doi:10.1038/nature11003

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). 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/) 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 where 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 10^6 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 ^{o}C or overnight at room temperature and stored at 4 ^{o}C. For long-term storage, all samples were stored at -20 ^{o}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 10^6 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 ^{o}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 ^{o}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 ^{o}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 ^{o}C and stored at -80 ^{o}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[1]. 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[1] 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[2] genotyping calls, and appropriate segmentation of the copy-number profiles. Finally, the Genomic Identification of Significant Targets in Cancer (GISTIC) algorithm[3] 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[4]; and background correction was accomplished using RMA (Robust Multichip Average)[5] and quantile normalization[6]. Quality assessment was performed to identify low performing microarrays, using the R package affyPLM[7]. 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[8]. 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[8] (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[1,9-13], (B) occurrence in 2 – 3 instances from the aforementioned studies and also present in a significantly amplified or deleted focal peak in primary tumors[14] or cell lines (this study), (C) membership in the Cancer Gene Census[15], or (D) significant mutation frequency across 441 tumors[16]; (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[17] 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[17,18]. Barcoded exon capture libraries were then pooled and sequenced on Illumina instruments (76 bp paired-end reads)[17]. 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[19], but with the following modifications in the "Alignment to the genome" step: Alignment was performed using BWA[20] (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[21]. 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[19,22]. 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[2] 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)[23] 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[14] 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[24]) 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 exp*O* (http://www.intgen.org/expo/) dataset (GEO accession GSE2109, consisting of 2,158 arrays from solid tumors), MILE[25,26] (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)[27] 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[28]. 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[29]. In contrast, the liver tumors to which they were compared were mostly associated with hepatitis C virus[30]. 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[31]. 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 ºC and 5% CO_2. Prior to sub-culturing, 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 ºC 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)[32]. 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). 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:

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

Alternatively, a constant model $y = A_{\inf}$ 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_{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_{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.

8.1. Sensitivity prediction using regression analysis

We applied an elastic net regression algorithm[33,34] 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. $p >> N$). 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[35] and R 2.13.1[32]to solve the following optimization problem:

$$\min_{(\beta_0,\beta)\in P^{p+1}} R_\lambda(\beta_0,\beta) = \min_{(\beta_0,\beta)\in P^{p+1}} \left[\frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=i}^{p} \left(x_{i,j}\beta_j \right) \right)^2 + \lambda \left((1-\alpha)\frac{1}{2}\|\beta\|_{l_2}^2 + \alpha\|\beta\|_{l_1} \right) \right]$$

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, α controls the relative strength of the L1 and L2 penalty terms, and λ controls the overall strength of the regularized regression penalty. The optimal setting for α and λ is chosen to minimize the root mean squared error using 10 leave-group-out cross-validations with 90% / 10% training/test splits for each (α,λ), 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, $\left(X^{BS_i}, Y^{BS_i}\right)_{i=1,\ldots,200}$, 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 α and λ settings, to generate a matrix of regression coefficients $\beta^{BS} \in P^{p,200}$, where each column of β^{BS} represents the solution for one bootstrap dataset. For each predictive feature, $j \in \{1,\ldots,p\}$, the percentage of bootstrap datasets in which it was inferred as significant is calculated as $r_j = \sum_{k=1}^{200} \left(1_{P\setminus 0}\left(\beta_{j,k}^{BS} \right) \right) / 200$, 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 τ.

8.2. Sensitivity prediction using categorical analysis

We also applied a naive Bayes classifier (R package e1071[36]) 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[37] 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[38]. 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/). Three independent shRNAs targeting AHR, two targeting *SLFN11*, and one targeting Luciferase were used: shLuc (TRCN0000072243, 5'-CTTCGAAATGTCCGTTCGGTT-3'), shAHR (hp1) (TRCN0000021254, 5'-CCCACAACAATATAATGTCTT-3'), shAHR (hp2) (TRCN0000021255, 5'- GCTTCTTTGATGTTGCATTAA- 3'), shAHR (hp4) (TRCN0000021257, 5'-CCATAATAACTCCTCAGACAT-3'), shSLFN11 (hp2) (TRCN0000155578, 5'- CCGATAACCTTCACACTCAAA-3'), and shSLFN11 (hp4) (TRCN0000152057, 5'- CAGTCTTTGAGAGAGCTTATT). 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 ºC, thawed on ice and centrifuged 10 minutes at 4 ºC 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 ºC 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'-CAAATCCTTCCAAGCGGCATA-3'; AHR reverse 5'-CGCTGAGCCTAAGAACTGAAAG-3'; CYP1A1 forward 5'-TCGGCCACGGAGTTTCTTC-3'; CYP1A1 reverse 5'-TCTTGAGGCCCTGATTACCCA-3'; GAPDH forward 5'-AAGGTGAAGGTCGGAGTCAAC-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 x 10^{-6} μM for PD-0325901, irinotecan, and topotecan; 150 μM to 1 x 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. In addition, the website offers direct links to data visualization tools such as IGV[39], as well as genepattern-based[40] 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 Sjoblom, 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).