Supplementary Information

Supplementary Methods

Supplementary Methods

1. Experimental procedures

1.1 Cell culture

CCLE cell lines were grown according to vendor recommendations as previously described[1] (Supplementary Table 1).

1.2 Whole Genome and Whole Exome Sequencing

Overview

Whole genome sequencing (WGS) for 329 cell lines and whole exome sequencing (WES) for 326 cell lines were performed at the Broad Institute Genomics Platform. Libraries were constructed and sequenced on either an Illumina HiSeq 2000 or Illumina GAIIX, with the use of 101-bp paired-end reads for whole-genome sequencing and 76-bp paired-end reads for whole-exome sequencing. Output from Illumina software was processed by the Picard data-processing pipeline to yield BAM files containing well-calibrated, aligned reads. All sample information tracking was performed by automated LIMS messaging.

Library construction

Starting with 3µg of genomic DNA, library construction in a subset of samples was performed as described by Fisher et al.[36]. Other samples, however, were prepared using minor modifications of the Fisher et al. protocol. Specifically, initial genomic DNA input into shearing was reduced from 3µg to 100ng in 50µL of solution, and for adapter ligation, Illumina paired end adapters were replaced with palindromic forked adapters with unique 8 base index sequences embedded within the adapter.

In-solution hybrid selection (for targeted sequencing libraries)

In-solution hybrid selection was performed as described by Fisher et al.[36].

Size selection (for whole genome shotgun libraries)

For a subset of samples, size selection was performed using gel electrophoresis with a target insert size of either 340bp or 370bp +/- 10%. Multiple gel cuts were taken for libraries that required high sequencing coverage. For another subset of samples, size selection was performed using Sage's Pippin Prep.

Preparation of libraries for cluster amplification and sequencing

After the above sample preparation, libraries were quantified using quantitative PCR (KAPA Biosystems) with probes specific to the ends of the adapters. This assay was automated using the Agilent Bravo liquid handling platform. Based on qPCR quantification, libraries were normalized to 2nM and then denatured using 0.1 N NaOH using Perkin-Elmer's MultiProbe liquid handling platform. The subset of the samples prepared using forked, indexed adapters was quantified using qPCR, normalized to 2nM using Perkin-Elmer's Mini-Janus liquid handling platform, and pooled by equal volume using an Agilent

Bravo Automated Liquid Handling Platform. Pools were then denatured using 0.1 N NaOH. Denatured samples were diluted into strip tubes using a Perkin-Elmer MultiProbe Robotic Liquid Handling System.

Cluster amplification and sequencing

Cluster amplification of denatured templates was performed according to manufacturer's protocol (Illumina), using either Genome Analyzer v3, Genome Analyzer v4, HiSeq 2000 v2, or HiSeq v3 cluster chemistry and flowcells. For a subset of samples, SYBR Green dye was added to all flowcell lanes following cluster amplification, and a portion of each lane was visualized using a light microscope in order to confirm target cluster density. Flowcells were sequenced either on a Genome Analyzer IIX using v3 or v4 Sequencing-by-Synthesis Kits and analyzed using RTA v1.7.48; or on an Illumina HiSeq 2000 using HiSeq 2000 v2 or v3 Sequencing-by-Synthesis Kits and analyzed using RTA v1.10.15 or RTA v.1.12.4.2. 101-bp paired-end reads were used for whole-genome sequencing, and 76-bp paired-end reads were used for whole-exome sequencing. For pooled libraries prepared using forked, indexed adapters, the Illumina Multiplexing Sequencing Primer Kit was used and a third 8-bp sequencing read was performed to read molecular indices.

1.3 RainDance targeted sequencing

For 950 cell lines, genomic loci with inadequate coverage by targeted hybrid capture sequencing were enriched using RainDance Technologies (RDT) platform to generate barcoded libraries of amplicons suitable for Illumina sequencing followed by massively parallel sequencing at the Broad Institute (Supplementary Table 2).

Per the RDT protocol, samples containing a minimum of 5 μg of high quality DNA were provided to RDT. Adaptor primers were designed to be used in the secondary amplification that contained Broad's required sample indexing and adaptor sequences. RDT provided enriched DNA to Broad containing a minimum of 100ng of amplified and Qiagen Min-elute purified DNA that had undergone the RDT enrichment process using the Primer Library and that had gone through a secondary PCR of 10 cycles with Adaptor Primers.

1.4 RNAseq profiling

RNA sequencing and analysis were performed for 1,019 cell lines as previously described[5]. In summary, non-strand specific RNA sequencing was performed using large-scale, automated method of the Illumina TruSeq RNA Sample Preparation protocol. Oligo dT beads were used to select polyadenylated mRNA. The selected RNA was then heat fragmented and randomly primed before cDNA synthesis. To maximize power to detect fusions, the insert size of fragments was set to 400nt. The resultant cDNA then went through Illumina library preparation (end-repair, base 'A' addition, adaptor ligation, and enrichment) using Broad-designed indexed adapters for multiplexing. Sequencing was performed on the Illumina HiSeq 2000 or HiSeq 2500 instruments with sequence coverage of no less than 100 million paired 101 nucleotides-long reads per sample.

1.5 microRNA profiling

Expression profiling of a panel of 734 microRNAs across 954 cell lines was performed using the Nanostring platform. All sample preparation and processing were performed according to the manufacturer's protocol. Hybridized probes were purified and counted on the nCounter Prep Station and Digital Analyzer (NanoString), following the manufacturer's instructions.

1.6 Global chromatin profiling

Histone modification profiling was performed as described previously for a total of 897 cell lines[15,16]. Briefly, the mass spectrometry-based method profiles relative changes in the levels of almost all common post-translational modifications on histone H3.1 and/or H3.2. This includes methylation and acetylation modifications on H3K4, H3K9, H3K14, H3K18, H3K23, H3K27, H3K36, H3K56, and H3K79. Phosphorylation is also profiled on H3S10, and ubiquityl marks were profiled on H3K18 and H3K23. Importantly, the marks are frequently profiled as combinations (i.e., H3K27me2K36me2) which is generally not possible with antibody-based methods. Some marks are omitted from visualizations for clarity. The changes observed are relative to other cell lines in the CCLE, with appropriate batch normalization. Common internal standards are used across all experiments.

1.7 Reverse Phase Protein Array (RPPA)

RPPA procedure

Cellular proteins were denatured by 1% SDS (with beta-mercaptoethanol) and diluted in five 2-fold serial dilutions in dilution lysis buffer. Serial diluted lysates were arrayed on nitrocellulose-coated slides (from Grace Bio-Labs) using an Aushon 2470 Arrayer (from Aushon BioSystems). A total of 5,808 array spots were arranged on each slide including the spots corresponding to serial diluted: 1) “Standard Lysates”; and 2) positive and negative controls prepared from mixed cell lysates or dilution buffer.

Each slide was probed with a primary antibody and a biotin-conjugated secondary antibody. Only antibodies with a Pearson correlation coefficient between RPPA and western blotting of greater than 0.7 were used. Antibodies with a single or dominant band on western blotting were further assessed by direct comparison to RPPA using cell lines with differential protein expression or modulated with ligands/inhibitors or siRNA for phospho- or structural proteins, respectively.

The signal obtained was amplified using a Dako Cytomation–Catalyzed system (Dako) and visualized by DAB colorimetric reaction. The slides were scanned, analyzed, and quantified using custom software to generate spot intensity.

Each dilution curve was fitted with a logistic model (“Supercurve Fitting” developed by the Department of Bioinformatics and Computational Biology in MD Anderson Cancer Center, “http://bioinformatics.mdanderson.org/OOMPA”). This fits a single curve using all the samples (i.e., dilution series) on a slide with the signal intensity as the response variable and the dilution step as the independent variable. The fitted curve is plotted with both the observed and fitted signal intensities on the y-axis and the $\log_2$- concentration of proteins on the x-axis for diagnostic purposes. The protein concentrations of each set of slides were then normalized for protein loading. Correction factor was calculated by first median-centering across samples of all antibody experiments and then median-centering across antibodies for each sample.

RPPA technical and biological controls

RPPA profiling was performed in two batches, with 422 samples in batch one and 544 samples in batch two. To evaluate the data reproducibility between the two batches, frozen lysates from 30 samples generated for batch one were profiled in batch two as technical controls. To evaluate the reproducibility between biological replicates, 6 cell lines were grown two times independently and profiled in batch two as biological replicates (Supplementary Table 14). Five of these cell lines were also grown and profiled in batch one independently.

In vitro validation of ponatinib/pSHP2 association

A total of 21 cell lines were used to validate the observed correlation between pSHP2 level and sensitivity to ponatinib. This included two BCR-ABL fusion containing CML cell lines (MEG01 and LAMA84) that were expected to be sensitive to ponatinib and 19 AML cell lines (CMK, HEL9217, THP1, NOMO1, HL60, HEL, KO52, P31FUJ, OCIAML2, SIGM5, GDM1, NKM1, KG1, MonoMAC6, KASUMI1, MonoMAC1, CTV1, MV411, and EOL1). These included all AML cell lines in the overlap between CCLE RPPA and GDSC drug sensitivity datasets and five additional cell lines to test the hypothesis. Based on their sensitivity to ponatinib, CTV1 and NKM1 were the two non-CCLE cell lines that were selected. EOL1, HEL9217 and MonoMAC1 were non-GDSC cell lines, selected based on their high pSHP2 level (EOL1, HEL9217) and FLT3 mutation and overexpression (MonoMAC1). CCLE cell lines were obtained through the CCLE project, NKM1 was obtained through the Japanese Collection of Bioresources, and CTV1 was obtained from Leibniz-Institut DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen). Cell lines were grown according to respective vendors' recommendations.

Whole cell extracts were prepared using a 1% NP40 lysis buffer and blotted with total and phosphorylated SHP2 antibodies (Cell Signaling Technology) as previously described [37]. p-SHP2 levels were quantified relative to total SHP2 using a Licor Odyssey imager.

Cellular sensitivity was determined by seeding cells in growth media in 96 well plates and treating with indicated small molecules for 96 hours in 6-8 replicates. Cell viabilities were quantified using CellTiterGlo and values were normalized to DMSO-treated cells as previously described[37].

1.8 *Reduced* representation bisulfite sequencing (RRBS)

For 843 cell lines, the RRBS method was used as previously described by (Boyle et al., Genome Biology 2012)[38].

1.9 *TERT* promoter mutation sequencing

Targeted sequencing of the TERT promoter was performed as described previously for 190 cell lines[39,40]. Paired-end sequencing with a 150 bp read length was performed on PCR amplicons of length 273 bp to high depth on an Illumina MiSeq instrument. We then combined this with variant calls for the *TERT* promoter from WGS dataset of 329 cell lines previously described[8]. Alternate allele fractions > 10% were called as mutant for pre-specified sites: chr5:1295161, chr5:1295228-1295229, chr 5:1295228, chr5:1295242-1295243, and chr5:129525 using MuTect v1.1.6[41] (Supplementary Table 5).

1.10 Quantitative PCR detection of MDM4 isoforms

Cell lines were processed using Trizol RNA extraction (Life Technologies)[1]. cDNA was reverse transcribed using the iScript cDNA synthesis kit (BioRad) with no reverse transcriptase samples serving as a negative control. Gene expression was quantified using the Power Sybr Green Master Mix (Applied Biosystems) and normalized to GAPDH. Quantitation of the MDM4-FL/MDM4-S ratio was determined by calculating the fold change of MDM4-FL and MDM4-S for each technical replicate relative to the TOV21G universal reference standard cell line using the ΔΔCt method. For each cell line, the mean and standard deviation of the log MDM4-FL/MDM4-S ratio was calculated across technical replicates (See Supplementary Table 11 for primer sequences).

1.11 *In* vivo primagraft experiment

14 AML primagrafts from the Public Repository of Xenografts (PRoXe.org) were first tested by RPPA for pSHP2 levels. Two of the highest pSHP2-expressing primagrafts (CBAM-87679, NVAM-

61786) and one low pSHP2-expressing primagraft (DFAM-68555) were selected for xenotransplantation to test for sensitivity to ponatinib treatment. Each primagraft was xenotransplanted into twenty female 7-week-old NOD scid gamma (NSG) mice from Jackson Laboratory (Bar Harbor, ME). Mice were intravenously injected with 0.15-1.0 x 10^6 cells via the lateral tail vein. Engraftment of human leukemia cells in mice was followed using FACS analysis of human CD45+ CD33+ or CD34+ cells in the peripheral mouse blood. Once leukemia was established with an average 0.4% human cells in the peripheral blood from the sentinel bleed mice, animals were randomized into 2 treatment groups of 10 mice each: ponatinib (40mg/kg oral once daily) and vehicle (25 mM citrate buffer, pH 2.75). For primagraft CBAM-87679, ponatinib dosing started two weeks after injection given a rapid progression of disease. Mice were treated with ponatinib for 3 weeks. Mice were euthanized once morbidity and/or stage 3 hind limb paralysis due to disease burden was observed. All animal studies were approved by the Dana-Farber Cancer Institute's Animal Care and Use Committee.

To assess the pharmacodynamic efficacy of treatments, three mice from each group were analyzed after 3 day of treatment. 2–4 h after the day 3 drug or vehicle dose, mice were euthanized and tissues collected. Spleen (1/4 of total spleen), one femur, and liver were fixed in 10% neutral-buffered formalin for immunohistochemistry and other studies. The remaining spleen was crushed, and bone marrow cells flushed from the 3 remaining leg bones were viably cryopreserved in 10% DMSO / 90% FBS.

The remaining mice (7 per group) were treated for a total of 21 days. Survival analysis based on these 7 mice per group was performed using the log-rank (Mantle-Cox) test (GraphPad Prism 7).

2. Computational analysis

2.1 *Variant* calling and filtering germline variants for WES, WGS, hybrid capture, and RainDance

A variant calling pipeline was designed to process all sequencing data generated in the CCLE. Mutation analysis for single nucleotide variants (SNVs) was performed using MuTect v1.1.6[41] in single sample mode with default parameters. Short indels were detected using Indelocator (http://archive.broadinstitute.org/cancer/cga/indelocator) in single sample mode with the default parameters. To ensure high quality variant calls, we required a minimum coverage of 4 reads with minimum two reads supporting the alternate allele. Variants with allelic fraction below 0.1 and variants outside the protein-coding region were excluded. To remove germline-like variants, any variant with a normal allelic frequency greater than 10^{-5} as described in the Exome Aggregation Consortium (ExAC) project[42] was excluded with the exception of any cancer-recurrent variants defined by a minimum frequency of 3 in TCGA or a frequency of 10 in COSMIC[42].

We also further filtered out sequencing artifacts and germline variants using a panel of normals (PoN). For each genomic position, we encoded the distribution of alt read counts across ~8,000 TCGA normals. For each mutation call, we computed a score indicating whether or not its observed read counts are at or below counts across the PoN. We flagged sites with a corresponding score above a certain threshold (PoN log-likelihood > -2.5). Thus, if a site recurrently harbors moderate sequencing noise in the PoN and is called at a low-to-moderate allelic fraction, it is flagged. Likewise, a call with many supporting reads at the same locus would not be. A common germline site would have recurrently high allelic fractions across the PoN, but any call at that site with an allelic fraction below germline levels would be flagged.

Whole exome sequencing data in the form of bam files from the GDSC was downloaded from the Sanger Institute (http://cancer.sanger.ac.uk/cell_lines, EGA accession number: EGAD00001001039) GDSC dataset and processed with the same pipeline[3].

2.2 Variant calling and filtering germline variants for RNAseq data

We applied a similar variant calling pipeline described in 2.1 to RNAseq data with some modifications. Instead of using indelocator for calling indels; we used the GATK best practices pipeline[43] (outlined in https://gatkforums.broadinstitute.org/gatk/discussion/3892/the-gatk-best-practices-for-variant-calling-on-rnaseq-in-full-detail) to call mutations and indels in STAR realigned RNAseq samples. We also ran MuTect v1.1.6[41] on Tophat 1.4 aligned samples to call SNVs. We then kept only the intersection of SNVs that were called by GATK and MuTect v1.1.6. We further called SNVs using MuTect v1.1.6 in 200 additional normal samples from the Genotype-Tissue Expression (GTEx) program. We used this list to exclude common artifacts and germline variants before running the passing variants through the same germline filtering process described earlier for WES and WGS. For 3 cell lines (HUH7_LIVER, FUOV1_OVARY, 2313287_STOMACH) the GATK pipeline failed to produce mutation calls, so we only used RNAseq-based mutation calls for the remaining 1,016 cell lines (Extended Data Fig. 2a).

2.3 Comparison with Sanger GDSC WES

To compare variant calls for CCLE cell lines and Sanger GDSC WES data, we applied MuTect to force call the germline filtered SNVs that were detected in either CCLE or GDSC cell lines. We also used a panel of ~100,000 common SNVs for comparing the germline variants. For each SNV, we calculated the allelic fraction as the ratio of number of reads supporting the alternate allele to total number of reads covering the locus (AF= N_alt/ (N_alt+N_ref)), where N_alt is the number of reads supporting alternative allele and N_ref is the number of reads supporting reference allele for each variant in each cell line. We included only variants that had a coverage of 10 or more reads in both datasets and allelic fraction of at least 0.1 in minimum one of the datasets. We then compared the CCLE and GDSC samples by calculating the Pearson Correlation between the allelic fractions for all variants (global comparison) and for each cell line (individual cell line comparison). This was done using both CCLE WES and CCLE hybrid capture data. We obtained highly comparable results between CCLE_WES_vs_Sanger_WES and CCLE_HC_vs_Sanger_WES (Extended Data Fig. 2f,g). We used correlation between CCLE_HC and Sanger WES to annotate the genetic drift in each cell line (Supplementary Table 3). For the merged mutational calls, we excluded 65 Sanger cell lines with Pearson r< 0.75 for somatic variants allelic fractions. For cancer hotspot mutations, we only included the subset of variants that were highly recurrently observed in TCGA (in 6 or more TCGA samples). We excluded the three germline mismatching cell lines (DOV13_OVARY, PC3_PROSTATE, ISHIKAWAHERAKLIO02ER_ENDOMETRIUM) in the global comparisons.

2.4 Structural variant analysis

932 whole genomes aligned to human genome reference GRCh37 available from Genomic Data Commons as part of the TCGA and 329 new whole genomes from the CCLE cell lines were run through the SvABA[44] structural variant caller using default settings with each tumor genome paired with its corresponding normal genome. For CCLE WGS, we used HCC1143BL as the normal, and further filtered out more possible germline SV with a structural variant blacklist constructed from the set of all germline structural variants detected as part of the SvABA structural variant calling pipeline.

2.5 Fusions analysis

Fusions detection and filtering

For gene fusion detection, we used STAR-Fusion v0.7.1 (https://github.com/STAR-Fusion/STAR-Fusion)[45] which identifies fusion transcripts from RNA-seq data and outputs all supporting data discovered during alignment. We used a cutoff of 5 reads (either spanning or crossing the fusion) to call the presence of a translocation. To reduce artifacts, we removed any fusions detected in more than one sample in GTEx or in 20 or more samples in CCLE and removed fusions involving mitochondrial chromosomes, or HLA genes, or immunoglobulin genes, or with (SpliceType=" INCL_NON_REF_SPLICE" and LargeAnchorSupport="No" and minFAF<0.02), or (sumFFPM<0.1 and minFAF<0.02). We further filtered fusions by fusion allelic fractions (FAF_left^2 + FAF_right^2 > 0.0225 and minFAF > 0.03, excluding fusions detected in TCGA). Here FAF_left is fusion allelic fraction for the left fusion partner reported by STAR-Fusion, FAF_right is the fusion allelic fraction for the right fusion partner, and minFAF is the minimum of the two.

Comparison of fusions with gene dependencies

To investigate the association between fusions and gene dependencies, for each of the gene dependency datasets (Achilles RNAi, Achilles CRISPR, and DRIVE RNAi), and for each of the two genes in the fusion gene pair, we divided cell lines into two groups based on the presence of the fusion, and applied two-sided t-test to compare the distribution of gene dependencies in the two groups. We used Benjamini & Hochberg procedure to obtain adjusted p-values. We used the difference between the mean dependencies in the two groups to calculate the effect size (Extended Data Fig. 3c, Supplementary Table 4).

2.6 Mutational Signature Analysis

Datasets

TCGA MC3 mutations calls were downloaded from https://gdc.cancer.gov/about-data/publications/mc3-2017 and filtered to keep only mutations with "PASS" or "wga" in "FILTER" column. Based on the mapping of CCLE cell lines to TCGA cancer types we only considered 19 cancer types having at least 20 cell lines; BLCA (n=29), BRCA (n=60), COAD.READ (n=72), DLBC (n=56), ESCA (n=38), GBM (n=45), HNSC (n=62), KIRC (n=55), LAML (n=46), LIHC (n=28), LUAD (n=84), LUSC (n=24), OV (n=60), PAAD (n=48), SARC (n=38), SKCM (n=79), STAD (n=46), and UCEC (n=29). All single nucleotide variants (SNVs) in both TCGA and CCLE cohorts were classified into 96 base substitutions in tri-nucleotide sequence contexts.

De-novo extraction

For each cancer type we combined TCGA and CCLE data and first performed de-novo signature discovery in each combined cohort exploiting a Bayesian variant of non-negative matrix factorization, "*SignatureAnalyzer*" (http://archive.broadinstitute.org/cancer/cga/msp)[46,47], inferring an optimal number of signatures best explaining observed mutations. In each de-novo extraction, we enforced a pure "C>T at CpG" signature as a default, which is profiled from the COSMIC1 signature (https://cancer.sanger.ac.uk/cosmic/signatures) after removing all other components except for C>T at ACG, CCG, GCG, and TCG. The separation of C>T_CpG components from the conventional COSMIC1 was aimed to minimize a possible interference between the background, residual components in COSMIC1 and COSMIC5, which are highly overlapping each other. Based on the manual inspection and the cosine similarity of extracted signatures to 30 COSMIC signatures we identified a set of active signatures in each cancer type (Supplementary Table 6) and exploited this information in the following projection step to infer the activity of COSMIC signatures in both TCGA and CCLE cohorts. Based on prior knowledge and literature we only allowed COSMIC3 (BRCA signature) in BRCA, OV, PAAD, SARC, STAD, and UCEC.

Projection

The comparison of signature attributions across different cancer types or different cohorts needs the use of the same signature profiles. Since the signature profiles from a de-novo extraction varied across cancer types, depending on the number of samples or mutations, here we performed a projection approach to infer sample-specific attributions based on 30 COSMIC signature profiles by modifying "*SignatureAnalyzer*". The pure "C>T at CpG" signature was used instead of COSMIC1. More specifically, the projection was done by minimizing the Kulbeck-Leibler divergence between the mutation count matrix, $\boldsymbol{X}$ (96 × N), N being a number of samples in each combined cohort of TCGA and CCLE, and a product of the signature-loading matrix $\boldsymbol{W}$ (96 × K) and the activity-loading matrix $\boldsymbol{H}$ (30 × K). During the optimization the signature-loading matrix $\boldsymbol{W}$, comprised of the normalized signature profiles of corresponding K COSMIC signatures, were strictly frozen and the activity-loading matrix $\boldsymbol{H}$ was iteratively refined through the multiplication update scheme to best approximate the mutation count matrix $\boldsymbol{X} \sim \boldsymbol{WH}$. The resulting row vectors in $\boldsymbol{H}$ represent de-convoluted signature activities across samples[48]. In each projection we restricted the usage of signatures only to the active ones identified from the de-novo extraction step (Supplementary Table 6; K being the number of active signatures). Due to the multiple MSI signatures (common signatures through most MSI samples - COSMIC6, 15, 21, 26, POLE+MSI – COSMIC14, POLD+MSI – COSMIC20)[49] all common MSI signatures were allowed when a de-novo extraction identified at least one of six MSI signatures, while COSMIC14 and COSMIC20, unique to POLE+MSI and POLD+MSI, respectively, were strictly allowed only when there is an evidence for the corresponding signature in de-novo extraction.

Signature Comparison between CCLE and TCGA

For each cancer type we first calculated the normalized activity of each individual signature across tumors and cell lines (number of mutations attributed to each signature / number of mutations in each sample), and compared the mean of normalized activities between the TCGA and CCLE cohorts.

2.7 MSI annotations

For each cell line profiled by sequencing, we inferred microsatellite instability (MSI) status by counting the total number of filtered deletions called by Indelocator (http://archive.broadinstitute.org/cancer/cga/indelocator) and the fraction of these deletions that were located in microsatellite regions as defined by three consecutive repeats of a sequence of less than five nucleotides in length. Based on the distributions of these values in each of the sequencing datasets (CCLE Hybrid Capture, CCLE WGS, CCLE WES, and Sanger WES), we specified a threshold value for the number of MS deletions (N_MS_del) and two threshold values for the percentage of microsatellite deletions (P_MS_del_1 and P_MS_del_2, see Supplementary Table 7). Cell lines were annotated as inferred-MSI if the number of MS deletions was greater than N_MS_del and the percentage of MS deletions was greater than P_MS_del_2. Similarly, cell lines were annotated as inferred-MSS if the number of MS deletions was less than N_MS_del and the percentage of MS deletions was less than P_MS_del_1 in any of the four datasets (Extended Data Fig. 5a, Supplementary Table 7).

2.8 ABSOLUTE copy number analysis

Allelic copy number, whole genome doubling, subclonality, purity and ploidy estimates were generated by the ABSOLUTE algorithm[50]. Somatic copy number used in ABSOLUTE analysis were derived either from SNP arrays or whole exome sequencing. Allelic fractions of mutation were derived from either Hybrid Capture sequencing or whole exome sequencing data.

2.9 DNA methylation analysis

Annotation of DNA methylation for promoters, enhancers, and CpG islands

Short reads from the Reduced Representation Bisulfite Sequencing (RRBS) data were aligned using Bismark 0.7.12[51] for 843 cell lines. CpG methylation was estimated using the read.bismark tool in the R MethylKit package[1,52] with parameters mincov = 5 and minqual = 20. To estimate gene promoter level methylations, we used RefSeq transcription start site (TSS) information for Hg19 downloaded from the UCSC genome browser. To define promoter regions, we used two approaches. First, for the global analysis of correlation between methylation and mRNA expression (Extended Data Fig. 6c), we used a fixed window size of 1000bp upstream of the transcription start site (TSS) for each gene and calculated a coverage-weighted average of CpG methylations for CpG sites within this region as previously described in (Ziller et al., Nature 2013)[53]. We found 17,182 genes with average coverage greater than 5 reads in the RRBS dataset. For most genes, we observed that the 1kb upstream TSS region contains the promoter methylation changes. However, for some genes, (e.g. VHL), we observed downstream methylation changes relative to the TSS. Therefore, we used an alternative approach to capture gene level methylation signal for the remainder of the analyses in the paper. For each TSS, using data for all cell lines, we first clustered CpG sites within (-3000, 2000) nucleotides of the TSS using the hclust function in R and cut the hierarchical clustering tree to form three clusters. This approach grouped together the CpG sites with similar methylation changes across samples, and these clusters usually represented the CpG sites in the promoter, upstream, and downstream regions. We used the same weighted averaging approach described above to calculate the methylation signal for each cluster in each sample.

To annotate the CpG island and enhancer methylations in the cell lines, we downloaded CpG island and VISTA enhancer coordinates from UCSC genome browser and applied the above unsupervised clustering to a window (coordinate start – 2000, coordinate end + 2000) to determine the methylation for each enhancer and CpG island sequence. For sequences with length greater than 5000, we first divided them to sections of length 5000, and then performed the same clustering process.

tSNE plots for DNA methylation data

To visualize the high dimensional DNA methylation data, we used the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm implemented in the Rtsne package in R with default parameters[54]. We used all the promoter methylation values for CpG clusters with a proper coverage (average CpG coverage > 25 reads) as input features for a two-dimensional embedding for visualization.

Comparison of DNA methylation and mRNA

To compare mRNA expression and promoter methylation, for each gene, we first calculated z-scores for its mRNA expression (log RPKM) and promoter methylation. We then calculated the linear regression coefficient associating expression to methylation while correcting for cancer type using the R function lm(expr~meth+cancer_type). For the null distribution, we permutated the gene labels for mRNA expression dataset and repeated the same procedure.

Comparison of DNA methylation and dependency

To investigate the association between promoter methylation and gene dependencies, for 2,776 genes with significant negative correlations between promoter methylation and mRNA expression (Pearson correlation < -0.5), we calculated Pearson correlations between promoter methylations and dependencies for all pairs of genes connected in the STRING dataset (string-db.org)[55]. Here, for each gene, we considered up to 100 top connected genes in STRING with a connectivity score above or equal to 800. For robust correlations, we excluded the top three cell lines with highest sums of squares of normalized dependency and methylation scores and calculated Pearson correlations using the remaining samples. This analysis was performed separately on the Achilles RNAi[5], Achilles CRISPR[7], and Project DRIVE[6] gene dependency datasets. For each correlation coefficient value, we assigned an estimated p-value by fitting a normal distribution to all correlation coefficients calculated within the respective dataset. We then used the p.adjust function in R to calculate the false discovery rate (q-value) for each methylation-dependency correlation (Fig. 2a and Supplementary Table 8).

LDHA, LDHB, and RPP25 promoter methylation in TCGA

We examined methylation-expression relationships for LDHA, LDHB, and RPP25 in 22 TCGA tumor types. Methylation profiling (Illumina HM450 BeadChip beta-values) and RNA-seq expression (log2 RPKM) data were sourced from the TCGA provisional datasets hosted at cBioPortal (cbioportal.org/datasets.jsp)[56,57]. We excluded tumor types with less than 100 samples with both methylation and expression annotations. Correlation values for methylation vs. expression of the same gene were then computed and are shown in order of magnitude (Extended Data Fig. 6i).

2.10 Global chromatin profiling analysis

Unsupervised clustering and heatmap

The 897 cell lines with available global chromatin data were clustered based on the 38 (of 42) chromatin modifications that were detected in more than 98% of the cell lines using *pheatmap* R function (Pretty Heatmaps v1.0.10) with parameters clustering_method = 'ward.D', clustering_distance_cols = 'euclidean', and cutree_cols=19.

CREBBP TAZ2/CH3 specific truncating mutations were annotated as the truncating mutations in CREBBP occurring between amino acids 1745 and 1846 (affecting TAZ2/CH3 domain but not ZZ domain). Similarly, for EP300 TAZ2/CH3 specific truncating mutations we included any truncating mutation in EP300 occurring between amino acids 1708 and 1809 (Fig. 3, Extended Data Fig. 7a).

EP300 and CREBBP enrichment volcano plot

Fisher test (two-sided) was used to evaluate enrichment of truncating mutations in the newly identified high H3K18/K3K27 acetylation cluster. For truncating mutations, we included any nonsense mutations, splice site mutations, or frameshift indels affecting any part of the gene. For the analysis in Extended Data Fig. 7b, only genes with at least 20 affected cell lines (N=684) were included. We used fisher.test function in R to estimate the odds ratios and P-values. Adjusted P-values were obtained using p.adjust function in R.

2.11 RNA-seq analysis

Short read alignment and calculation of gene expression

RNA-seq reads were aligned to the GRCh37 build of the human genome reference using STAR 2.4.2a[58]. The GENCODE v19 annotation was used for the STAR alignment and all other quantifications. Gene level RPKM and read count values were calculated using RNA-SeQC v1.1.8[59]. Exon-exon junction read counts were obtained from STAR. Isoform-level expression in TPM was quantified using RSEM v.1.2.22. All methods were run as part of the pipeline developed for the GTEx Consortium (https://gtexportal.org)[60].

CCLE comparison to GTEx and TCGA

We compiled log2(TPM + 1) gene expression data for 1,019 CCLE cancer cell lines, 10,535 TCGA primary tumor samples, and 11,688 GTEx normal tissue samples. TCGA Pan-Cancer TOIL RSEM TPM data was obtained from Xena Browser (https://xenabrowser.net/) and GTEx v7 TPM data was accessed from the GTEx Portal (https://gtexportal.org/home/datasets). We compared CCLE and TCGA data using a subset of 5,000 genes that were highly variable in the CCLE and TCGA data and 22 cancer types that were common to both the TCGA and CCLE datasets. In each dataset, we averaged the gene expression data across all samples per cancer type, then mean subtracted per gene. We calculated the pairwise Pearson correlation between the averaged CCLE gene expression and the averaged TCGA gene expression. We compared CCLE and GTEx data using a subset of 5,000 genes that were highly variable in the CCLE and GTEx data. We averaged the CCLE and GTEx gene expression data across all samples per cancer type or primary site, respectively, mean subtracted per gene, and calculated the pairwise Pearson correlation between the averaged CCLE gene expression and the averaged GTEx gene expression. We also compared individual CCLE cell lines to TCGA and GTEx average profiles. The gene expression data for individual cell lines were mean subtracted per gene using the same vector of means as the averaged CCLE expression. We calculated the pairwise Pearson correlation between the gene expression for these cell lines and the averaged TCGA and GTEx gene expression (Supplementary Table 9).

Exon inclusion ratios

To quantify alternative splicing in cell lines, we used the STAR junction read counts to estimate the fraction of times each exon was spliced in. For both ends of each exon, we calculated the total number of junction reads supporting inclusion of that exon (n_i) and the total number of junction reads supporting skipping of the exon (n_j). We estimated the inclusion ratio as r = $\frac{n_i}{n_i+n_s}$. We required each exon ratio to be supported by at least 10 reads ($n_i + n_s \geq 10$).

Splicing vs dependency

To investigate if some gene dependencies were more strongly correlated with exon splicing instead of total mRNA expression, we correlated exon inclusion ratios produced using the above method with Achilles RNAi gene dependency data and compared the results to a similar analysis based on mRNA expression. For each exon, we calculated the Pearson correlation between exon inclusion and the DEMETER dependency score of the same gene (x-axis on Fig. 4a) and compared that correlation with the respective Pearson correlation between the total mRNA expression and dependency of the same gene (y-axis on Fig. 4a). In this analysis, we only included exons quantified in at least 200 cell lines with Achilles data to obtain robust correlation estimates.

2.12 microRNA analysis

Nanostring data QC and Normalization

Samples were divided into 14 batches, and two replicates of the K-562 cell line were included in each batch as a control. Internal positive and negative controls were used for normalization as recommended by NanoString using NanoString nSolver software. We excluded samples that failed NanoString nSolver quality control as well as one sample based on low positive control signal (normalization coefficient > 6) and another sample based on high background signal (with second ranked negative control value > 80). To estimate the background signal, we sorted the values for the negative controls within each sample and picked the second highest value as the background estimate. The median background estimate across all cell lines was 26.1. We used log(50 + N), where N is the nSolver normalized value to reduce the effect of the background signal in the downstream analyses.

Comparison of microRNA and dependency

To identify the strongest specific associations between microRNA expression and gene dependencies, we calculated the Pearson correlation between the expression of each microRNA and each gene dependency score in the Achilles RNAi dataset. We then normalized the Pearson correlations for each microRNA (z1, x-axis on Extended Data Fig. 10b) and for each gene dependency (z2, y-axis on Extended Data Fig. 10b). Several gene dependency/ microRNA pairs showed outlier correlations (with $|z1|>6$ or $|z2|>6$). We chose the top scoring association (CTNNB1/mir-215) for further investigation and comparison with data from The Cancer Genome Atlas (TCGA) (Extended Data Fig. 10c-j; Supplementary Table 13).

2.13 RPPA analysis

Batch effect correction and quality control

RPPA data was normalized within each batch as described in (1.7), and the log-transformed values were merged and corrected for batch effect using the removeBatchEffect method in Limma package in Bioconductor[61,62].

Out of the 925 cell lines that were profiled, 26 lines were excluded. These were comprised of 19 lines with low total protein content and 7 lines with poor overall mRNA-protein correlation. For the 6 cell lines with biological replicates, the average of the two replicates in batch two were used.

Correlation of mRNA and protein

For 154 RPPA antibodies against single gene total proteins, Pearson correlations for mRNA (RNAseq log2 RPKM) and protein levels were obtained. For null distribution, gene labels were randomly permuted (Extended Data Fig. 11a).

Effect of RPPA dynamic range on protein-mRNA correlation

For 154 RPPA antibodies against single gene total proteins, dynamic range was calculated as the difference between the third highest and the third lowest values across all cell lines. Dynamic range was plotted against mRNA-protein correlations (Extended Data Fig. 11b). Statistical significance was determined using two-sided Pearson correlation test.

Effect of Antibody type and antibody quality of protein-mRNA correlation

For 154 RPPA antibodies against single gene total proteins, Wilcoxon rank sum test was used to evaluate difference between validated antibodies (N=96) and those annotated as "with caution" (N=58) as provided by MD Anderson Cancer Center Reverse Phase Protein Array (RPPA) Core Facility (S11C, left, Supplementary Table 14). Similarly, we compared the protein-mRNA correlations of antibodies against single gene total protein (N=154) with antibodies against single gene phospho-proteins (N=50).

Comparison of mRNA-protein correlations between CCLE and TCGA

mRNA and protein correlations for 181 antibodies across 3,467 TCGA samples from 11 tumor types were calculated for each antibody and compared with CCLE mRNA-protein correlations[63]. Two-sided Pearson correlation test was used to evaluate statistical significance. (Extended Data Fig. 11d).

RPPA elastic net analysis

An elastic net regression analysis similar to the one used in Barretina et al.[1] was run to find genomic features that predict drug sensitivities as measured by area under the dose response curve (AUC). The feature set included mutations, DNA copy number, mRNA expression and RPPA protein data. These features were used to predict sensitivities to 24 compounds profiled in the CCLE and 138 compounds from Genomics of Drug Sensitivity in Cancer (GDSC) project.

Features with an absolute Pearson correlation of greater than 0.1 with the target drug sensitivity profile were selected. Optimal values for the alpha and lambda parameters were found by a 10-fold cross validation using cv.glmnet function in the glmnet R package[64]. A 200-fold bootstrapping was then performed using the optimal parameter values. We calculated the frequency of selection and average weight for each feature.

The above analysis was performed twice for each drug, once using all features and another time using all features with the exclusion of RPPA values. The model prediction errors for the two models were compared to estimate the accuracy gained by adding the RPPA data.

3. Data Availability

3.1 CCLE datasets

All the CCLE processed datasets are available at the CCLE portal (www.broadinstitute.org/ccle) and depMap portal (http://www.depmap.org). Raw sequencing data is available at Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra); SRA Data accession number: PRJNA523380.

3.2 Gene Dependency Datasets

Achilles RNAi data (DEMETER scores) was downloaded from https://portals.broadinstitute.org/achilles.

Achilles CRISPR Avana 18Q3 public dataset (gene effects, CERES scores) was downloaded from https://figshare.com/articles/DepMap_Achilles_18Q3_public/6931364/1.

Novartis Project DRIVE RNAi dataset (ATARiS scores) was obtained from the Project DRIVE authors.

3.3 Drug Sensitivity Datasets

Cancer Therapeutics Response Portal (CTRP) Area Under the Dose-Response Curve (AUC) scores was downloaded from NCI website (ftp://caftpd.nci.nih.gov/pub/OCG-DCC/CTD2/Broad/CTRPv2.0_2015_ctd2_ExpandedDataset).

Sanger Genomics of Drug Sensitivity in Cancer (GDSC) drug sensitivity (AUC and IC50 scores) were downloaded from Sanger website (https://www.cancerrxgene.org/downloads).

4. Code Availability

Most of the statistical analyses were performed in R (version 3.5.2). Source codes are available upon request.

Supplementary References:

36. Fisher S, Barry A, Abreu J, et al. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome biology.* 2011;12(1):R1.
37. Johannessen CM, Johnson LA, Piccioni F, et al. A melanocyte lineage program confers resistance to MAP kinase pathway inhibition. *Nature.* 2013;504(7478):138-142.
38. Boyle P, Clement K, Gu H, et al. Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. *Genome biology.* 2012;13(10):R92.
39. Brat DJ, Verhaak RG, Aldape KD, et al. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *The New England journal of medicine.* 2015;372(26):2481-2498.
40. Integrated genomic characterization of papillary thyroid carcinoma. *Cell.* 2014;159(3):676-690.
41. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology.* 2013;31(3):213-219.
42. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536(7616):285-291.
43. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics.* 2013;43:11.10.11-33.
44. Wala JA, Bandopadhayay P, Greenwald NF, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome research.* 2018;28(4):581-591.
45. Haas B, Dobin A, Stransky N, et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv.* 2017.
46. Kasar S, Kim J, Improgo R, et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nature Communications.* 2015;6:8866.

47. Kim J, Mouw KW, Polak P, et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature genetics.* 2016;48(6):600-606.

48. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell.* 2017;171(4):950-965.e928.

49. Haradhvala NJ, Kim J, Maruvka YE, et al. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat Commun.* 2018;9(1):1746.

50. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology.* 2012;30(5):413-421.

51. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics (Oxford, England).* 2011;27(11):1571-1572.

52. Akalin A, Kormaksson M, Li S, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome biology.* 2012;13(10):R87.

53. Ziller MJ, Gu H, Muller F, et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature.* 2013;500(7463):477-481.

54. Van der Maaten LJPaH, G.E. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research.* 2008;9:2579-2605.

55. Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research.* 2015;43(Database issue):D447-452.

56. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery.* 2012;2(5):401-404.

57. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling.* 2013;6(269):pl1.

58. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England).* 2013;29(1):15-21.

59. DeLuca DS, Levin JZ, Sivachenko A, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics (Oxford, England).* 2012;28(11):1530-1532.

60. Consortium GT, Aguet F, Brown AA, et al. Genetic effects on gene expression across human tissues. *Nature.* 2017;550:204.

61. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research.* 2015;43(7):e47.

62. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology.* 2004;3:Article3.

63. Akbani R, Ng PK, Werner HM, et al. A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat Commun.* 2014;5:3887.

64. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software.* 2010;33(1):1-22.