; Fig. 3k) but not in neuroblastoma cells (CHP-212; Fig. 3k), indicating that other factors may govern *CYP1A1* expression in the latter lineage. Together, these results suggest that *AHR* dependency may co-occur with MAP kinase activation in some *NRAS*-mutant cancer cells, and that elevated *AHR* may serve as a

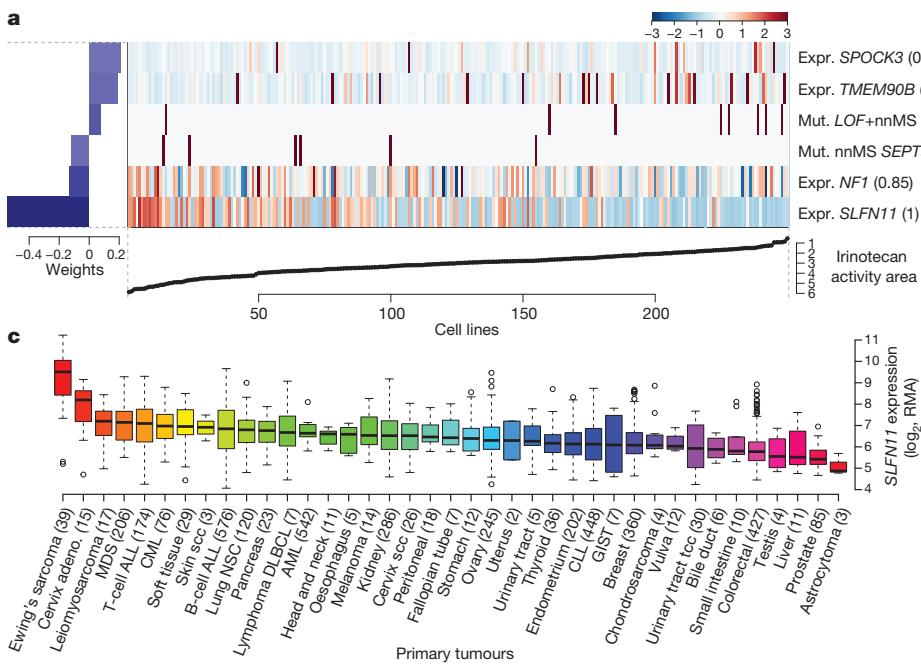


**Figure 3 | AHR expression may denote a tumour dependency targeted by MEK inhibitors in NRAS-mutant cell lines.** **a**, Predictive features for PD-0325901 sensitivity (using the 'varying baseline' activity area) in validated NRAS-mutant cell lines. **b**, Growth inhibition curves for NRAS-mutant cell lines expressing high (red) or low (blue) levels of AHR mRNA in the presence of the MEK inhibitor PD-0325901. **c**, Relative AHR mRNA expression across a panel of NRAS-mutant cell lines (arrows indicate cell lines where AHR dependency was analysed). **d–h**, Proliferation of NRAS-mutant cell lines displaying high (**d–f**) and low (**g, h**) AHR mRNA expression, after introduction of shRNAs against mechanistic biomarker for enhanced MEK inhibitor sensitivity in this setting.

We also looked for markers predictive of response to several conventional chemotherapeutic agents (Supplementary Fig. 7 and Supplementary Table 6) and identified *SLFN11* expression as the top correlate of sensitivity to irinotecan (Fig. 4a), a camptothecin analogue that inhibits the topoisomerase I (TOP1) enzyme. *SLFN11* expression

AHR (red lines) or luciferase (blue lines). **i**, Left: proliferation of IPC-298 cells (high AHR) after introduction of additional shRNAs against AHR (shAHR\_1 and shAHR\_4; green and purple lines, respectively) or luciferase (control shLuc; blue line). Right: corresponding immunoblot analysis of AHR protein. **j**, Equivalent studies as in **i** using SK-MEL-2 cells (high AHR). **k**, Endogenous CYP1A1 mRNA expression in the neuroblastoma line CHP-212 or the melanoma lines IPC-298 and SK-MEL-2 after exposure to vehicle (blue) or MEK inhibitors (PD-0325901, green or PD-98059, purple). Error bars indicate standard deviation between replicates, with  $n = 12$  (**b**),  $n = 3$  (**c**),  $n = 6$  (**d–k**).

also emerged as the top predictor of topotecan sensitivity (another TOP1 inhibitor; Supplementary Figs 8 and 14). Overall, 12 of 16 lineages showed significant *SLFN11* associations for topotecan or irinotecan sensitivity (Pearson's  $r \geq 0.2$ , Supplementary Fig. 14b). This finding was independently validated using data from the NCI-60 collection (Supplementary Fig. 15). *SLFN11* knockdown did not affect steady-state growth sensitivity profiles (Supplementary Fig. 14d–f).



**Figure 4 | Predicting sensitivity to topoisomerase I inhibitors.** **a**, Elastic net regression analysis of genomic correlates of irinotecan sensitivity is shown for 250 cell lines. **b**, Dose-response curves for three Ewing's sarcoma cell lines (MSS-ES-1, SK-ES-1 and TC-71) and two control cell lines with low *SLFN11* expression (HCC-56 and SK-HEP-1). Grey vertical bars, standard deviation of

the mean growth inhibition ( $n = 2$ ). **c**, *SLFN11* expression across 4,103 primary tumours. Box-and-whisker plots show the distribution of mRNA expression for each subtype, ordered by the median *SLFN11* expression level (line), the inter-quartile range (box) and up to 1.5× the inter-quartile range (bars). Sample numbers ( $n$ ) are indicated in parentheses.

All three Ewing's sarcoma cell lines screened showed both high *SLFN11* expression and sensitivity to irinotecan (Fig. 4b and Supplementary Fig. 14). Ewing's sarcomas also exhibited the highest *SLFN11* expression among 4,103 primary tumour samples spanning 39 lineages (Fig. 4c), suggesting that TOP1 inhibitors might offer an effective treatment option for this cancer type. Towards this end, several ongoing trials in Ewing's sarcoma are examining irinotecan-based combinations, or the addition of topotecan to standard regimens<sup>26</sup>. For some lineages with high *SLFN11* expression (for example, cervical adenocarcinoma), topoisomerase inhibitors already comprise a standard chemotherapy regimen. In other tumours where topoisomerase inhibitors are commonly used (for example, colorectal and ovarian cancers), a range of *SLFN11* expression was observed, raising the possibility that high *SLFN11* expression might enrich for tumours more likely to respond. If confirmed in correlative clinical studies, *SLFN11* expression may offer a means to stratify patients for topoisomerase inhibitor treatment.

By assembling the CCLE, we have expanded the process of detailed annotation of preclinical human cancer models (<http://www.broadinstitute.org/cCLE>). Genomic predictors of drug sensitivity revealed both known and novel candidate biomarkers of response. Even within genetically defined sub-populations—or when agents were broadly active without clear genetic targets—elastic net modelling studies identified key predictors or mechanistic effectors of drug response. Additional efforts that increase the scale and provide complementary types of information (for example, whole-genome/transcriptome sequencing, epigenetic studies, metabolic profiling or proteomic/phosphoproteomic analysis) should enable additional insights. In the future, comprehensive and tractable cell-line systems provided through this and other efforts<sup>27</sup> may facilitate numerous advances in cancer biology and drug discovery.

## METHODS SUMMARY

A total of 947 independent cancer cell lines were profiled at the genomic level (data available at <http://www.broadinstitute.org/cCLE> and Gene Expression Omnibus (GEO) using accession number GSE36139) and compound sensitivity data were obtained for 479 lines (Supplementary Table 11). Mutation information was obtained both by using massively parallel sequencing of >1,600 genes (Supplementary Table 12) and by mass spectrometric genotyping (OncoMap), which interrogated 492 mutations in 33 known oncogenes and tumour suppressors. Genotyping/copy number analysis was performed using Affymetrix Genome-Wide Human SNP Array 6.0 and expression analysis using the GeneChip Human Genome U133 Plus 2.0 Array. Eight-point dose–response curves were generated for 24 anticancer drugs using an automated compound-screening platform. Compound sensitivity data were used for two types of predictive models that used the naive Bayes classifier or the elastic net regression algorithm. The effects of *AHR* expression silencing on cell viability were assessed by stable expression of shRNA lentiviral vectors targeting either this gene or luciferase as control. The effect of compound treatment on *AHR* target gene expression was assessed by quantitative RT–PCR. A full description of the Methods is included in Supplementary Information.

Received 25 July 2011; accepted 1 March 2012.

- Caponigro, G. & Sellers, W. R. Advances in the preclinical testing of cancer therapeutic hypotheses. *Nature Rev. Drug Discov.* **10**, 179–187 (2011).
- MacConaill, L. E. & Garraway, L. A. Clinical implications of the cancer genome. *J. Clin. Oncol.* **28**, 5219–5228 (2010).
- Lin, W. M. et al. Modeling genomic diversity and tumor dependency in malignant melanoma. *Cancer Res.* **68**, 664–673 (2008).
- Neve, R. M. et al. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10**, 515–527 (2006).
- Sos, M. L. et al. Predicting drug susceptibility of non-small cell lung cancers based on genetic lesions. *J. Clin. Invest.* **119**, 1727–1740 (2009).
- Dry, J. R. et al. Transcriptional pathway signatures predict MEK addiction and response to selumetinib (AZD6244). *Cancer Res.* **70**, 2264–2273 (2010).
- Garraway, L. A. et al. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117–122 (2005).
- Greshock, J. et al. Molecular target class is predictive of *in vitro* response profile. *Cancer Res.* **70**, 3677–3686 (2010).
- McDermott, U. et al. Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. *Proc. Natl. Acad. Sci. USA* **104**, 19936–19941 (2007).

- Solit, D. B. et al. BRAF mutation predicts sensitivity to MEK inhibition. *Nature* **439**, 358–362 (2006).
- Staunton, J. E. et al. Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. USA* **98**, 10787–10792 (2001).
- Weinstein, J. N. et al. An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–349 (1997).
- Thomas, R. K. et al. High-throughput oncogene mutation profiling in human cancer. *Nature Genet.* **39**, 347–351 (2007).
- Beroukhim, R. et al. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
- Ross, D. T. et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.* **24**, 227–235 (2000).
- Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* **67**, 301–320 (2005).
- Konecný, G. E. et al. Activity of the dual kinase inhibitor lapatinib (GW572016) against HER-2-overexpressing and trastuzumab-treated breast cancer cells. *Cancer Res.* **66**, 1630–1639 (2006).
- Tsai, J. et al. Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity. *Proc. Natl. Acad. Sci. USA* **105**, 3041–3046 (2008).
- Zou, H. Y. et al. An orally available small-molecule inhibitor of c-Met, PF-2341066, exhibits cytoreductive antitumor efficacy through antiproliferative and antiangiogenic mechanisms. *Cancer Res.* **67**, 4408–4417 (2007).
- Müller, C. R. et al. Potential for treatment of liposarcomas with the MDM2 antagonist Nutlin-3A. *Int. J. Cancer* **121**, 199–205 (2007).
- Nishio, M. et al. Serum heparan sulfate concentration is correlated with the failure of epidermal growth factor receptor tyrosine kinase inhibitor treatment in patients with lung adenocarcinoma. *J. Thorac. Oncol.* **6**, 1889–1894 (2011).
- Guo, W. et al. Formation of 17-allylamino-demethoxygeldanamycin (17-AAG) hydroquinone by NAD(P)H:quinone oxidoreductase 1: role of 17-AAG hydroquinone in heat shock protein 90 inhibition. *Cancer Res.* **65**, 10006–10015 (2005).
- Kelland, L. R., Sharp, S. Y., Rogers, P. M., Myers, T. G. & Workman, P. DT-Diaphorase expression and tumor cell sensitivity to 17-allylamino, 17-demethoxygeldanamycin, an inhibitor of heat shock protein 90. *J. Natl. Cancer Inst.* **91**, 1940–1949 (1999).
- Moreau, P. et al. Phase I study of the anti insulin-like growth factor 1 receptor (IGF-1R) monoclonal antibody, AVE1642, as single agent and in combination with bortezomib in patients with relapsed multiple myeloma. *Leukemia* **25**, 872–874 (2011).
- Reiners, J. J. Jr, Lee, J. Y., Clift, R. E., Dudley, D. T. & Myrand, S. P. PD98059 is an equipotent antagonist of the aryl hydrocarbon receptor and inhibitor of mitogen-activated protein kinase kinase. *Mol. Pharmacol.* **53**, 438–445 (1998).
- Wagner, L. M. et al. Temozolamide and intravenous irinotecan for treatment of advanced Ewing sarcoma. *Pediatr. Blood Cancer* **48**, 132–139 (2007).
- Garnett, M. J. et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* <http://dx.doi.org/10.1038/nature11005> (this issue).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank the staff of the Biological Samples Platform, the Genetic Analysis Platform and the Sequencing Platform at the Broad Institute. We thank S. Banerji, J. Che, C. M. Johannessen, A. Su and N. Wagle for advice and discussion. We are grateful for the technical assistance and support of G. Bonamy, R. Brusch III, E. Gelfand, K. Gravelin, T. Huynh, S. Kehoe, K. Matthews, J. Nedzel, L. Niu, R. Pinchback, D. Roby, J. Slind, T. R. Smith, L. Tan, V. Trinh, C. Vickers, G. Yang, Y. Yao and X. Zhang. The Cancer Cell Line Encyclopedia project was enabled by a grant from the Novartis Institutes for Biomedical Research. Additional funding support was provided by the National Cancer Institute (M.M., L.A.G.), the Starr Cancer Consortium (M.F.B., L.A.G.), and the NIH Director's New Innovator Award (L.A.G.).

**Author Contributions** For the work described herein, J.B. and G.C. were the lead research scientists; N.S., K.V. and A.M.M. were the lead computational biologists; M.P.M., W.R.S., R.S. and L.A.G. were the senior authors. J.B., G.C., S.K., P.M., J.M., J.T., A.S., N.L. and K.A. performed cell-line procurement and processing; P.M. and K.A. performed or directed nucleic acid extraction and quality control; S.G., W.W. and S.B.G. performed or directed genomic data generation; C.J.W., F.A.M., E.B.-F., I.H.E., P.A., M.d.S., K.J. and V.E.M. performed pharmacological data generation; N.S., K.V., G.V.K., A.R., M.F.B., J.C., G.K.Y., M.D.J., T.L., M.R. and G.G. contributed to software development; N.S., K.V., A.A.M., J.L., G.V.K., D.S., A.R., M.L., M.F.B., A.K., P.R., J.C., G.K.Y., J.Y., M.D.J., L.W., C.H., E.P., J.P.M., V.C. and M.P.M. performed computational biology and bioinformatics analysis; J.B., G.C., N.S., L.M., J.E.M., J.J.-V., M.P.M., W.R.S., R.S. and L.A.G. performed biological analysis and interpretation; N.S., K.V., A.A.M., J.L., A.R., M.L., L.M., A.K., J.J.-V., J.C., G.K.Y. and J.Y. prepared figures and tables for the main text and Supplementary Information; J.B., G.C., N.S., K.V., A.A.M., J.L., G.V.K., J.J.-V., M.P.M. and L.A.G. wrote and edited the main text and Supplementary Information; J.B., G.C., N.S., K.V., S.K., C.J.W., J.L., S.M., C.S., R.C.O., T.L., L.M.C., W.W., M.R., N.L., S.B.G., K.A. and V.C. performed project management; J.P.M., V.E.M., B.L.W., J.P., M.W., P.F., J.L.H., M.M. and T.R.G. contributed project oversight and advisory roles; and M.P.M., W.R.S., R.S. and L.A.G. provided overall project leadership.

**Author Information** Data have been deposited in the Gene Expression Omnibus (GEO) using accession number GSE36139 and are also available at <http://www.broadinstitute.org/cCLE>. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details accompany the full-text HTML version of the paper at [www.nature.com/nature](http://www.nature.com/nature). Readers are welcome to comment on the online version of this article at [www.nature.com/nature](http://www.nature.com/nature). Correspondence and requests for materials should be addressed to L.A.G. (Levi\_Garraway@dfci.harvard.edu) or R.S. (robert.schlegel@novartis.com).

# CORRECTIONS & AMENDMENTS

## ADDENDUM

doi:10.1038/nature11735

### Addendum: The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity

Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, Joseph Lehár, Gregory V. Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F. Berger, John E. Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa A. Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H. Engels, Jill Cheng, Guoying K. Yu, Jianjun Yu, Peter Aspesi Jr, Melanie de Silva, Kalpana Jagtap, Michael D. Jones, Li Wang, Charles Hatton, Emanuele Palescandolo, Supriya Gupta, Scott Mahan, Carrie Sougnez, Robert C. Onofrio, Ted Liefeld, Laura MacConaill, Wendy Winckler, Michael Reich, Nanxin Li, Jill P. Mesirov, Stacey B. Gabriel, Gad Getz, Kristin Ardlie, Vivien Chan, Vic E. Myer, Barbara L. Weber, Jeff Porter, Markus Warmuth, Peter Finan, Jennifer L. Harris, Matthew Meyerson, Todd R. Golub, Michael P. Morrissey, William R. Sellers, Robert Schlegel & Levi A. Garraway

Nature 483, 603–607 (2012); doi:10.1038/nature11003

In the Supplementary Information of this Letter, the use of distinct data normalization and directionality methods for pharmacological response calculations caused minor inconsistencies. We have therefore updated Supplementary Table 11 and some of the Supplementary Figures to resolve any confusion (see the Supplementary Information to this Addendum). We also wish to describe the relevant drug sensitivity normalization and response score calculations more completely.

Two versions of the drug response data were generated. First, raw activity values were calculated at each dose as  $A = 100(T/U - 1)$ , in

which  $T$  represents the Cell Titer Glo (CTG) level measured for the compound-treated well, and  $U$  is the median level of the untreated wells across the plate. This raw  $A$  is 0% with no drug and 100% for fully active compounds, when no CTG is detected. Second, the data were adjusted to a plate surface pattern and normalized to the MG132 positive control, as described in the Supplementary Methods. This normalized  $A$  is also 0% with no drug, but 100% corresponds to the median MG132 response on that plate. Although normalized drug responses were used to determine  $EC_{50}$ ,  $IC_{50}$  and  $A_{max}$  values, we used the raw drug responses for calculating the activity area (AA). This distinction is now clear in the corrected Supplementary Table 11 (the two AA measures, derived from raw or normalized data, correlate closely:  $r = 0.98$ ).

The activity is the sum of differences between the measured  $A_i$  at concentration  $i$  and  $A = 0$ , excluding positive  $A$  values:  $AA = \sum i\{0 - \min(0, A_i/100)\}$ . This AA has a value of 0 with no drug, and +8 for a compound inhibiting at  $A = 100\%$  at all eight drug concentrations, as illustrated in Fig. 2b of the original Letter. We hope that this definition eliminates any confusion that may have existed in the original Supplementary Methods (page 13) and enables others to reproduce our AA results starting from the raw drug sensitivity data. As a further means of clarification, we have added three columns to Supplementary Table 11 showing the raw (non-normalized) response data necessary to calculate AA, MG132 activity, and AA derived from normalized response data.

In addition, although all computational analyses used the above AA formula, a few Supplementary Figures (Supplementary Figs 6, 11, 9 and 14b) used a scale showing  $8 - AA$ . This value was used for display purposes, so that low values corresponded to sensitive cell lines and the visualization remained consistent with other sensitivity metrics ( $IC_{50}$ ,  $A_{max}$ ). This specification was noted in Supplementary Fig. 8 but had been inadvertently cut off the Supplementary Fig. 9 legend. We have therefore updated the Supplementary Fig. 9 legend to clarify where an inverted scale was used, and updated the scale of Supplementary Figs 6, 11 and 14b to reflect our definition of AA (noted above).

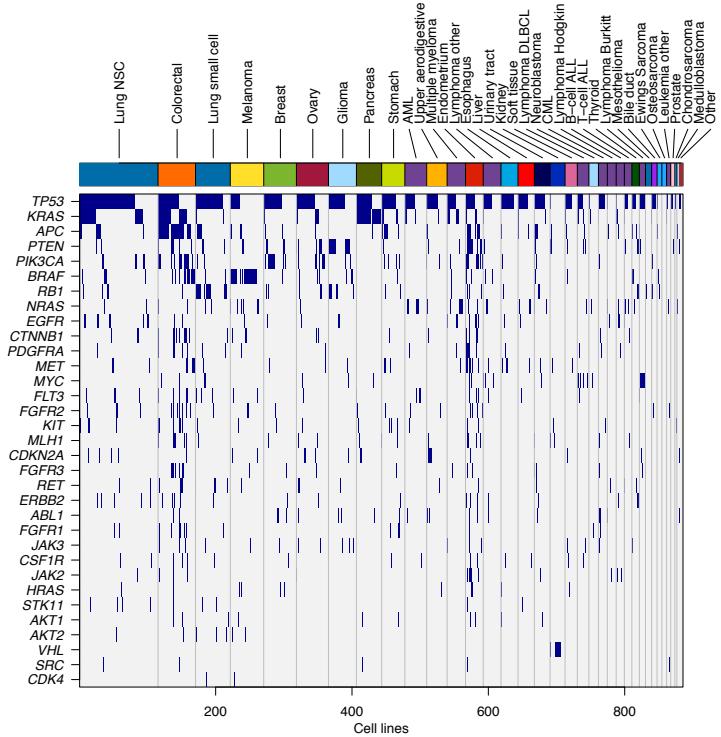
These changes do not affect the analyses, results or scientific conclusions presented in the paper. The authors are indebted to B. Yadav, who alerted them to these inconsistencies.

**Supplementary Information** is available in the online version of the Addendum.

# SUPPLEMENTARY INFORMATION

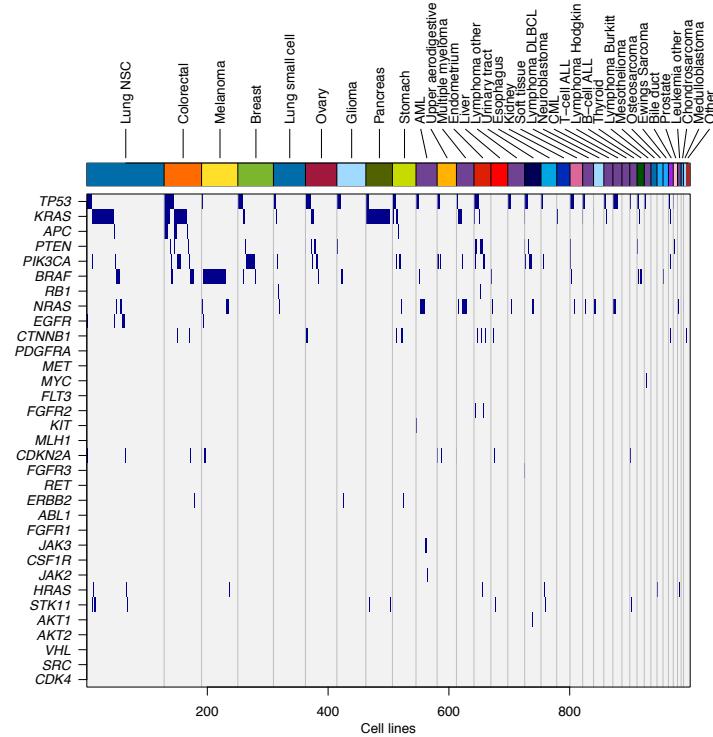
doi:10.1038/nature11003

a



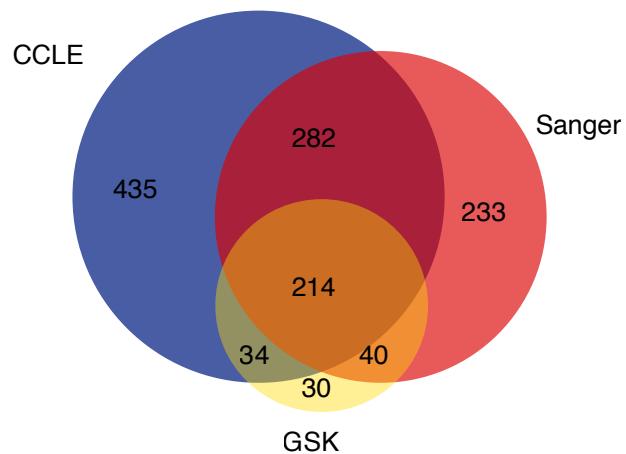
Mutation status derived from the hybrid capture data

b

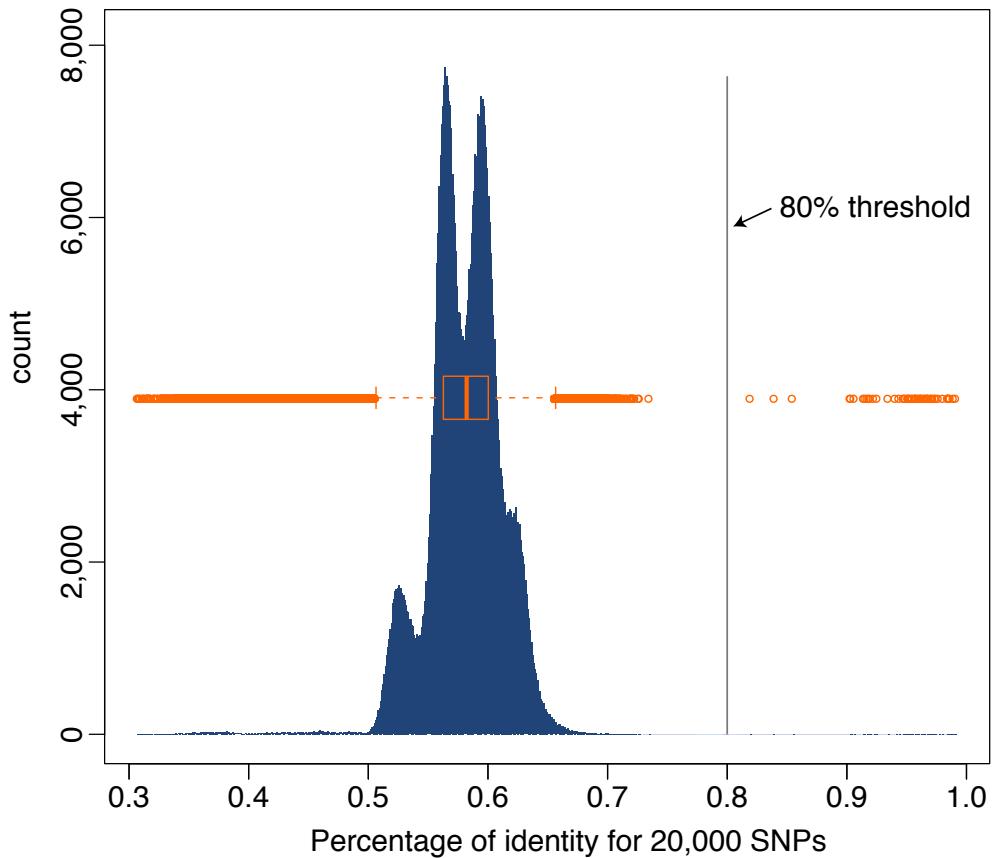


Mutation status derived from the Oncomap data (392 mutations assayed)

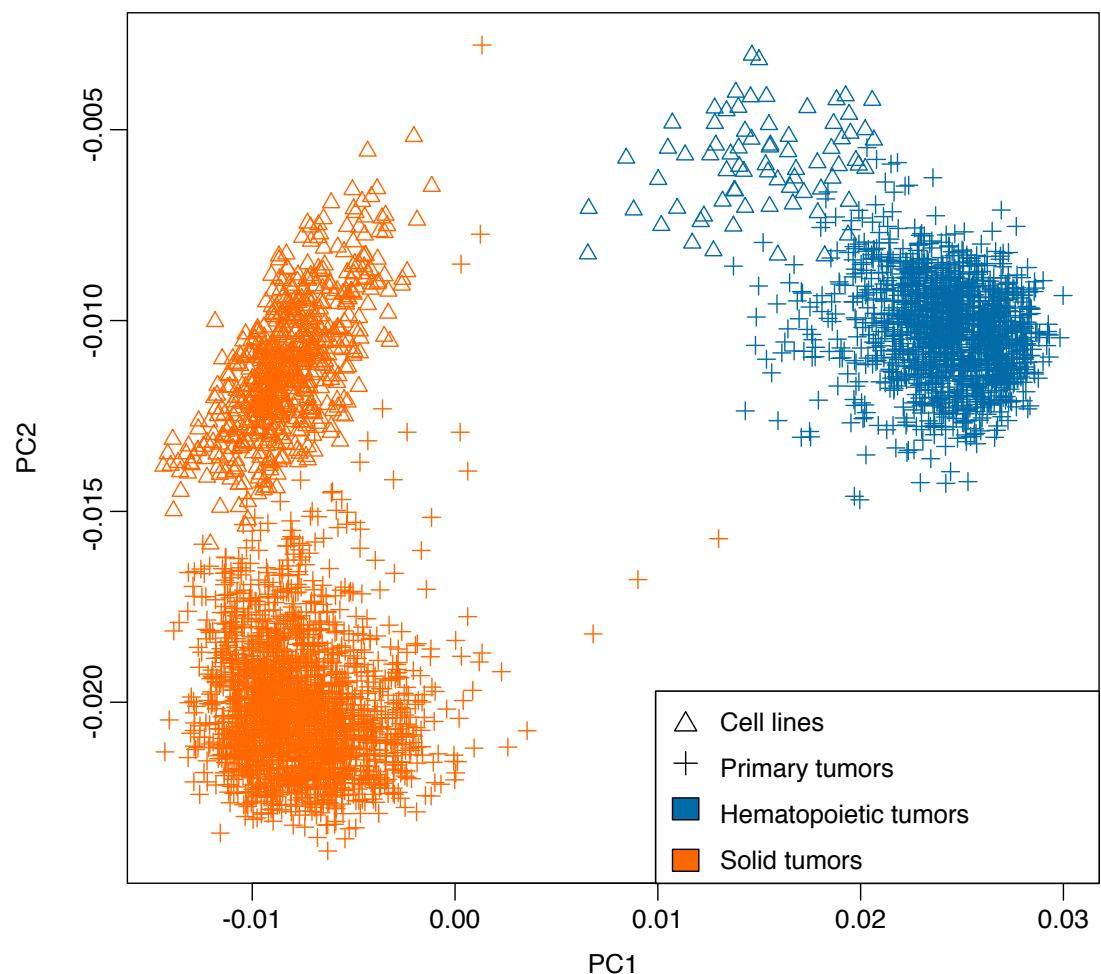
**Supplementary Figure 1:** Matrix-representation of point mutation data derived from hybrid capture/sequencing (a) or OncoMap analysis (b) across all CCLE cell lines. Blue bars indicate detection of at least one mutation in the relevant gene. Cell lines are sorted by lineage and cancer gene mutation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



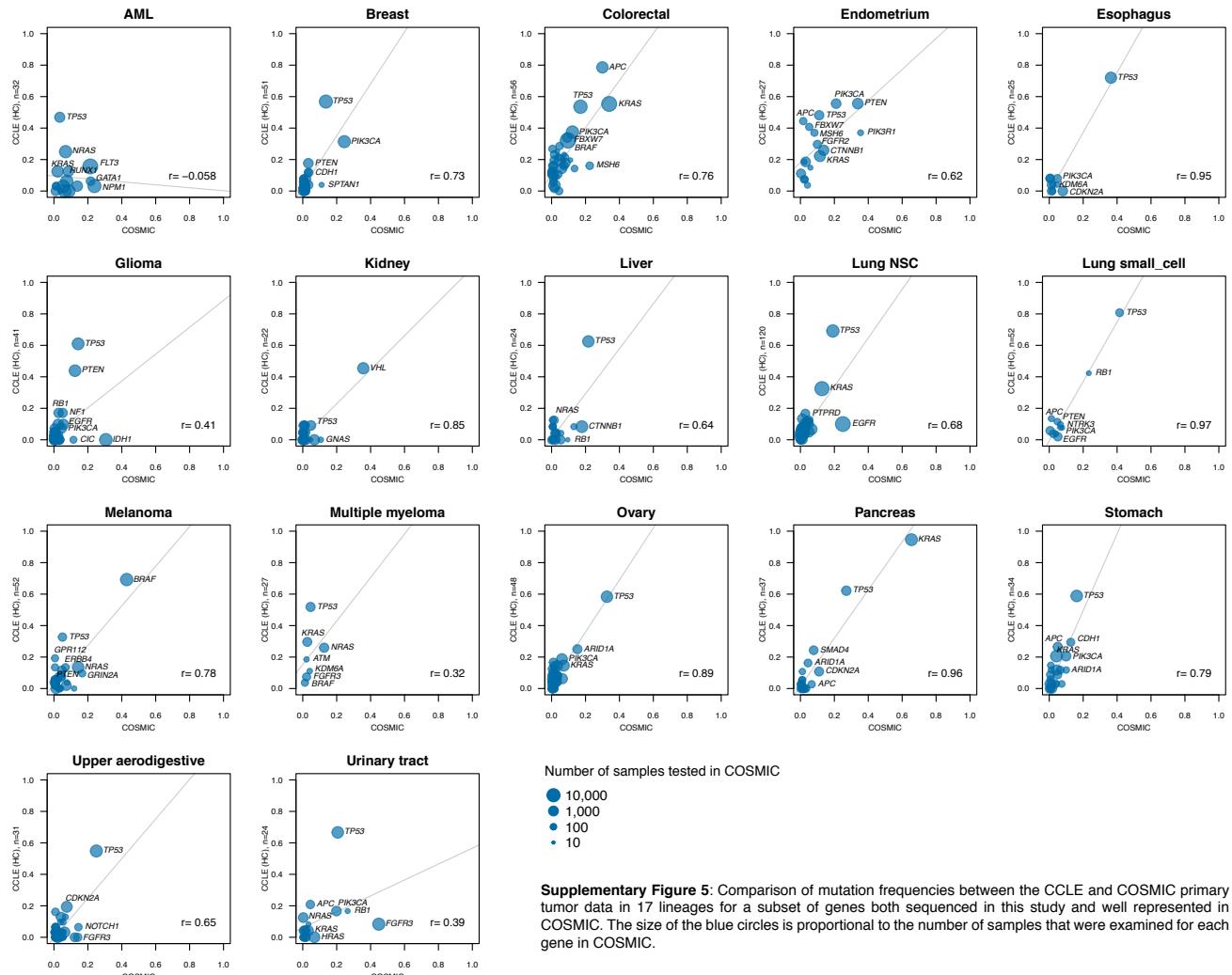
**Supplementary Figure 2:** Venn diagram of overlap between the cell lines in the CCLE, the Sanger Institute's Cancer Cell Line Project (<http://www.sanger.ac.uk/genetics/CGP/CellLines/>) and the GlaxoSmithKline (GSK) Cancer Cell Line Genomic Profiling dataset ([https://cabig.nci.nih.gov/caArray\\_GSKdata/](https://cabig.nci.nih.gov/caArray_GSKdata/)).

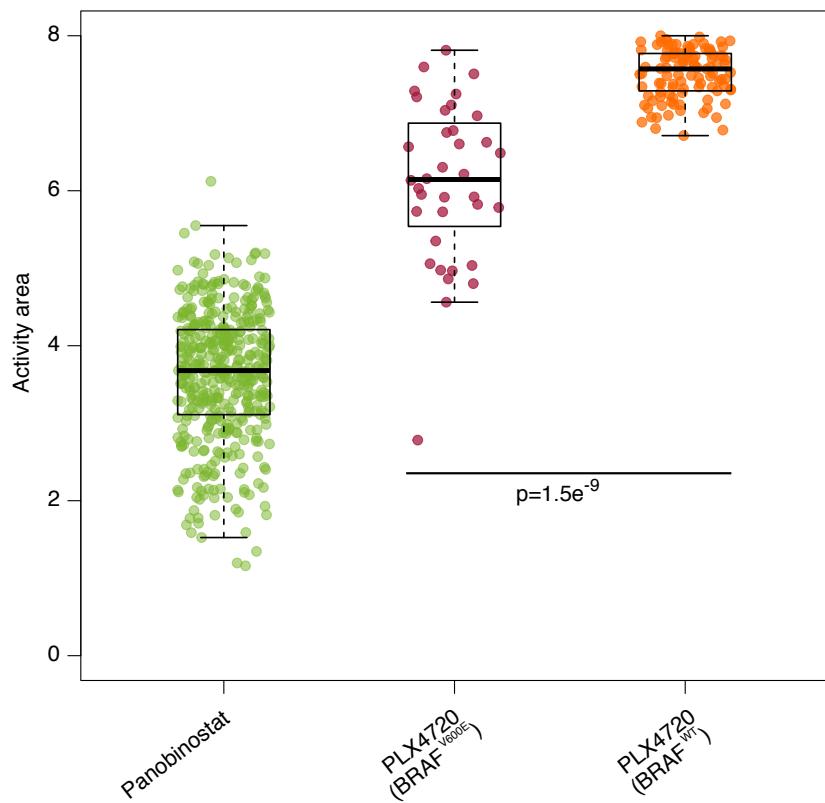


**Supplementary Figure 3:** Distribution of the percentages of SNP identity across cell lines. The boxplot (orange) corresponds to the same values shown as a histogram in blue. Values above 80% identity denote cell line pairs presumed to derive from the same individual.

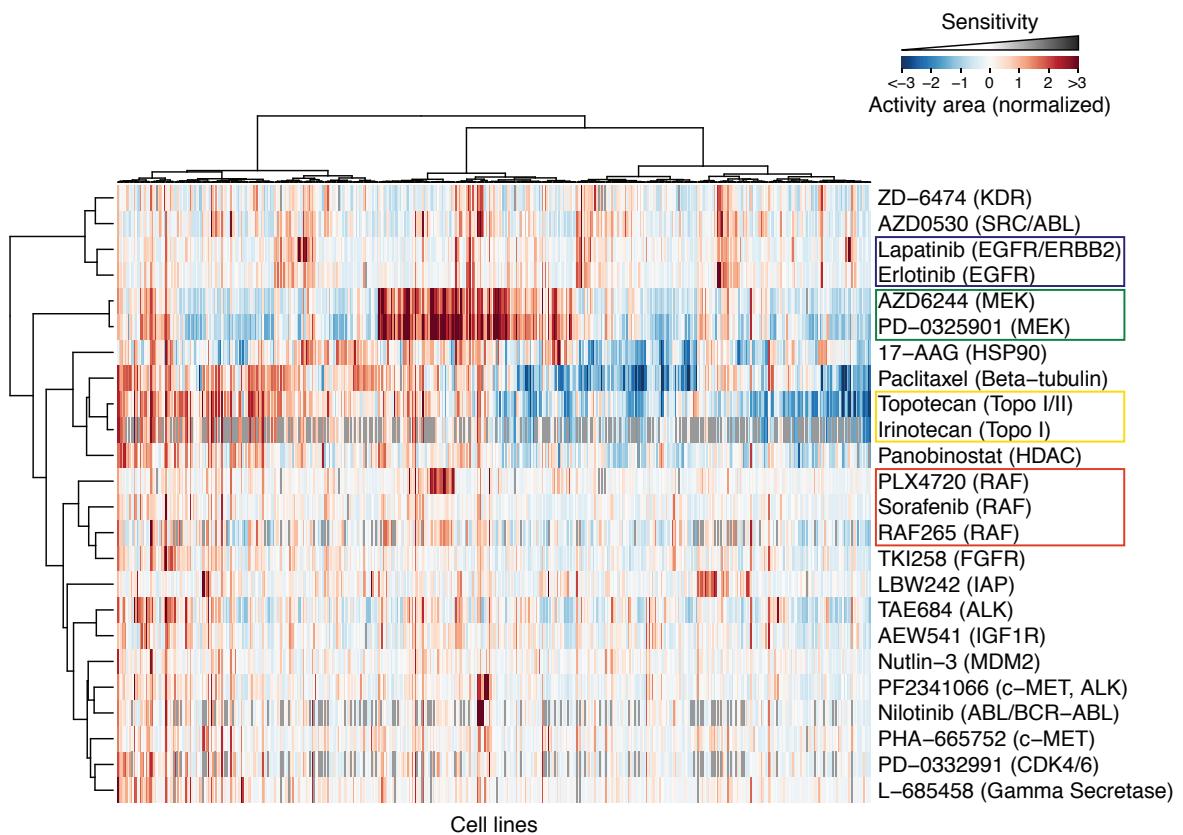


**Supplementary Figure 4:** Principal component analysis of expression data from cell lines and primary tumors, for the 1,000 most varying genes. Those derived from solid tumors are shown in orange, and those corresponding to hematopoietic lineages are shown in blue. This figure indicates that there 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.

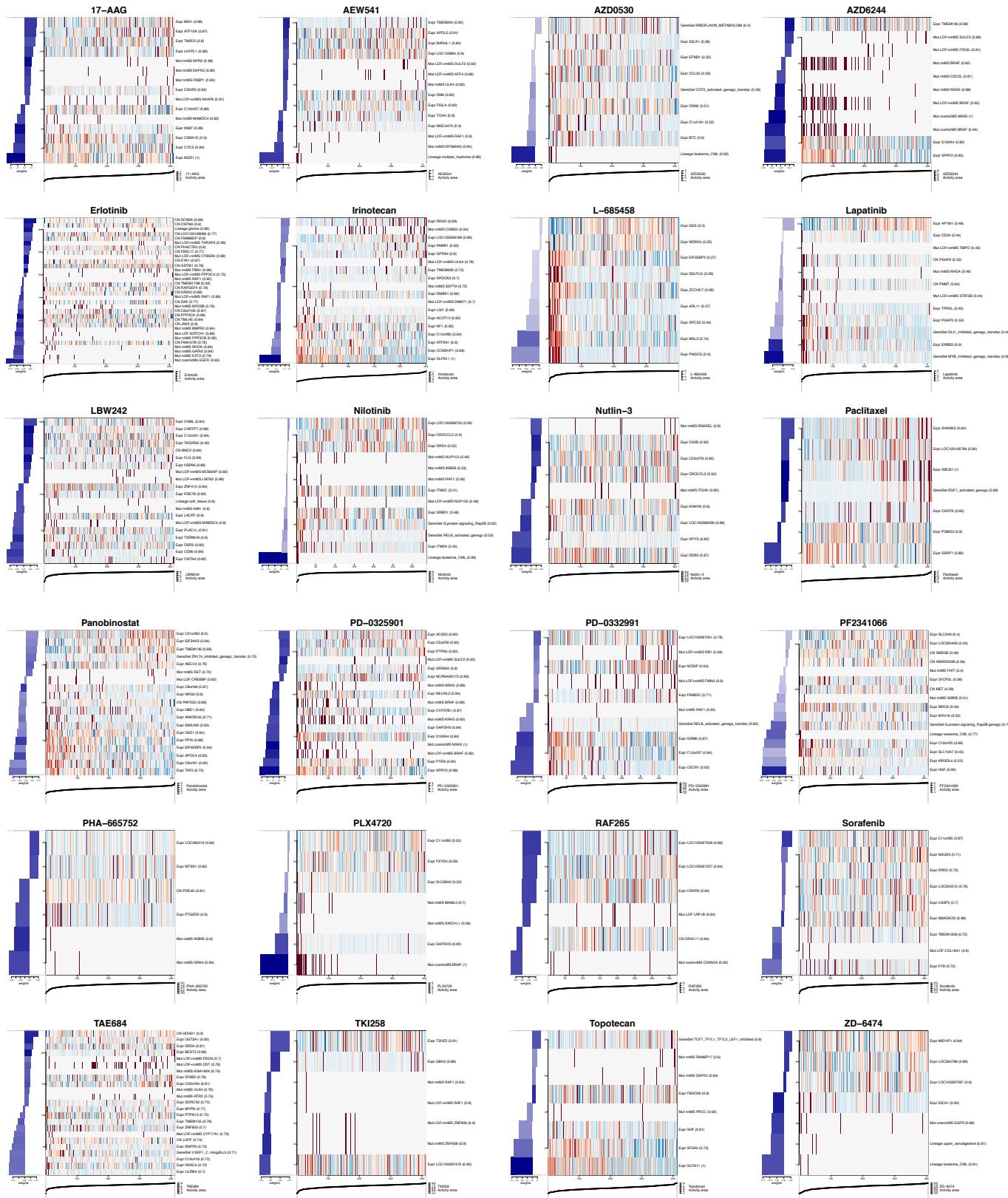




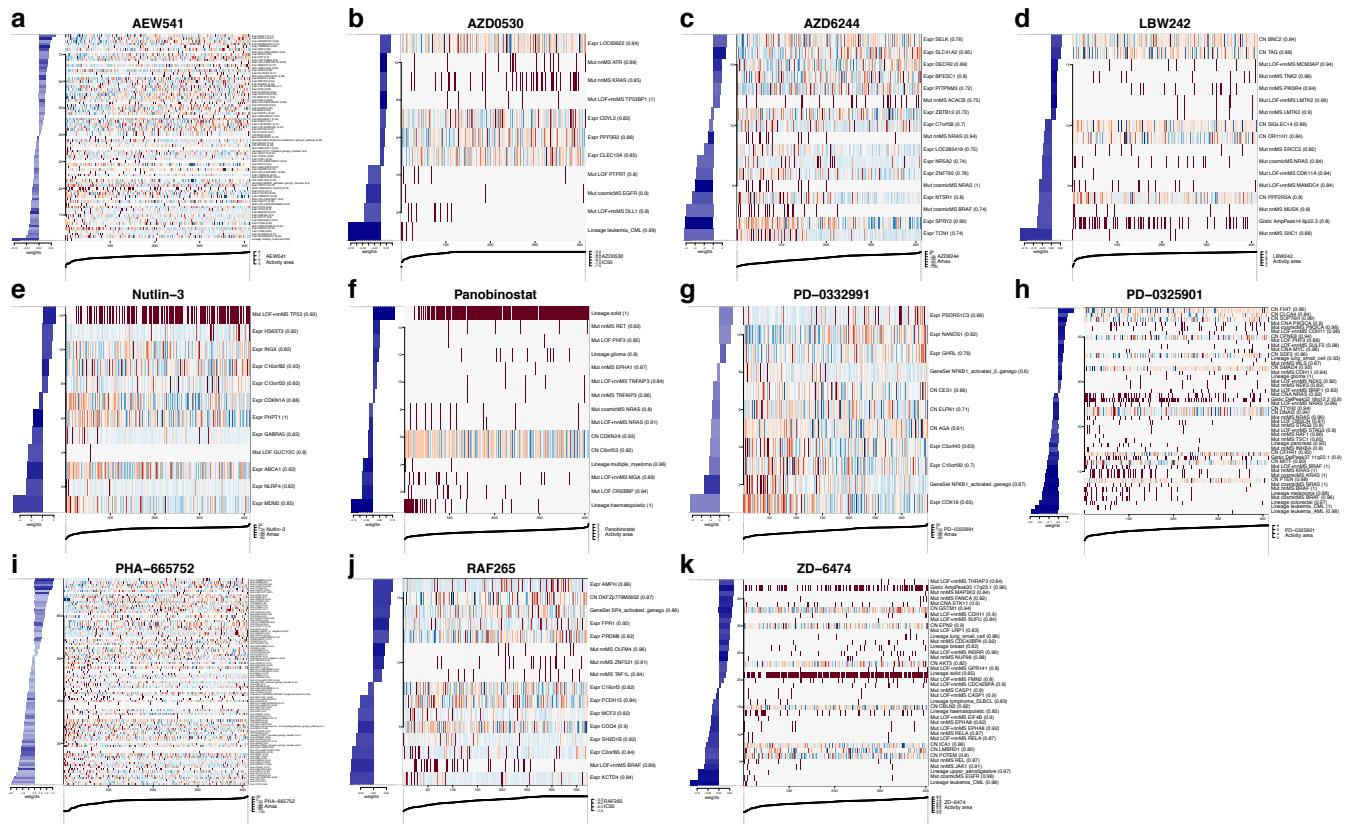
**Supplementary Figure 6:** Drug responses for Panobinostat (green) and PLX4720 (orange/purple), as measured by the activity area.



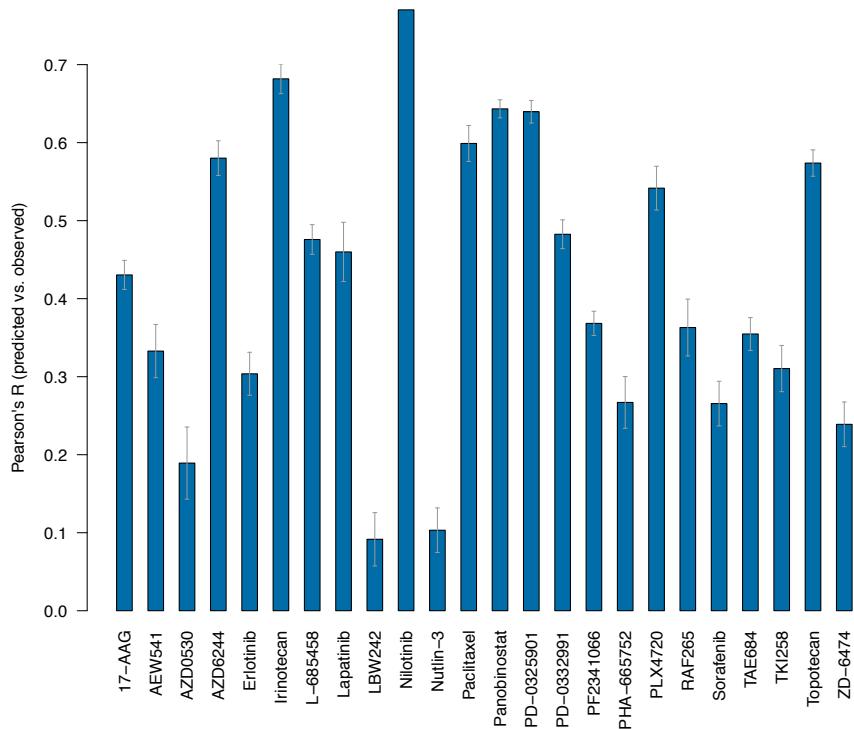
**Supplementary Figure 7:** Clustered response profiles across all cell lines for all the tested compounds, based on the activity area between the response curve and the zero dose effect level, up to the top tested concentration of 8  $\mu$ M. Each row was median-normalized. Colored boxes denote compounds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



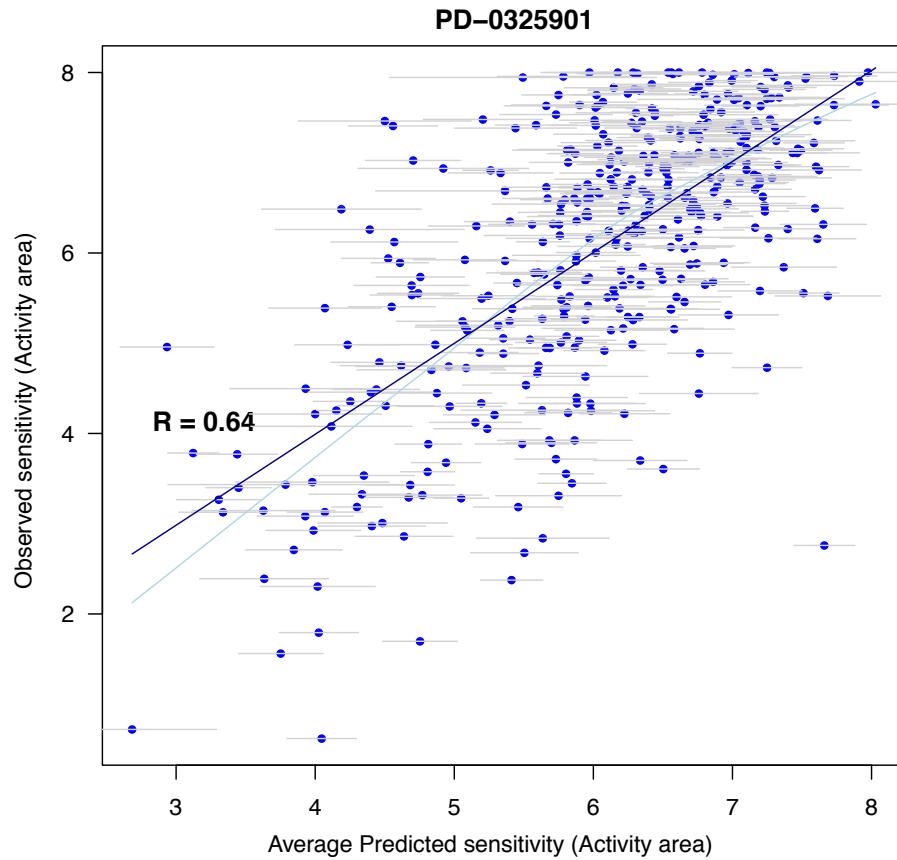
**Supplementary Figure 8:** Elastic net-based predictive modeling results for 24 compounds using all features as input and activity area as the sensitivity metric. In this figure (and some panels in supplementary figure 9),  $\delta$  – activity area was used as the vector of response values, so that 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.



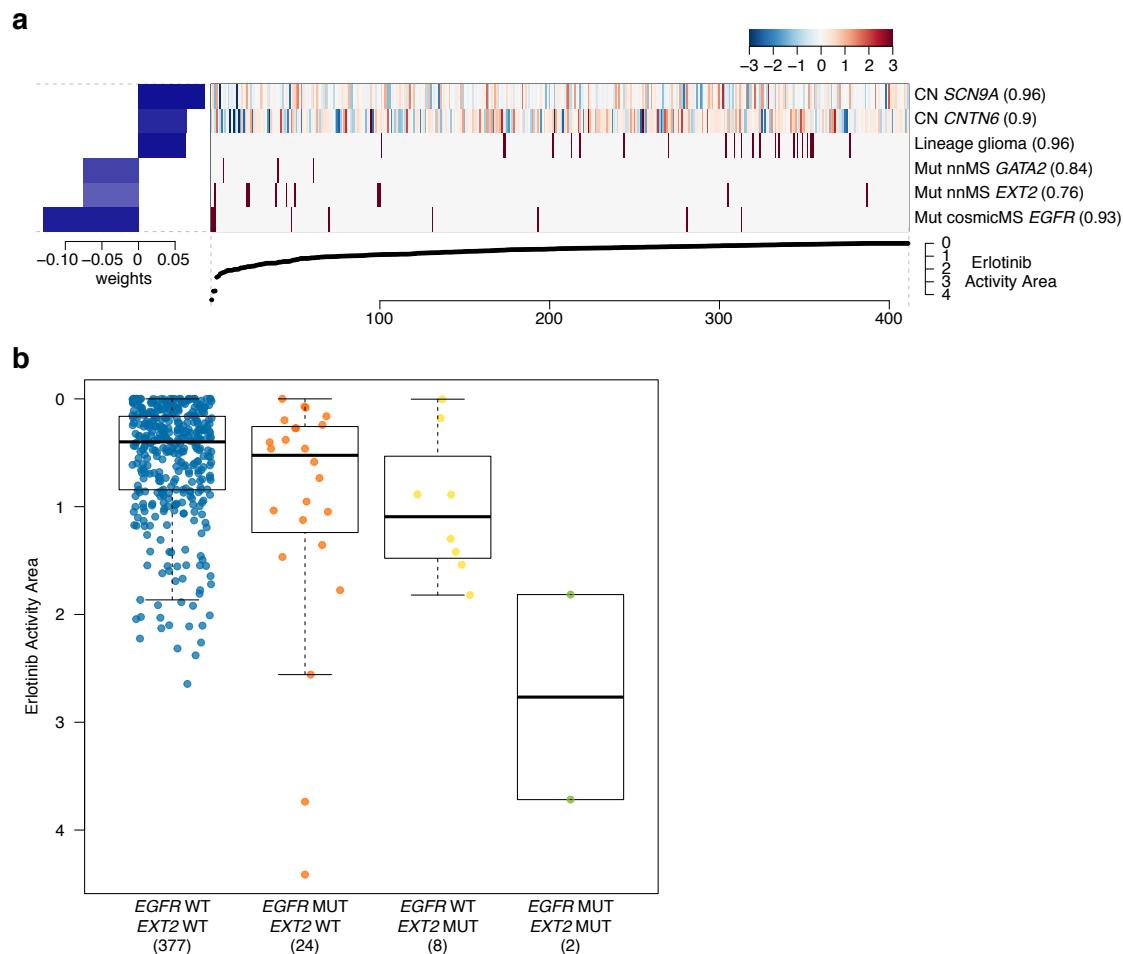
**Supplementary Figure 9:** Elastic net models built using a reduced feature set and/or utilizing  $A_{\max}$  or  $IC_{50}$  as the sensitivity metric and/or with a lower bootstrapping cutoff so as to show more features. Panel d, f, h and k were built using copy number, mutation, and lineage data only, 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



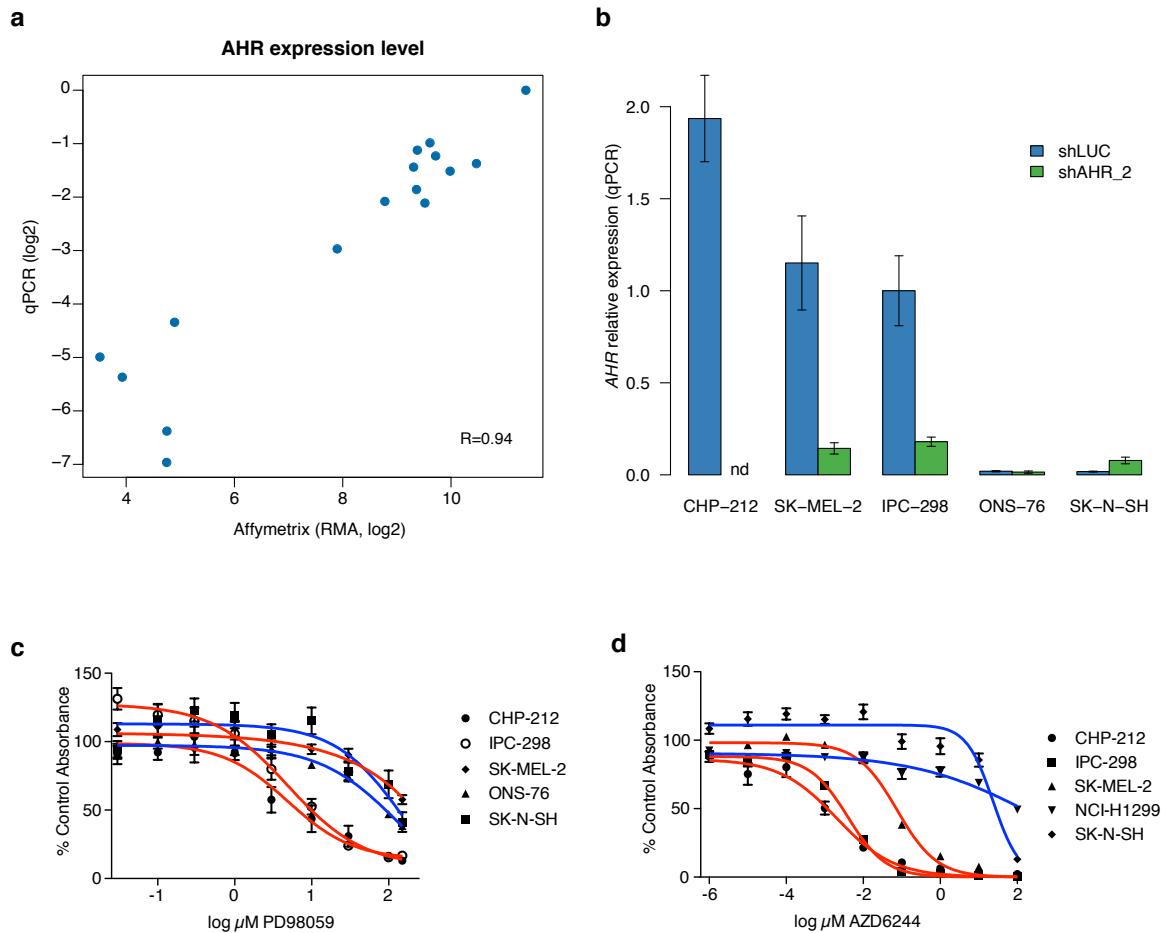
**Supplementary Figure 10:** Performance of the predictive models in Supplementary Figure 8 as measured by Pearson's correlation coefficients between observed and predicted sensitivity values (mean of 10 iterations of 10-fold cross-validations).



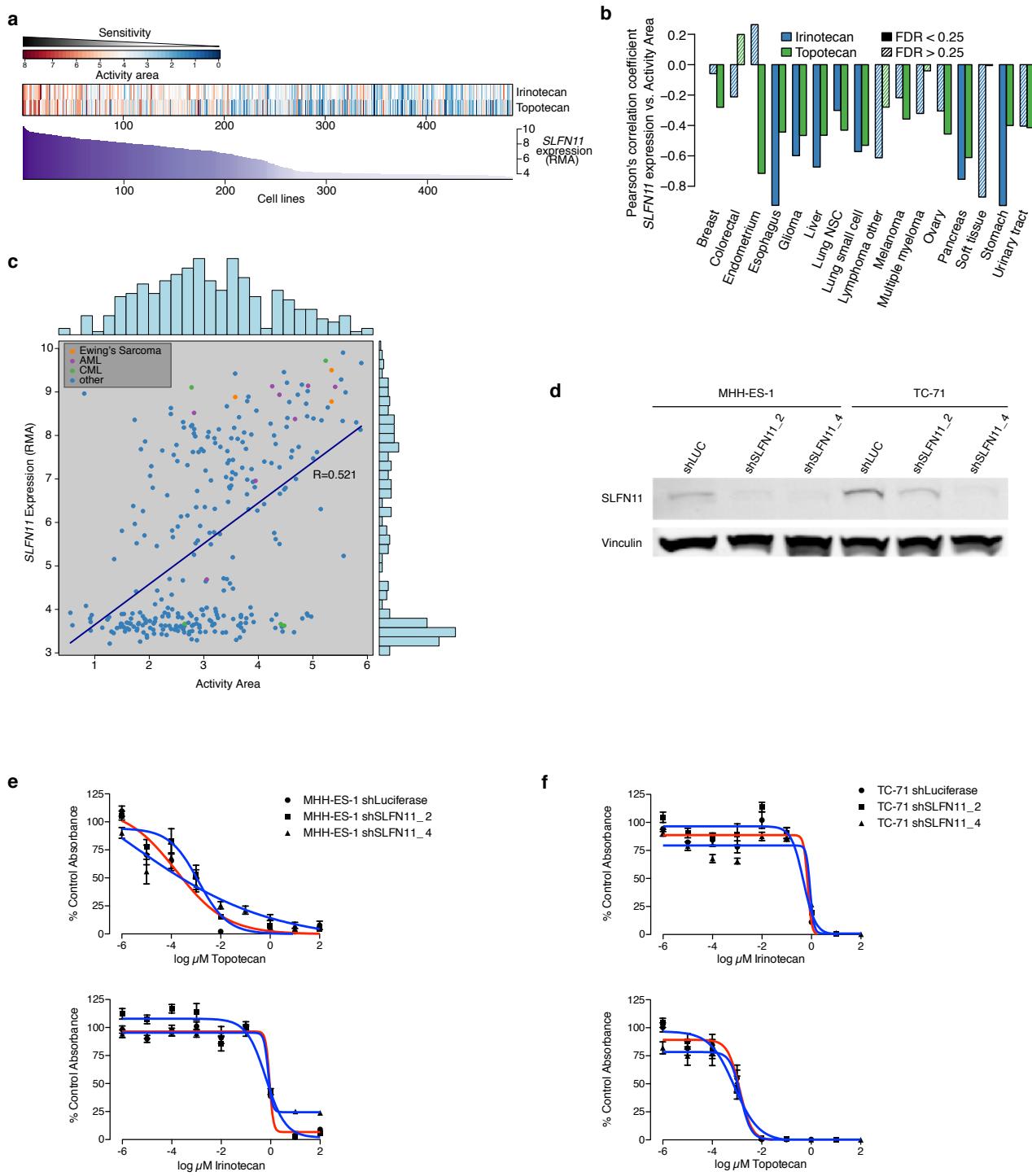
**Supplementary Figure 11:** Predicted sensitivity values for PD-0325901 (Activity area), versus observed sensitivities. Horizontal grey bars represent the standard deviation of the predicted values across 10 iterations of cross-validation.



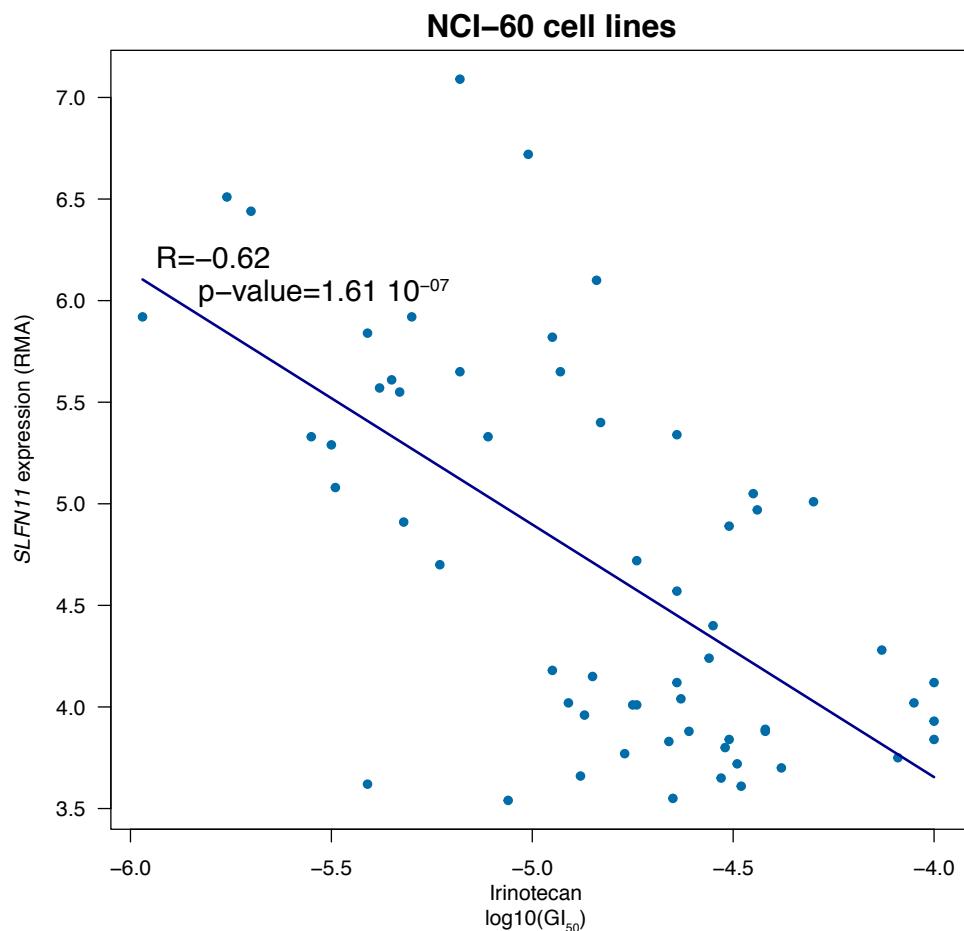
**Supplementary Figure 12:** *EXT2* variations may enrich for sensitivity to EGFR inhibitors. **a.** Elastic net-based predictive modeling results for Erlotinib using DNA copy number, mutations and lineages as input, and activity area as the sensitivity metric. *EGFR* mutations at or near recurrent COSMIC 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<sup>41</sup>. 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.**

## Supplementary Methods

### 1. Cell line selection and annotation

After curating a list of 1461 cancer cell lines available from public repositories, we targeted 1021 cell lines for possible acquisition. A total of 947 human cancer cell lines were obtained, cultured and processed from commercial vendors in the U.S., Germany, the U.K., Japan, Italy, and South Korea. These include:

ATCC (<http://www.atcc.org/>)  
DSMZ (<http://www.dsmz.de/>),  
ECACC (<http://www.hpacultures.org.uk/collections/ecacc.jsp>)  
HSRRB (<http://www.jhsf.or.jp/English/hsrrb.html>)  
RIKEN (<http://www.brc.riken.jp/lab/cell/english/>)  
ICLC (<http://www.iclc.it/Listanuova.html>)  
KCLB (<http://cellbank.snu.ac.kr/english/index.php>)

Vendors are listed in **Supplementary Table 1**. A small number of lines were obtained from academic labs. The final cell line collection spans 36 cancer types. Representation of cell lines for each cancer type was mainly driven by cancer mortality in the United States, as a surrogate of unmet medical need, as well as availability. Briefly, for cancer types with >7,000 deaths/year, a maximum of 60 cell lines were obtained; for the other types, the minimum number of cell lines per cancer type was set to 15 whenever possible.

We created an annotation pipeline that would take cell line names as input to generate “COSMIC-compatible” anatomic and histologic annotations for the collection ([http://www.sanger.ac.uk/genetics/CGP/cosmic/data/cosmic\\_classification\\_alias\\_list\\_27\\_01\\_11.xls](http://www.sanger.ac.uk/genetics/CGP/cosmic/data/cosmic_classification_alias_list_27_01_11.xls)). In some cases the COSMIC classification was known, but in others the annotation was inferred by keywords derived from scientific or commercial literature. If no prior internal consensus decision on the annotation could be gleaned, then the CosmicMint ([ftp://ftp.sanger.ac.uk/pub/CGP/cosmic/cell\\_line\\_export/](ftp://ftp.sanger.ac.uk/pub/CGP/cosmic/cell_line_export/)) was queried to find a putative assignment. If an assignment could not be pulled from the CosmicMint, then keywords for primary anatomic site and histology were used to infer the most likely COSMIC-compatible annotation. Gender, race, presence of metastases, and anatomic site from which the tumor cells were obtained were also assigned based on literature or vendor information when available. In cases of CosmicMint-assigned or keyword-inferred annotations, the results were manually reviewed by internal pathologists and biologists.

Additional steps were taken to ensure that cell line names were consistent in instances where syntax or punctuation differences occurred between names or aliases. Here, we verified that the “punctuation-free” name and primary site combination was uniquely and correctly identified. Consider the “T.T” and “TT” cell lines as an example:

their “clean” primary site disambiguated names became TT\_OESOPHAGUS and TT\_THYROID respectively.

Multiple quality control steps were incorporated at each stage of cell culture and data production to ensure consistency of all datasets. We confirmed the identity of the cell lines at multiple steps using SNP fingerprinting. The possibility of cross-contamination was ruled out by comparing SNP genotypes derived from the SNP arrays generated by the CCLE effort as well as external cell line characterization studies (see “Genomic characterization” and “Cell lines identity validation by SNP genotyping”).

## 2. DNA and RNA extraction

Cells were cultured according to vendors’ instructions for propagation and preservation. DNA was isolated from frozen cell pellets that contained 1-10 x 10<sup>6</sup> cells (average cell count ~4.5 million cells) using the Qiagen Gentra Puregene method. Briefly, Cell Lysis and RNAase A solutions were added to each frozen pellet and vortexed, and samples were incubated until fully homogenized. Protein Precipitation solution was added and samples were vortexed vigorously for 20 seconds. Samples were centrifuged at 3000 g for 15 minutes, and the supernatant was poured into a new tube. 1 mL of 100% isopropanol was added and mixed by inversion to precipitate the DNA. Samples were centrifuged at 3000 g for 5 minutes to pellet the DNA. The supernatant was discarded and the pellet was dried for 1 minute. Next, 70% ethanol was added (1 mL) and the sample was pelleted by centrifugation at 3000 g for 2 minutes. The supernatant was discarded and pellet was allowed to dry for 10 minutes before hydration with 100-200 µL of DNA Hydration solution (Qiagen). DNA pellets were rehydrated for 1 hour at 65 °C or overnight at room temperature and stored at 4 °C. For long-term storage, all samples were stored at -20 °C. DNA samples were quantified using Picogreen (ThermoScientific Varioskan Flash instrument). Additionally an aliquot of each sample was also run on a quality gel (Invitrogen 1% Agarose E-gel) to assess the quality of the material.

RNA was isolated from frozen cell pellets containing 1-10 x 10<sup>6</sup> cells using Trizol (Invitrogen). Briefly, 1 mL of Trizol was added to each pellet and the pellet was resuspended by pipetting. The lysate suspension was transferred to a 1.5 mL Eppendorf tube and incubated at room temperature for 5 minutes. Next, chloroform (200 µL) was added to each sample and mixed by pipetting. Samples were centrifuged at 13,000 rpm for 10 minutes at 4 °C. The aqueous (upper) phase was transferred to a fresh 1.5 mL Eppendorf tube, and isopropanol (500 µL) was added. Samples were mixed and incubated at room temperature for 10 minutes, followed by centrifugation at 13,000 rpm for 10 minutes at 4 °C. The supernatant was discarded. The remaining RNA pellet was

washed with 75% ethanol and centrifuged at 5,000 rpm for 5 minutes at 4 °C. Ethanol was removed carefully so as not to dislodge the pellet, and the RNA pellets were then allowed to dry for 10 minutes (room air). RNA pellets were rehydrated in 60 µL of DEPC treated water. Samples were incubated at 5 minutes at 55 °C and stored at -80 °C. RNA samples were quantified using both the Nanodrop 8000 spectrophotometer (ThermoScientific) and via Agilent 2100 Bioanalyzer. Only samples with a RIN of 7.0 or higher were considered passing. Samples that failed quality control were re-isolated from a second frozen pellet.

### 3. Genomic characterization

#### *SNP Arrays:*

Cell line genomic DNA was hybridized to the genome-wide human Affymetrix SNP Array 6.0 according to the manufacturer's instructions and analyzed as described previously<sup>1</sup>. Briefly, DNA was digested with NspI and StyI enzymes (New England Biolabs), ligated to the respective Affymetrix adapters using T4 DNA ligase (New England Biolabs), amplified (Clontech), purified using magnetic beads (Agencourt), labeled, fragmented, and hybridized to the arrays. Following hybridization, the arrays were washed and stained with streptavidin-phycoerythrin (Invitrogen). Array preparation and scanning was performed by the Genetics Analysis Platform (GAP) at the Broad Institute.

The raw CEL files were normalized to copy number estimates using a GenePattern pipeline, as described previously<sup>1</sup> and hg18 Affymetrix probe annotations. Normalized copy number estimates ( $\log_2$  ratios) were segmented using the Circular Binary Segmentation (CBS) algorithm, followed by median centering of the segment values to a value of zero in each sample. Next, quality checking of each array was performed, including visual inspection of the array pseudo-images, probe-to-probe noise variation between copy-number values, confidence levels of Birdseed<sup>2</sup> genotyping calls, and appropriate segmentation of the copy-number profiles. Finally, the Genomic Identification of Significant Targets in Cancer (GISTIC) algorithm<sup>3</sup> was used to identify focal regions of copy number alterations in individual samples. A gene-level copy-number was also generated, defined as the maximum absolute segmented value between the gene's genomic coordinates, and calculated for all genes and miRNA using the hg18 coordinates provided by the refFlat and wgRna databases from UCSC Genome Browser (<http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/>).

#### *Expression arrays:*

mRNA expression data was obtained using Affymetrix Human Genome U133

Plus 2.0 arrays according to the manufacturer's instructions. Array preparation and scanning was performed by the Genomics Analysis Platform at the Broad Institute. Gene-centric expression values were obtained using updated Affymetrix probe set definition files (CDF files) from Brainarray<sup>4</sup>; and background correction was accomplished using RMA (Robust Multichip Average)<sup>5</sup> and quantile normalization<sup>6</sup>. Quality assessment was performed to identify low performing microarrays, using the R package affyPLM<sup>7</sup>. Outliers in the distribution of NUSe, RLE, background signal and percentage of "present" genes were flagged to be re-processed. In addition, all microarray pseudoimages were checked visually.

*Mass spectrometric mutation detection:*

Mutation data was generated for specific cancer gene loci using the mass spectrometric genotyping based OncoMap platform, and data analysis was performed as previously described<sup>8</sup>. In brief, we used a panel of 456 genotyping assays representing 392 mutations (380 unique amino-acid substitutions) in 33 genes (**Supplementary Table 2**). In a first phase, we identified candidate mutations in a multiplexed fashion, using 19 pools of 24 assays, and with relaxed detection thresholds. All positive mutation calls were subsequently validated using non-multiplexed assays and hME chemistry<sup>8</sup> (Sequenom).

*Solution phase hybrid capture and massively parallel sequencing:*

**Sequencing data generation:** 1,651 protein-coding genes were selected for sequencing based on their known or potential involvement in tumor biology, according to one or more of the following criteria: (1) genes identified as somatically altered in cancer based on (A) occurrence in at least 4 instances, collectively, from recently published literature<sup>1,9-13</sup>, (B) occurrence in 2 – 3 instances from the aforementioned studies and also present in a significantly amplified or deleted focal peak in primary tumors<sup>14</sup> or cell lines (this study), (C) membership in the Cancer Gene Census<sup>15</sup>, or (D) significant mutation frequency across 441 tumors<sup>16</sup>; (2) genes identified in either the literature, or meeting abstracts and presentations, as putative oncogenes, tumor suppressors, members of cancer related pathways, or having a cancer-related function(s); or (3) protein kinases.

Multiplexed libraries for exome capture sequencing were constructed as previously described<sup>17</sup> utilizing the custom SureSelect Target Enrichment System (Agilent Technologies). Cell line genomic DNA was sheared and ligated to Illumina sequencing adapters, including 8 bp indexes. Adaptor ligated DNA was then size-selected for lengths between 200-350 bp and hybridized with an excess of bait in solution phase, as described previously<sup>17,18</sup>. Barcoded exon capture libraries were then pooled and sequenced on Illumina instruments (76 bp paired-end reads)<sup>17</sup>. The 8 bp index was used to assign sequencing reads to a particular sample in the downstream data aggregation

pipeline.

The median value for the average coverage across cell lines was 121x, while the median value for the fraction of targeted sequences with depth of coverage equal to or higher than 10x reached 84.3%.

**The sequencing data-processing pipeline (“Picard pipeline”):** We generated a BAM file for each sample using the sequencing data processing pipeline known as “Picard” (<http://picard.sourceforge.net>). Picard consists of four steps, described in detail in<sup>19</sup>, but with the following modifications in the “Alignment to the genome” step: Alignment was performed using BWA<sup>20</sup> (<http://bio-bwa.sourceforge.net>) to the NCBI Human Reference Genome GRCh37.

**Quality control:** We verified concordance between genotypes detected by sequencing and SNP arrays to ensure that there were no mix-ups between samples. In addition, sequencing reads aggregated from different barcoded pools were checked for genotype concordance, to ensure sample identity.

**Local realignment:** Sequence reads corresponding to genomic regions that may harbor small insertions or deletions (indels) were jointly realigned to improve detection of indels and to decrease the number of false positive single nucleotide variations caused by misaligned reads, particularly at the 3' end<sup>21</sup>. Sites that are likely to contain indels were defined as sites of known germline indel variation from dbSNP, sites containing reads initially aligned by BWA with indels and sites adjacent to the cluster of detected nucleotide substitutions.

**Variant calling and annotation:** Nucleotide substitutions were detected with MuTect (<http://www.broadinstitute.org/cancer/cga/MuTect>) and short indels were called with Indelocator (<http://www.broadinstitute.org/cancer/cga/Indelocator>), as described in Supplementary Material for prior studies<sup>19,22</sup>. Both programs were applied using a mode that does not require matching normal DNA and thus identifies all variants that differ from a reference genome. Variants were annotated using the Oncotator (<http://www.broadinstitute.org/cancer/cga/Oncotator>) software.

**Variant filtration by exclusion of variants with low allelic fraction:** The allelic fraction was calculated for each detected variant per cell line as a fraction of reads that supported an alternative allele (e.g., different from the reference) among reads overlapping the position. Only reads with allelic fractions above 0.25 were used in the downstream sensitivity prediction analysis.

**Variant filtration by exclusion of common germline variants:** Variants for which the global allele frequency (GAF) in dbSNP134 or allele frequency in the NHLBI

Exome Sequencing Project (<http://evs.gs.washington.edu/EVS>, data release ESP2500) was higher than 0.1% were excluded from further analysis.

**Variant filtration by exclusion of variants observed in a panel of normals:** Variants detected in a panel of 278 whole exomes sequenced at the Broad as part of the 1000 Genomes Project were excluded from further analysis. Beyond removal of additional germline variation, this step also allowed elimination of common false positives that originate predominantly from alignment artifacts.

*DNA identity analysis by mass spectrometric SNP genotyping:*

The identity of all DNA samples was assessed by mass spectrometric genotyping of two multiplexed panels of 24 SNPs (Sequenom, San Diego, CA).

#### 4. Cell line identity validation by SNP genotyping

The Birdseed<sup>2</sup> algorithm was used to call the genotypes from CCLE Affymetrix SNP 6.0 array data (processed as described above). Next, 20,000 SNPs were randomly chosen from among those interrogated by the arrays. The percentage of identity between the genotype calls of any two cell lines was then calculated in order to identify pairs that may have derived from the same individual. This analysis yielded a bimodal distribution, with a small number of cell line pairs showing a score > 80% (suggestive of genetic “identity”; **Fig. S11**). SNP array data for 14 replicates of a series of 15 Hapmap samples were also included to confirm that this threshold was consistent with identity percentages observed between known identical (or different) individuals across replicates (not shown).

All cell line pairs showing more than 80% genotypic identity were subject to more detailed review. In some cases, they corresponded to pairs known to have been sampled from the same individual (e.g. a primary tumor- and a metastasis-derived cell line). Other unexpected cases may reflect heretofore unrecognized cross-contamination during in vitro cultivation at some point during the cell lines’ history. In ambiguous cases, cell lines were re-purchased from the original vendor and all data derived from the initial stock was discarded.

As an additional confirmation of cell line identity, CCLE SNP array data was compared to publicly available data from the CGP cell line project (<http://www.sanger.ac.uk/genetics/CGP/Archive/>), after processing this data as described above. Cell lines with matching names but non-matching SNP genotypes—or cell lines with matching genotypes but non-matching names/aliases—were manually reviewed. In cases of doubt, cell lines were re-purchased from the original vendor and all ambiguous

data was discarded, as described above.

## 5. Gene set activity scores

Gene expression values were Z-normalized and additively combined into “pathway scores” for gene sets derived from 1) Molecular Signatures Database (MSigDB)<sup>23</sup> version 2.5 and 2) MetaBase from GeneGo Incl (<http://www.genego.com>). In particular, we used the following subsets: 386 positional gene sets (C1), 630 canonical pathways (C2) and 837 motif gene sets (C3) from MSigDB; 570 canonical pathways, and 716 transcription factors directional gene sets from GeneGo.

## 6. Cell line-to-primary tumor comparison

The cell lines and primary tumors were compared by measuring the feature correlations from one sample type to the other, broken down by cancer lineages. Because there is no direct correspondence between cell lines and tumors, the feature sets for each cell line were correlated with the average across tumors in that lineage, and vice-versa. Thus, the resulting correlation matrices are asymmetric: the top left showing how well the tumor features correlate with the average of the cell lines in a lineage, and the bottom right showing the converse. The diagonal shows the agreement between sample types within each cancer lineage.

### *Copy-number comparison:*

To compare chromosomal copy number alterations between cell lines and primary tumors samples, we used segmented DNA copy number profiles from the Tumorscape<sup>14</sup> website (<http://www.broadinstitute.org/tumorscape/>). Here, 12 tumor types were selected that were common to both CCLE and Tumorscape and contained at least 15 samples in each tumor type. The resulting dataset spanned 452 cell lines and 1,515 primary tumors. We then gathered the breakpoints of all cell lines and primary tumors from the segmented profiles and recorded the copy-number values for each sample at positions proximal to each breakpoint. That is, for breakpoint  $B_j$  at genomic position  $P_j$ , the copy-number value recorded for each sample was the copy-number value at  $P_j - 1$ . Regions of known germline copy-number variation were removed from the data. Next, we calculated the means of the copy-number values ( $G$  scores<sup>24</sup>) for each tumor type separately for the cell lines and primary tumors to obtain two matrices of  $G$  scores at all breakpoints and across all tumor types. Finally, we calculated the pair-wise Pearson correlation coefficient of copy-number profiles between the two matrices.

### *Expression comparison:*

We assembled a set of Affymetrix U133+2.0 expression arrays from public repositories using the *expO* (<http://www.intgen.org/expo/>) dataset (GEO accession GSE2109, consisting of 2,158 arrays from solid tumors), MILE<sup>25,26</sup> (GEO accession GSE13159, consisting of 2,096 arrays from hematopoietic tumors), other datasets for primary tumors (GEO accession GSE12102 – 37 samples), and 679 expression arrays from the CCLE. All arrays were normalized as a single collection using the approach described above. All cancer types with at least 10 primary tumor samples and 7 cell lines were retained for downstream analysis (these represented 18 tumor types in total). Next, we restricted the dataset to the 5,000 genes with the largest interquartile range (IQR).

The following functions were then performed for cell lines and primary tumors separately: for each cancer type, we fitted a linear model using Linear Models for Microarray Data (LIMMA)<sup>27</sup> and calculated the average fold-change for each gene between that cancer type and a sampling of all other cancer types. This was done with  $n$  arrays by tissue-type (for cell lines,  $n \geq 15$  and for primary tumors,  $n \geq 20$ ) to ensure homogeneous tissue-type representation in the reference set. Finally, we calculated the pairwise Pearson's correlation coefficient between the fold-change values obtained for tumors and cell lines. The final correlation matrix represents the average of 10 iterations of the above procedure, each with different samplings.

#### *Mutation-rate comparison*

Primary tumor mutation data were downloaded from the COSMIC database v56 (<http://www.sanger.ac.uk/genetics/CGP/cosmic/> using the file name CosmicCompleteExport\_v56\_151111.tsv). Cell lines and primary samples were annotated as belonging to tumor types displayed in **Fig. 1d** of the main text. 17 tumor types with more than 20 cell lines in the CCLE and 20 primary samples in COSMIC were kept for the analysis. The COSMIC dataset was filtered 1/ to consider only primary tumor data and exclude cell lines, 2/ for coding mutations in genes common with the CCLE hybrid capture set of 1,651 genes, 3/ for genes that had more than 90% of target bases covered in more than 75% of the CCLE cell lines in the sequencing data described above and 4/ for genes mutated in more than 4% of the primary samples in at least one tumor type and with a synonymous to non-synonymous ratio higher than 9:1. In the end, 62 genes were retained for the analysis. For each tumor type and each tested gene, we calculated the percentage of primary samples or cell lines with reported coding mutations. Finally, we determined the pairwise Pearson's correlations between the matrices of mutation frequencies across tumor types, for primary samples and cell lines, in all common genes.

For some lineages (e.g., urinary tract, liver, pancreas, and thyroid cancer) the correlations were weaker in one or more comparisons (**Fig. 1b-d and Supplementary Fig. 5**). For example, urinary tract cancer cell lines and tumors matched well in

expression space, but they differed in oncogene mutation frequencies. This discordance was driven primarily by a paucity of *FGFR3*-mutant CCLE lines (**Supplementary Fig. 5**), and likely reflects the fact that most urinary tract cancer cell lines derive from high-grade, invasive carcinomas as opposed to low-grade tumors in which *FGFR3* mutations are more prevalent<sup>28</sup>. Liver cancer cell lines and tumors differed across all three comparisons, possibly because many hepatocellular carcinoma lines were derived from patients exposed to hepatitis B virus<sup>29</sup>. In contrast, the liver tumors to which they were compared were mostly associated with hepatitis C virus<sup>30</sup>. Surprisingly, none of the glioma cell lines contained *IDH1* mutations despite its high mutation frequency in primary tumors. Nonetheless, glioma cell lines generally correlated positively with their primary tumor counterparts, despite published reports to the contrary<sup>31</sup>. This pattern may reflect increased representation of this cancer type in the CCLE as compared to previous studies. As expected, the mutation frequencies of lineages where *TP53* mutations are predominant (e.g. esophagus, liver, head & neck) correlated most strongly when *TP53* was included ( $r=0.95, 0.64, 0.65$ , respectively, **Supplementary Fig. 5**).

## 7. Pharmacological characterization

Cells lines (504 total, constituting 480 unique lines) were chosen for profiling based on ease of in vitro cultivation under assay conditions. Suspension cell lines were generally grouped together to facilitate process flow.

All cell lines were cultured in RPMI or DMEM with 10% fetal bovine serum (FBS; Invitrogen). Cells lines were cultured in T-175 or 3 layer T-175 “triple” flasks using standard tissue culture techniques performed robotically (CompacT, The Automation Partnership). Cell lines were incubated at 37 °C and 5% CO<sub>2</sub>. Prior to subculturing, adherent cells lines were dislodged using TrypLE (Invitrogen). From frozen stocks, cells were expanded through at least 1 passage (1:3 dilution) and usually 2 to 3 passages before being added to 1,536-well assay micro-titer plates. Cell count and viability was measured using Trypan dye exclusion with a ViCell counter (Beckman-Coulter). All cell lines were tested for and shown to be free of mycoplasma using a PCR-based detection methodology (<http://www.radil.missouri.edu>).

Our initial set of compounds (termed NP24) included both targeted therapeutics and cytotoxic drugs (**Supplementary Table 6**). Compounds were dissolved in 90% DMSO/10% water at 2 mM and stored at -20 °C until use. Prior to screening, the stock solutions were arrayed in microtiter plates and serially diluted 3.16 fold, yielding a concentration range of 2 mM to 636 nM. Purity and integrity of all compounds and solutions was checked using standard liquid chromatography-mass spectrometry, verifying UV adsorption and mass of the major UV peaks.

All assays were automated and performed with an ultra-high throughput screening system built by the Genomics Institute of the Novartis Research Foundation (<http://www.gnfsystems.com>)<sup>32</sup>. Cell lines were dispensed into 1,536-well plates (optimized for tissue culture) with a final volume of 5 µL and a concentration of 250 cells per well. 12 to 24 hours after plating, 20 nL of each compound dilution series were transferred to the 1,536-well plates (containing the tumor cells) using slotted pins (V&P Scientific <http://www.vp-scientific.com/index.html>). This yielded final drug concentration ranges of 8 µM to 2.5 nM (8 point dose response assays) by 3.16-fold dilutions, and a final DMSO concentration of just under 0.4%. The cell-compound mixtures were incubated for 72 to 84 hours; afterwards, cell numbers were determined by measuring the amount of ATP per well using Cell Titer Glo (Promega). Luminescence/well was measured using a ViewLux plate reader (Perkin Elmer). Within a cell line plating day, compounds were tested in duplicate; occasionally, lines were assayed multiple times (weeks to months apart), yielding additional replicate values. On all plates, wells containing vehicle only or the positive control compound MG132 (a proteasome inhibitor toxic to most cell lines at 1 µM) were also included. Raw values were normalized on a plate-by-plate basis such that 0% was equivalent to the median of vehicle wells and -100% equivalent to the median of the MG132 positive control. The normalized data was further corrected using a surface pattern model to remove edge and region effects.

All dose-response data was reduced to a fitted model using a decision tree methodology based on the NIH/NCGC assay guidelines ([http://assay.nih.gov/assay/index.php/Table\\_of\\_Contents](http://assay.nih.gov/assay/index.php/Table_of_Contents)). Models were generated for the duplicate data points generated for each cell line run day. In brief, dose-response data was fitted to one of three models depending on the statistical quality of the fits measured using a Chi-squared test. One approach was the 4 parameter sigmoid model shown below:

$$y = A_{\text{inf}} + \left( \frac{A_0 - A_{\text{inf}}}{1 + \left( \frac{x}{EC_{50}} \right)^Hill} \right)$$

Alternatively, a constant model  $y = A_{\text{inf}}$  was employed; or a non-parametric spline interpolation of the data points was performed (note that this last model represents less than 5% of models). In these models,  $A_0$  and  $A_{\text{inf}}$  are the top and bottom asymptotes of the response;  $EC_{50}$  is the inflection point of the curve; and  $Hill$  is the Hill slope, which describes the steepness of the curve. Other key parameters derived from the models include the  $IC_{50}$ , the concentration where the fitted curve crosses -50%; and  $A_{\text{max}}$ , which is the maximal activity value reached within a model. For the spline interpolation model,

$IC_{50}$  and  $EC_{50}$  parameters were both set to the concentration where the fitted model first crosses -50%. Additionally, we calculated two forms of the *Activity area* for each curve, defined as the area between the response curve and a fixed reference  $A_{ref} = 0$  or a variable reference  $A_{ref} = \max(0, A_{low})$  where  $A_{low}$  is the activity at the lowest concentration, up to the maximum tested concentration. In practice, the *Activity area* was calculated as the sum of differences between the measured  $A_i$  at concentration  $i$  and the reference level. Thus, using the fixed reference, *Activity area* = 0 corresponds to an inactive compound, and 8 corresponds to a compound which had  $A = -100\%$  at all eight concentrations points. The variable reference form was introduced to adjust for curves with large positive activities close to zero concentration, which are usually artifacts of imperfectly corrected variations on the assay plate. For this measure, the median of all replicate activity values was used regardless of cell line run day. To prevent confusion, the *Activity Area* was calculated using  $A_{ref} = 0$  unless otherwise noted.

For inactive compounds it is formally impossible to derive an  $IC_{50}$ ; however, the analytical algorithms require a value for all cell lines examined. In this instance, we simply used the maximum tested concentration as the default value—which serves primarily as a placeholder to allow algorithms to work on all samples. It should be noted that another sensitivity value that we have used, the activity area, does not suffer from this limitation, as it is possible to derive a value for all dose-response curves.

## 8. Prediction of drug response

Two approaches were used: a discrete or “categorical” classifier based on the naive Bayes algorithm, and a regression analysis based on the elastic net algorithm. Both methods produce a set of genomic predictors of response. Inputs, methodological steps and outputs were made consistent between the two approaches for ease of comparison. Several parameters from the dose-response curves were used, including log-transformed  $IC_{50}$ ,  $A_{max}$ , or *Activity area*. Also, different subsets of the feature data were used (all features or genomic features only, excluding gene expression data), and the models were run both within specific lineages and across all cancer types.

For predicting response to each compound, we used the following two inputs in each of the approaches above:

(1) a vector,  $Y \in P^{N,1}$ , where  $N$  is the number of cell lines treated by that compound, and the values represent the responses across the panel of cell lines, computed either as the area over the dose-response curve (“*Activity Area*”),  $A_{max}$  or  $IC_{50}$  using the curve-fitting procedure described above.

(2) a matrix of genomic features  $X \in P^{N,p}$ , where  $N$  is the number of cell lines, and  $p$  is the number of predictive features (e.g. gene expression, gene copy number, gene

mutation values, lineage, pathway activity scores derived from gene expression data (described above), or regions of recurrent copy-number gain or loss derived from GISTIC). Mutation data is represented as a binary value (pre-normalization) for each gene and summarized in different vectors as described below:

**Strongly damaging mutations:** These include nonsense, frame-shift and splice-site mutations that have been observed in less than 5 samples have been combined in a list of “highly damaging” or “loss of function” mutations. Since such variants are not expected to be highly recurrent, this frequency cut-off served as an additional filter to remove germline variations. These mutations were named “Mut LOF” in the feature matrix.

**Non-neutral missense variants:** These are missense substitutions that created amino acid observed at the same position in homologous proteins from two or more warm-blooded vertebrates were considered likely to be neutral, and excluded from further analysis. Multiple amino acid alignments for 46 vertebrates’ proteomes were obtained from UCSC Genome Browser repository. Remaining missense variants were aggregated in a list of “non-neutral missenses”. These mutations were named “Mut nnMS” in the feature matrix. In addition, we also built vectors of mutations containing both the non-neutral missenses and the damaging mutations (“Mut LOF+nnMS”).

**Missense mutations at COSMIC recurrent positions:** Subsets of missense mutations that have defined genomic positions have been selected from COSMIC database v55. Amino acid positions at which mutations were described in three or more non-cell line unique samples were considered to be sites of recurrent mutations. Missense variants observed in our cell lines sequencing data within a 3 amino acid residues distance from recurrent COSMIC sites were aggregated into “Recurrent COSMIC missenses” feature list. These mutations were named “Mut cosmicMS” in the feature matrix.

The data utilized for the analyses in this study included 947 cell lines profiled with SNP arrays, 917 with expression arrays, 860 cell lines had hybrid capture/sequencing data and 479 had been profiled with pharmacologic compounds. A total of 435 cell lines had all data types and were used for sensitivity prediction (**Supplementary table 1**). For up-to-date lists of cell lines and associated data from the CCLE project, please refer to [www.broadinstitute.org/ccle](http://www.broadinstitute.org/ccle).

### 8.1. Sensitivity prediction using regression analysis

We applied an elastic net regression algorithm<sup>33,34</sup> combined with a bootstrapping procedure to derive predictive models that explained the drug sensitivity profiles based

on genetic features of the cell lines. The elastic net algorithm is particularly well suited to inference in this domain because it is designed to work in settings where the number of features is far greater than the number of observations (i.e.  $p \gg N$ ). The algorithm also combines  $L1$  and  $L2$  regularized regression penalty terms in order to strike a balance between obtaining a parsimonious model (through the  $L1$  term), while retaining groups of correlated features (through the  $L2$  term), such as co-expressed genes or copy number of genes situated within the same amplicon.

As input to the algorithm, we used a prediction matrix  $X \in P^{N,p}$ , as described above, where each column of  $X$  is normalized to have zero mean and unit standard deviation. For each compound, we generated a vector,  $Y \in P^{N,1}$ , with either the *Activity Area*,  $A_{max}$  or the log-transformed  $IC_{50}$ . We used the *glmnet* 1.7 software package<sup>35</sup> and R 2.13.1<sup>32</sup> to solve the following optimization problem:

$$\min_{(\beta_0, \beta) \in P^{p+1}} R_\lambda(\beta_0, \beta) = \min_{(\beta_0, \beta) \in P^{p+1}} \left[ \frac{1}{2N} \sum_{i=1}^N \left( y_i - \beta_0 - \sum_{j=1}^p (x_{i,j} \beta_j) \right)^2 + \lambda \left( (1-\alpha) \frac{1}{2} \|\beta\|_{L_2}^2 + \alpha \|\beta\|_{L_1} \right) \right]$$

For computational efficiency, we solved the optimization problem using only features that were correlated with the response vector with  $R > 0.1$  based on Pearson correlation, regardless of  $N$ . In the elastic net equation,  $\alpha$  controls the relative strength of the L1 and L2 penalty terms, and  $\lambda$  controls the overall strength of the regularized regression penalty. The optimal setting for  $\alpha$  and  $\lambda$  is chosen to minimize the root mean squared error using 10 leave-group-out cross-validations with 90% / 10% training/test splits for each  $(\alpha, \lambda)$ , with 10 values of  $\alpha \in [0.2, 1.0]$  and 250 values of  $\lambda = e^\gamma$  with  $\gamma \in [-6, 5]$ .

After parameter optimization, a bootstrapping procedure is used to generate 200 resampled datasets,  $(X^{BS_i}, Y^{BS_i})_{i=1,\dots,200}$ , where  $X^{BS_i} \in P^{N,p}$ ,  $Y^{BS_i} \in P^{N,1}$ , and the samples (i.e. cell lines) constituting each bootstrap dataset are obtained by sampling with replacement from the complete set of samples. The elastic net equation is solved for each bootstrap dataset, using the optimal  $\alpha$  and  $\lambda$  settings, to generate a matrix of regression coefficients  $\beta^{BS} \in P^{p,200}$ , where each column of  $\beta^{BS}$  represents the solution for one bootstrap dataset. For each predictive feature,  $j \in \{1, \dots, p\}$ , the percentage of bootstrap datasets in which it was inferred as significant is calculated as  $r_j = \sum_{k=1}^{200} (1_{P \setminus 0}(\beta_{j,k}^{BS})) / 200$ , where  $1_{P \setminus 0}$  is the indicator function defined as  $1_{P \setminus 0}(x) = \begin{cases} 0 & \text{if } x=0, \\ 1 & \text{otherwise.} \end{cases}$  Using this procedure,  $r_j$  provides a robust measure of the predictive value of feature  $j$ .

In order to evaluate the prediction performance of each model, we performed 10 iterations of a 10-fold cross-validation of the entire procedure described above, excluding the bootstrapping. This allowed calculation of an average predicted response  $Y' \in P^{N,1}$  where each value was the mean of 10 cross-validated predictions. The prediction performance of each model was then estimated using Pearson's correlation coefficient  $r$  and Kendall's  $\tau$ .

## 8.2. Sensitivity prediction using categorical analysis

We also applied a naive Bayes classifier (R package e1071<sup>36</sup>) combined with statistical feature selection, to derive predictive models that explain the drug sensitivity profiles based on cell line genomic features. An advantage of the naive Bayes classifier is that it assumes independence of the input features. Thus, only a small amount of training data is needed to estimate the parameters necessary for classification (*i.e.*, means and variances of the features). This classifier can therefore handle cases where the number of predictive features is significantly larger than the number of samples used for classification.

For each compound, starting from the vector of responses  $Y$  above, we considered the shape of the rank-ordered plot of response values (for  $A_{max}$ , log-transformed  $IC_{50}$  or *Activity Area*) in order to assign cell lines into sensitive, intermediate and refractory classes. Assignments were made before any modeling work was undertaken. (Additional details of an automated sensitivity calling method will be presented in a manuscript under preparation.)

Next, using either the non-parametric Wilcoxon Sum Rank Test (for continuous features such as gene expression) or a Fisher's Exact Test (for discrete features such as gene mutation), we selected features whose profile was significantly different between the “sensitive” and “refractory” populations of cell lines. A feature type-specific correction of  $P$ -values (e.g., for a given gene expression or GISTIC peak) was performed to account for false discovery rate, and a feature was considered statistically significant if its FDR corrected  $P$ -value<sup>37</sup> was less than 0.25. The top 30 such features ordered by  $P$ -value were automatically selected for inclusion into the naive Bayes model, with the fraction of features selected within a given feature type reflecting the proportion of that feature type in the entire feature matrix.

To reduce the number of correlated features included in the naive Bayes model, the remaining significant features within each feature type were clustered using a message-passing algorithm<sup>38</sup>. Features identified as “cluster representatives” by that procedure were included together with the top 30 features above as predictive features into the naive Bayes model. We evaluated model performance using five iterations of ten-fold cross-validation (e.g., 90%/10% training/test splits) and we computed the model

performance according to the area under the ROC curve (AUC), sensitivity, specificity, positive predictive value, and negative predictive value. The set of predictive features considered in the model, their statistical significance, and the effect size (mean fold-change or odds ratio) was also obtained.

## 9. AHR validation experiments

**Cell lines and cell culture conditions:** CHP-212, SK-MEL-2, NCI-H1299 and SK-N-SH cells were obtained from ATCC, IPC-298, TC-71, and MHH-ES-1 were obtained from DSMZ, and ONS-76 from HSRRB. IPC-298, SK-MEL-2, NCI-H1299, ONS-76, and MHH-ES-1 were maintained in RPMI growth media consisting of RPMI 1640 plus L-glutamine (Mediatech) with 10% fetal bovine serum (Gemini Bio-Products) and 1% penicillin/streptomycin (Invitrogen). SK-N-SH was maintained in MEM growth media consisting of MEM plus L-glutamine (Mediatech) with 10% fetal bovine serum (Gemini Bio-Products) and 1% penicillin/streptomycin (Invitrogen). CHP-212 was maintained in a 1:1 mixture of MEM with L-glutamine (Mediatech) and F12 medium with L-Glutamine (GIBCO) with 10% fetal bovine serum (Gemini Bio-Products) and 1% penicillin/streptomycin (Invitrogen). TC-71 was maintained in IMDM growth media consisting of IMDM plus L-glutamine (GIBCO) with 10% fetal bovine serum (Gemini Bio-Products) and 1% penicillin/streptomycin (Invitrogen).

**Lentivirally delivered short hairpin RNA:** The pLKO1-puromycin lentiviral vector carrying shRNAs specific for *AHR*, *SLFN11*, or Luciferase (Luc) sequences were obtained from the Broad Institute RNAi Consortium ([http://www.broadinstitute.org/genome\\_bio/trc/](http://www.broadinstitute.org/genome_bio/trc/)). Three independent shRNAs targeting *AHR*, two targeting *SLFN11*, and one targeting Luciferase were used: shLuc (TRCN0000072243, 5'-CTTCGAAATGTCCGTTGGT-3'), shAHR (hp1) (TRCN0000021254, 5'-CCCACAAACAATAATGTCTT-3'), shAHR (hp2) (TRCN0000021255, 5'- GCTTCTTGATGTTGCATTAA- 3'), shAHR (hp4) (TRCN0000021257, 5'-CCATAATAACTCCTCAGACAT-3'), shSLFN11 (hp2) (TRCN0000155578, 5'- CCGATAAACCTTCACACTCAAA-3'), and shSLFN11 (hp4) (TRCN0000152057, 5'- CAGTCTTGAGAGAGCTTATT). To perform lentiviral infections, cells were plated at 50-60% confluence and incubated overnight. The following day, the medium was replaced with virus diluted in fresh medium with 8 µg/mL polybrene, and cells were incubated at 37 °C for 24 hours. Subsequently, the medium was removed and replaced with fresh medium containing puromycin (2 µg/mL) for selection. Cells were grown in the presence of puromycin for 4 days before they were seeded for growth curve experiments and protein/mRNA analysis to determine knockdown.

**Immunoblot analysis:** Cells were harvested, washed with PBS, and lysed on ice with 1% NP-40 buffer [150 mM NaCl, 50 mM Tris pH 7.4, 2 mM EDTA pH 8, 25 mM NaF and 1% NP-40] containing protease inhibitors (Roche) and Phosphatase Inhibitor Cocktails I and II (CalBioChem). After 30 min incubation, lysates were frozen at -20 °C, thawed on ice and centrifuged 10 minutes at 4 °C at full speed. Protein concentrations were measured by the BCA method (Pierce). Equal amounts of total protein were subjected to SDS gel electrophoresis and transferred to PVDF membrane. The membranes were blocked for 1 hour at room temperature with Blocking Buffer (Licor), and incubated overnight at 4 °C with the primary antibody in buffer supplemented with 0.1% Tween-20. Subsequently, the membranes were washed three times with Tween-TBS buffer and incubated with secondary antibody diluted in Blocking Buffer with 0.1% Tween-20 in dark for 2 hours at room temperature. The membranes were washed two times with Tween-TBS for 10 minutes each and a final wash with PBS for 10 minutes. Results were obtained using the Odyssey Infrared Imager (Licor). Rabbit polyclonal anti-SLFN11 antibody was obtained from Sigma-Aldrich (HPA023030) and used at a 1:500 dilution; mouse monoclonal anti-AHR antibody was obtained from Abcam (ab2770) and used at a 1:1000 dilution; mouse monoclonal anti-vinculin was obtained from Sigma-Aldrich (V9131) and used at a 1:20000 dilution; goat polyclonal anti-actin was obtained from Sigma-Aldrich and used at a 1:2000 dilution; goat anti-mouse secondary IRDye 800 CW antibody was obtained from Licor and used at 1:15000 dilution; donkey anti-goat secondary IRDye 800 CW antibody was obtained from Licor and used at 1:20000; goat anti-rabbit secondary IRDye 800 CW was obtained from Licor and used at 1:15000 dilution.

**Analysis of mRNA expression by quantitative RT-PCR (qRT-PCR):** Cellular RNA was extracted using the RNeasy Mini Kit (Qiagen) following manufacturer's protocol. For qRT-PCR, RNA was reverse-transcribed using SuperScript III First-Strand Synthesis SuperMix for qRT-PCR kit (Invitrogen). qRT-PCR was performed in 384-well format using LightCycler 480 SYBR Green I Master (Roche) and the 7900HT Fast Real-Time PCR System (Applied Biosystems). Data presented are the average of two individual experiments; within each experiment, technical triplicates were performed. Fold changes were calculated relative to control by the  $2^{-\Delta\Delta Ct}$  method. The following primer sequences were used: AHR forward 5'-CAAATCCTCCAAGCGGCATA-3'; AHR reverse 5'-CGCTGAGCCTAACGAACTGAAAG-3'; CYP1A1 forward 5'-TCGGCCACGGAGTTCTTC-3'; CYP1A1 reverse 5'-TCTTGAGGCCCTGATTACCCA-3'; GAPDH forward 5'-AAGGTGAAGGTGGAGTCAAC-3'; GAPDH reverse 5'-GGGGTCATTGATGGCAACAATA-3'.

**Growth curves:** CHP-212, IPC-298, SK-MEL-2, ONS-76, and SK-N-SH cell lines were seeded into 96-well plates at densities of 5,000, 4,000, 5,000, 2,000, and 5,000

cells per well, respectively, in the appropriate culture medium and each with 6 replicates. Proliferation rates were measured either on days 5, 6, 7, and 8 post-infection or on days 4, 5, 6, 7 using WST-1 (Roche). WST-1 reagent was diluted in medium to a final concentration of 10%, incubated with the cells at 37 °C for 2 hours, and the plates were read at 440 nm using a SpectraMax 190 microplate reader. Absorbance on each day was displayed after background subtraction.

**Pharmacologic growth inhibition curves:** CHP-212, IPC-298, SK-MEL-2, NCI-H1299, ONS-76, SK-N-SH, TC-71, and MHH-ES-1 cell lines were seeded into 96-well plates at densities of 10,000, 1,000, 5,000, 5,000, 1,000, 4,000, 3,000, and 8,000 cells per well, respectively, in the appropriate culture medium. Twenty-four hours after seeding, serial dilutions of the relevant compound were prepared in DMSO and added to cells, yielding final drug concentrations ranging from 100 μM to  $1 \times 10^{-6}$  μM for PD-0325901, irinotecan, and topotecan; 150 μM to  $1 \times 10^{-2}$  μM for PD-98059, with the final volume of DMSO not exceeding 1%. Cells were incubated for 96 hours following addition of drug. Cell viability was measured using the WST-1 viability assay (Roche). Viability was calculated as a percentage of control (untreated cells) after background subtraction. Six replicates were performed for each cell line and drug combination. Data from growth-inhibition assays were modeled using a nonlinear regression curve fit with a sigmoid dose-response. These curves were displayed using GraphPad Prism 5 (GraphPad).

## 10. Data sharing/release

All raw and processed data are available at the CCLE website: [www.broadinstitute.org/ccle](http://www.broadinstitute.org/ccle). In addition, the website offers direct links to data visualization tools such as IGV<sup>39</sup>, as well as genepattern-based<sup>40</sup> analysis tools for expression and copy-number class comparison analyses.

**Supplementary References:**

- 1 Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061-1068, (2008).
- 2 Korn, J. M. et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* 40, 1253-1260, (2008).
- 3 Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 12, R41, (2011).
- 4 Dai, M. et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 33, e175, (2005).
- 5 Irizarry, R. A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249-264, (2003).
- 6 Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185-193, (2003).
- 7 Brettschneider, J., Collin, F. o., Bolstad, B. M. & Speed, T. P. Quality Assessment for Short Oligonucleotide Microarray Data. *Technometrics* 50, 241-264, (2008).
- 8 MacConaill, L. E. et al. Profiling critical cancer gene mutations in clinical tumor samples. *PLoS One* 4, e7887, (2009).
- 9 Ding, L. et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 455, 1069-1075, (2008).
- 10 Forbes, S. A. et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* 39, D945-950, (2011).
- 11 Parsons, D. W. et al. An integrated genomic analysis of human glioblastoma multiforme. *Science* 321, 1807-1812, (2008).
- 12 Sjöblom, T. et al. The consensus coding sequences of human breast and colorectal cancers. *Science* 314, 268-274, (2006).
- 13 Wood, L. D. et al. The genomic landscapes of human breast and colorectal cancers. *Science* 318, 1108-1113, (2007).
- 14 Beroukhim, R. et al. The landscape of somatic copy-number alteration across human cancers. *Nature* 463, 899-905, (2010).
- 15 Futreal, P. A. et al. A census of human cancer genes. *Nat Rev Cancer* 4, 177-183, (2004).
- 16 Kan, Z. et al. Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* 466, 869-873, (2010).
- 17 Gnirke, A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 27, 182-189, (2009).
- 18 Fisher, S. et al. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol* 12, R1, (2011).
- 19 Chapman, M. A. et al. Initial genome sequencing and analysis of multiple myeloma. *Nature* 471, 467-472, (2011).
- 20 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589-595, (2010).
- 21 Depristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43, 491-498, (2011).

- 22 Stransky, N. et al. The mutational landscape of head and neck squamous cell carcinoma. *Science* 333, 1157-1160, (2011).
- 23 Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102, 15545-15550, (2005).
- 24 Beroukhim, R. et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A* 104, 20007-20012, (2007).
- 25 Haferlach, T. et al. Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the International Microarray Innovations in Leukemia Study Group. *J Clin Oncol* 28, 2529-2537, (2010).
- 26 Kohlmann, A. et al. An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the Microarray Innovations in LEukemia study prephase. *Br J Haematol* 142, 802-807, (2008).
- 27 Smyth, G. K. in *Bioinformatics and computational biology solutions using R and Bioconductor* Ch. Limma: linear models for microarray data, xix, 473 p. (Springer Science+Business Media, 2005).
- 28 Cappellen, D. et al. Frequent activating mutations of FGFR3 in human bladder and cervix carcinomas. *Nat Genet* 23, 18-20, (1999).
- 29 Laurent-Puig, P. et al. Genetic alterations associated with hepatocellular carcinomas define distinct pathways of hepatocarcinogenesis. *Gastroenterology* 120, 1763-1773, (2001).
- 30 Chiang, D. Y. et al. Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. *Cancer Res* 68, 6779-6788, (2008).
- 31 Li, A. et al. Genomic changes and gene expression profiles reveal that established glioma cell lines are poorly representative of primary human gliomas. *Mol Cancer Res* 6, 21-30, (2008).
- 32 Melnick, J. S. et al. An efficient rapid system for profiling the cellular activities of molecular libraries. *Proc Natl Acad Sci U S A* 103, 3153-3158, (2006).
- 33 Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J Roy Stat Soc B* 67, 301-320, (2005).
- 34 Zou, H. & Zhang, H. H. On the Adaptive Elastic-Net with a Diverging Number of Parameters. *Ann Stat* 37, 1733-1751, (2009).
- 35 Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432-441, (2008).
- 36 R Development Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2010).
- 37 Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. *Stat Med* 9, 811-818, (1990).
- 38 Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* 315, 972-976, (2007).
- 39 Robinson, J. T. et al. Integrative genomics viewer. *Nat Biotechnol* 29, 24-26, (2011).
- 40 Reich, M. et al. GenePattern 2.0. *Nat Genet* 38, 500-501, (2006).

- 41 Shankavaram, U. T. et al. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics* 10, 277, (2009).